COMPUTING edge

- AI Trust
- Quantum
- Software Development
- Cybersecurity Mesh

DECEMBER 2025

www.computer.org

IEEE COMPUTER SOCIETY

IEEE

COMPUTING edge

IEEE COMPUTER SOCIETY IEEE

IEEE COMPUTER SOCIETY computer.org

Printed with inks containing soy and/or vegetable oils

STAFF

Editor
Lucy Holden

Periodicals Portfolio Senior Managers
Carrie Clark and Kimberly Sperka

Director, Publications and Special Projects
Robin Baldwin

Production & Design Artist
Carmen Flores-Garvey

Periodicals Operations Project Specialists
Priscilla An and Christine Shaughnessy

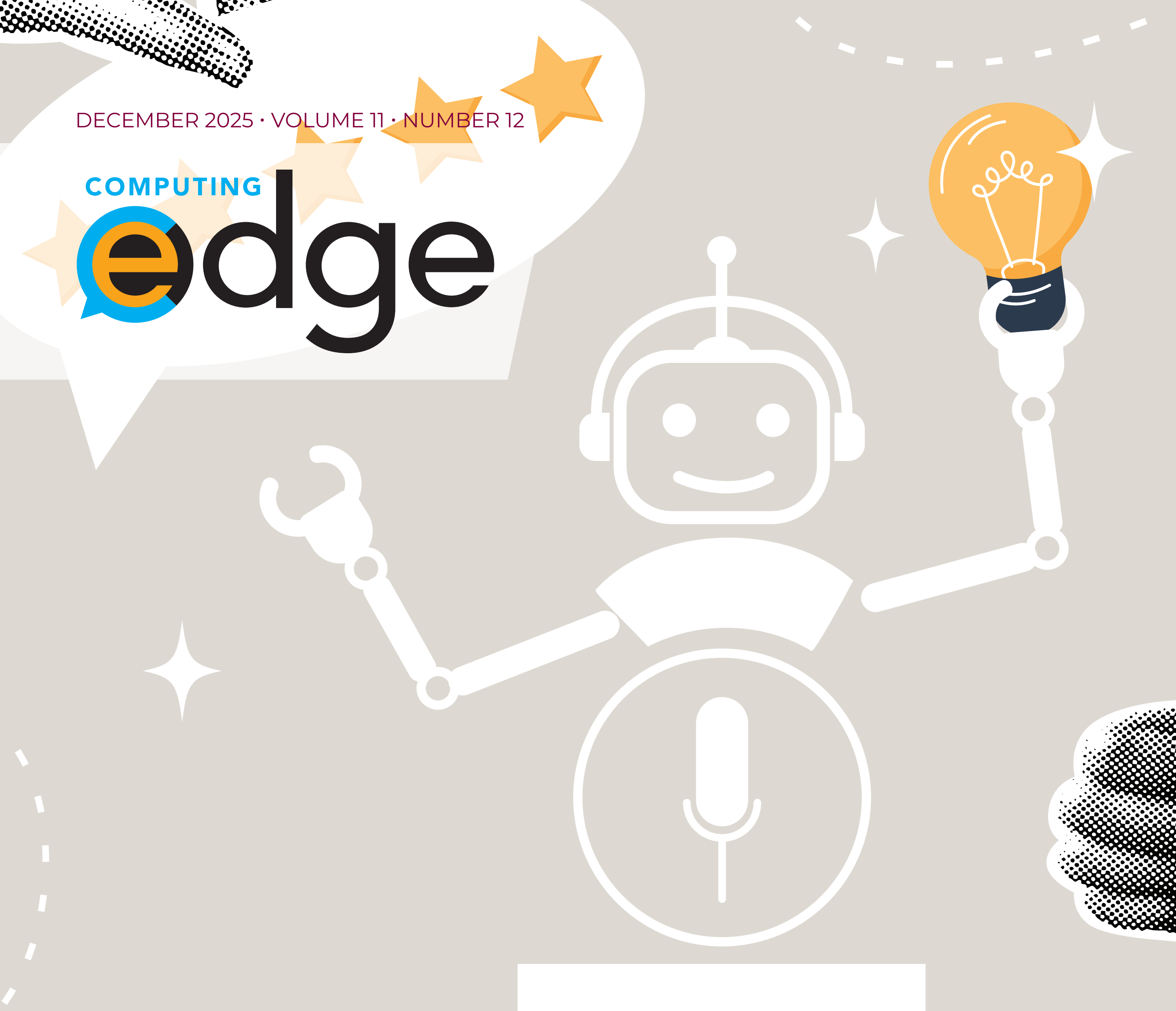
Senior Advertising Coordinator
Debbie Sims

Circulation: *ComputingEdge* (ISSN 2469-7087) 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.

Postmaster: Send address changes to *ComputingEdge*-IEEE Membership Processing Dept., 445 Hoes Lane, Piscataway, NJ 08855. Periodicals Postage Paid at New York, New York, and at additional mailing offices. Printed in USA.

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

2025 IEEE Computer Society Magazine Editors in Chief



DECEMBER 2025 · VOLUME 11 · NUMBER 12

COMPUTING edge

AI Trust

Quantum

Software Development

Cybersecurity Mesh

Departments

Magazine Roundup

The IEEE Computer Society's lineup of 12 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.

Computer

AI That Learns, Thinks, and Acts: The Next Frontier of Generative AI

This article, featured in the October 2025 issue of *Computer*, dives into the recent transformation of artificial intelligence (AI) systems from passive content generators into autonomous decision makers and executors through the advancements of self-improving retrieval-augmented generation, generative AI-native agents, and large action models, which highlight their potential to automate workflows and enhance real-world applicability.

Computing in Science & Engineering

A Probabilistic Analysis of the Mexican Lotería Game

Lotería is a Bingo-like game and a typical Mexican pastime. It allows, in its entertainment mode (without bets), young and old to compete at the same level while having fun. However, despite its great popularity and relevance, this game of chance has not been analyzed from a probabilistic point of view. The authors of this July–September 2025 *Computing in Science & Engineering* article deduce the probabilistic distributions of the Mexican Lotería game. They do this by viewing the game as a process that has a well-defined beginning, but whose end is indeterminate.

IEEE Annals of the History of Computing

Information Technology Pioneers of Aotearoa New Zealand

This article, featured in the April–June 2025 issue of *IEEE Annals of the History of Computing*, traces the history of New Zealand's information technology innovators from the late 19th century until the arrival of the first stored-program computers in 1960. This history parallels the country's evolution toward a modern economy.

IEEE Computer Graphics and Applications

Circuit Mining in Transcriptomics Data

A central goal in neuropharmacological research is to alter brain function by targeting genes whose expression is specific to the corresponding brain circuit. Identifying such genes in large spatially resolved transcriptomics data requires the expertise of bioinformaticians for handling data complexity and to perform statistical tests. This time-consuming process is often decoupled from the routine workflow of neuroscientists, inhibiting fast target discovery. The authors of this September/October 2025 *IEEE Computer Graphics and Applications* article present a visual analytics approach to mining expression data in the context of meso-scale brain circuits for potential target genes tailored to domain experts with limited technical background.

IEEE Intelligent Systems

A Blockchain-Driven Framework for Deep Reinforcement Learning-Controlled UAV Bridge Health Monitoring

Structural health monitoring (SHM) of bridges is essential for maintaining safe and reliable transportation networks. This study, which was featured in the July/August

2025 issue of *IEEE Intelligent Systems*, integrates the EOSIO blockchain with unmanned aerial vehicles (UAVs) and deep reinforcement learning using the proximal policy optimization algorithm to optimize UAV-based bridge inspections and crack detection. The blockchain ensures secure and tamper-proof crack data storage, enhancing transparency and data integrity.

Reliable Cloud Operations Using Transformers

Managing large-scale public clouds presents significant challenges, particularly in analyzing the risk of commands executed by operators to guarantee that they do not cause damage to the infrastructure. Traditional rule-based systems have been used but fall short in scalability when managing operations at a global scale. Although machine learning techniques have been proposed as alternatives, they have typically been applied in small-scale environments. This article from the May/June 2025 issue of *IEEE Internet Computing* presents a novel approach to address the complexities of global-scale command risk analysis and standard operating procedure verification using large language models.

Assessing Processor Sustainability Using the First-Order FOCAL Carbon Model

To overcome the inherent data uncertainty regarding the sustainability of computing devices in general and processors in particular, this article, featured in the July/August 2025 issue of *IEEE Micro*, proposes the parameterized First-Order analytical CArbon modeL (FOCAL) to assess processor sustainability from first principles. The author uses FOCAL to analyze and categorize a broad set of archetypal processor mechanisms into strongly, weakly, or less sustainable design choices, providing insight and intuition into how to reduce a processor's environmental footprint with implications to both hardware and software.

Exploring the Privacy Protection Capabilities of Chinese Large Language Models

Large language models (LLMs) are renowned for their advanced capabilities, significantly enhancing artificial intelligence. However, these advancements have also raised increasing concerns about privacy and security. To address these issues, the authors of this April–June 2025 *IEEE MultiMedia* article developed a three-tiered framework designed to evaluate privacy in language systems through progressively complex tests. Their findings suggest that current Chinese LLMs show widespread privacy protection issues, indicating that this challenge remains common and may pose corresponding privacy risks in applications based on these models.

Characterization and Feasibility of Wearable Spectroscopic Tracking of Nutrition Biomarkers

Disorders related to nutrition and metabolism remain a critical public health challenge globally, affecting a significant portion of the population. Wearable biosensing technologies present a promising avenue for more precise, noninvasive monitoring of nutritional intake, providing individuals with real-time feedback to improve their dietary habits. In this study, featured in the July–September 2025 *IEEE Pervasive Computing* issue, the authors explore the feasibility of using an off-the-shelf wearable device to track nutrients in the blood.

IEEE SECURITY&PRIVACY

Swiping Safely in Online Dating Platforms

Online dating platforms dominate intimate relationship initiation but entail risks. In this article, featured in the July/August 2025 issue of *IEEE Security & Privacy*, the authors assess whether the European Union Digital Services Act protects users from these risks and provide recommendations to improve the experience and safety of online dating.

IEEE Software

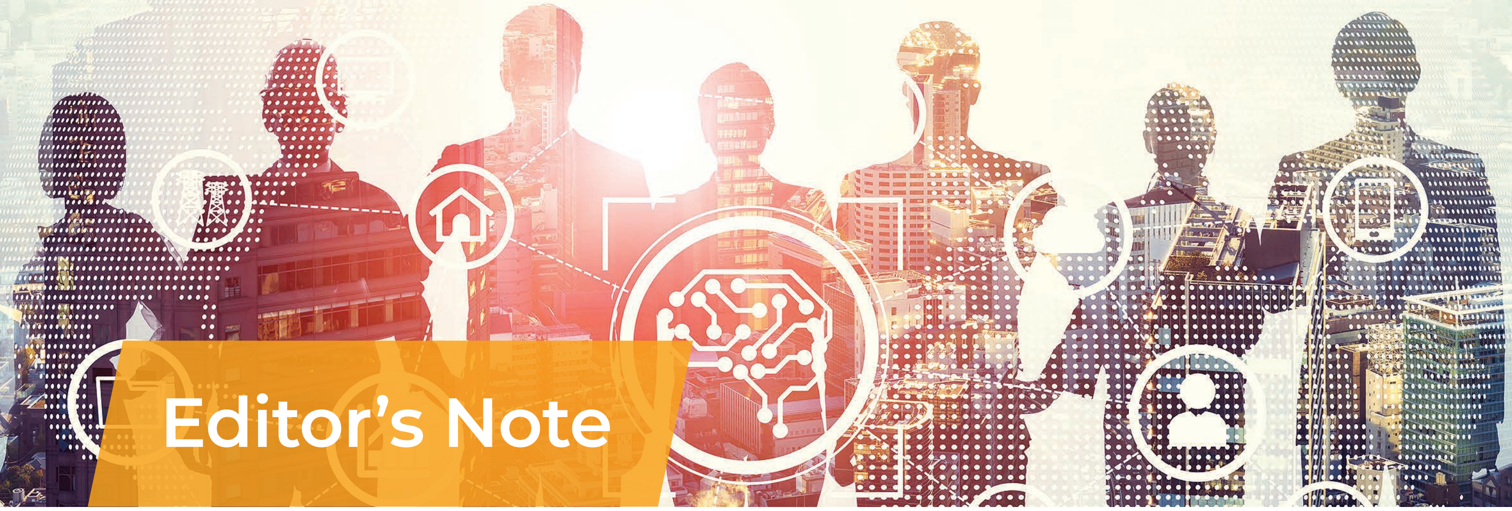
Static Analysis and Transformation for Quantum Programming Languages

Despite progress in quantum research, there is a notable lack of development tools and technologies that assist developers in designing and analyzing quantum programs. This article from the September/October 2025 issue of *IEEE Software* aims to address this gap using a software engineering approach.

IT Professional

A Scheme to Enhance Dense Internet of Things Networks Performance: An Early Evaluation

Ultradense networks (UDNs) provide enhanced coverage and improve capacity to meet the ever-growing demand of data. In UDNs, users experience high data rates, low latency, and many other significant features; however, with the continuous growth of the Internet of Things (IoT), the IoT-UDNs in the future will face multiple significant hurdles. In this article, featured in the July/August 2025 issue of *IT Professional*, the authors particularly focus on higher packet drop rate issue in the IoT-UDNs environment. They propose a simple and effective approach to alleviate packet drop rate.



Can We Trust AI?

Despite the enormous potential of artificial intelligence (AI), its trustworthiness remains a significant concern. Is it secure, and can it be governed? Can it give accurate responses or complete tasks correctly? How can we ensure its accountability in making objective and fair decisions? This issue of *ComputingEdge* explores concerns around AI trust and improper use, as well as its applications in reliability engineering, systems engineering, and software development. The articles also explain how to teach and secure quantum computing. The issue concludes with a discussion of two distinct components of cybersecurity mesh.

Organizations and engineers need to be careful when using AI to maintain security and quality. In “Governing Agentic AI: Security, Identity, and Oversight in the Age of Autonomous Intelligent Systems” from *Computer*, the author outlines the security challenges and need for strategies to mitigate risks when deploying agentic AI systems in organizations. The authors of “General and Agentic AI, and the Challenges of Xplainable Reliability” from *Computer*, highlight the benefits and risks of using AI in reliability engineering.

Quantum computing is a highly abstract discipline that needs to be made more approachable for students and more secure against future cryptanalysis. *IEEE Computer Graphics and Applications* article “Quantum Computing for All: Online Courses Built Around an Interactive Visual Quantum Circuit Simulator” introduces an online course designed to teach quantum computing for diverse students with no prior knowledge of quantum physics. In “Post-Quantum Adversarial Modeling: A User’s Perspective” from *Computer*, the authors offer recommendations to mitigate quantum attacks by predicting the risks before they happen.

It is important for engineers to keep up with the constantly evolving field of software development, particularly when it comes to AI and human influence on software. *IEEE Software* article “Bringing Software Engineering Discipline to the Development of AI-Enabled Systems” presents a series of papers that reflect the current development of the field of AI systems engineering and AI software development. The *IEEE Software* article “Human Aspects and Security in Software Development,” compares human and technical aspects of software development.

Decentralized identity and distributed key generation are two components of a robust cybersecurity mesh system. *Computer* article “The Road to Decentralized Identity: The Techniques, Promises, and Challenges of Tomorrow’s Digital Identity” posits decentralized identity as a way for users to control and protect their digital identities and personal data. The author of “Threshold Signatures,” from *IEEE Security & Privacy*, proposes threshold signatures as a technique to ensure redundancy and distribution of trust for a secret signing key.

2469-7087/25 © 2025 IEEE

EDITOR: Nir Kshetri, The University of North Carolina at Greensboro, nbkshetr@uncg.edu

This article originally appeared in
Computer
vol. 58, no. 8, 2025

DEPARTMENT: COMPUTING'S ECONOMICS

Governing Agentic AI: Security, Identity, and Oversight in the Age of Autonomous Intelligent Systems

Nir Kshetri, *The University of North Carolina at Greensboro*

This article examines the governance and security challenges of deploying agentic artificial intelligence systems in organizations. It highlights the need for strategies to ensure compliance and mitigate risks.

Nvidia CEO Jensen Huang has called artificial intelligence (AI) agents a ˮmultitrillion-dollar opportunity," highlighting their potential to transform industries and economies.[1] In 2025, the application programming interface (API) management firm Gravitee examined the adoption of Agentic AI systems and large language models (LLMs) among large and midsize companies and found that 72% of respondents reported their organizations were actively using agentic AI systems at the time.[2] According to the PagerDuty Agentic AI Survey 2025, conducted by Wakefield Research with 1,000 IT and business executives from the United States, United Kingdom, Australia, and Japan, more than 60% expect AI agents to deliver a return on investment (ROI) of over 100%, with U.S. companies forecasting an average ROI of 192%. Furthermore, nearly 45% believe agentic AI will surpass generative AI in impact, and over 90% plan to adopt agents at a faster pace than they did generative AI.[3]

Despite its economic potential, agentic AI's autonomy, adaptability, and complexity make it difficult to govern. Operating independently through nonhuman identities, these systems often lack transparency, raising concerns about accountability, bias, and fairness in high-stakes environments like finance and health care.[4] Increased AI autonomy and accessibility heighten risks if governance and security are lacking. More users mean greater chances of unauthorized or improper use, including inputting sensitive data into unapproved tools, leading to potential leaks. Unsupervised AI agents in critical systems amplify risks from shadow configurations. These agents also face issues like hallucination, bias, and model drift, which can undermine reliability and safety.[5]

Unsurprisingly many organizations have recognized the need for robust frameworks to ensure these systems operate responsibly and align with ethical, security, and regulatory standards. In a 2025 survey by independent authors and UiPath of over 250 U.S. IT executives, the top concern (cited by 56%) was IT security, especially as AI agents, which derive value from acting on company data, become more integrated with enterprise systems.[6] Similarly, a Deloitte survey found that 58% of early adopters expressed concerns about managing sensitive data, underscoring the need for robust data governance and cybersecurity frameworks to address the growing challenges as AI agents become further embedded within enterprise systems.[7] Likewise, a recent survey of 300 tech leaders by API management and security platform Gravitee found that while businesses are rapidly adopting agentic AI and LLMs for efficiency gains, governance has become a top priority, with 76% rating it "extremely important" due to concerns over system integration, data security, and managing LLM costs.[8]

Beyond security, organizations must navigate regulatory uncertainty, as laws often lag behind AI's complexity and pace.[4] National and global bodies are likely to introduce agent-specific rules soon. Some jurisdictions are already advancing regulatory initiatives to

Digital Object Identifier 10.1109/MC.2025.3572173
Date of current version: 28 July 2025

address the rising deployment of agentic AI systems that influence or replace human decision-making processes. The California Privacy Protection Agency (CPPA) has proposed national standards to regulate automated decision-making technology (ADMT) under the CPPA, defining it as technology that processes personal data to "execute a decision, replace human decision making, or substantially facilitate human decision making," including when its output is a key factor in human decisions.[9] The European Union (EU) AI Act, while not explicitly naming agentic AI, provides a risk-based framework that ensures its ethical, transparent, and accountable use. Agentic AI often falls under high-risk categories, particularly in areas like customer service, finance, and rights-based decision making. The Act requires clear disclosure when users interact with AI, and mandates understandable explanations for AI-driven decisions, such as dynamic pricing or credit assessments.[10]

This article explores the governance challenges and security concerns associated with the integration of agentic AI into enterprise workflows. It highlights the need for effective frameworks to manage AI agents' autonomous behaviors while ensuring compliance with ethical, security, and regulatory standards.

GOVERNANCE OF NONHUMAN IDENTITIES

As AI agents become more embedded in enterprise workflows, they rely on nonhuman identities (NHIs).[11] NHIs are digital credentials—such as bots, API keys, service accounts, and Open Authorization (OAuth) tokens—used by applications, services, and machines to authenticate, access resources, and interact within enterprise systems.[12] These machine credentials now vastly outnumber human identities and form a critical yet often overlooked attack surface.[11] AI agents amplify existing NHI security challenges by operating at machine speed and scale, chaining unpredictable tools and permissions, running continuously without clear session limits, requiring broad system access, and introducing new attack vectors in multiagent setups.[11]

Organizations are embracing automation to streamline processes and boost efficiency—a trend intensified by agentic AI.[12] Estimates of nonhuman identities per human vary widely—one study cited a 20:1 ratio in September 2024,[13] another reported 45:1 in April 2025,[14] and a Dark Reading analysis estimated 50:1 in December 2024.[12] With NHIs vastly outnumbering human identities, organizations must automate lifecycle controls—spanning creation to expiration—to ensure continuous oversight and minimize risk.[13]

DESPITE ITS ECONOMIC POTENTIAL, AGENTIC AI'S AUTONOMY, ADAPTABILITY, AND COMPLEXITY MAKE IT DIFFICULT TO GOVERN.

In 2024, 23.77 million secrets—such as API keys and credentials—were leaked on GitHub, a 25% rise from the previous year, underscoring how the rapid growth of NHIs like service accounts and AI agents is expanding the attack surface for cyber threats.[15] Cybercriminals target NHIs—especially autonomous Internet of Things entities—by exploiting weak authentication, misconfigurations, and poor monitoring, leading to breaches and system disruptions.[12] Securing AI agents requires rigorous governance of their NHIs—restricting access and monitoring permissions to prevent credential misuse and data exposure.[11]

Traditional AI governance practices—such as data governance, risk assessments, explainability, and continuous monitoring—remain essential, but governing agentic systems requires going further to address their autonomy and dynamic behavior.[4] For instance, some key issues in NHI management in agentic systems include controlling what actions identities can perform—such as sending or receiving data, allowed destinations, volumes, and formats—restricting access to external or sensitive resources, and continuously verifying any changes in an identity's permissions or software state since its last activity.[12] Centralized oversight enables rapid detection, isolation, and response to violations, reducing fragmentation and enhancing security posture. NHIs rely on tokens and certificates—unlike humans who use passwords or biometrics—requiring distinct management practices, such as credential rotation, access monitoring, and privilege adjustments to maintain secure authentication. Despite the critical importance of managing NHIs, fewer than 20% of organizations have formal processes to offboard and rotate API keys, leaving these identities exposed to prolonged security risks.[13]

> *AS AI AGENTS TAKE ON MORE AUTONOMOUS ROLES, ORGANIZATIONS MUST ALSO ESTABLISH GOVERNANCE STRUCTURES TO MANAGE THEIR BEHAVIOR AND OUTCOMES.*

Because these identities create serious vulnerabilities, enterprises are shifting to Zero Trust security models that emphasize verification and restricted access. But access control is just one piece of the puzzle. As AI agents take on more autonomous roles, organizations must also establish governance structures to manage their behavior and outcomes. This has prompted growing interest in controlled agency, a model that enforces accountability while allowing agents to act independently within defined limits.

