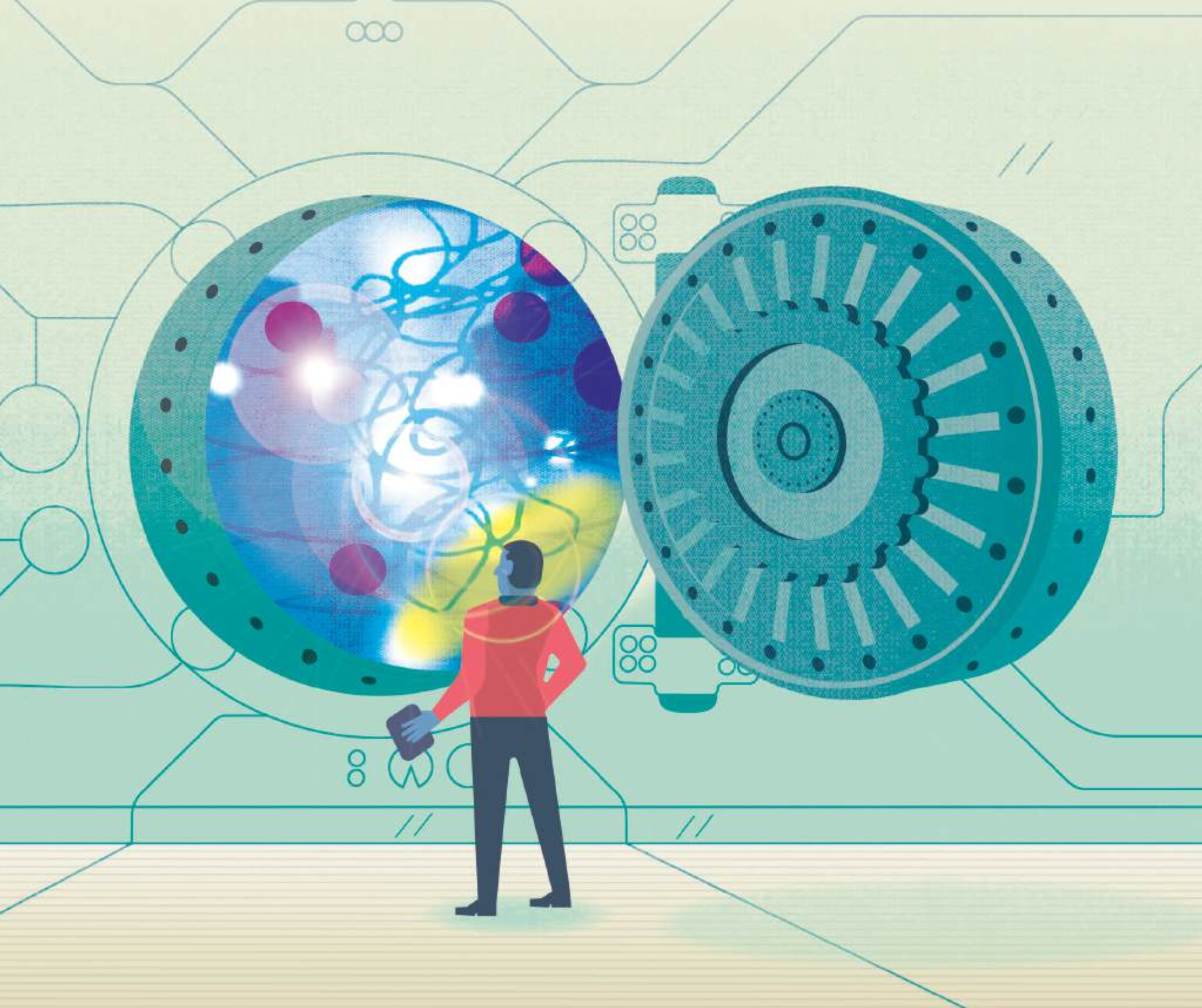
04.25

Computer

SECURE AND INTELLIGENT SYSTEMS



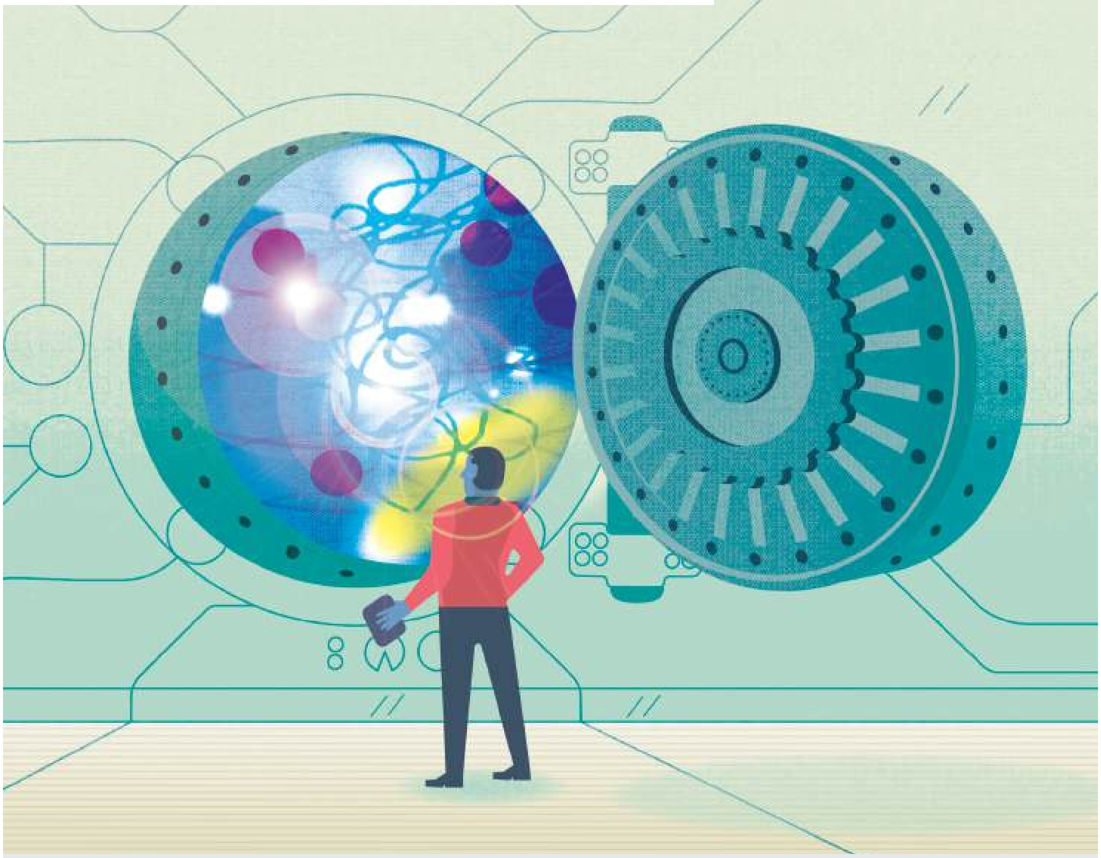
IEEE
IEEE COMPUTER SOCIETY

vol. 58 no. 4

www.computer.org/computer

Digital Object Identifier 10.1109/MC.2025.3546116

Computer