SECURITY RISKS AND THE SHIFT TO ZERO TRUST

Traditional identity systems, such as OAuth and Security Assertion Markup Language (SAML) are widely adopted industry standards for secure identity and access management in web applications and services.[16] However, they were designed for static human users—not for the dynamic, autonomous workflows of AI agents. These agents often shift from human to nonhuman identities to perform tasks, requiring adaptive, fine-grained access controls that maintain security, accountability, and policy enforcement. OAuth and SAML, traditional frameworks for authentication and authorization, are thus ill-suited for AI agents. OAuth's token-based model, designed for static human user permissions, lacks the granularity needed for AI agents that require dynamic, context-driven access control. SAML, relying on static session-based authentication, also fails to meet AI agents' need for continuous, real-time privilege adjustments. Both systems assume trust once authenticated, but AI agents need ongoing validation due to evolving contexts and security risks. This highlights the need for dynamic identity management tailored to AI environments.[17]

Given these limitations, Zero Trust principles are essential to secure AI agents operating in dynamic and autonomous environments.[17] They focus on three key principles: explicitly verifying all requests, limiting access based on least privilege, and assuming breaches to minimize risk. This approach challenges the traditional belief that internal networks are safe, instead treating all requests as untrusted and requiring verification before granting access. It emphasizes continuous authentication, data protection, and proactive threat detection to safeguard enterprise systems, particularly as AI agents become more integrated.[18]

By enforcing continuous verification, AI agents must be authenticated and authorized in real time to ensure that only trusted entities access resources. Implementing least privilege access ensures agents receive only the minimal permissions needed to perform their tasks, helping to prevent privilege escalation. Micro-segmentation limits lateral movement within AI-driven environments, reducing the risk of compromised agents accessing unrelated systems. Additionally, anomaly detection and automated responses help to identify and mitigate unusual behavior, further strengthening security. Adopting these practices enhances the security of AI systems and reduces the potential for adversarial attacks and unauthorized access.[17]

Zero Trust–based solutions are being launched to address a growing concern: AI agents are increasingly accessing sensitive systems, creating new identity risks. In March 2025, Microsoft announced six new agentic Security Copilot solutions designed to autonomously manage high-volume security tasks. Built within a Zero Trust framework, these agents integrate with Microsoft Security tools, adapt to workflows, and enhance threat response and risk prioritization while keeping teams in full control.[19] In April 2025, identity security company CyberArk announced plans to integrate its Identity Security Platform with Accenture's AI Refinery to enforce Zero Trust protections tailored to AI agents—securing them as rigorously as human and machine users.[20]

CONTROLLED AGENCY AND HUMAN OVERSIGHT

Controlled agency in agentic automation refers to the combination of tools and practices that allow AI agents to operate independently while keeping their actions aligned with enterprise standards. These agents are given guardrails that preserve their creativity and problem-solving capabilities but ensure outcomes remain trustworthy.[21] By regulating their autonomy, accuracy, and resilience, controlled agency mitigates security risks and maintains human oversight. Rather than replacing entire workflows, AI agents function within a coordinated system that upholds data protection, trust, and alignment with organizational priorities.[6]

One example of a company promoting controlled agency is UiPath, which claims its next-generation platform for agentic automation enables secure collaboration between AI agents, robots, and people. At the core of its system is a controlled agency model, designed to ensure AI agents operate within strict guardrails that support security, predictability, and compliance. UiPath's Maestro orchestration layer provides centralized oversight and integrates process intelligence and key performance indicator monitoring to manage agents safely at scale. The platform also includes governance features, such as real-time vulnerability assessments and stringent data access controls to reinforce enterprise trust and accountability.[22]

It should, however, be noted that controlled agency is not, so far, an established or widely used term in mainstream academic or policy literature on AI governance; it does not appear in key frameworks from organizations, such as the National Institute of Standards and Technology, the Organization for Economic Cooperation and Development, and Human-Centered Artificial Intelligence, and seems to be a conceptual or marketing term used primarily in select industry sources like UiPath and Deloitte.

EMERGING TOOLS AND ORGANIZATIONAL GOVERNANCE PRACTICES

New approaches like adversarial testing and guardian agents are being used to monitor agentic AI, but they must align with organizational governance on ethics, security, and risk.[23] These strategies must be integrated with existing governance frameworks to ensure they effectively mitigate potential risks and maintain compliance with evolving regulatory standards.

Table 1 presents an overview of leading companies advancing agentic AI solutions, with a focus on their governance frameworks and core capabilities. It compares the approaches of SAS, NVIDIA, IBM, and Microsoft in ensuring responsible AI development—emphasizing human oversight, embedded governance, decision transparency, and integration with secure, scalable technologies. These company-specific strategies demonstrate how governance is being embedded into the infrastructure of agentic AI, ensuring that autonomy and adaptability are balanced with accountability, compliance, and ethical safeguards.

SAS

Analytics company SAS has extended its capabilities with a more flexible agentic AI framework that emphasizes human oversight, embedded governance, and decision explainability—built on its Viya platform, which supports auditability, bias detection, and compliance across AI processes.[24] In May 2025, SAS introduced Intelligent Decisioning within its Viya analytics platform to support responsible development and scaling of agentic AI. Announced at the SAS Innovate conference, the update also includes Viya Copilot, a preview of the synthetic data tool Data Maker (following its November 2024 acquisition of Hazy), and new cloud-based coding tools in Viya Workbench. Intelligent Decisioning integrates rule-based analytics with LLM reasoning, supports customizable human-AI collaboration, and includes built-in governance to ensure ethical and regulatory compliance.[25]

NVIDIA

NVIDIA's Neural Modules (NeMo) Guardrails enables developers to efficiently define and update behavioral rules for AI agents, supporting safe, consistent, and adaptable deployment in real-world settings. Major

TABLE 1. Key players supporting governance in agentic AI.

Company	Key capabilities and focus	Governance practices
SAS	SAS extends agentic AI with Intelligent Decisioning on its Viya platform, integrating rule-based analytics with LLM reasoning to enhance human-AI collaboration and scalability	Built-in governance to ensure ethical AI development, with a focus on auditability, bias detection, and regulatory compliance.
NVIDIA	NVIDIA provides NeMo Guardrails to support safe and adaptable deployment of AI agents and foundational tools like Blueprints for AI applications.	NeMo Guardrails filter content to ensure AI agents comply with policies, preventing unauthorized data processing even in case of prompt injection. Nutanix leverages NVIDIA tools, such as Blueprints and NeMo Guardrails, to ensure policy compliance in AI models.
IBM	IBM watsonx Orchestrate on AWS streamlines HR workflows, forecasts workforce needs, and provides personalized data-driven insights.	Integrates IBM watsonx.governance for explainable, auditable HR decisions, ensuring compliance with standards like GDPR and EEOC, and enhancing security and scalability through AWS.
Microsoft	Microsoft is supporting the development of agentic AI by combining Azure AI Foundry and Microsoft Fabric to offer tools for monitoring, governance, and continuous improvement.	Azure AI Studio tracks model performance, drift, and responsible AI metrics, while Fabric's integration with Purview ensures data governance and lineage. Microsoft also unifies operational, analytical, and real-time data under consistent security and governance via SQL databases in Fabric—featuring encryption, Private Link, and high availability by default.

GDPR: General Data Protection Regulation; AWS: Amazon Web Services.

> *WHILE THE POTENTIAL OF AGENTIC AI IS UNDENIABLE, OFFERING TRANSFORMATIVE ECONOMIC OPPORTUNITIES AND EFFICIENCIES, IT PRESENTS SIGNIFICANT CHALLENGES IN TERMS OF GOVERNANCE AND SECURITY.*

firms, including Amdocs, Cerence AI, and Palo Alto Networks are utilizing NeMo Guardrails to enhance governance and ensure more controlled, transparent, and reliable agentic AI interactions.[26] Nutanix, a cloud software company that offers a unified platform to run applications and manage data across on-premises datacenters, public clouds, and edge environments, leverages NVIDIA's AI tools to address key enterprise AI challenges. It uses NVIDIA Blueprints as foundational components for building AI applications and employs NeMo Guardrails to enforce content and policy compliance, ensuring that large language models do not process or generate unauthorized data—even in the event of prompt injection attempts.[27]

IBM

IBM's watsonx.governance is an automated governance tool which supports responsible AI deployment and regulatory compliance across the model lifecycle.[23] IBM watsonx Orchestrate on Amazon Web Services (AWS) enables organizations to use AI agents that streamline human resources (HR) workflows, forecast workforce demands, and boost employee engagement, while seamlessly integrating with existing systems to provide scalable and personalized, data-driven insights. watsonx Orchestrate emphasizes responsible AI practices for HR automation. It integrates watsonx.governance to support explainable and auditable decision making, particularly important in regulated environments. Running on AWS enhances security and compliance, while built-in governance aims to help meet standards like those of General Data Protection Regulation (GDPR) and the Equal Employment Opportunity Commission (EEOC). The system uses adaptive AI to adjust to workforce data and trends, distinguishing it from more static automation tools.[28]

MICROSOFT

Microsoft is supporting the development of agentic AI by combining Azure AI Foundry and Microsoft Fabric to offer tools for monitoring, governance, and continuous improvement. Azure AI Studio tracks model performance, drift, and responsible AI metrics, while Fabric's integration with Purview ensures data governance and lineage—helping organizations understand and manage how data influences agent behavior.[29]

At Microsoft Ignite 2024, which is Microsoft's annual flagship event, the company enhanced governance capabilities in agentic AI development by integrating operational SQL databases into Microsoft Fabric. This unification allows consistent security and governance policies across analytical, real-time, and operational data. The system is secure by default—featuring encryption, Private Link networking, and automated high availability—supporting trustworthy, policy-compliant AI agent deployments.[30]

While the potential of agentic AI is undeniable, offering transformative economic opportunities and efficiencies, it presents significant challenges in terms of governance and security. As organizations rapidly adopt these technologies, the focus on securing AI agents, managing nonhuman identities, and enforcing robust data governance frameworks becomes paramount. The shift toward Zero Trust security models and controlled agency highlights the need for continuous oversight, ensuring that AI systems operate within safe, transparent, and accountable boundaries. Regulatory initiatives, such as the EU AI Act and California's proposed standards, are also evolving to address these concerns, underscoring the importance of maintaining ethical, secure, and compliant practices. As AI agents continue to evolve, organizations must balance their autonomy with human oversight to mitigate risks and harness their full potential.

REFERENCES

1. B. Seipel, "Nvidia's Jensen Huang says AI agents are 'a multi-trillion-dollar opportunity' and 'the age of AI Agentics is here'," *Fortune*, Jan. 6, 2025. Accessed: May 15, 2025. [Online]. Available: https://fortune.com/2025/01/06/nvidias-jensen-huang-agentics-ai-robots-blackwell-gpu-ces-2025-toyota-autonomous-vehicles/
2. I. Barker, "Governance is top priority for agentic AI users," *BetaNews*, May 1, 2025. Accessed: May 15, 2025. [Online]. Available: https://betanews.com/2025/05/01/governance-is-top-priority-for-agentic-ai-users/
3. L. Wilkinson, "Enterprises have high ROI hopes for agentic AI," *CIO Dive*, Apr. 2, 2025. Accessed: May 10, 2025. [Online]. Available: https://www. ciodive.com/news/enterprise-agentic-AI-adoption-ROI-expectations-PagerDuty/744265/
4. C. Stryker. "AI agent governance: Big challenges, big opportunities." IBM, Apr. 2024. Accessed: May 15, 2025. [Online]. Available: https://www.ibm.com/think/insights/ai-agent-governance
5. H. Gentile, "How governance and security can drive agentic AI adoption," *CIO Dive*, Apr. 23, 2025. Accessed: May 15, 2025. [Online]. Available: https://www.ciodive.com/news/agentic-ai-governance-security/745161/
6. C. Mehin, "93% of IT execs are eager to implement agentic AI – But have they considered governance?" *Diginomica*, May 6, 2025. Accessed: May 15, 2025. [Online]. Available https://diginomica.com/it-execs-eager-implement-agentic-ai-considered-governance
7. J. Loucks, "Autonomous generative AI agents are coming: 4 ways to prepare," *Deloitte Insights*, May 10, 2025. [Online]. Accessed: May 15, 2025. Available https://deloitte.wsj.com/cmo/autonomous-generative-ai-agents-are-coming-4-ways-to-prepare-20b607b2
8. S. Evans, "Governance is top priority for companies using agentic AI: Survey," *AI Business*, May 8, 2025. Accessed: May 15, 2025. [Online]. Available: https://aibusiness.com/generative-ai/governance-is-top-priority-for-companies-using-agentic-ai-survey
9. C. T. Howell and A. J. Liederman, "The intersection of agentic AI and emerging legal frameworks." Foley & Lardner LLP, Dec. 19, 2024. Accessed: May 15, 2025. [Online]. Available: https://www.foley.com/insights/publications/2024/12/intersection-agentic-ai-emerging-legal-frameworks/
10. M. C. Borrelli and S. Musch, "How to use agentic AI in line with the EU AI Act," *CX Network*, Feb. 11, 2025. Accessed: May 15, 2025. [Online]. Available: https://www.cxnetwork.com/artificial-intelligence/articles/how-to-use-agentic-ai-in-line-with-the-eu-ai-act
11. "The identities behind AI agents: A deep dive into AI & NHI," *The Hacker News*, Apr. 10, 2025. Accessed: May 15, 2025. [Online]. Available: https://thehackernews.com/2025/04/the-identities-behind-ai-agents-deep.html
12. D. Tait, "Non-human identities gain momentum, requires both management, security," *Dark Reading*, Dec. 24, 2024. Accessed: May 13, 2025. [Online]. Available: https://www.darkreading.com/cybersecurity-operations/non-human-identities-gain-momentum-requires-both-management-security
13. M. Zorz, "Securing non-human identities: Why fragmented strategies fail," *Help Net Security*, Sep. 25, 2024. [Online]. Accessed: May 15, 2025. Available:

https://www.helpnetsecurity.com/2024/09/25/john-yeoh-csa-nhi-security/

14. "Securing non-human identities in the age of AI agents | CSA Summit 2025 at RSAC." Cloud Security Alliance, Apr. 29, 2025. Accessed: May 15, 2025. [Online]. Available https://cloudsecurityalliance.org/artifacts/securing-non-human-identities-in-the-age-of-ai-agents-rsac-2025
15. "Explosive growth of non-human identities creating massive security blind spots," *The Hacker News*, Apr. 9, 2025. Accessed: May 15, 2025. [Online]. Available: https://thehackernews.com/2025/04/explosive-growth-of-non-human.html
16. R. Bhargava, "SAML vs OAuth: What's the difference?" *Auth Thoughts*, Apr. 20, 2023. Accessed: May 15, 2025. [Online]. Available: https://www.descope.com/blog/post/saml-vs-oauth
17. K. Huang, "Agentic AI identity management approach," *Cloud Security Alliance*, Mar. 11, 2025. Accessed: May 9, 2025. [Online]. Available: https://cloudsecurityalliance.org/blog/2025/03/11/agentic-ai-identity-management-approach
18. "What is zero trust?" *Learn Microsoft*, Feb. 27, 2025. Accessed: May 10, 2025. [Online]. Available: https://learn.microsoft.com/en-us/security/zero-trust/zero-trust-overview
19. V. Jakkal, "Microsoft unveils Microsoft Security Copilot agents and new protections for AI," *Microsoft Security Blog*, Mar. 24, 2025. Accessed: May 15, 2025. [Online]. Available: https://www.microsoft.com/en-us/security/blog/2025/03/24/microsoft-unveils-microsoft-security-copilot-agents-and-new-protections-for-ai/
20. S. C. Bhasin, "CyberArk and Accenture partner to secure enterprise AI agents," *MSSP Alert*, Apr. 10, 2025. [Online]. Accessed: May 15, 2025. Available: https://www.msspalert.com/news/cyberark-and-accenture-partner-to-secure-enterprise-ai-agents
21. Y. Broustas, "Agentic automation: The path to an orchestrated enterprise," *UiPath*, Dec. 11, 2024. Accessed: May 15, 2025. [Online]. Available: https://www.uipath.com/blog/ai/agentic-automation-path-to-orchestrated-enterprise
22. G. Peters, "UiPath launches the 'first enterprise-grade platform' for agentic automation," *iTWire*, May 1, 2025. Accessed: May 15, 2025. [Online]. Available https://itwire.com/it-industry-news/strategy/uipath-launches-the-%e2%80% 98first-enterprise-grade-platform%e2%80%99-for-agentic-automation.html
23. P. Boinodiris and J. Parker, "The evolving ethics and governance landscape of agentic AI," *IBM Think*, Mar. 21, 2025. Accessed: May 15, 2025. [Online]. Available: https://www.ibm.com/think/insights/ethics-governance-agentic-a
24. G. Lawton, "SAS extends business rules governance experience to agentic AI. Here's why," *Diginomica*, May 8, 2025. Accessed: May 15, 2025. [Online]. Available https://diginomica.com/sas-extends-business-rules-governance-experience-agentic-ai-heres-why
25. E. Avidon, "SAS update targets responsible agentic AI development," *TechTarget*, May 7, 2025. Accessed: May 15, 2025. [Online]. Available https://www.techtarget.com/searchbusinessanalytics/news/366623773/SAS-update-targets-responsible-agentic-AI-development
26. D. Reber Jr., "How agentic AI enables the next leap in cybersecurity," *NVIDIA Blog*, Apr. 28, 2025. Accessed: May 15, 2025. [Online]. Available: https://blogs.nvidia.com/blog/agentic-ai-cybersecurity/
27. E. van Klinken, "New version of Nutanix Enterprise AI makes agentic AI manageable," *Techzine*, May 7, 2025. Accessed: May 15, 2025. [Online]. Available https://www.techzine.eu/news/infrastructure/131190/new-version-of-nutanix-enterprise-ai-makes-agentic-ai-manageable/
28. K. Sachdeva, "How watsonx Orchestrate on AWS delivers agentic AI to transform Enterprise HR," *IBM Blog*, Mar. 14, 2025. Accessed: May 15, 2025. [Online]. Available: https://www.ibm.com/products/blog/watsonx-orchestrate-aws-delivers-agentic-ai-transform-enterprise-hr
29. U. T. Malaviarachchi, "Empowering agentic AI: The symbiotic power of Microsoft fabric and Azure AI Foundry," *Dev.to*, May 10, 2025. Accessed: May 15, 2025. [Online]. Available: https://dev.to/umeshtharukaofficial/empowering-agentic-ai-the-symbiotic-power-of-microsoft-fabric-and-azure-ai-foundry-1hdb
30. W. McKelvey, "The art of simplifying the complex: Microsoft Fabric's superpower," *Microsoft Fabric Blog*, Feb. 24, 2025. Accessed: May 15, 2025. [Online]. Available: https://www.microsoft.com/en-us/microsoft-fabric/blog/2025/02/24/the-art-of-simplifying-the-complex-microsoft-fabrics-superpower/

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

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

IEEE

revised 2 September 2025

IEEE CS EXECUTIVE COMMITTEE

IEEE CS BOARD OF GOVERNORS

IEEE EXECUTIVE STAFF

IEEE OFFICERS

DEPARTMENT: NOTES FROM THE FIELD

This article originally appeared in
Computer
vol. 58, no. 6, 2025

General and Agentic AI, and the Challenges of Xplainable 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 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

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

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]

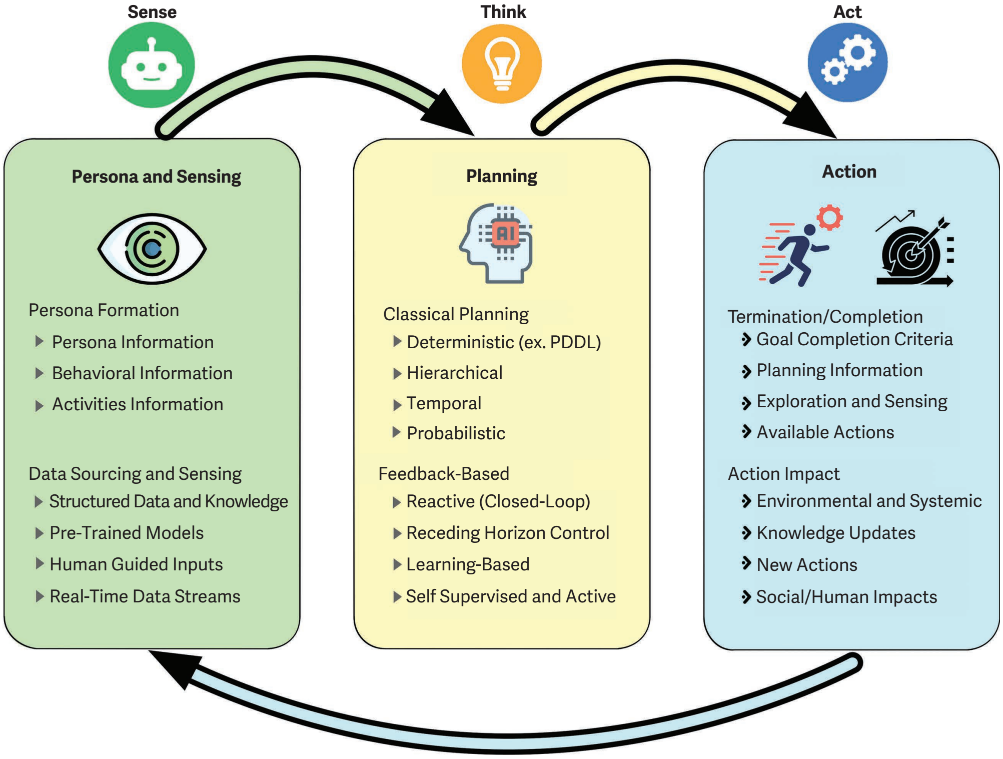


FIGURE 1. Agentic AI Sense, Think, Act lifecycle.

- *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 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

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

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.

EDITORS: Beatriz Sousa Santos, bss@ua.pt
Alejandra J. Magana, admagana@purdue.edu
Rafael Bidarra, R.Bidarra@tudelft.nl

This article originally appeared in IEEE Computer Graphics and Applications vol. 44, no. 5, 2024

DEPARTMENT: EDUCATION

Quantum Computing for All: Online Courses Built Around an Interactive Visual Quantum Circuit Simulator

Juha Reinikainen, Vlad Stirbu, Teiko Heinosaari, Vesa Lappalainen, and Tommi Mikkonen,
University of Jyväskylä, 40014, Jyväskylä, Finland

Quantum computing is a highly abstract scientific discipline, which, however, is expected to have great practical relevance in future information technology. This forces educators to seek new methods to teach quantum computing for students with diverse backgrounds and with no prior knowledge of quantum physics. We have developed an online course built around an interactive quantum circuit simulator designed to enable easy creation and maintenance of course material with ranging difficulty. The immediate feedback and automatically evaluated tasks lower the entry barrier to quantum computing for all students, regardless of their background.

Quantum computing is a rapidly developing field that has the potential to transform the way we approach complex computational problems. By harnessing the principles of quantum mechanics, quantum computers may be able to solve problems that are currently unsolvable or require an impractical amount of time to solve using classical computers.

Quantum computing is a highly mathematical and abstract field that requires a strong foundation in physics, computer science, and mathematics. The concepts are often difficult to grasp, especially for students who have not had prior exposure to quantum mechanics or linear algebra. Therefore, it is often seen as a purely theoretical subject, and it can be challenging to relate the concepts to real-world applications or to explain how they can be used to solve practical problems.

Since quantum computing is becoming a practical field, it requires students to gain hands-on experience with both quantum software and hardware. However, the learning curve for quantum computing should be smoother and not require deep programming expertise from the start. Further, even if there are already some quantum computers available for online access, the computing time may be limited, especially for students who are not enrolled in specialized programs. These reasons have motivated the University of Jyväskylä to develop an interactive quantum circuit simulator as part of the existing learning platform.

The rest of this article is organized as follows, aligning with the phases of the Design Science Research methodology.[9] The "Motivation" section introduces the problem identification and motivation, followed by the "Learning Environment" section in which the solution is integrated. Then, the "Artifact Design" section presents the quantum simulator implementation, and the "Demonstration" section describes how it is used to conduct three specific tasks within the educational process. Finally, the "Discussion" section discusses how the visual editor meets the teaching goals for the target student audience, and we finish by exploring the feedback received from the first group of users that completed the course in which the system was used.

Digital Object Identifier 10.1109/MCG.2024.3444952
Date of current version 25 October 2024.

TABLE 1. Student personas, their knowledge, expertise, and expected results from using the educational system to learn quantum computing (QC).

Student persona	Background	Expectations
Physics	Knows the basics of quantum mechanics	QC is not only a paper exercise
Software engineer	Knows programming and development processes	Gets familiar with the theoretical background of the field
Business	Knows how to use technology in commercial context	Understands the challenges of QC

MOTIVATION

Quantum computing is an interdisciplinary topic that cannot be put under a single scientific field. For that reason, also students from various study tracks are eager to learn about quantum computing. However, their different background and prerequisite knowledge poses an obvious challenge. A physics student knows about quantum mechanics, but may not be fluent in programming. A computer science student knows programming, but has not the needed theoretical background of quantum physics. Further, a third case is, e.g., a student in business school, who understands how technology relates to business, and would hence keen to know the basics of quantum computing. These student personas, their educational background and expertise, as well as their expectations are summarized in Table 1. When developing the new teaching material, the different personas have served as imaginary student profiles and helped us to take into account their diverse viewpoints.

To lower the entry barrier for both the students and teaching staff, we decided to extend the learning platform, the interactive material (TIM), for teaching also the quantum computing course. As the TIM platform is used at the university for various courses, especially programming courses, the staff and major part of the students are already familiar with it. The TIM platform needed an extension to be able to have interactive quantum circuit exercises and examples. The interactive quantum circuit simulator is, in fact, the most important part of the online course. It can also be used as a stand-alone tool when demonstrating quantum algorithms or other quantum information protocols in various courses or events.

LEARNING ENVIRONMENT

TIM[8] is a massive open online course (MOOC) learning environment. The development of TIM began in 2014 when there was no ready-made tool suitable for programming courses that would make it easy to produce and maintain long interactive book-like learning materials. TIM has been an open-source application under the MIT license from the beginning.[a] The development of TIM is the responsibility of the staff at the Information Technology faculty. A large part of TIM has been developed as theses in computer science education.

The basic idea of TIM is document based. Everything done in TIM is essentially documents. Documents consist of blocks. A block can be one or more normal text paragraphs or an interactive element. An interactive element can be, for example, a multiple-choice task, a drawing task, a free text field, a programming code task, an image, a series of images, a video, a mathematically handwritten task returned as an image, an assisted LaTeX-written answer, or a new component. The appearance of TIM is adjusted with CSS styles, so users can be offered ready-made appearance styles or users can create their own style sheet. Teachers can choose the appearance from ready-made styles or create their own style for their course.

TIM can be used to create lecture materials according to the original plan, but in the 10 years, TIM has developed so that it can handle all the necessary course bookkeeping from individual tasks done in TIM to the final grade. The instructor can see all the attempts the student has made on a specific task and can use it to try to figure out why the student may initially go in the wrong direction. The instructor can write feedback on the student's answers. This way, TIM provides a lot of data for learning analytics. Students can comment on the material and the instructor can respond to the comments. In fact, this is currently the most common way to communicate with students using TIM. The information model that supports this interaction is presented in Figure 1.

TIM documents can be used to borrow blocks from one document to another, and thus with a single maintenance, smaller documents suitable for different needs can be made from one document, or vice versa, smaller documents can be combined into cohesive entities. Thanks to macros, parts that depend on years or even tool version numbers can be combined into one place in TIM, thus speeding

[a] [Online]. Available: https://github.com/TIM-JYU/TIM

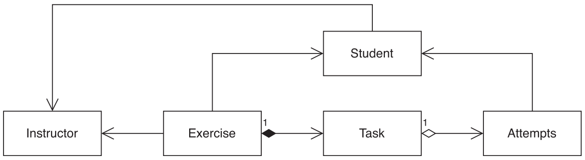
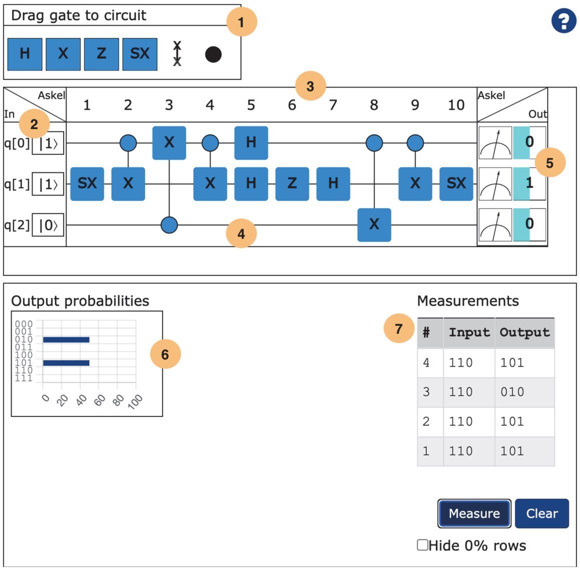


FIGURE 1. Information model supporting instructor–student interactions.



#	Input	Output
4	110	101
3	110	010
2	110	101
1	110	101

FIGURE 2. Quantum circuit simulator interface tour: (1) available gates toolbar, (2) input qubits and values, (3) maximum circuit steps, (4) circuit layout grid, (5) qubit output measurements—value and probability, (6) results probability distribution, and (7) individual execution results table—inputs and output measurements.

up maintenance as things change. The TIM document can be translated into several different languages using the automatic translation provided by DeepL.[b] Naturally, the human must check the result produced by the translator. If changes are made to the original document, the translation administrator will be notified and new blocks or a note of the changed section will be added to the translated document.

When using TIM, the teacher and student can do everything in one system without the need to switch from one system to another. TIM is suitable and has been used in all kinds of teaching methods from classroom teaching to MOOC courses. TIM is at its best in teaching where there are performers in many different ways. Conditional blocks allow different groups to be offered slightly different instructions if necessary.

ARTIFACT DESIGN

We implemented the quantum circuit simulator in the TIM platform, so that it can be directly used for developing the quantum technology curriculum.

At the start of the project, TIM already had all the essential features for lecture-oriented course as well as for a MOOC. Our purpose was to design and build a tool that is functional in both types of courses. The quantum circuit simulator is implemented as a TIM plugin, and it can be customized to create different types of exercises. The teacher has options to customize the different components to be visible or not according to the context it is used in.

The simulator is usable by different kind of demographics. Using the simulator does not require neither physics knowledge nor programming knowledge, such as Qiskit.[10] On the other hand, the simulator allows the user to export the circuit to the Qiskit equivalent, or the final quantum state and output probabilities for further analysis if needed.

Circuit Simulator Interface

The inspiration for our quantum circuit simulator was drawn from the existing graphical quantum circuit simulators, such as IBM Quantum Composer,[1] Quantum Länd Circuit Simulator,[2] Quantum JavaScript,[3] Quirck,[4] or Uranium.[5] One of the design goals of our circuit simulator was adaptability so that it could be modified to fit different types of exercises *within* the TIM platform. For this reason, we chose to create a new simulator that incorporates the relevant visual elements for designing circuits and visualize the results present in existing simulators.

The circuit simulator contains an *editor* area that allows the user to construct the circuit—toolbar with available gates, input qubits names and values (e.g., $|0\rangle$ or $|1\rangle$), maximum number of circuit steps, circuit layout grid, and qubit output measurements with probability distribution, and a *results* area for visualizing the simulation results—probability distribution, or individual results table. The annotated quantum circuit simulator user interface is depicted in Figure 2.

The editor functionality of the simulator allows the user to construct the circuit from the gates available in the toolbar, by dragging and dropping gates in the circuit layout grid. The controlled gates can be created

[b] [Online]. Available: https://www.deepl.com/translator

by dragging control nodes to same step as target gate. The controls are connected by vertical wires to the target gate, while the anticontrols—e.g., controls that are activated with zero instead of one—are represented with unfilled circles. The multiqubit gates are represented as rectangles covering adjacent lines. The default configuration of the simulator contains a set of common gates, which can be extended with own custom gates.

The size of the circuit layout grid is determined by input qubit count for rows and step count for columns, and is immutable. This way user knows exactly how many qubits and steps the answer consists of, which can help to narrow down their choices on how to implement the desired circuit. Even on simple exercises, there are many possible ways to arrange the available gates in the circuit so this predetermined circuit size might make the exercises more manageable for the user.

The user can edit the circuit and see the results in real time, allowing them to interactively experiment with the behavior of gates to understand how quantum computation works. The progression of the computation is represented as discrete steps going from left to right as gates modify the quantum state. The measurement is done on all qubits at the end in the $|0\rangle$, $|1\rangle$ basis.

The input state of circuit is represented using bits that user can toggle between 0 and 1. Being able to change input bit values makes it possible to explore the behavior of the circuit especially when working with simple gates, such as X and CX. The instructor can decide whether to use braket or bit notation for inputs and give initial values and names for the qubits. The values of the qubits can be locked to a specific value if changing the value is not needed in the exercise.

The graphical presentation of the circuit is similar to what is used in the literature.[6,7] For example, the Control-X gate is typically represented as $\oplus$. Since our simulator allows to add a control dot to any quantum gate, we chose to use a more unified notation, see Figure 3. First, every gate is represented by a rectangle with its name inside it. A control dot, or several of them, can be dragged and connected to the gate. The only exception to this rule is the swap gate, which is represented by two X connected by wire.

The results part of the simulator interface provides several widgets that allow the visualization of the quantum computation. The bar chart widget shows the probability distribution of the output states. For larger circuits, rows where output probability is zero can be hidden from the chart to make the results easier to understand. The measurements widget allows the user to perform shot measurements and to visualize the sampled output value in a table containing the measurement number, and the corresponding input and output.

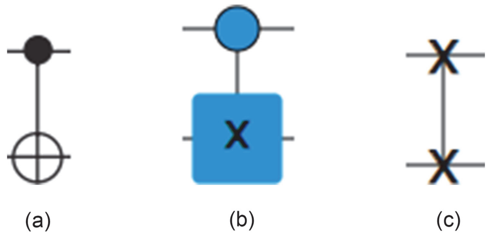


FIGURE 3. Example gate representations in the visual simulator's toolbar. (a) Common notation for a CX-gate. (b) CX-gate using our unified notation. (c) SWAP gate.

The simulator computes the full state vector allowing for statistical sampling of the outputs. Samples can be drawn from the theoretical output probability distribution computed from the final state vector. The output probabilities shown in the chart can be these exact probabilities or the probabilities can be based on automatic sampling of certain sample size or from the measurements made by the user.

The visibility of the elements of the circuit simulator interface is configurable, therefore the user can first be shown the basic version and then additional features can be gradually introduced making the simulator more approachable for new users.

System Architecture

The quantum circuit simulator implementation consists of a client component and server component. The client component, written in JavaScript, using Angular[c] framework, integrates the circuit visualization into the TIM exercise user interface—executed into the web browser main thread, and the client-side simulator that integrates into the TIM application engine—executed in a web worker. The server component implements the server-side circuit simulator, using the Qulacs library,[11] and the exercise assessment. The interaction between the client-side and server-side simulators is realized via a REST interface. The decomposition of the system is depicted in Figure 4.

When opening a new task in the exercise, the user interface renders the circuit editor and

[c] [Online]. Available: https://angular.io

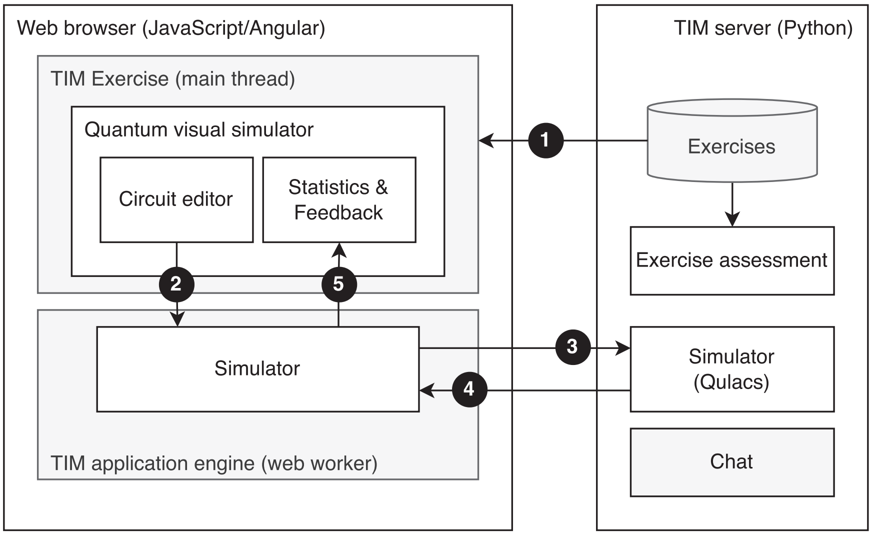


FIGURE 4. Visual quantum simulator architecture overview (existing TIM components in gray), and interaction between the components: (1) loading the exercise, (2) initialize client-side simulator, (3) initialize server-side simulator if circuit is large, (4) return the results, and (5) render the results.

initializes the client-side simulator. If the circuit is too large to be handled locally, the server-side simulator is initialized. After initialization, the circuit is executed on the simulator and the results are collected, send to the browser if necessary, then rendered in the user interface. A new execution process is performed whenever the user modifies the circuit or changes the input values, allowing the user interface to be always up to date. To maintain the responsiveness of the user interface, the client-side simulator is implemented in a web worker. Therefore, even under intense load, the browser main thread is not blocked.

When there are a lot of simulation exercises on one web page, the load time increases making it quite slow especially on slower devices, such as tablets. This was mitigated by lazy loading the simulator components. The simulator is fully rendered and the simulation computation is done when user clicks on the exercise.

The exercises can be automatically checked for correctness by specifying a model circuit to compare the user's circuit to. The circuit equality is checked by iterating over all 2^{nQubits} input combinations and checking that both circuit result in same probability vector. The number of checked inputs can be reduced by giving the regular expressions patterns to match each input bit-strings against, and only running the simulation against these matching bit-string inputs.

MOOC Integration

To support the teaching activities, the visual quantum circuit simulator is seamlessly integrated into TIM's MOOC environment. Each task instance is visualized by an instance of the circuit simulator. When creating a task, the instructor can control the appearance and behavior of the simulator using a YAML-based configuration, depicted in Listing 1. The configuration options can be grouped into the following categories: the task description (lines 1 and 2), the appearance of the simulator's user interface elements (lines 4–10), the circuit properties—number of qubits, circuit steps, qubits initial state and editability, initial and target circuit state, and special gate conditions (lines 12–35), and the feedback provided to the student (lines 37–40). The text-based format of the configuration enables the instructor to reuse and adapt circuits for different learning tasks, lowering the effort to create or maintain the course material.

The student attempt to solve a task is saved on the server. This allows the instructor to assess the progress for every student. The instructor can engage in a conversation with the student using the TIM's chat function, both sharing the same attempt to solve a task context. As the visual simulator is implemented as a TIM plugin, it leverages the interactive capabilities of the MOOC platform to provide the teaching experience that is not possible with individual study tools, such as IBM Quantum Composer.

```
header: "Tehtävä"
stem: "Käytä kahta CX-porttia ja etsi niille
↪   sellainen muodostelma, että ensimmäisen bitin
↪   arvo q[0] kopioituu kahdelle seuraavalle
↪   bitille kun nämä ovat alkutilassa 0. Kun olet
↪   saanut piirin valmiiksi, paina 'Tallenna' ja
↪   ratkaisusi tarkistetaan."

qubitNotation: "bit"
showChart: false
showOutputBits: true
middleAxisLabel: "Askel"
leftAxisLabel: "In"
rightAxisLabel: "Out"

nQubits: 2
nMoments: 4
gates: ["X","control"]
samplingMode: matrix
qubits:
  - value: 0
    editable: true
  - value: 0
    editable: false
modelCircuit:
  - controls:
      - 0
    editable: true
    name: X
    target: 1
    time: 0
  - controls:
      - 1
    editable: true
    name: X
    target: 2
    time: 1
modelConditions: ["C1X == 2"]
  - pointsRule: # tämän alle säännöt miten pisteitä
  ↪   saa
      multiplier: 5.0 # Millä luvulla kerrotaan
      ↪   tehtävästä saadut pisteet

feedbackText:
    correct: "Oikein"
    conditionWrong: "Väärin. Vastauksessa pitää
    ↪   olla kaksi CX-porttia."
feedbackShowTable: false
```

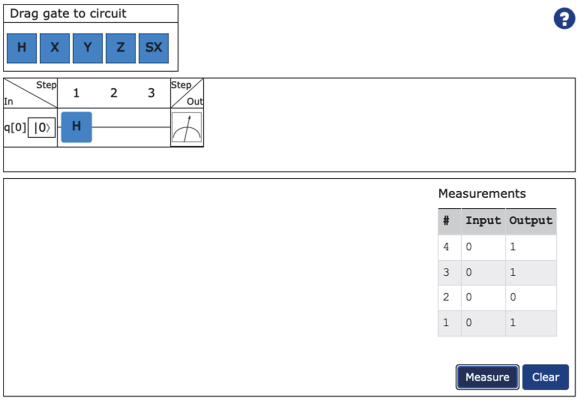
Listing 1. YAML representation of a circuit in TIM, with localization properties in Finnish.

DEMONSTRATION

The quantum circuit simulator is a flexible tool to construct various exercises. In the following, we demonstrate three different kind of exercises that are used in our online courses.

Understanding the Probabilistic Nature of Quantum Gates

Quantum gates, such as Hadamard gates and square-root-X, do not lead to determined measurement outcomes. They are the simplest examples of nonclassical building blocks of quantum circuits and for that reason natural first elements in a quantum computing course. As one of the first demonstrations of the difference of standard gate-based computation and quantum gate computation, a student is given an exercise where she needs to find what gates lead to deterministic outcomes. In that kind of exercise, the simulator is given with the settings as in Figure 5. In that setting, the circuit gives individual measurement outcomes and the nondeterministic nature of quantum gates is transparent. Pressing "measure" several times and getting different outcomes is very concrete and better for this task than a static visualization of a probability distribution. Furthermore, both Hadamard and square-root-X gates offer intriguing demonstration of quantum randomness as individually used they lead to random outcomes, but two similar gates to deterministic outcomes.



Measurements

#	Input	Output
4	0	1
3	0	1
2	0	0
1	0	1

FIGURE 5. Task 1—understanding probabilistic gates.

Working With Quantum Circuits

The action of a single quantum gate is simple. To achieve certain task, one may have to use a circuit composed of several gates. Their collective performance may be hard to infer. The circuit simulator offers a fantastic workaround as the student is encouraged to experiment with gate constellations. The simulator gives an instant visualization of the output in the form of measurement outcome distribution.

A real quantum computer has certain native gates and all the other gates must be build up from those. For that reason, it is important to learn to recognize circuits with equivalent performance. The simulator enables to give complicated gate decompositions (see Figure 6), and the task for a student is to find an equivalent circuit with less gates. In the depicted case, one can replace the whole circuit by just one controlled X-gate. The simulator has an automatic system to check answers, and in this way, a student gets an immediate feedback if her trial is the correct one. A typical exercise does not have a unique answer, but the automatic evaluator is capable of recognizing all correct answers from their input–output behavior. It

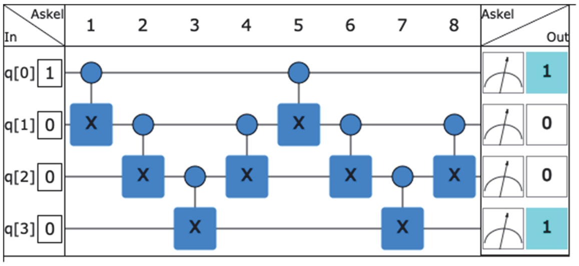


FIGURE 6. Task 2—working with circuits.

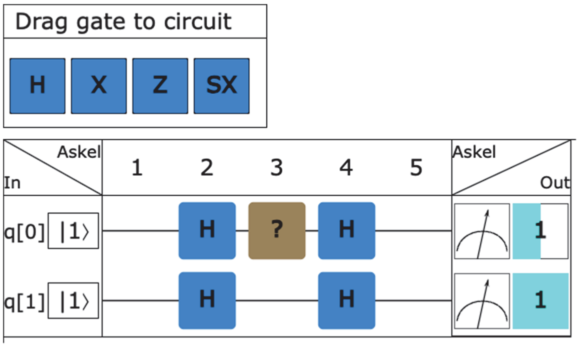


FIGURE 7. Task 3—identifying unknown quantum gate.

is, in fact, common that students find novel solutions that a teacher was not able to foresee.