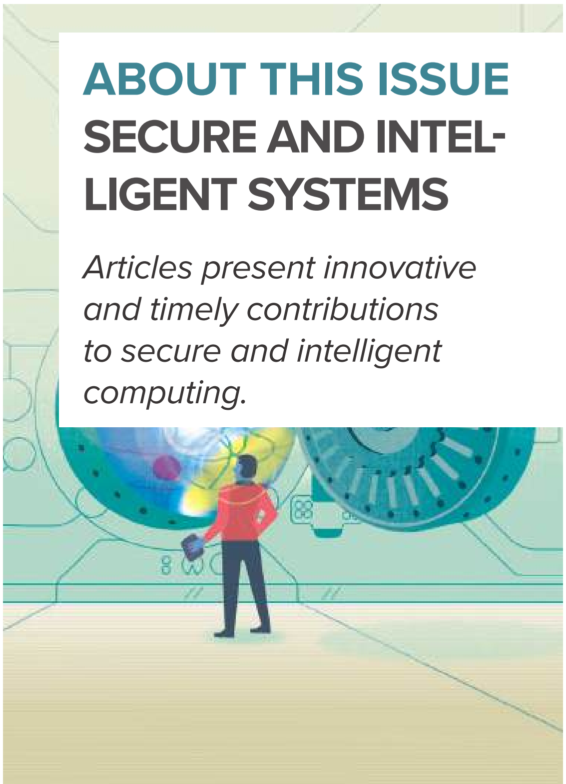


APRIL 2025
FEATURES

Digital Object Identifier 10.1109/MC.2025.3533810

APRIL 2025

CONTENTS



ABOUT THIS ISSUE

SECURE AND INTEL-LIGENT SYSTEMS

Articles present innovative and timely contributions to secure and intelligent computing.