Identifying an Unknown Quantum Gate

Once a student has reached some familiarity of quantum gates, she can proceed to more interesting exercises. With the simulator, the teacher can create arbitrary gates and label them in any desired way. This makes it possible to form exercises where a student has to identify an unknown gate, or their combination, by using it in concatenation with other gates (see Figure 7). These kind of exercises are almost like entertaining puzzles. As the gate can be any of the previously introduced gates, a student necessarily has to review previously learned material. Some quantum gates, such as the Z-gate, do nothing in the computational basis unless they are coupled with some other, appropriately chosen, quantum gates. This quantum feature makes additional excitement to suitably planned exercises.

DISCUSSION

Positioning Among Development Tools

The spectrum of experiences that can be used by quantum technology practitioners (excluding hardware), can be divided into the following areas, as depicted in Figure 8: paper, paper-like, notebook, and integrated development environment (IDE). The paper is used mainly by practitioners with physics and mathematical background and does not necessitate computer systems for experimenting besides paper or a whiteboard. The paper-like provides a low threshold for using a computer system that offers an experience close to the paper, but with instant feedback aided by tools, such as TIM or IBM Quantum Composer. The notebook experience is familiar to scientists who are using Python programming language and quantum development toolkits (e.g., Qiskit, Cirq[d] or Pennylane[e]), and use tools, such as Jupyter and IBM Quantum Lab, to develop algorithms and perform experiments. The IDE experience is tailored for software developers who use advanced software development tools, such as Visual Studio Code[f] or PyCharm.[g]

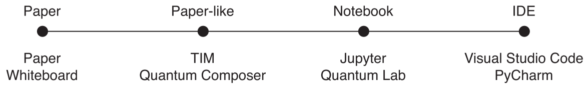


FIGURE 8. Spectrum of tools that can be used by quantum information and computing practitioners.

As the visual appearance of the TIM circuit simulator was inspired by the IBM Quantum Composer, we can say that their core functionality is similar, allowing the users to design and execute quantum circuits using metaphors specific to graphic user interface, and are able to explore their essential characteristics with specialized widgets. However, as the purpose of each tool varies—TIM is a tool optimized to facilitate learning in a formal education environment, while Quantum Composer is a commercial tool designed to smooth the onboard process of developers into IBM's quantum platform—we expect the latter to add new visual programming capabilities and integrations with advanced simulators and actual quantum hardware at a much higher pace. Nevertheless, by being familiarized with the same metaphors and core capabilities, TIM's users will be able to easy transition to more advanced commercial tools.

User Feedback

Since February 2024, we have been running the first MOOC for a group of 60 early adopters (see the

[d] [Online]. Available: https://quantumai.google/cirq
[e] [Online]. Available: https://pennylane.ai
[f] [Online]. Available: https://code.visualstudio.com
[g] [Online]. Available: https://www.jetbrains.com/pycharm/

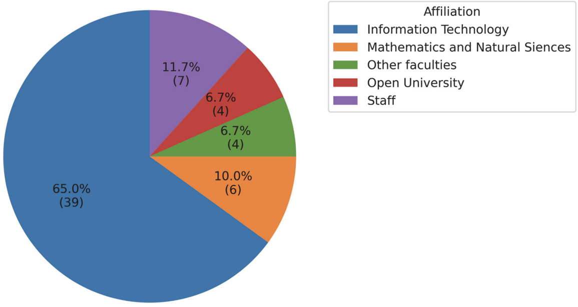


FIGURE 9. Participants breakdown by department affiliation.

details in Figure 9). The Quantum Computing Essentials course is the first in a series of three courses planned at the University of Jyväskylä. The language used is Finnish, and participants receive two study credits upon completion. The largest group is formed of students from the Information Technology, and Mathematics and Natural Sciences faculties that study quantum computing as part of their curriculum. Another large segment was formed by the university staff, followed by students from faculties that do not mandate the study of quantum computing, and participants from the open university. The study group covers two from the three student personas described earlier (e.g., physicist and software engineer), allowing us to draw some preliminary observations.

Given the diverse backgrounds of the students for whom the MOOC is designed, we anticipated that the difficulty of tasks would vary among students. In order to prevent any task from becoming a significant bottleneck, we included additional hints in the course material that effectively served their purpose. To measure the difficulty of a task, we have used the average number of attempts made by the student before reaching the correct answer. Since the number of attempts does not impact the course grade, students are motivated by the opportunity to tackle the task challenges without the pressure of exam failure or low grades. In this regard, the MOOC format is functioning as intended, fostering a supportive environment for learning.

We recall that the simulator's primary advantage over similar products is its seamless integration with the existing TIM learning platform. It's not confined to a single course but can be utilized across various courses, allowing for the easy generation of exercises with diverse difficulty levels. The integration with TIM provides access to all necessary tools for a university course, including discussion forums, straightforward registration, and grading. The value of these features has been demonstrated, and the simulator has fulfilled its intended purpose. Among our target student groups, the paper-like experience provided by the TIM's quantum circuit simulator offers the optimal experience, as it does not require familiarity neither with programming languages, nor with mathematical concepts, instead leading students to experiment with circuits by drag-and-dropping gates and instantly see the results.

Very recently, the MOOC was started in the open university, where it is offered for free. This will be a real-world test of our primary objective: making quantum computing accessible to everyone.

ACKNOWLEDGMENTS

This work was supported by Business Finland Project TORQS 8582/31/2022.

REFERENCES

1. IBM quantum composer. 2024. Accessed: Feb. 22, 2024. [Online]. Available: https://quantum.ibm.com/composer/files/new
2. Quantum circuit simulator—The Quantum Länd. 2024. Accessed: Feb. 22, 2024. [Online]. Available: https://thequantumlaend.de/quantum-circuit-designer/
3. Quantum JavaScript (q.js). 2024. Accessed: Feb. 22, 2024. [Online]. Available: https://github.com/stewdio/q.js
4. Quirk: Quantum circuit simulator. 2024. Accessed: Feb. 22, 2024. [Online]. Available: https://algassert.com/quirk
5. Uranium circuit-editor. 2024. Accessed: Feb. 22, 2024. https://uranium.transilvania-quantum.org/circuit-editor/
6. K. Chen et al., "VeriQBench: A benchmark for multiple types of quantum circuits," 2022, *arXiv:2206.10880*.
7. Z.-Y. Chen, Q. Zhou, C. Xue, X. Yang, G.-C. Guo, and G.-P. Guo, "64-qubit quantum circuit simulation," *Sci. Bull.*, vol. 63, no. 15, pp. 964–971, 2018.
8. V. Isomöttönen, A.-J. Lakanen, and V. Lappalainen, "Less is more! preliminary evaluation of multi-functional document-based online learning environment," in *2019 IEEE Front. Educ. Conf.*, 2019, pp. 1–5.
9. M. A. Rothenberger, K. Peffers, T. Tuunanen, and S. Chatterjee, "A design science research methodology for information systems research," *J. Manage. Inf. Syst.*, vol. 24, no. 3, pp. 45–77, 2007.

10. A. Javadi-Abhari et al., "Quantum computing with Qiskit," 2024, *arXiv:2405.08810*.
11. Y. Suzuki et al., "Qulacs: A fast and versatile quantum circuit simulator for research purpose," *Quantum*, vol. 5, 2021, Art. no. 559.

JUHA REINIKAINEN is currently working toward the master's degree in mathematical information technology with the University of Jyväskylä, 40014, Jyväskylä, Finland. His research interests include quantum computing, simulation, and software development. Contact him at juha.a.reinikainen@jyu.fi.

VLAD STIRBU is currently a postdoctoral researcher at the University of Jyväskylä, 40014, Jyväskylä, Finland. His research interests include software engineering, software architecture, quantum software, and software development in regulated industries. He is a member of IEEE Computer Society. He is the corresponding author of this article. Contact him at vlad.a.stirbu@jyu.fi.

TEIKO HEINOSAARI is currently a professor of quantum computing at the University of Jyväskylä, 40014, Jyväskylä, Finland. His research interests include quantum information theory and hybrid classical–quantum computing. He is developing novel ways to teach quantum computing to students who have no prior knowledge of quantum physics. Contact him at teiko.heinosaari@jyu.fi.

VESA LAPPALAINEN is currently a senior lecturer at the University of Jyväskylä, 40014, Jyväskylä, Finland, with more than 40 years of teaching experience. He is the creator of the TIM-platform. His current research focuses on e-learning. Contact him at vesa.t.lappalainen@jyu.fi.

TOMMI MIKKONEN is currently a professor of software engineering at the University of Jyväskylä, 40014, Jyväskylä, Finland. Contact him at tommi.j.mikkonen@jyu.fi.

Contact department editor Beatriz Sousa Santos at bss@ua.pt or department editor Alejandra J. Magana at admagana@purdue.edu or department editor Rafael Bidarra at R.Bidarra@tudelft.nl.

IEEE COMPUTER SOCIETY
Call for Papers
Enhance the credibility and prestige of your research by publishing with a globally recognized and respected organization.
GET PUBLISHED
www.computer.org/cfp
IEEE COMPUTER SOCIETY
IEEE

DEPARTMENT: PERSPECTIVES

This article originally appeared in
Computer
vol. 56, no. 8, 2023

Post-Quantum Adversarial Modeling: A User's Perspective

Teik Guan Tan and Jianying Zhou, *Singapore University of Technology and Design*
Vishal Sharma, *Queen's University Belfast*
Saraju P. Mohanty, *University of North Texas*

What are adversaries doing to prepare for the impending arrival of quantum computers? We take an adversarial modeling approach to identify various personas and flesh out passive and active actions such adversaries can take to gain an advantage, before providing recommendations on suitable mitigation strategies.

Quantum computers are coming, and they make computer security systems vulnerable to cryptanalysis. Shor's algorithm[1] running on a quantum computer can break asymmetric key cryptosystems such as Rivest-Shamir-Adleman (RSA) and elliptic curve cryptography (ECC). Additionally, Grover's algorithm[2] will weaken symmetric key encryption, hashing, and password systems to half their designed security strength. This means that many existing security deployments such as transport layer security (TLS) used to secure Internet browsing, PDF signing used to protect electronic documents, code signing used for automatic system updates, end-to-end encryption used to provide data privacy, and so on, will no longer be deemed secure and instead be rendered untrustworthy with the advent of quantum computers. The good news, though, is that the current noisy intermediate-state quantum computers are not ready to break industrial-strength cryptography. For quantum computers to be cryptanalysis-capable, the number of qubits, depth of circuits, duration of coherence, and accuracy of error correction all need to improve several hundred or thousand-fold, and this is not likely to happen for at least another ten years.[3]

Much like how Hurlburt[4] has succinctly described the cybersecurity perils of technology due to untimely prevention and intervention, we view the quantum roadmap as a race between how fast quantum technology is invented versus how soon existing applications can be adapted to shield against its side effects. On the one hand, researchers are working on building larger, more fault-tolerant quantum technologies which can be used for new applications in simulations,[5] machine learning,[6] and even entertainment.[7] On the other hand, system owners need to embark on a migration or renewal plan for their applications to defend against possible quantum attacks. As depicted in Figure 1, instead of holding back on quantum technology advancements, we should strive to reduce the impact and likelihood of quantum attacks so that quantum benefits significantly outweigh the risks.

From an adversarial standpoint, quantum computers present a golden opportunity to exploit the vulnerabilities promised. But since fault-tolerant quantum technology is not yet ready, are the adversaries simply waiting or are they already carrying out activities to maximize their impact when cryptanalysis-capable quantum computers become available? In this article, we explore the question:

What are adversaries doing today to prepare for the impending availability of cryptanalysis-capable quantum computers? To answer this question comprehensively, we model the adversary in a three-step approach:

Step 1. Profile the different adversarial personas, their resources, and motivations. Assuming

Digital Object Identifier 10.1109/MC.2022.3218046
Date of current version: 26 July 2023

that each adversary has access to a cryptanalysis-capable quantum computer today, flesh out the attacks with which the adversary can use the quantum computer to compromise their targets. This is covered in the section "Profiling Adversarial Personas."

Step 2. Abstract relevant preprocessing activities that each adversary performs prior to using the quantum computer. We expect these are likely passive work being carried out by adversaries today since they are not bottle-necked by the availability of quantum computers and yet maximize the adversaries' advantage against their peers and targets. We describe these passive actions in the section "Abstracting Passive Actions."

Step 3. Identify various points in the targets' quantum roadmap and their dependencies that can be affected by an adversary. Besides passive steps, we expect some adversaries to take active steps in increasing their targets' exposure to quantum computers. This would be in the form of disrupting the quantum readiness of the vulnerable targets, with the aim of leaving them still vulnerable when cryptanalysis-capable quantum computers become available. This is analyzed in the section "Identifying Active Interventions." From this adversary model, we can then derive recommendations on concrete actions that can be taken today to prevent exposure to or mitigate such adversarial activities.

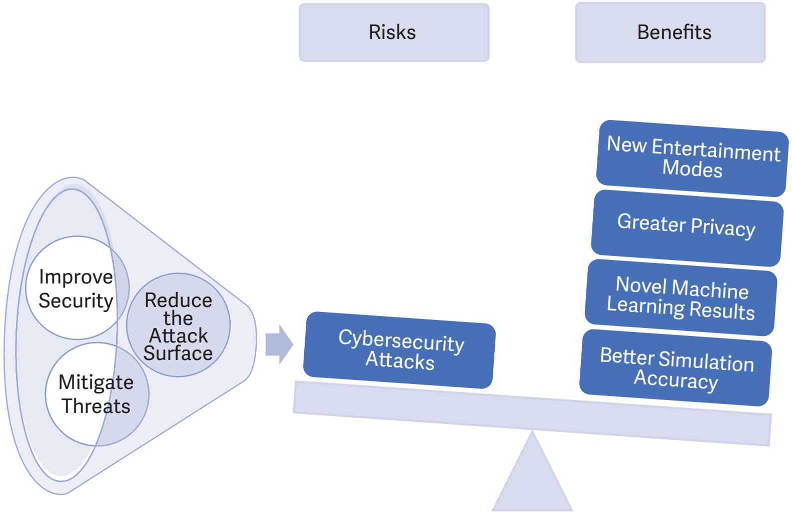


FIGURE 1. Can quantum benefits outweigh the risks?

PROFILING ADVERSARIAL PERSONAS

We take reference from Rocchetto and Tippenhauer[9] who perform a comprehensive study on various adversary profiles. We bucket them into five personas next in increasing level of resources:

- *Basic User*: The Basic User is a technology-literate individual with limited resources who carries out attacks out of curiosity, personal glory, fun, and are typically non-malicious. A "Script Kiddy" is an example of a Basic User who uses readily available hacking tools to attack known vulnerabilities. More sophisticated users include white-hat hackers who participate in bug-bounty challenges to uncover new vulnerabilities.
- *Insider (or Disgruntled Employee)*: The Insider is also an individual but festers malicious intent in sabotaging an entity for revenge purposes. The Insider would typically have access to protected

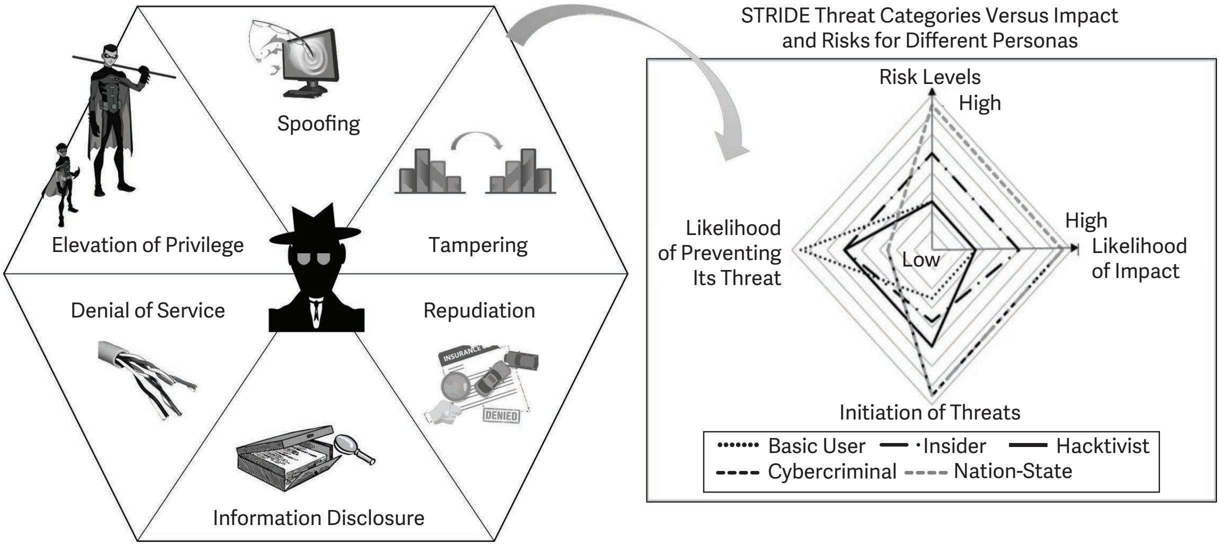


FIGURE 2. The STRIDE threat categories[8] along with impact and risk range for different users.

information and privileged access rights which allow for such adversarial activities to easily take place.

- *Hacktivist*: The Hacktivist may operate solo but more commonly in small, decentralized groups where they use cyber-hacking activities to drive specific agendas such as environmental awareness, data emancipation, and so on. The "Anonymous" group is a famous Hacktivist example fighting against governmental oppression. In its extreme form, Hacktivists engage in terrorist activities that spread fear and cause widespread disruption.
- *Cybercriminal (or Hacker for hire)*: The Cybercriminal is usually part of organized crime syndicates that turn their knowledge and skills of compromising systems into financial gain. Beyond making a statement, the Cybercriminal would engage in corporate espionage, blackmail, ransom, and other lucrative activities.
- *Nation-state*: The Nation-state adversary is often pictured as an elusive underground organization with vast resources (sponsored by the state) to carry out nefarious activities on behalf of the government. The targets are typically public infrastructure projects, including public utilities, transport, and the financial system, which, if compromised, can significantly disrupt the lives of the citizens residing in the target country.

Based on these personas and their motivation, we identify the threats by asking how these adversaries will use quantum computers if they are available today. The threat model we have chosen here is STRIDE[8] as it is an appropriate framework for examining software and data threats that quantum computers present. The six categories of threats are enumerated using the acronym that STRIDE represents, namely 1) **S**poofing (when authentication is violated), 2) **T**ampering (when integrity is violated), 3) **R**epudiation (when proof of confirmation is violated), 4) **I**nformation disclosure (when confidentiality is violated), 5) **D**enial of service (when availability is violated), and 6) **E**levation of privilege (when access rights is violated). Figure 2 illustrates each of the threat categories pictorially.

Table 1 shows the range of attacks that can be carried out using a quantum computer by different adversaries. They mainly involve the compromise of authentication credentials (passwords, Wi-Fi, access credentials, wallets), protected data (phone data, emails, documents, trade secrets, classified messages), or the protocol (hijacking Internet sessions, modifying transactions, disrupting mobile, and satellite communications). Based on this information, we then abstract the actions that can already be performed prior to the availability of cryptanalysis-capable quantum computers and cover this in the next two sections.

ABSTRACTING PASSIVE ACTIONS

We define passive actions as nondisruptive actions (akin to eavesdropping) that the adversary can take without affecting the quantum roadmap, both from the invention of cryptanalysis-capable quantum computers and from the migration to quantum-secure cryptography by system owners.

Data harvesting

An obvious passive action that adversaries can do is to identify and collect authentication credentials and protected data for subsequent cryptanalysis in a "harvest-then-decrypt" attack.[10] From Table 1, these include the following:

TABLE 1. Adversarial personas[9] and how they can use quantum computers to achieve their goals.

Persona	Motivation	STRIDE threats
Basic user	Curiosity, personal glory	**S:** Cracking passwords **T:** Defacing websites **R:** Faking crypto-currency payments **I:** Recovering private data on phones **D:** Hijacking Wi-Fi connections **E:** Sending commands to appliances
Insider	Revenge	**S:** Phishing for supervisor accounts **T:** Changing employee records **R:** Erasing audit evidence **I:** Reading company email **D:** Deleting databases **E:** Creating backdoors
Hacktivist	Publicity, agenda, emotions	**S:** Impersonating others on Internet **T:** Spreading false text messages **R:** Forging documents and contracts **I:** Revealing private documents **D:** Shutting down mobile networks **E:** Getting root access to firewalls
Cyber-criminal	Financial return	**S:** Forging fake passports **T:** Manipulating financial records **R:** Creating fake payment cards **I:** Stealing trade secrets **D:** Taking over payment wallets **E:** Sending illegal trade instructions
Nation-state	Disruption, erosion of trust	**S:** Faking as official news media **T:** Planting backdoor applications **R:** Issuing fake certificates **I:** Tapping on classified messages **D:** Disrupting satellite communications **E:** Taking control of public utilities

- *Password hashes*: Passwords are stored as one-way hashes in the backend database so that they can only be used to verify the users presenting the passwords but not expose the values. However, using Grover's algorithm,[2] these protection mechanisms are weakened, and the actual passwords may be revealed.
- *Certification authority (CA) certificates*: Certificates issued by a CA are used to assert the identities of the users associated with the certificates. However, using Shor's algorithm, the CA's private key can be computed from the certificates, thereby allowing forged certificates (possibly with different identities that can be backdated) to be generated.
- *Biometric minutiae*: Similar to password hashes, biometric minutiae are used to verify the identity of the users with the right biometric features. Grover's algorithm could be used to perform a brute-force search to reverse the minutiae to reveal the features, violating privacy, and impersonation concerns. The alarming problem here is that while passwords can be updated to mitigate the threat, biometric features on a person cannot be changed.
- *Electronic Contracts*: Electronic contracts are digitally signed using an asymmetric key cryptosystem to ensure the integrity, nonrepudiation and time stamp of the contents. Using Shor's algorithm, the private signing key can be computed from the public verification key, which renders the contract contents untrustworthy since the relying party can no longer prove the difference between a real or fake contract.
- *Trade secrets*: Companies use trade secrets as a means of maintaining a business advantage over their competitors and store such secrets encrypted using hardware vaults. Since many

of these vault implementations employ asymmetric key cryptography to protect the encryption keys and their own backups, Shor's algorithm can quickly allow an adversary to decrypt the trade secrets without needing to break the vault.

- *Confidential data exchange*: Secure emails, peer-to-peer messaging, mobile, Internet, and satellite communications rely on asymmetric key encryption to exchange session keys used to protect the communication. Using Shor's algorithm, the key exchange protocol can be cryptanalysed to reveal the session keys, which then allows the adversary to see the communication in the clear.

From our analysis, what becomes apparent besides the sheer amount of data mentioned previously that can be harvested is that there would be an emergence of a data marketplace or broker of sorts to facilitate the demand-generation, collection, storage, sale, and delivery of such data.

Target prioritization

Even when cryptanalysis-capable quantum computers do become available, we expect the supply of such quantum computing resources to be scarcer compared to the demand. Hence, adversaries will need to figure out which targets they will go after first, and this can be done now. The prioritization can be done based on the following criteria:

- *By value of target*: Depending on the motivations and available resources behind the adversary, each target may be valued differently. Naturally, we expect a rational adversary to use a cost-benefit ratio to choose targets where *benefit derived from the compromise >> cost to carry out the compromise*.
- *By ease of break*: Since Shor's algorithm[1] compromises asymmetric key algorithms while Grover's algorithm[2] merely weakens symmetric key and hash algorithms, it is clear that applications using asymmetric key algorithms will be first targeted. The key sizes of the asymmetric key algorithms do matter in sizing the capacity of the quantum computer to carry out the compromise. The estimated number of fault-tolerant qubits needed on a quantum computer to break the RSA and ECC cryptosystems are $2n + 2$[11] and $6n$[12], respectively, where n is the key size. This roughly translates to a 4,000+ qubit quantum computer to break an RSA-2048 application and a 1,500+ qubit quantum computer to break the ECC-256 application.
- *By the scope of break*: There is a difference when it comes to compromising RSA and ECC used in key exchange protocols. When RSA is used in TLS for key exchange, the same private RSA key is used repeatedly to decrypt different session keys for the lifetime of the RSA key. This means that an adversary can get access to all session keys once the common private RSA key is compromised by a quantum computer. On the other hand, ECC can be used with perfect-forward secrecy in TLS v1.3 where a different ECC key is used each time a session is established. This means that each quantum computer cryptanalysis only reveals one session key, which is "annoying" to adversaries and limits the scope of the compromise. Such a difference does not exist for digital signature applications using RSA or ECC.

A result that we can derive from the previous criteria is that adversaries will likely prioritize 1) ECC-based digital signature applications, 2) RSA-based digital signature and key exchange applications, 3) ECC-based key exchange applications, and 4) symmetric key and hash-based applications assuming the value of the targets are similar, and the data prerequisites discussed in the section "Data Harvesting" can be met.

IDENTIFYING ACTIVE INTERVENTIONS

To further increase their advantage, we expect some adversaries may choose to carry out active interventions to affect their targets' quantum roadmap. From a modeling perspective, Mosca[3] has prepared a quantum readiness metric which defines three parameters $\mathcal{X}$, $\mathcal{Y}$, and $\mathcal{Z}$ as follows:

- $\mathcal{X}$: the duration of time that the cryptographic secrets need to be kept secret.

- $\mathcal{Y}$: the time needed to deploy tools that are quantum-secure.
- $\mathcal{Z}$: the time duration before a quantum computer breaks the algorithm or reveals the secrets.

Mosca's theorem states that if $\mathcal{X} + \mathcal{Y} > \mathcal{Z}$, then the target should be worried about the vulnerability exposure. Mosca then attempts to assign absolute values to $\mathcal{Z}$, where he estimates a 1/7 chance of RSA-2048 being broken in 2026, with the chance increasing to 1/2 by 2031.

When using this model from an adversary's point of view, the difference of $(\mathcal{X} + \mathcal{Y}) - \mathcal{Z}$, represents the time advantage the adversary has on the target. The actions of adversaries can thus be classified into extending $\mathcal{X}$, extending $\mathcal{Y}$ and reducing $\mathcal{Z}$ as discussed next.

Extending $\mathcal{X}$

We need to first recognize that $\mathcal{X}$ is different depending on the threats posed:

- For **S**poofing, **D**enial-of-Service, and **E**levation of privilege, the value of $\mathcal{X}$ is close to zero. This is because the duration in which authentication credentials need to be protected is only at the point of usage. If they have been compromised by quantum computers, the credentials just need to be changed to mitigate the threat. The only exception is biometric authentication credentials which cannot be easily changed.
- For **T**ampering and **R**epudiation, the value of $\mathcal{X}$ is the duration in which the information is relied on. If an agreement is tampered with or the origin of the agreement is cast into doubt only after the expiry of the agreement, the risk impact is minimized. We expect long-term contracts (extending for more than five years), financial records and audit logs to have the longest $\mathcal{X}$.
- For **I**nformation disclosure, the value of $\mathcal{X}$ may be exceedingly long since an adversary may still derive benefit from exposing such information long past the validity of the information. An example is wikileaks.org, whose published documents impacted the image and reputation of several persons and their positions, despite them no longer holding public office. In fact, it may be futile to migrate existing quantum-vulnerable encrypted data to quantum-secure encryption since multiple copies of the existing encrypted data may exist and remain vulnerable.

To the adversary, focusing on vulnerabilities related to **T**ampering **R**epudiation and **I**nformation disclosure has an effect of extending $\mathcal{X}$. The actions are similar to the data harvesting actions discussed in the section "Data Harvesting."

Extending $\mathcal{Y}$

In examining $\mathcal{Y}$, we see that $\mathcal{Y}$ can be further broken down into three subcomponents that require to run consecutively. We define them here as follows:

- $\mathcal{Y}_1$: time needed to design/select a viable quantum-secure solution
- $\mathcal{Y}_2$: time needed to upgrade the application code and protocol to support the quantum-secure solution
- $\mathcal{Y}_3$: time needed to migrate the users, system, keys, data, and processes to use the quantum-secure solution.

The National Institute of Science and Technology (NIST) is working with the industry on a postquantum cryptography (PQC) standardization process. The process is to select asymmetric key algorithms both for key exchange, where a session key used for confidentiality is mutually established and for digital signing, where the integrity and nonrepudiation of signed data are ensured. Now into the fourth round of evaluation,[13] a total of one key-exchange and three digital signature algorithms have been selected for standardization. What remains is the selection of nonlattice key-exchange algorithms to complete the portfolio of general-purpose PQC algorithms. NIST expects the draft standards to be finalized by 2024. These coordinated efforts in evaluating each of the candidate algorithms and putting them through various evaluation criteria represent a top-down approach that directly impacts $\mathcal{Y}_1$ with identified preferences (e.g., drop-in replacement) that may help reduce the duration for $\mathcal{Y}_2$ for some applications.

TABLE 2. Actions taken by adversaries to affect $\mathcal{Y}$.

Persona	Actions	Impact
Basic user	Provide inappropriate consulting and advice on how to do quantum migration. Affects $\mathcal{Y}_2$.	Low
Insider	Capture wrong system requirements or omit some systems, resulting in gaps and delays in the migration. Affects $\mathcal{Y}_3$.	Medium
Hacktivist	Drive the importance of other agenda to deflect focus from quantum migration. Affects $\mathcal{Y}_2$.	Low
Cyber-criminal	Work with equipment vendors to deliver nonquantum-secure systems. Affects $\mathcal{Y}_2$ and $\mathcal{Y}_3$.	High
Nation-state	Disrupt the PQC standardization process. Affects $\mathcal{Y}_1$.	High

In Table 2, we take a bottom-up implementation approach and run through various actions that adversaries can take to affect $\mathcal{Y}$ and their potential impact.

When evaluating the impact of the actions, we see that the *Cybercriminal* and *Nation-state* can cause a systemic failure to quantum migration efforts across the industry while the actions by *Basic User*, *Insider*, and *Hacktivist* are generally isolated to individual companies or organizations.

Reducing $\mathcal{Z}$

For the well-resourced *Nation-state*, they may be able to channel more money and engineers in speeding up research into building a cryptanalysis-capable quantum computer, thus actually reducing $\mathcal{Z}$.

For medium-resourced adversaries such as *Hacktivist* or *Cybercriminal*, they may adopt a misinformation campaign instead to give the perception of either a shorter or longer $\mathcal{Z}$. If $\mathcal{Z}$ is perceived to arrive quicker, it may force targets to divert resources from the original quantum migration plans and adopt short-term parameter defenses instead, thus delaying $\mathcal{Y}$. Separately, an illusion of a longer $\mathcal{Z}$ may loosen the targets' vigilance and inadvertently also lengthen $\mathcal{Y}$. We do not expect the *Basic User* or *Insider* to impact $\mathcal{Z}$ in any meaningful way.

RECOMMENDATIONS AND NEXT STEPS

In this section, we use the analysis done in the earlier section to derive possible actions that can be taken to either mitigate the threats or reduce the adversaries' advantage. Considering that organizations may take several years of planning and execution to migrate cryptographic algorithms,[14] we have categorized the recommendations (see Figure 3) into what an organization can start doing now versus what should be done within the next two to three years while awaiting NIST's postquantum cryptographic standards to become established.

- To start now:
  - *Know your exposure:* Organizations should perform team-based threat modeling exercises[15] to understand their current risk and reduce exposure to insider threats. This will allow planning and allocation of resources and budget for quantum migration.
  - *Use quantum annoyance:* Begin the migration process by upgrading session encryption and other communications to use TLS v1.3 or other protocols which have perfect-forward secrecy to limit the scope of any possible compromise to within each session.
  - *Deter data harvesting:* In order to mitigate against data harvesting, use a minimum of a 256-bit symmetric encryption key, over and above existing security protection, to protect data and files in storage. This can be in the form of encrypted databases or encrypted storage devices that perform blanket encryption of all data.
- To do over the next two to three years:
  - *Update organizational practices*: Increase organizational awareness of potential quantum threats and start to include quantum-secure compliance or requirements into procurement and implementation practices.
  - *Prioritize integrity and non-repudiation*: Based on our adversarial model, the first threats to be mitigated are tampering and repudiation. Long-dated documents, contracts, logs, and

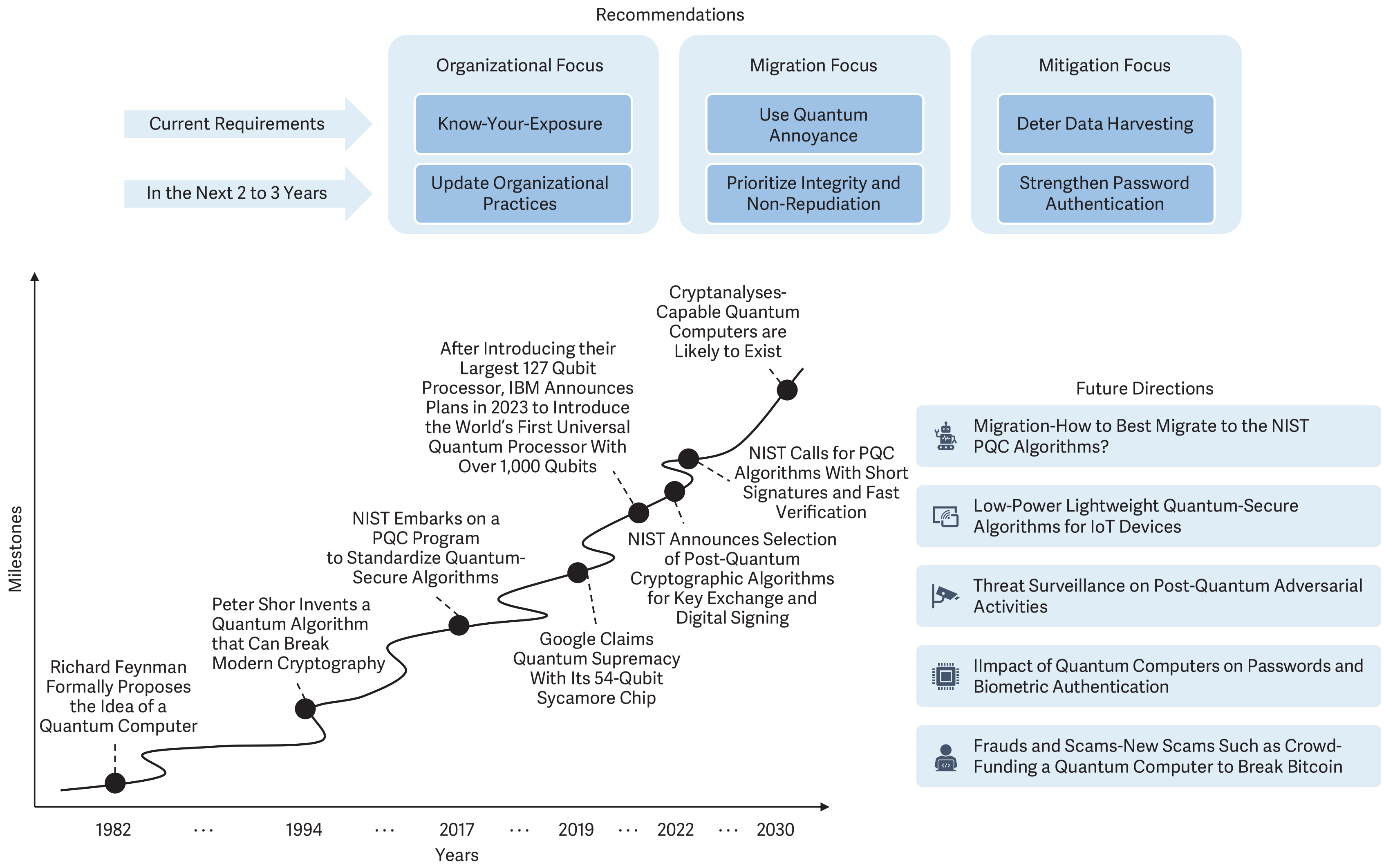

Recommendations
Current Requirements
In the Next 2 to 3 Years
Organizational Focus
Know-Your-Exposure
Update Organizational Practices
Migration Focus
Use Quantum Annoyance
Prioritize Integrity and Non-Repudiation
Mitigation Focus
Deter Data Harvesting
Strengthen Password Authentication
Milestones
Richard Feynman Formally Proposes the Idea of a Quantum Computer
Peter Shor Invents a Quantum Algorithm that Can Break Modern Cryptography
NIST Embarks on a PQC Program to Standardize Quantum-Secure Algorithms
Google Claims Quantum Supremacy With Its 54-Qubit Sycamore Chip
After Introducing their Largest 127 Qubit Processor, IBM Announces Plans in 2023 to Introduce the World's First Universal Quantum Processor With Over 1,000 Qubits
NIST Announces Selection of Post-Quantum Cryptographic Algorithms for Key Exchange and Digital Signing
NIST Calls for PQC Algorithms With Short Signatures and Fast Verification
Cryptanalyses-Capable Quantum Computers are Likely to Exist
1982
...
1994
...
2017
...
2019
...
2022
...
2030
Years
Future Directions
Migration-How to Best Migrate to the NIST PQC Algorithms?
Low-Power Lightweight Quantum-Secure Algorithms for IoT Devices
Threat Surveillance on Post-Quantum Adversarial Activities
IImpact of Quantum Computers on Passwords and Biometric Authentication
Frauds and Scams-New Scams Such as Crowd-Funding a Quantum Computer to Break Bitcoin

FIGURE 3. An illustration of the milestones for quantum-research, recommendations, and future directions.

other data should be time-stamped with a quantum-secure mechanism prior to the availability of cryptanalysis-capable quantum computers to ensure their integrity and nonrepudiation status. Such mechanisms include using blockchains, time-stamping digital signatures that are quantum-secure,[16] or using stateful hash-based signature scheme[17] by NIST.

» *Strengthen password authentication*: Despite many issues related to users' passwords being phished, passwords as an authentication mechanism are relatively resistant to quantum cryptanalysis as compared to ECC or RSA-based authentication. Critical systems should have multifactor authentication implemented where one of the factors is password authentication with password entropy of up to 256-bits to ensure authentication remains secure.

For combating more broad-based threats such as misinformation and the emergence of data marketplaces/brokers, this work is beyond the scope of a single organization.

When comparing with other relevant works,[13,18] we adopt a broader perspective to dispel the notion that quantum computers simply pose a cryptographic threat and that a replacement of algorithms would address the problem. This is similarly echoed by Mashatan and Turetken.[19] We also note that our recommendations are more comprehensive as compared to the infographic guide[20] prepared by the U.S. Department of Homeland Security. More work still needs to be done to prepare for quantum readiness. We have highlighted the timeline and future directions in Figure 3. Besides looking for an appropriate replacement cryptographic algorithm to RSA and ECC, work must now focus on how these algorithms can be properly migrated to, and how other applications such as IoT devices and biometric authentication may need different algorithms. This direction is clearly demonstrated by NIST who has started a new call for a specific PQC algorithm with short signatures and fast verification, soon after announcing the general-purpose PQC algorithms for standardization. Developing possible frameworks on how the industry can establish additional norms in transparent surveillance and coordinated responses to deny adversaries any advantage is the crucial future direction of this study.

Tracing the quantum timeline, it has been almost 40 years since the idea of a quantum computer was floated. The industry's developments in the past few years in achieving quantum supremacy are nothing short of awe. While we look forward to breakthroughs facilitated by the power of quantum computing, we are mindful of the security and privacy threats that quantum computers pose. Adversaries are not sitting on their hands, and we highlight various approaches and scenarios where they can already gain an advantage. We have provided recommendations and included additional areas where research work is needed to close the gap.

REFERENCES

1. P. W. Shor, "Polynomial-time algorithms for prime factorization and discrete logarithms on a quantum computer," *SIAM Rev.*, vol. 41, no. 2, pp. 303–332, 1999, doi: 10.1137/S0036144598347011.
2. L. K. Grover, "Quantum mechanics helps in searching for a needle in a haystack," *Phys. Rev. Lett.*, vol. 79, no. 2, p. 325, 1997, doi: 10.1103/PhysRevLett.79.325.
3. M. Mosca, "Cybersecurity in an era with quantum computers: Will we be ready?" *IEEE Security Privacy*, vol. 16, no. 5, pp. 38–41, Sep./Oct. 2018, doi: 10.1109/MSP.2018.3761723.
4. G. F. Hurlburt, "The tyranny of urgency," *Computer*, vol. 52, no. 6, pp. 68–72, Jun. 2019, doi: 10.1109/MC.2019.2905652.
5. C. C. McGeoch, R. Harris, S. P. Reinhardt, and P. I. Bunyk, "Practical annealing-based quantum computing," *Computer*, vol. 52, no. 6, pp. 38–46, Jun. 2019, doi: 10.1109/MC.2019.2908836.
6. M. Roetteler and K. M. Svore, "Quantum computing: Codebreaking and beyond," *IEEE Security Privacy*, vol. 16, no. 5, pp. 22–36, Sep./Oct. 2018, doi: 10.1109/MSP.2018.3761710.
7. T. Humble, "Consumer applications of quantum computing: A promising approach for secure computation, trusted data storage, and efficient applications," *IEEE Consum. Electron. Mag.*, vol. 7, no. 6, pp. 8–14, Nov. 2018, doi: 10.1109/MCE.2017.2755298.

8. L. Kohnfelder and P. Garg, "The threats to our products," Microsoft Corp., Redmond, WA, USA, 1999. Accessed: Aug. 2022. [Online]. Available: https://adam.shostack.org/microsoft/The-Threats-To-Our-Products.docx
9. M. Rocchetto and N. O. Tippenhauer, "On attacker models and profiles for cyber-physical systems," in *Proc. Eur. Symp. Res. Comput. Security*, Springer-Verlag, 2016, pp. 427–449, doi: 10.1007/978-3-319-45741-3_22.
10. A. Mashatan and D. Heintzman, "The complex path to quantum resistance," *Commun. ACM*, vol. 64, no. 9, pp. 46–53, 2021, doi: 10.1145/3464905.
11. Y. Takahashi and N. Kunihiro, "A quantum circuit for Shor's factoring algorithm using 2n+2 qubits," *Quantum Inf. Comput.*, vol. 6, no. 2, pp. 184–192, 2006, doi: 10.26421/QIC6.2-4.
12. J. Proos and C. Zalka, "Shor's discrete logarithm quantum algorithm for elliptic curves," 2003, *arXiv Preprint: Quant-ph/0301141*.
13. G. Alagic et al., "Status report on the third round of the NIST post-quantum cryptography standardization process," National Institute of Standards and Technology, Gaithersburg, MD, USA, Tech. Rep. 8413, 2022.
14. M. Gardiner, and A. Truskovsky, and G. Neville-Neil, and A. Mashatan, "Quantum-safe trust for vehicles: The race is already on," *Commun. ACM*, vol. 64, no. 9, pp. 54–61, 2021, doi: 10.1145/3466174.
15. C. C. Lee, T. G. Tan, V. Sharma, and J. Zhou, "Quantum computing threat modelling on a generic CPS setup," in *Proc. Int. Conf. Appl. Cryptography Netw. Secur.*, Springer-Verlag, 2021, pp. 171–190, doi: 10.1007/978-3-030-81645-2_11.
16. T. G. Tan and J. Zhou, "Layering quantum-resistance into classical digital signature algorithms," in *Proc. 24th Int. Secur. Conf. (ISC 2021)*, Springer-Verlag, 2021, pp. 26–41, doi: 10.1007/978-3-030-91356-4_2.
17. D. A. Cooper, D. C. Apon, Q. H. Dang, M. S. Davidson, M. J. Dworkin, and C. A. Miller, Recommendation for stateful hash-based signature schemes," National Institute of Standards and Technology Gaithersburg, MD, USA, NIST Special Publication 800-208, 2020.
18. G. Mone, "The quantum threat," *Commun. ACM*, vol. 63, no. 7, pp. 12–14, 2020, doi: 10.1145/3398388.
19. A. Mashatan and O. Turetken, "Preparing for the information security threat from quantum computers," *MIS Quart. Executive*, vol. 19, no. 2, pp. 157–163, 2020.
20. "Preparing for post-quantum cryptography," Department of Homeland Security, Washington, DC, USA, 2021. Accessed: Aug. 2022. [Online]. Available: https://www.dhs.gov/publication/preparing-post-quantum-cryptography-infographic

TEIK GUAN TAN is cofounder and CEO of pQCee Pte Ltd, Singapore 038987. His research interests lie at the intersection of authentication, applied cryptography and quantum algorithms. Tan received a Ph.D. in cybersecurity from the Singapore University of Technology and Design, Singapore 138682. He is a Member of IEEE. Contact him at teikguan@pqcee.com.

JIANYING ZHOU is a professor and co-center director for iTrust at Singapore University of Technology and Design, Singapore 138682. His research interests include applied cryptography and network security, cyber-physical system security, and mobile and wireless security. Zhou received a Ph.D. in information security from Royal Holloway, University of London. He is a Senior Member of IEEE. Contact him at: jianying_zhou@sutd.edu.sg.