FEATURES CONTINUED

COLUMNS

Department

Membership News

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

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

／# Computer

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

Digital Object Identifier 10.1109/MC.2025.3533812

ELSEWHERE IN THE CS

Computer Highlights Society Magazines

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

Computing in Science & Engineering

Predicting Links in Knowledge Graphs With the Canonical Correlation Analysis and Fusing Tensor Model

Relation prediction in knowledge graphs is critical for uncovering missing links between entities. Previous models mostly focused on learning the distance of entities and relation within each triplet. However, they relied heavily on linear metric learning-based methods to evaluate the connections between them, which ignore high-level complex interactions. To address these problems, the authors of this October–December 2024 *Computing in Science & Engineering* article introduce a canonical correlation analysis and fusing tensor model (CAFT) for relation prediction.

IEEE Annals of the History of Computing

Developing and Using CAD/CAM/CAE Systems in Boeing

Application programs to improve the quality and performance of its aerospace products are a critical part of Boeing's computing environment. This article, featured in the October–December 2024 issue of *IEEE Annals of the History of Computing*, focuses on how the company developed its own modeling, manufacturing, and engineering programs and built custom software to address shortcomings in commercial, off-the-shelf systems. It also details Boeing's attempt to produce its own computer-aided design system.

IEEE Computer Graphics and Applications

Enhancing Virtual Reality Training Through Artificial Intelligence: A Case Study

In this November/December 2024 *IEEE Computer Graphics and Applications* article, the authors propose an architecture that aims to facilitate the integration of artificial intelligence (AI) assistance into virtual reality training environments to improve user engagement and reduce authoring effort. The proposed architecture was tested in a study that compared a virtual training session with and without a digital assistant powered by AI.

IEEE Intelligent Systems

Regulated Federated Learning Against the Effects of Heterogeneity and Client Attacks

Federated learning (FL) can complete a learning task without compromising user privacy. However, the FL mechanism, where clients train models using personal data locally and exchange model updates instead of raw data, gives rise to new challenges. The problems caused by data heterogeneity and malicious client behaviors are universal in practical applications. The authors of this November/December 2024 *IEEE Intelligent Systems* article propose a regulated FL that introduces a generator and uses weighted aggregations to regulate client model training and complete federated aggregation.

Digital Object Identifier 10.1109/MC.2025.3536507
Date of current version: 27 March 2025

IEEE Internet Computing

AI Design: A Responsible Artificial Intelligence Framework for Prefilling Impact Assessment Reports

Impact assessment reports for high-risk artificial intelligence (AI) systems will be legally required but challenging to complete, especially for smaller companies. That is because the current process is complex, costly, and relies on guidebooks with limited assistance. The authors of this article from the September/October 2024 issue of *Internet Computing* propose AI Design, a semiautomatic framework for prefilling these reports.

IEEE micro

AMD XDNA NPU in Ryzen AI Processors

The authors of this article featured in the November/December 2024 issue of *IEEE* Micro discuss the AMD Ryzen 7040 series, the first x86 processor with an integrated neural processing unit (NPU). The artificial intelligence (AI)-optimized capabilities of the Ryzen 7040 NPU enable new AI experiences that are not possible without XDNA, making it a fundamental component in today's Ryzen-AI-powered devices and setting the foundation for an exciting roadmap toward future AI capabilities in mobile PCs.

IEEE MultiMedia

Multimodal Agents: From Vision to Reality

This October–December 2024 *IEEE MultiMedia* article explores the evolution of multimodal agents, highlighting their ability to transcend the limitations of single-modality systems and deliver results based on a comprehensive, context-aware understanding of their environment.

IEEE pervasive COMPUTING

The Future of Consumer Edge-AI Computing

In the last decade, deep learning has rapidly infiltrated the consumer end, due to hardware acceleration across devices. The authors of the July–September 2024 issue of *IEEE Pervasive Computing* introduce a novel paradigm centered around EdgeAI-Hub devices, designed to reorganize and optimize compute resources and data access at the consumer edge.