VISHAL SHARMA is an assistant professor in the School of Electronics, Electrical Engineering, and Computer Science (EEECS), Queen's University Belfast (QUB), BT7 1NN Belfast, U.K. His research interests include autonomous systems, unmanned aerial vehicle communications, network behavior modeling, 5G and beyond, blockchain, cyber-physical systems, and Internet of Things. Sharma received the Ph.D. in computer science and engineering from Thapar University. He is a Senior Member of IEEE. Contact him at v.sharma@qub.ac.uk.

SARAJU P. MOHANTY is a professor in the Department of Computer Science and Engineering, University of North Texas, Denton, TX 76203 USA. His research interests include "smart electronic systems" which has been funded by the National Science Foundations, Semiconductor Research Corporation, U.S. Air Force, Indo-U.S. Science and Technology Forum (IUSSTF), and Mission Innovation. Mohanty received a Ph.D. in computer science and engineering from the University of South Florida. He is a Senior Member of IEEE. Contact him at saraju.mohanty@unt.edu.

EDITORS: Silvia Abrahão, Universitat Politècnica de València, sabrahao@disc.upv.es
Miroslaw Staron, Chalmers University of Technology and University of Gothenburg, miroslaw.staron@cse.gu.se

This article originally appeared in
Software
vol. 41, no. 5, 2024

DEPARTMENT: PRACTITIONERS' DIGEST

Bringing Software Engineering Discipline to the Development of AI-Enabled Systems

Miroslaw Staron, Silvia Abrahão, Grace Lewis, Henry Muccini, and Chetan Honnenahalli

Engineering AI software systems is starting to evolve from the pure development of machine learning (ML) models to a more structured discipline that treats ML components as part of much larger software systems. As such, more structured principles are required for their development, such as established design principles, established development models, and safeguards for deployed ML models. This column focuses on papers presented at the Third International Conference on AI Engineering—Software Engineering for AI (CAIN 2024). The selected papers reflect the current development of the field of AI systems engineering and AI software development, taking it to the next level of maturity. Feedback or suggestions are welcome. In addition, if you try or adopt any of the practices included in the column, please send us and the authors of the paper(s) a note about your experiences.

SOFTWARE DESIGN PRINCIPLES IN ML SOFTWARE

ML components are often developed by data scientists with diverse educational backgrounds, which often leads to code that does not adhere to best practices for software design. This may pose challenges for the maintenance and longevity of the software product. The paper "Investigating the Impact of SOLID Design Principles on Machine Learning Code Understanding" by Raphael Cabral, Marcos Kalinowski, Maria Teresa Baldassarre, Hugo Villamizar, and Helio Côrtes Vieira Lopes—winner of the Distinguished Paper Award—presents the results of a study to understand the impact of SOLID design principles on ML code understanding. SOLID stands for: Single responsibility, Open-closed, Liskov substitution, Interface substitution, and Dependency-inversion.

The study was a controlled experiment in which the authors restructured real industrial ML code that did not use SOLID principles. They conducted three trials, in which, within each trial, one group was presented with the original code, and the other was presented with the code restructured following SOLID principles. Each group was asked to perform some code analysis tasks and then fill out a questionnaire related to their understanding of the code and agreements related to the impact of each SOLID principle. The results show statistically significant evidence that adopting SOLID design principles improves ML code understanding. Therefore, the recommendation is to include software design best practices in the education and training of data scientists, which will lead to more robust and maintainable ML code. The paper was presented at CAIN 2024 and is available at https://tinyurl.com/ypcxubjc.

V-MODEL FOR DEVELOPING ML SYSTEMS

ML-enabled systems are typically built by multidisciplinary teams, which often generates collaboration challenges, especially when moving ML prototype models into production environments. Because of this, it becomes difficult to use traditional software lifecycle models such as waterfall, spiral, or agile models. The paper "Exploratory Study of V-Model in Building ML-Enabled Software: A Systems Engineering

Digital Object Identifier 10.1109/MS.2024.3408388
Date of current version 13 August 2024

Perspective" by Jie JW Wu explores the use of the V-Model—a process commonly used in systems engineering—to address these interdisciplinary collaboration challenges.

Practitioners from nine software companies were interviewed to understand their collaboration challenges and discuss how the V-Model could address these challenges. The result was a set of eight recommendations for using the V-Model to manage interdisciplinary collaborations when building ML-enabled systems, which include the involvement of ML model developers in system requirement discussions, clear responsibility and definition of interfaces for both ML and non-ML components, management of risks associated to the inherent uncertainty of ML models, and the consideration of data and infrastructure as first-class system components, among others. However, there is also the recognition of the additional effort required by the V-Model, as well its limited flexibility, agility, and adaptability to changes, which warrants further research on the tradeoffs and benefits. The paper was presented at CAIN 2024 and is available at https://tinyurl.com/y8wcayxa.

DOMAIN-SPECIFIC AI ASSISTANTS

General-purpose AI assistants often provide generic or inaccurate responses. The paper "Developer Experiences with a Contextualized AI Coding Assistant: Usability, Expectations, and Outcomes" by Gustavo Pinto, Cleidson de Souza, Thayssa Rocha, Igor Steinmacher, Alberto de Souza, and Edward Monteiro analyzes how contextualized, domain-specific, AI coding assistants may produce effective solutions that are tailored to an organization's unique needs.

The authors selected the domain-specific AI tool StackSpot AI, a highly contextualized AI coding assistant that considers individual developers' specific needs and the complexities of specific projects. StackSpot AI leverages the retrieval augmented generation (RAG) mechanism. Sixty-two practitioners within Zup Innovation, a large software development company, were involved in a controlled online experiment to evaluate user experience and the correctness of the generated solution using StackSpot AI. The study aimed to introduce practitioners to StackSpot AI and collect representative feedback to enhance the product's quality and usability.

The study showed that StackSpot AI boosts productivity and time efficiency, with participants valuing its quick, accurate code generation. However, its effectiveness depends on precise prompts and proper configuration. Challenges include technical limitations, better IDE support, and inherent large language model issues requiring further research. The paper was presented at CAIN 2024, and its preprint is available at https://tinyurl.com/4f5ktpr7.

SAFEGUARDING ML SYSTEMS

Deploying ML models is often a complex task, not only because of performance but especially when it comes to ensuring safety, security, and transparency. The paper "ML-On-Rails: Safeguarding Machine Learning Models in Software Systems: A Case Study" by Hala Abdelkader, Mohamed Abdelrazek, Scott Barnett, Jean-Guy Schneider, Priya Rani, and Rajesh Vasa discusses these challenges and the importance of addressing them. The paper introduces ML-On-Rails, a protocol aimed at safeguarding ML models and facilitating communication between ML model providers and software engineers. Key components of ML-On-Rails include safeguards such as adversarial attack detection, out-of-distribution data detection, model explainability, input validation, and integration of NeMo Guardrails for generative models.

ML COMPONENTS ARE OFTEN DEVELOPED BY DATA SCIENTISTS WITH DIVERSE EDUCATIONAL BACKGROUNDS, WHICH OFTEN LEADS TO CODE THAT DOES NOT ADHERE TO BEST PRACTICES FOR SOFTWARE DESIGN.

The protocol utilizes HTTP status codes for reporting successful outcomes and errors, enhancing transparency and interpretability. A case study of the MoveReminder application illustrates the implementation of ML-On-Rails, highlighting its role in ensuring the reliability of activity recommendations and handling failures. The paper concludes by emphasizing the need for continuous evolution of the protocol and outlines future work including a qualitative evaluation through a targeted survey involving professionals. The paper was presented at CAIN 2024 and is available at https://tinyurl.com/45wmhbkx.

MEASURING ACCOUNTABILITY OF ML SYSTEMS

As AI evolves into more advanced forms, particularly with the advent of large-scale generative models (GenAI) such as large language models (LLMs), it brings not only technological innovations but also significant safety challenges. These challenges include data privacy concerns, lack of transparency, and the dissemination of misinformation. Such issues are particularly critical in sectors where the misuse of AI, especially advanced AI, can lead to unfavored results such as biased decision-making and dual-use technology concerns.

Addressing these challenges necessitates the operationalization of responsible AI (RAI). Accountability, a cornerstone of RAI, serves as the backbone for enhancing AI safety. While the proliferation of RAI principles worldwide highlights a growing recognition of their importance, legislative efforts, such as the EU AI Act, underscore the urgency of developing more concrete guidelines to enforce these principles, particularly accountability.

The study presented in the paper "Towards a Responsible AI Metrics Catalogue: A Collection of Metrics for AI Accountability" by Boming Xia, Qinghua Lu, Liming Zhu, Sung Une (Sunny) Lee, Yue Liu, and Zhenchang Xing advocates for a process-oriented approach to AI accountability, integrating both technical and socio-technical dimensions. Through a multivocal literature review, the study develops a system-level metrics framework tailored to operationalize AI accountability. This framework includes process metrics (procedural guidelines), resource metrics (necessary tools and frameworks), and product metrics (resultant artifacts). The main contributions are the creation of a catalog of process-centric metrics for AI accountability, the categorization of these metrics, and the foundation for a comprehensive RAI framework. The paper was presented at CAIN 2024 and is available at https://tinyurl.com/4nrpfr7z.

CONCEPT DRIFT MONITORING FOR ANOMALY DETECTION

Various model adaptation techniques can be used to mitigate the effects of concept drift in anomaly detection in AIOps solutions. From the retraining data perspective, previous work proposed two model retraining techniques, namely the full-history approach and the sliding-window approach. From the retraining frequency perspective, there are two retraining techniques proposed in the literature, namely periodic retraining and retraining based on concept drift detection. However, the performance of those retraining strategies for anomaly detection models on operational data has not been extensively analyzed. The paper "Is Your Anomaly Detector Ready for Change? Adapting AIOps Solutions to the Real World" by Lorena Poenaru-Olaru, Natalia Karpova, Luís Cruz, Jan S. Rellermeyer, and Arie van Deursen first assesses the performance of state-of-the-art anomaly detection models on operational data, then analyzes the impact of the aforementioned model retraining techniques.

The empirical study was conducted by selecting the Yahoo A1 and the Numenta Anomaly Benchmark (NAB) popular operational datasets; the FFT, SR, PCI, LSTM-AE, and SR-CNN, five popular unsupervised anomaly detection models; and the FEDD concept drift detector.

The study shows that updated anomaly detectors generally perform better than those never updated. Models using a sliding-window approach benefit from identifying anomalies in the original time series, while the full-history approach benefits models transforming time series into another domain. Additionally, retraining anomaly detectors based on drift detector outputs improves performance compared to detectors that are never updated. The paper was presented at CAIN 2024 and is available at https://tinyurl.com/kwpn9pmk.

MIROSLAW STARON is a full professor in the Interaction Design and Software Engineering Division, Department of Computer Science and Engineering, Chalmers University of Technology and University of Gothenburg, 412 96 Gothenburg, Sweden. Contact him at https://www.staron.nu or miroslaw.staron@cse.gu.se.

SILVIA ABRAHÃO is a full professor of software engineering in the Department of Computer Systems and Computation, Universitat Politècnica de València, 46022 Valencia, Spain. Contact her at https://sabrahao.wixsite.com/dsic-upv or sabrahao@dsic.upv.es.

GRACE LEWIS is a principal researcher and the lead of the Tactical and AI-enabled Systems (TAS) initiative at the Carnegie Mellon Software Engineering Institute (SEI), Pittsburgh, PA 15213 USA. Contact her at glewis@sei.cmu.edu.

HENRY MUCCINI is a full professor of software engineering at the University of L'Aquila, 67100 L'Aquila, Italy. Contact him at henry.muccini@univaq.it.

CHETAN HONNENAHALLI is an engineering manager at Meta, Menlo Park, CA 94025 USA. Contact him at https://www.linkedin.com/in/hschetan/.

EDITORS: Silvia Abrahão, Universitat Politècnica de València, sabrahao@disc.upv.es
Miroslaw Staron, Chalmers University of Technology and University of Gothenburg, miroslaw.staron@cse.gu.se

This article originally appeared in
IEEE Software
vol. 41, no. 4, 2024

DEPARTMENT: PRACTITIONERS' DIGEST

Human Aspects and Security in Software Development

Miroslaw Staron, Silvia Abrahão, Birgit Penzenstaler, and Alexander Serebrenik

This edition of the "Practitioners' Digest" brings you recent papers on approaches to addressing selected human and technical aspects of software development, from finding security vulnerabilities and system-level testing, to understanding the impact of personality on requirements engineering activities, from the 46th ACM/IEEE International Conference on Software Engineering (ICSE 2024) and the 17th International Conference on Cooperative and Human Aspects of Software Engineering (CHASE 2024). Feedback or suggestions are welcome. In addition, if you try or adopt any of the practices included in the column, please send us and the authors of the paper(s) a note about your experiences.

IMPACT OF PERSONALITY ON REQUIREMENTS ACTIVITIES

Personality is a human aspect that influences the success of software projects, but few empirical studies have focused on the study of personality in requirements engineering (RE) activities. In the paper "What's Personality Got to Do With It? A Case Study on the Impact of Personality on Requirements Engineering-Related Activities," Dulaji Hidellaarachchi, John Grundy, Rashina Hoda, and Ingo Mueller conducted an exploratory case study on an 11-member software development team, observing 28 team meetings, conducting follow-up interviews, and analyzing team members' personality profiles using the IPIP-NEO 120 assessment tool developed based on the standard five-factor model of personality.

The findings of their study contribute to understanding the potential impact of the personalities of software team members on the way they carry out RE-related activities. Most team members exhibit high traits of agreeableness, conscientiousness, and openness to experience and medium traits of extraversion and neuroticism. Through six weeks of observations and six follow-up interviews, they investigated how different personality characteristics appear to impact RE-related activities.

They identified that software team members' personality characteristics related to agreeableness, conscientiousness, and openness to experience appear to positively influence the successful conduction of requirements validation, prioritization, requirement change management, obtaining clarity on requirements, and timely completion of allocated tasks. Some team members' reluctance to speak out or their desire to work alone, which can be identified as opposites of extraversion and agreeableness, appear to create difficulties in completing allocated tasks.

The authors identified a set of strategies followed by the team members to overcome perceived challenges that occurred due to the personality differences. The implications for practice are as follows:

- Software practitioners' responses to changes in requirements depend on their diverse personalities.
- Detailed discussions in story walk-through meetings are helpful in clarifying requirements.
- Collaboration is the key to successful completion of RE-related tasks.
- Assertiveness is helpful in managing stakeholder demands.

The paper was published as part of CHASE 2024. Access it at https://tinyurl.com/4uypjzvk.

Digital Object Identifier 10.1109/MS.2024.3385455
Date of current version: 12 June 2024

EXTRACTING TAINT SPECIFICATIONS FROM SOFTWARE DOCUMENTATION

Finding security vulnerabilities is an important, yet underestimated, task for software developers. It is especially important in products such as Android, where users store and process private information. Vulnerability detection is often performed using static analyzers or runtime profilers, which may require both access to source code and good test scenarios. However, in the paper "DocFlow: Extracting Taint Specifications From Software Documentation," Marcos Tileria, Jorge Blasco, and Santanu Kumar Dash offer an alternative. The starting point of this paper is that Android documentation contains a lot of syntactic and semantic information about application programming interface (API) classes, information that can be used to create information flows. Thanks to the advantages of natural language processing methods, the authors can parse the documentation and identify sensitive API methods.

The paper presents a tool capable of automatically identifying sensitive methods, which are those that store, process, or analyze user-sensitive information, such as geolocation or sensor readings. DocFlow is compared with two other tools, which use source code information, and the results show that it performs equally well or better, with an accuracy of 0.89 or higher, depending on the task. DocFlow is capable of handling software evolution by effectively modeling changes in the API documentation. Additionally, DocFlow is adaptable for performing more semantic analysis tasks (in addition to security vulnerability identification). The paper was published as part of the research track of ICSE 2024. Access it at https://tinyurl.com/4sayv6j4.

USING CLIPBOARD IN ANDROID APPS

"When in doubt, copy-paste" is a motto of many programmers, but copying and pasting is one of the activities everyone does on a daily basis; it is how we often store, process, and even share information. In the paper "Attention! Your Copied Data Is Under Monitoring: A Systematic Study of Clipboard Usage in Android Apps," Yongliang Chen, Ruoqin Tang, Chaoshun Zuo, Xiaokuan Zhang, Lei Xue, Xiapu Luo, and Qingchuan Zhao systematically study how we use the clipboard in Android apps. The paper presents a tool, ClipboardScope, that uses static program analysis to explore how clipboard data are used in mobile applications. The tool classifies each clipboard operation in one of four categories: "spot-on," which is the normal use of the data when the data are just copied and displayed on the screen; "grand-slam," which directly stores or sends out the data without examination; "selective," which stores or sends the data if they satisfy specific formats; and "cherry-pick," which stores or sends the extracted part of the clipboard data after content validation.

The tool was evaluated on 2.2 million apps from the Google Play repository. The study identified 2,253 apps that could transfer data or store them internally. In addition, the paper found that many programmers use the SharedPreferences object to store historical data, which can become an "unnoticeable privacy leakage channel." The paper was published as part of the research track of ICSE 2024. Access it at https://tinyurl.com/mtx2rnzx.

ENSURING THE QUALITY OF AUTONOMOUS VEHICLES

"Safety is our priority" is a guiding principle for many software engineers. We design software systems that control our cars, airplanes, heavy machinery, or even power grids. As the use cases in these domains become more complex, our software becomes larger, more complex, and less consistent. Current safety standards often prescribe white-box analysis for safety systems, such as the use of test coverage metrics during quality assurance. In the era of artificial intelligence and machine learning, these metrics may not be adequate. In the paper "Towards Reliable AI: Adequacy Metrics for Ensuring the Quality of System-Level Testing of Autonomous Vehicles," Neelofar Neelofar and Aldeida Aleti discuss the challenges of ensuring the reliability and safety of AI-powered systems, particularly autonomous vehicles (AVs), due to their complex nature.

They propose a new metric suite, Test Suite Instance Space Adequacy (TISA), that measures the adequacy of the test suite from a coverage and diversity perspective. For example, the area of instance space is a metric that measures the diversity of the test suite, the area of buggy region captures the ability of the test suite to detect defects, and the coverage of instance space measures how much of the theoretically calculated test space is covered by the test suite. These metrics are evaluated on two datasets from previous studies on search-based AV test selection, with over 30,000 test scenarios in total. The results show that all TISA metrics have a positive correlation with defects, and the area of buggy region stands out as the best. Although these metrics are evaluated on only two datasets, they present great promise for advancing the state of the art in AV testing. The paper was published as part of the research track of ICSE 2024. Access it at https://tinyurl.com/mpzmcrrw.

BILLS OF MATERIALS FOR SOFTWARE SYSTEMS

Modern software systems are constructed by combining existing libraries and frameworks with new functionality. This means that the supply chain of software systems is much longer than it used to be, and therefore, software development companies must prepare a software bill of materials (SBOM) when they ship their software. Such SBOMs are often long and require significant effort to prepare. In the paper "BOMs Away! Inside the Minds of Stakeholders: A Comprehensive Study of Bills of Materials for Software System" Trevor Stalnaker, Nathan Wintersgill, Oscar Chaparro, Massimiliano Di Penta, Daniel M. German, and Denys Poshyvanyk shed some light on the process of creating and maintaining SBOMs.

The paper studied 138 practitioners who identified twelve different challenges, such as the complexity of SBOM specifications, knowledge of which data fields to include, incompatibility among SBOM standards, keeping them up to date, or verifying the accuracy and completeness of SBOMs (to name just a few). In addition to the challenges, the study identifies solutions to the challenges. It also maps these challenges/solutions to roles in software development organizations. The study shows that in this important area, much remains to be done to ensure that SBOMs truly live up to the rationale behind them. The paper was published as part of the research track of ICSE 2024. Access it at https://tinyurl.com/4a6myxkk.

PYTHON OPEN SOURCE ECOSYSTEM

One of the most renowned cartoons dedicated to open source depicts an intricate collection of blocks representing all facets of modern digital architecture, resting precariously upon a slender bar of a project that "some random person in Nebraska has been thanklessly maintaining since 2003" (https://xkcd.com/2347/). This xkcd cartoon presents a puzzling paradox: why do many open source projects, despite their popularity and widespread adoption, rely on a very limited number of developers to maintain them? In the paper "Novelty Begets Long-Term Popularity, but Curbs Participation: A Macroscopic View of the Python Open-Source Ecosystem," Hongbo Fang, James Herbsleb, and Bogan Vasilescu provide a compelling insight into this paradox, suggesting that the innovativeness of a project is a critical factor. On one hand, more innovative, and hence atypical, projects attract more attention from users, possibly due to their competitive advantage over other readily available projects. On the other hand, maintaining less typical projects proves to be more challenging and attracts fewer developers capable of or interested in undertaking the task.

This conclusion stems from an empirical study of 1,055 open source Python packages hosted on GitHub. The packages were analyzed in terms of their typical usage of other packages, GitHub stars, number of downloads (both reflecting popularity), and developer count. The statistical analysis revealed several key insights:

- Project atypicality is positively associated with higher GitHub star counts in the long term but may result in a lower number of stars in the short term.
- No significant association was observed between project atypicality and project download counts.
- When controlling for project size, novel projects tend to attract fewer developers. However, this correlation is reversed when project size is not included as a control variable.
- A one-standard-deviation increase in the typicality score reduces the probability of project abandonment to 95% of its original value.

The findings of this study are important both for open source projects and for closed source projects that depend on them. For the latter, the dilemma arises: when selecting among several open source alternatives, should one opt for the most popular option, risking its sustainability, or choose an alternative with a larger development base, potentially less popular or innovative? For the former, the challenge lies in mitigating the negative consequences of being more innovative, such as fewer contributors available for maintenance tasks, while preserving the positive outcomes, such as higher popularity. The paper was published as part of the research track of ICSE 2024. Access it at https://tinyurl.com/563e7b8u.

MIROSLAW STARON is a professor in the Interaction Design and Software Engineering Division, Computer Science and Engineering Department, Chalmers University of Technology and the University of Gothenburg, SE-412 96 Gothenburg, Sweden. Contact him at https://www.staron.nu or miroslaw.staron@cse.gu.se.

SILVIA ABRAHÃO is a full professor in the Department of Computer Science, Universitat Politècnica de València, 46022 Valencia, Spain. Contact her at https://sabrahao.wixsite.com/dsic-upv or sabrahao@dsic.upv.es.

BIRGIT PENZENSTALER is an associate professor at Chalmers University of Technology and the University of Gothenburg, 412 96 Gothenburg, Sweden. Contact her at https://www.chalmers.se/en/staff/pages/birgit.aspx or birgitp@chalmers.se.

ALEXANDER SEREBRENIK is a professor at Eindhoven University of Technology, 5600 MB Eindhoven, The Netherlands. Contact him at a.serebrenik@tue.nl.

EDITOR: Mark Campbell, EVOTEK, mark@evotek.com

This article originally appeared in
Computer
vol. 56, no. 6, 2023

DEPARTMENT: IT INNOVATION

The Road to Decentralized Identity: The Techniques, Promises, and Challenges of Tomorrow's Digital Identity

Mark Campbell, *EVOTEK*

As society increasingly relies on digital services, identity management becomes increasingly vital. Decentralized identity offers a novel approach to address today's identity challenges, putting users in control of their own digital identities and personal data.

As society increasingly relies on digital services, secure and reliable identity management becomes increasingly vital. Traditional centralized identity (CID) systems have been plagued by data breaches, identity theft and loosely controlled usage of users' personal information. Decentralized identity (DID) offers a novel approach to address these challenges by putting users in control of their own digital identities.

DIGITAL SERVICES AND DIGITAL IDENTITY

Online transactions via digital services have become ubiquitous and virtually invisible in today's online world, at least until they fail or are misused by bad actors. To securely complete a digital transaction, a user (that is, a person, system, application, or device) and the service provider must complete an intricate handshake to verify user identity and digital service permission.

At the core of this interaction is the digital identity, which ideally must be[1]:

- *Convenient:* It must be easy to use, easy to remember, portable between devices, and frictionless to the user.
- *Private:* Identity information must be kept private from parties outside the transaction.
- *Secure:* Identity verification must be secure from eavesdropping, breaches, spoofing, and unauthorized replication, and leave no transaction traces for bad actors to exploit.
- *Scalable:* Identity verification must be fast and highly scalable.

A digital identity can be verified by something we[2]:

- know (for example, password, mother's maiden name)
- have (for example, physical key, mobile device)
- are (for example, our fingerprint, other biometrics)
- are temporarily granted (for example, expiring one-time passcode).

The username-password model is today's most common identity verification technique, but it's often augmented with other technologies, such as single sign on (SSO), one-time passcode, multifactor authentication, and passwordless techniques, to reduce user friction and increase security.

CID

Today's standard form of identity management is CID, in which digital services are controlled by a single central service provider platform. A user registers their identity (for example, username) and proof

Digital Object Identifier 10.1109/MC.2023.3263020
Date of current version: 31 May 2023

(for example, password) with the service provider, who maps them to the services they are allowed to execute.[3] When a user application requests a service from the service provider, they assert their digital identity and provide proof of this assertion. The service provider authenticates the identity with the proof and verifies if the requested service use is authorized. If so, the service provider executes the service and provides the results back to the user application (Figure 1).

CID frameworks have existed since the early days of digital computation but have evolved over the decades.

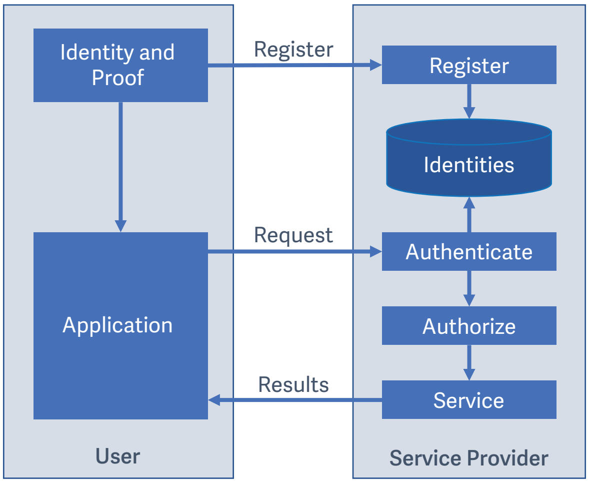


FIGURE 1. CID

Federated identity

Multiple service providers can form a "federation" of trust by linking their authentication systems to allow users to access multiple services with a single set of login credentials, thus eliminating multiple usernames and passwords.[1] This "federated identity" is often built on a hub-and-spoke model with a central identity hub (for example, Google, Meta, Microsoft) responsible for identity authentication, and other "spoke" federation members maintaining their own authorization maps. A big advantage is that new federation members need only plug into the existing hub via a published application programming interface (API) to join a federation.[4]

Passkeys

A "passkey" system, formally called a *fast identity online* or *FIDO* multidevice credential and built on the FIDO2 *webauthn* standard, is a cryptographic entity—invisible to the user—and used in place of a password.[5] A passkey uses a public key registered with the service provider and a private key held only by the user. Using industry-standard public key infrastructure (PKI), users get a seamless identity verification process on multiple devices much more securely than traditional passwords. Passkeys can be stored and verified by a central federated identity hub like those hosted by Apple, Google, and Microsoft.[6]

ID as a service

An ID as a service (IDaaS) has a central identity server manage and verify identities of subscribing service providers. IDaaS solutions are commonly hosted as a multitenant framework in a public cloud.

A CID example: India's unified payments interface

Launched in 2016, India's unified payments interface (UPI) merges traditional payment applications and paper processes into a single platform. The UPI system integrates banks, payment systems, and merchants, allowing users to access a wide range of payment services with a single interface. Easy integrating with UPI has spurred 50% digital payment volume growth per year for the past 5 years.[7] UPI is based on a collection of uniquely Indian innovations, including:

- *The India Stack:* This is a collection of open APIs allowing institutions to integrate with UPI systems.
- *No-frills accounts:* Introduced in 2005 by the Reserve Bank of India, these basic bank accounts charge no fees for most transactions

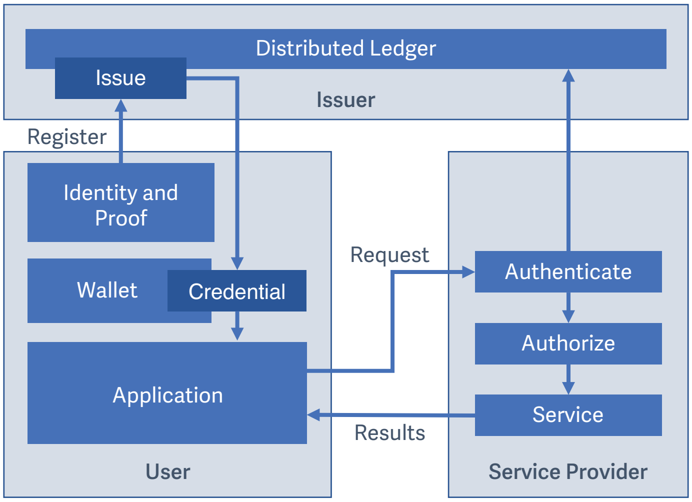


FIGURE 2. DID

and require no minimum balance, thereby granting financial access to low-income populations.[8]
- *Aadhaar:* Hindi for *foundation*, Aadhaar is a free unique 12-digit random number issued by the Unique Identification Authority of India. Residents need only minimal demographic and biometric information to receive their Aadhaar, which, once coupled with their bank, enables payment transactions anywhere from stock markets to street stalls.[1,9]

Rapid growth of the UPI system has raised concerns about security, privacy, and fraud, with reports of phishing attacks, unauthorized transactions, and data breaches.[10]

CID shortcomings

While CID solutions have been deployed for decades and handle nearly all current digital transactions, they have several inherent shortcomings:

- *Trust and data breaches:* Users must trust the central service provider's availability, integrity, and confidential treatment of personal information.[3] Unfortunately, most CID platforms (for example, Facebook, Google) have been victims of serious data breaches, rendering their best security intentions moot.[11]
- *Privacy and fraud:* CID approaches require users to surrender aspects of their privacy.[11] Should this information be divulged through a data breach, internal bad actor, or accidental disclosure, the user's privacy is forfeit. On the user side, credentials can be stolen, replicated, or phished by fraudsters.
- *Convenience:* CID systems require users to authenticate identity for each accessed service provider. While passkeys, federation, and SSO solutions ease user friction, multiple identity credential must be created and remembered for each identity federation accessed.[12]

Passkeys, federated identity, IDaaS, passwordless, and SSO techniques are continually advancing to ensure the longevity of CID solutions. However, CIDs' shortcomings have prompted many to seek alternative DID platforms.

DID

DID is a user-centric identity management approach where individuals control their identity data and disclose only select information to specific service providers. The user registers personal information (for example, name, age, credit card number) along with proof (for example, private key) with an independent issuer. The issuer records this personal information in a distributed ledger (for example, blockchain) and returns a signed credential to the user to store in a digital wallet. When the user's application requests a service from the service provider, it passes the appropriate signed credential for only the identity items the service requires (for example, credit card number but not age). The service provider then verifies the credential with the distributed ledger, verifies the identity is authorized to use the service, and returns service results to the user (Figure 2).

Two central components of DID are the *digital wallet* and *distributed ledger*. The digital wallet stores user information (for example, name, age, address, citizenship, credit card number) in an unphishable cryptographic credential created and signed by the issuer.[12] The distributed ledger is most commonly built on blockchain, but increasingly other frameworks, like distributed file systems and hashgraphs. Emerging distributed ledger technology (DLT) solutions include Microsoft ION, Hyperledger Indy, and the Tangle Identity framework.[3] Today's thriving DID community includes Bitnation, Civic, EverID, IDchainZ, LifeID, Self-Key, ShoCard, Sovrin, THEKEy, and uPort.[13]

Several international organizations, such as the World Wide Web Consortium, Decentralized Identity Foundation, and European Digital Identity are striving to establish a new DID ecosystem through published standards and frameworks.[3]

Self-sovereign identity (SSI) is often used in connection with DID. Each describes similar but not equivalent identity perspectives. DID details identity management without relying on a centralized verification authority,[14] while SSI focuses on users retaining control over their identity and digital footprint, with or without DID.

DID benefits

DID can overcome major shortcomings of CID, such as:

- *Trust and data breaches:* DID eliminates the single point of compromise found in CID's central or federated service provider.[3] Should a DID service provider be breached, little to no personally identifiable data can be exploited. A growing number of users trust global distributed ledgers more than a service provider's assurance that proper security is in place.
- *Privacy and fraud:* Because only pertinent identity items are used for each transaction, user data privacy is more easily controlled and secured. With total control of their identity items, users can grant or revoke access to service providers as needed and minimize identity theft.
- *Convenience:* A user only registers identity items once, then grants access to service providers as needed, eliminating multiple identity registrations, usernames, and passwords.

DID challenges

Despite the technical solutions and promising characteristics of DID, widespread adoption faces significant challenges:

- *Interoperability:* The DID ecosystem is built on a common issuer and secure digital wallet. Today's proliferation of DID frameworks makes integration by service providers an arduous task.[12] In the coming years, consortiums and standards will coalesce into a common DID ecosystem much like India's UPI in the CID space.
- *Scalability:* Supporting a global population of identity credentials, service requests, and verifications requires a gargantuan underlying DLT. It's not certain if current DLT technologies can achieve this scale performantly and cost effectively.[13]
- *Governance:* Identity governance programs, like the European Union's General Data Protection Regulation (GDPR) are built on CID assumptions. DID will require many regulations be overhauled. For instance, GDPR regulates the user's "right to be forgotten," mandating service providers delete user data and identity on request. This is simply not possible or applicable to service providers in a DID regime.[2]
- *Service provider resistance:* Today's global service providers (for example, Meta, Google, Amazon, Microsoft) operate platforms dependent on direct user data or indirect user behavior. DID removes this asymmetric control and greatly restricts user data monetization, making global services providers resistive towards DID.

Like any disruptive technology, there is initial resistance to change. There will undoubtedly be security vulnerabilities, performance bottlenecks, scaling issues, fear, uncertainty, and doubt encountered with DID adoption. However, user benefits, technical advancement, deployment framework maturation, and societal pressure will make DID more palatable.

TOMORROW'S DIGITAL IDENTITY

Traditional CID remains the default identity approach and newer CID techniques, like passkeys, will continue to develop. However, DID and SSI solutions will soon mature, proliferate, and become the standard identity framework.

Beyond adopting DID platforms, other techniques and features will be developed to augment digital identity, including:

- *Zero-knowledge proofs (ZKP):* This cryptographic technique proves digital identity without the user revealing private information. When coupled with a DID approach, ZKP offers a novel alternative to passwords and maintains user control over private data.[15] Initial use cases are being explored in finance, health care, commerce, education, smart city, and travel industries.[1]
- *Nonrepudiation:* This technique encrypts a transaction so neither the sender nor recipient of a message can deny its creation, transmission, or receipt. However, in practice, nonrepudiation is left up to the DLT used by the issuer. DID implementations will increasingly build nonrepudiation into their digital transaction lifecycle.
- *Fault tolerance:* It's difficult for today's DID solutions to recover from breakdowns in the middle of a transaction. Future frameworks will be more resistant to mid-transaction failures and automate the continuation, reconstruction, or rollback of transactions interrupted in flight.
- *Multisource identity:* Future identity systems will incorporate multiple indirect user sources (for example, cell phone mast data, behavior patterns, digital idiosyncrasies) to more accurately verify identity and lower user friction.[2]
- *Quantum safe:* Today's DID platforms are built on PKI. However, quantum computing advancements could render PKI-based solutions readily transparent to decryption. Other encryption techniques—like lattice-based encryption, code-based encryption, multivariate polynomial cryptography, and hash-based signatures—are much more resistant to quantum computing.[16] These "postquantum" or "quantum-safe" techniques will replace PKI in DID solutions.
- *Regulations and governance:* Regulations lag greatly compared to technical advancements, but DID and SSI techniques will require governance bodies to rethink and redraft polices and regulations.

DID solutions can fundamentally transform how users interact with digital services. By offering enhanced security, privacy preservation, and self-sovereignty, DIDs can address many limitations inherent in traditional centralized identity systems. As DID technology matures and gains wider acceptance, it will have far-reaching implications across industries as it streamlines processes, increases security, and improves trust.

REFERENCES

1. M. Bajaj, private communications, Jan. 19, 2023.
2. M. Kennedy, private communications, Jan. 19, 2023.
3. Š. Čučko and M. Turkanović, "Decentralized and self-sovereign identity: Systematic mapping study," *IEEE Access*, vol. 9, pp. 139,009–139,027, Oct. 2021, doi: 10.1109/ACCESS.2021.3117588 [Online]. Available: https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=9558805
4. R. Broeckelmann, "Authentication vs. Federation vs. SSO," *Medium*, Sep. 2017. [Online]. Available: https://medium.com/@robert.broeckelmann/authentication-vs-federation-vs-sso-9586b06b1380>
5. V. Bertocci. "Our take on passkeys." Auth0. Accessed: Mar. 16, 2023. [Online]. Available: https://auth0.com/blog/our-take-on-passkeys/
6. B. Collins, "Why passkeys from Apple, Google, Microsoft may soon replace your passwords," *CNBC*, Feb. 2023. [Online]. Available: https://www.cnbc.com/2023/02/11/why-apple-google-microsoft-passkey-should-replace-your-own-password.html
7. J. Kearns and A. Mathew, "Explained: How India is outpacing the world in digital payments," *Int. Monetary Fund*, Oct. 2022. [Online]. Available: https://www.imf.org/en/News/Articles/2022/10/26/cf-how-indias-central-bank-helped-spur-a-digital-payments-boom
8. "What is no-frills account – Its eligibility, documents required and how to apply." Navi. Accessed:Mar. 16, 2023. [Online]. Available: https://navi.com/blog/no-frills-account/#:~:text=No%20Frills%20Account%20is%20a,who%20have%20low%2Dincome%20backgrounds
9. "About your Aadhaar." Government of India. Accessed: Mar. 16, 2023. [Online]. Available: https://uidai.gov.in/en/my-aadhaar/about-your-aadhaar.html
10. A. Mukherjee, "A little epoxy can unglue India's welfare system," *Bloomberg*, Jun. 2022. [Online]. Available: https://www.bloomberg.com/opinion/articles/2022-06-23/india-s-aadhaar-id-system-delivers-benefits-but-is-at-risk-of-widespread-fraud?leadSource=uverify%20wall

11. C. Schram, "A future built on decentralized identity," *Bloom*, Dec. 2021. [Online]. Available: https://bloom.co/blog/a-future-built-on-decentralized-identity/
12. D. Shou, "How decentralized identity is reshaping privacy for digital identities," *Forbes*, Dec. 2021. [Online]. Available: https://www.forbes.com/sites/forbestechcouncil/2021/12/10/how-decentralized-identity-is-reshaping-privacy-for-digital-identities/?sh=2893b82c3226
13. O. Dib and K. Toumi, "Decentralized identity systems: Architecture, challenges, solutions and future directions," *Ann. Emerg. Technol. Comput.*, vol. 4, no. 5, pp. 19–40, Dec. 2020, doi: 10.33166/AETiC.2020.05.002. [Online]. Available: https://www.researchgate.net/publication/347753956_Decentralized_Identity_Systems_Architecture_Challenges_Solutions_and_Future_Directions
14. G. Weston, "Self sovereign identity & decentralized identity – An unlimited guide," *101 Blockchains*, Jul. 2022. [Online]. Available: https://101blockchains.com/self-sovereign-identity-and-decentralized-identity/
15. "Zero knowledge identity proof," Identity Management Institute, Chatsworth, CA, USA, 2023. [Online]. Available: https://identitymanagementinstitute.org/zero-knowledge-identity-proof/
16. "Quantum computing and postquantum cryptography," National Security Agency, Washington, DC, USA, Aug. 2021. [Online]. Available: https://media.defense.gov/2021/Aug/04/2002821837/-1/-1/1/Quantum_FAQs_20210804.PDF
17. A. Sidana, private communications, Jan. 17, 2023.
18. C. Hughes, "IDaaS explained: How it compares to IAM," *CSO*, May 2022. [Online]. Available: https://www.csoonline.com/article/3660554/idaas-explained-how-it-compares-to-iam.html
19. J. S. Savariraj and S. De Simone, "An introduction to post-quantum public key cryptography," *InfoQ*, Feb. 2022. [Online]. Available: https://www.infoq.com/articles/post-quantum-cryptography-introduction/

MARK CAMPBELL is the chief innovation officer at EVOTEK, San Diego, CA 92121 USA. Contact him at mark@evotek.com.

EDITORS: Alfred John Menezes, ajmeneze@uwaterloo.ca
Douglas Stebila, dstebila@uwaterloo.ca

DEPARTMENT: CRYPTOGRAPHY

This article originally appeared in
IEEE SECURITY & PRIVACY
vol. 22, no. 6, 2024

Threshold Signatures

Chelsea Komlo, *University of Waterloo*

Threshold signatures are a helpful cryptographic primitive to ensure redundancy and distribution of trust for a secret signing key. In this article, we introduce the concept of threshold signatures, discuss practical use cases of threshold signatures, and review considerations for practitioners when deploying threshold signatures in the real world.

A threshold signature scheme is a special case of a multiparty computation (MPC) functionality. MPC is a generic mechanism to securely partition a general functionality among a set of players so that all players must participate to perform the operation, and up to some subset of players is assumed to be corrupted.

WHAT IS A THRESHOLD SIGNATURE SCHEME?

Importantly, an MPC operation can be performed without parties revealing their inputs to each other while still issuing a valid output created from each of their respective (private) inputs. A threshold signature scheme allows for partitioning the functionality of issuing a digital signature among a set of mutually untrusted parties.

Recall that a single-party signature scheme offers the security property of unforgeability; if a signature is issued, with overwhelming probability, the signature was generated by an entity that held possession of the private key corresponding to the public key. In addition, at a high level, a threshold signature scheme allows for two security properties: 1) redundancy and 2) corruption resilience. A threshold signature scheme offers resilience in that it allows a threshold t number of parties out of n possible parties to issue a signature under a joint public key. A threshold signature scheme offers corruption resilience in that if at most $(t - 1)$ parties are corrupted (such as by having their private keys stolen), then the threshold signature scheme remains unforgeable.

In a little more detail, a threshold signature scheme is composed of n total possible signers. At the time of key generation, each party is given a secret key share that corresponds to a share of a single joint secret key (which is unknown to any party). The set of signers is represented by a single joint public key. Then, at the time of signing, a coalition of at least t signers participates in a signing protocol. The output from this signing protocol is a joint signature over a message. Similarly to a single-party signature scheme, the joint signature is valid under the joint public key.

HOW IS KEY GENERATION PERFORMED FOR THRESHOLD SIGNATURES?

Key generation for threshold signatures can be performed in two different manners, differing in the required trust assumptions and resulting complexity. Recall that the output from key generation for a threshold signature is a joint public key and secret signing key shares for each participant.

Trusted Dealer

First, key generation can be performed by a trusted dealer, which is a single entity that is trusted as it is required to follow the protocol honestly and to delete secret key material after the dealer has performed its duties. The trusted dealer performs the key generation operation directly and sends each party their respective secret key share and the joint public key. Under the

Digital Object Identifier 10.1109/MSEC.2024.3463915
Date of current version: 11 November 2024

hood, many schemes simply require the trusted dealer to perform Shamir secret sharing, and so implementing this operation is fairly straightforward and can be done noninteractively. However, the use of a trusted dealer introduces a single point of failure (albeit only at the time of key generation) and so involves additional risks.

Distributed Key Generation

To move to a setting without a single point of failure, key generation can itself be distributed into a multiparty protocol, where each party is equally trusted. A multiparty protocol to generate key material for a threshold signature scheme is referred to as a *distributed key generation* (*DKG*) protocol. At a high level, many DKG protocols have each participant perform an instance of a trusted dealer operation and then involve an "aggregation" step at the end to aggregate key material from each instance into a single public key and set of secret key shares.

Note that in the DKG setting, no party learns the resulting secret key; each party learns only the public key and its respective secret key share. As such, so long as fewer than (t – 1) parties are corrupted; then, no party can recover the joint secret key share.

Several well-established DKGs exist for discrete-logarithm-based signature schemes.[1] The protocols generally require participants to be authenticated by a preexisting public-key infrastructure and to communicate over a broadcast channel, which is assumed to be authenticated and consistent. In practice, many implementations use "echo-broadcast" techniques to ensure the consistency of participant views.[2] Moreover, a "coordinator" role can be used to coordinate messages exchanged between participants to avoid the complexity of point-to-point communication.

HOW ARE THRESHOLD SIGNATURES USED IN PRACTICE?

Threshold signatures are important in practice due to their ability to offer improved protections for redundancy and corruption resilience. For example, in the setting for single-party signatures, if a machine that held a secret signing key were to fail, then the ability to issue signatures would no longer be available. If a threshold signature scheme were to be used, the signing capability could simply be rerouted to other signers.

While the previous example could also be accommodated in the single-party setting simply by copying the private key among a set of machines, doing so increases the exposure footprint of the private key and therefore the probability that it might be compromised. With this, the second property of threshold signatures comes into play. For a threshold signature scheme, even if (t – 1) machines holding (t – 1) secret key shares were to be compromised, the (t – 1) malicious machines would not have the ability to issue valid signatures without enlisting the help of another uncorrupted machine. Because a threshold signature scheme requires t participating signers, then a threshold signature scheme protects against possible corruptions, up to t corruptions.

> *A THRESHOLD SIGNATURE SCHEME ALLOWS FOR PARTITIONING THE FUNCTIONALITY OF ISSUING A DIGITAL SIGNATURE AMONG A SET OF MUTUALLY UNTRUSTED PARTIES.*

Settings where threshold signatures are particularly useful are where the signing authority is particularly trusted. For example

1. *Cryptocurrency Wallets:* In the setting of a cryptocurrency wallet, a single fatal signature can drain a user's funds completely. As such, threshold signatures allow for multifactor authentication at the cryptographic level so

that it is impossible to issue a valid signature without a sufficient number of participating signers. Furthermore, threshold signatures allow cryptocurrency users to maintain some amount of control over their funds while likewise allowing services such as cryptocurrency exchanges to help clients recover in the event that the client loses their secret key share.

2. *Certificate Authorities:* A signing key for a certificate authority is a highly valuable asset, and the compromise of a signing key can be both costly and difficult to remediate. A certificate authority could partition its signing key among many machines to help protect against attacks.
3. *Consensus Validators:* Distributed networks generally require the ability to perform consensus validation, where a set of parties can attest to the current state of the network. Threshold signatures allow such parties to issue a compact signature as opposed to each party issuing their own signature, which scales linearly in the number of validating parties.
4. *Code-Signing Packages:* When distributing software packages and updates, it is possible to sign the code or binary with a certificate to ensure that the binary was not tampered with before installation on a client's device. To distribute trust in the signing key, a threshold signature scheme can be used to partition the signing key among a set of signing machines.

CASE STUDY: THRESHOLD ELLIPTIC-CURVE DIGITAL SIGNATURE ALGORITHM AND THRESHOLD BLS

An ideal characteristic for threshold signatures is to be compatible with existing single-party verification algorithms. We will then review two practical examples of threshold signatures.

FROST: Flexible Round-Optimized Schnorr Threshold Signatures

One simple example we have in practice is FROST (Flexible Roud-Optimized Schnorr Threshold Signatures),[3] which allows for issuing a Schnorr signature using a threshold number of signers. Furthermore, because the Edwards-curve Digital Signature Algorithm (EdDSA) is a variant of Schnorr, FROST can issue signatures that are compatible with the EdDSA verification algorithm.

The signing protocol for FROST can be done in either two online rounds or one online round after performing a batched preprocessing phase. FROST protects against a class of concurrency attacks,[4] which prior threshold Schnorr signature schemes did not consider. Concurrency attacks are possible when signers can open many signing sessions in parallel, giving an adversary the option to choose which signing sessions to terminate and which sessions to complete.

Public keys for FROST are identical to public keys for EdDSA. However, each signer receives a Shamir secret share of the corresponding private key. Then, signatures can be issued using either a two-round interactive online protocol or a one-round noninteractive protocol after a batched preprocessing phase. Either participants can perform in a broadcast manner by sending all messages to each other, or a centralized untrusted coordinator can be used to route messages. The output signature is identical to a single-party EdDSA signature, and so verification of the signature is exactly the EdDSA verification algorithm.

FROST is not robust, meaning that if even one signer fails to produce a signature share, then the protocol must be rerun. For some applications, the probability of such a failure may be minimal, such as the setting where only a small number of signers participate. However, for other applications, such as when the number of signers is large, then robustness is a desirable property. In this setting, ROAST (Asynchronous Schnorr Threshold Signatures)[5] is a wrapper protocol to use in conjunction with FROST to allow for efficient robustness while treating the underlying signature scheme as a black box.

FROST has been specified in an Internet Engineering Task Force informational draft and is fast nearing completion. For those interested in implementing FROST, this is a good reference point to begin with.

Threshold BLS

The BLS signature scheme[6] allows for an even more efficient threshold scheme.[7] Key generation for threshold BLS is similar to FROST in that the public key remains the same as single-party BLS and each signer receives a Shamir secret share of the signing key.

Signing for threshold BLS, however, can be performed in a single noninteractive communication round. Each party, in fact, performs the single-party BLS signing operation but additionally adds a Lagrange coefficient to allow for interpolating the Shamir secret share. Again, similarly, the output signature is identical to a single-party BLS signature, and so verification is identical to single-party BLS.

Interestingly, threshold BLS is robust. If any party fails to produce a signature share, then as long as t parties respond to a signing query for a particular message, a valid signature can be produced by computing different Lagrange coefficients for that particular subset.

WHAT TO CONSIDER IN PRACTICE

While threshold signatures have many upsides, there are some additional factors to consider when adopting threshold signatures into applications. A couple of important questions to consider when moving an application to use threshold signatures include how to manage backups, key rotation, and what corruption threshold is in fact important for that particular application.

How Many Signers to Choose

First, one question to consider is the number of signing parties required. For some applications, a simple two-of-three settings is perfectly fine, with t = 2 being the number of required signers and n = 3 being the total number of possible signers. In this setting, one corrupted party is tolerated but no more. Such a configuration is common for cryptocurrency wallets.

However, other settings may wish to have a higher corruption threshold and can tolerate having more signers participating in a signing protocol. For example, an application that uses threshold signatures to validate a network consensus might wish to have a higher number of signers and corruption threshold.

How to Back Up Signing Key Shares

Another question to consider is how backups are maintained. While maintaining backups is an important practice, it is important that backups retain the same trust assumptions. For example, to retain the corruption-resilience property, backups of secret signing shares should also be partitioned. Otherwise, the corruption of where backups are maintained could recover all shares in a single attack, defeating the purpose of using a threshold scheme to partition signing shares.

How to Perform Key Rotation

Key rotation is an important defense-in-depth measure and helps protect against corruption attacks. A nice feature of threshold signatures is that secret signing key shares can be rotated while the public key remains the same, which is not possible in the single-signer setting. Such a feature is called *proactive security*. Key rotation has been documented extensively in prior literature[8] and implemented in practice.

> *AT A HIGH LEVEL, A THRESHOLD SIGNATURE SCHEME ALLOWS FOR TWO SECURITY PROPERTIES: 1) REDUNDANCY AND 2) CORRUPTION RESILIENCE.*

How to Manage the State

For multiround signature schemes such as FROST, signers must generate fresh randomness for each new signing session and then store this randomness within its state for the remainder of the session. After performing the final round of signing, it is critical that signers delete this state. Furthermore, it is important that signers do not reuse the state across different sessions. As such, implementers must be sure that the state is securely managed, a challenge that grows harder when signing sessions can be performed concurrently. To ensure safety even in a concurrent setting, managing the state can be performed using a secure data store with a locking mechanism.

FUTURE DEVELOPMENTS

It is an exciting time for threshold signatures, and we expect the research and use of threshold signatures to only grow. One exciting development to watch out for is the NIST effort[9] to formalize the existing threshold signature scheme and move toward a standardization process. This process will help produce concrete standards for a range of threshold signature schemes, including for both EdDSA and the Elliptic Curve Digital Signature Algorithm (ECDSA).

Furthermore, we have seen promising research emerge in the area of quantum-secure threshold signatures, and we expect to see even more. This is an open and evolving research area, so while it will be some time before it is clear which schemes are ready for use in practice, we expect a range of schemes to emerge over the next several years.

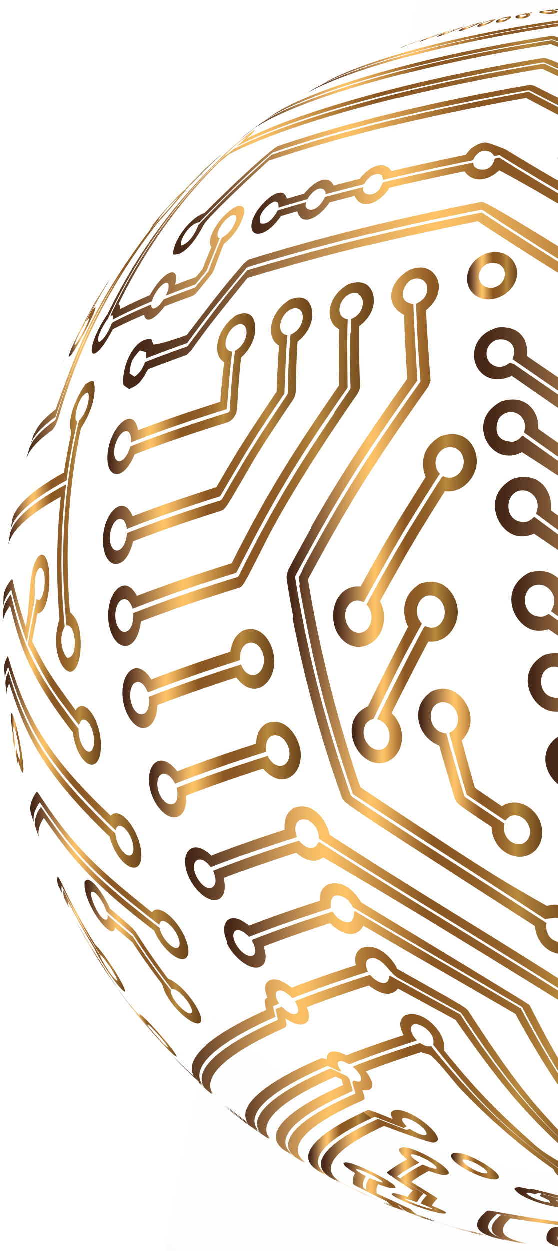
Overall, threshold signatures offer a practical mechanism to partition authority among a set of mutually untrusted entities in such a way that offers improved robustness and corruption resilience. While some threshold signatures do introduce additional performance overhead when performing signing, due to the requirement that signers cooperate to generate a signature, this latency can offer improved defense-in-depth measures against corruption. We hope to see threshold signatures continue to be adopted in implementations across a range of use cases.

REFERENCES

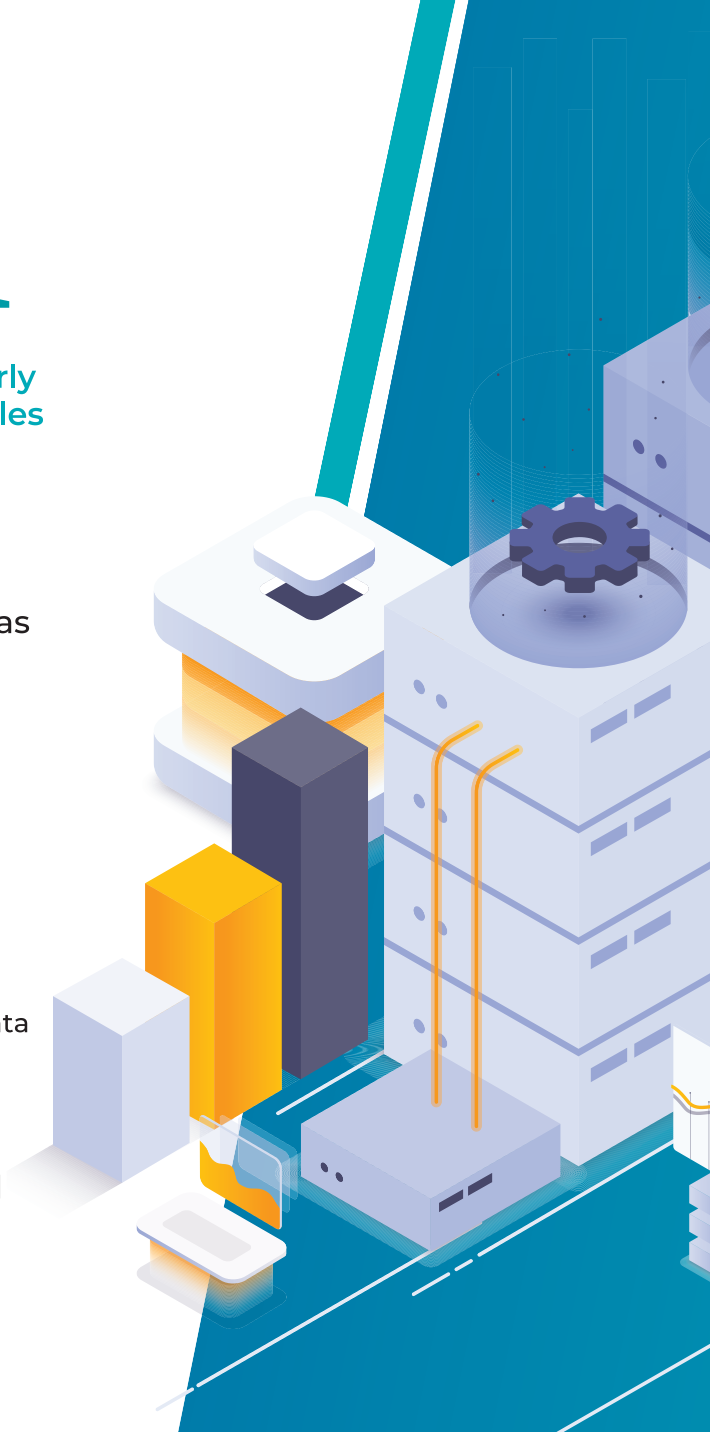
1. R. Gennaro, S. Jarecki, H. Krawczyk, and T. Rabin, “Secure distributed key generation for discrete-log based cryptosystems,” *J. Cryptol.*, vol. 20, no. 1, pp. 51–83, 2007, doi: 10.1007/s00145-006-0347-3.
2. S. Goldwasser and Y. Lindell, “Secure multi-party computation without agreement,” *J. Cryptol.*, vol. 18, no. 3, pp. 247–287, 2005, doi: 10.1007/s00145-005-0319-z.
3. C. Komlo and I. Goldberg, “FROST: Flexible round-optimized Schnorr threshold signatures,” in *Proc. Int. Conf. Sel. Areas Cryptogr.*, 2020, pp. 34–65.
4. F. Benhamouda, T. Lepoint, J. Loss, M. Orrù, and M. Raykova, “On the (in)security of ROS,” *J. Cryptol.*, vol. 35, no. 4, 2022, Art. no. 25, doi: 10.1007/s00145-022-09436-0.
5. T. Ruffing, V. Ronge, E. Jin, J. Schneider-Bensch, and D. Schröder, “ROAST: Robust asynchronous Schnorr threshold signatures,” in *Proc. ACM SIGSAC Conf. Comput. Commun. Security (CCS)*, 2022, pp. 2551–2564, doi: 10.1145/3548606.3560583.
6. D. Boneh, B. Lynn, and H. Shacham, “Short signatures from the Weil pairing,” *J. Cryptol.*, vol. 17, no. 4, pp. 297–319, 2004, doi: 10.1007/s00145-004-0314-9.
7. A. Boldyreva, “Threshold signatures, multi-signatures and blind signatures based on the gap-Diffie-Hellman-group signature scheme,” in *Public Key Cryptography*, Y. G. Desmedt, Ed., Berlin, Heidelberg: Springer, 2003, pp. 31–46.
8. T. M. Laing and D. R. Stinson, “A survey and refinement of repairable threshold schemes,” *J. Math. Cryptol.*, vol. 12, no. 1, pp. 57–81, 2018, doi: 10.1515/jmc-2017-0058.
9. L. Brandão and R. Peralta, “NIST first call for multi-party threshold schemes,” Nat. Inst. Standards Technol., Gaithersburg, MD, USA, 2023. [Online]. Available: https://nvlpubs.nist.gov/nistpubs/ir/2023/NIST.IR.8214C.ipd.pdf
10. D. Connolly, C. Komlo, I. Goldberg, and C. A. Wood, “Two-round threshold Schnorr signatures with FROST,” IETF Datatracker, RFC 9591, Jun. 2024. [Online]. Available: https://www.rfc-editor.org/rfc/rfc9591.html

CHELSEA KOMLO is a staff research scientist at the University of Waterloo, Waterloo, ON N2L3G1, Canada. Her research interests include threshold cryptography, public-key cryptography, secret sharing, post-quantum cryptography, and multiparty computation. Komlo received a Ph.D. in computer science from the University of Waterloo. Contact her at contact@chelseakomlo.com.

E-Currency and Fairness ▪ Ransomware Defense ▪ A National Cybersecurity Policy
IEEE SECURITY & PRIVACY
Building Dependability, Reliability, and Trust
SECURITY & PRIVACY
Digital Forensics Part 2
The NICE Challenge Project ▪ Improving Data Breach Notifications ▪ AI Models
IEEE SECURITY & PRIVACY
Building Dependability, Reliability, and Trust
Software and Cybersecurity ▪ Big Data: Privacy Versus Accessibility ▪ Resiliency in Cloud Computing
IEEE SECURITY & PRIVACY
Building Dependability, Reliability, and Trust



IEEE
COMPUTER ARCHITECTURE
LETTERS



IEEE TRANSACTIONS ON
BIG DATA
A publication of the IEEE Computer Society

Conference Calendar

IEEE Computer Society conferences are valuable forums for learning on broad and dynamically shifting topics from within the computing profession. With over 200 conferences featuring leading experts and thought leaders, we have an event that is right for you. Questions? Contact conferences@computer.org.

JANUARY

3 January

- VLSID (Int'l Conf. on VLSI Design & Int'l Conf. on Embedded Systems), Pune, India

14 January

- ICOIN (Int'l Conf. on Information Networking), Hanoi, Vietnam

26 January

- AIxVR (IEEE Int'l Conf. on Artificial Intelligence and eXtended and Virtual Reality), Osaka, Japan

31 January

- CGO (IEEE/ACM Int'l Symposium on Code Generation and Optimization), Sydney, Australia
- HPCA (IEEE Int'l Symposium on High Performance Computer Architecture), Sydney, Australia

FEBRUARY

2 February

- AIxDKE (Int'l Conf. on AI x Data and Knowledge Eng.), Laguna Hills, USA
- BigComp (IEEE Int'l Conf. on Big Data and Smart Computing), Guangzhou, China
- ICSC (Int'l Conf. on Semantic Computing), Laguna Hills, USA

16 February

- ICNC (Int'l Conf. on Computing, Networking and Communications), Maui, USA

MARCH

6 March

- WACV (IEEE/CVF Winter Conf. on Applications of Computer Vision), Tucson, USA

16 March

- PerCom (IEEE Int'l Conf. on Pervasive Computing and Communications), Pisa, Italy

17 March

- SANER (IEEE Int'l Conf. on Software Analysis, Evolution and Reengineering), Limassol, Cyprus

20 March

- 3DV (Int'l Conf. on 3D Vision), Vancouver, Canada

21 March

- VR (IEEE Conf. on Virtual Reality and 3D User Interfaces), Daegu, Korea

22 March

- SSIAI (IEEE Southwest Symposium on Image Analysis and Interpretation), Santa Fe, USA

23 March

- SaTML (IEEE Conf. on Secure and Trustworthy Machine Learning), Munich, Germany

APRIL

12 April

- AST (IEEE/ACM Int'l Conf. on Automation of Software Test), Rio de Janeiro, Brazil
- FormaliSE (IEEE/ACM Int'l Conf. on Formal Methods in Software Eng.), Rio de Janeiro, Brazil
- ICSE (IEEE/ACM Int'l Conf. on Software Eng.), Rio de Janeiro, Brazil
- MOBILESoft (IEEE/ACM Int'l Conf. on Mobile Software Eng. and Systems), Rio de Janeiro, Brazil
- MSR (IEEE/ACM Int'l Conf. on Mining Software Repositories), Rio de Janeiro, Brazil

15 April

- COOL CHIPS (IEEE Symposium on Low-Power and High-Speed Chips and Systems), Tokyo, Japan

20 April

- PacificVis (IEEE Pacific Visualization Conf.), Sydney, Australia

MAY

4 May

- HOST (IEEE Int'l Symposium on Hardware Oriented Security and Trust), Washington, DC, USA

2469-7087/25 © 2025 IEEE

8 May

- BigDataSecurity (IEEE Conf. on Big Data Security on Cloud), New York City, USA
- CAI (IEEE Int'l Conf. on Artificial Intelligence), Granada, Spain
- HPSC (IEEE Int'l Conf. on High Performance and Smart Computing), New York City, USA
- IDS (IEEE Int'l Conf. on Intelligent Data and Security), New York City, USA
- SmartCloud (IEEE Int'l Conf. on Smart Cloud), New York City, USA

11 May

- SenSys (ACM/IEEE Int'l Conf. on Embedded Artificial Intelligence and Sensing Systems), Saint Malo, France

12 May

- RTAS (IEEE Real-Time and Embedded Technology and Applications Symposium), Saint Malo, France

13 May

- FCCM (IEEE Annual Int'l Symposium on Field-Programmable Custom Computing Machines), Atlanta, USA

18 May

- ICFEC (IEEE Int'l Conf. on Fog and Edge Computing), Sydney, Australia
- ICST (IEEE Int'l Conf. on Software Testing, Verification and Validation), Daejeon, Korea
- SP (IEEE Symposium on Security and Privacy), San Francisco, USA

19 May

- ICDE (IEEE Int'l Conf. on Data Eng.), Hong Kong, China
- ISMVL (IEEE Int'l Symposium on Multiple-Valued Logic), Sendai, Japan

25 May

- FG (IEEE Int'l Conf. on Automatic Face and Gesture Recognition), Kyoto, Japan
- IPDPS (IEEE Int'l Parallel and Distributed Processing Symposium), New Orleans, USA https://www.ipdps.org/

JUNE

1 June

- ICHI (IEEE Int'l Conf. on Healthcare Informatics), Minneapolis, USA

3 June

- CBMS (IEEE Int'l Symposium on Computer-Based Medical Systems), Limassol, Cyprus

6 June

- CVPR (IEEE/CVF Conf. on Computer Vision and Pattern Recognition), Denver, USA

22 June

- DCOSS-IoT (Int'l Conf. on Distributed Computing in Smart Systems and the Internet of Things), Reykjavik, Iceland
- ICSA (IEEE Int'l Conf. on Software Architecture), Amsterdam, Netherlands

26 June

- IEEE Cloud Summit, Washington, DC, USA

27 June

- ISCA (ACM/IEEE Annual Int'l Symposium on Computer Architecture), Raleigh, USA

JULY

6 July

- EuroS&P (IEEE European Symposium on Security and Privacy), Lisbon, Portugal
- ICALT (IEEE Int'l Conf. on Advanced Learning Technologies), Hung Yen, Vietnam

7 July

- COMPSAC (IEEE Annual Computers, Software, and Applications Conf.), Madrid, Spain
- ISVLSI (IEEE Computer Society Annual Symposium on VLSI), Kolkata, India