IEEE SECURITY & PRIVACY

Android Permissions: Evolution, Attacks, and Best Practices

In this article, featured in the November/December 2024 issue of *IEEE Security & Privacy*, the authors study the evolution of Android permissions. They describe the rationale behind key changes in Android's permission model and disclose two permission-related security vulnerabilities they discovered. Finally, they provide developers actionable insights to proactively address permission-related security and privacy risks during development.

IEEE Software

Toward an Open Source MLOps Architecture

The authors of this article from the January/February 2025 issue of *IEEE Software* present a Kubernetes-based, open source MLOps framework to streamline the lifecycle management of machine learning models in production environments. They compare state-of-the-art MLOps tools and frameworks, demonstrating that their features meet the same features as proprietary options, such as Amazon SageMaker.

IT Professional

MetaDigiHuman: Haptic Interfaces for Digital Humans in the Metaverse

As technology continues to advance, the demand for sophisticated and immersive interfaces to interact with the metaverse has become increasingly crucial. This November/December 2024 *IT Professional* article introduces the concept of MetaDigiHuman, a groundbreaking framework that combines blended digital humans and haptic interfaces. By harnessing cutting-edge technologies, MetaDigiHuman enables seamless and immersive interaction within the metaverse.


IEEE COMPUTER SOCIETY
Call for Papers
Ensure your research is easily discoverable by being indexed in major databases and optimized for search engines.
GET PUBLISHED
www.computer.org/cfp
IEEE COMPUTER SOCIETY
IEEE
Digital object identifier 10.1109/Mc.2025.3546117

50 & 25 YEARS AGO

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

APRIL 1975

https://www.computer.org/csdl/magazine/co/1975/04

Guest Editor: Design Automation; Herbert M. Wall (p. 19): "The set of design tools and practices collectively known as 'design automation' had its origin twenty years ago. … But by far the widest application of design automation is in the noncreative, "cookbook" type clerical tasks of design recording, checking, data reformatting, and information transfer. … There has been a long standing interest in applying the formality of design automation procedures to the business of software generation." *[Editor's note: The papers below come from the 11th Design Automation Workshop.]*

The IPAD System: A Future Management/Engineering/Design Environment; Carlos A. Garrocq et al. (p. 23): "The design of a new aerospace vehicle is a complex, long-term process." (p. 24) "The overall goal of IPAD (*Editor's note: Integrated Programs for Aerospace Vehicle Design*) is the automation of appropriate sections of the design process to shorten design time, reduce cost, and improve the ultimate product." (p. 26) "The total automated capability of the engineering/science community is resident in a library of automated operational modules. … This bank stores all historical, statistical, and other data that has been accumulated from previous studies and is a vital part of the experience of a design team." *[Editor's note: A detailed description of both a method and an implementation of a system for the design of aerospace vehicles that also includes a cost saving analysis.]*

Automated Inspection of Electronic Assemblies; Charles A. Harlow et al. (p. 36): "Such advances in production greatly complicate the inspection of these devices by humans." (p. 37) "This paper describes some requirements for an automated inspection system for electronic assemblies and discusses in detail the system's components. This system will be referred to as AVIS (Automated Visual Inspection System)." (p. 38) "The image scanner and the digital processor replace the human inspectors in an ordinary production-line inspection system." *[Editor's note: The article then continues to describe in detail the mechanical and computer hardware aspects of their solution. In my mind, the article falls short in describing and analyzing the processes that lead to the actual detection of assembly mistakes in, for example, computer boards.]*

Computer-Aided Ship Design at MarAd *(Editor's note: Marine Administration)*; **Alan H. Woodyard** (p. 46): "Figure I illustrates the spiraling nature of the design verification from basic requirements to final design of the ship and indicates the sequential nature of the calculations." (p. 46) "Whereas a series of batch programs is computationally satisfactory, input preparation problems arise, both real (documentation inadequate, control cards confusing) and psychological. … The input files are stored on the timesharing system and sent to the batch system along with a control card file for execution by the user-selected batch program." *[Editor's note: A detailed description follows, but it's interesting to note that some aspects of the system read rather "old" even taking the 1975 state of the art into account.]*

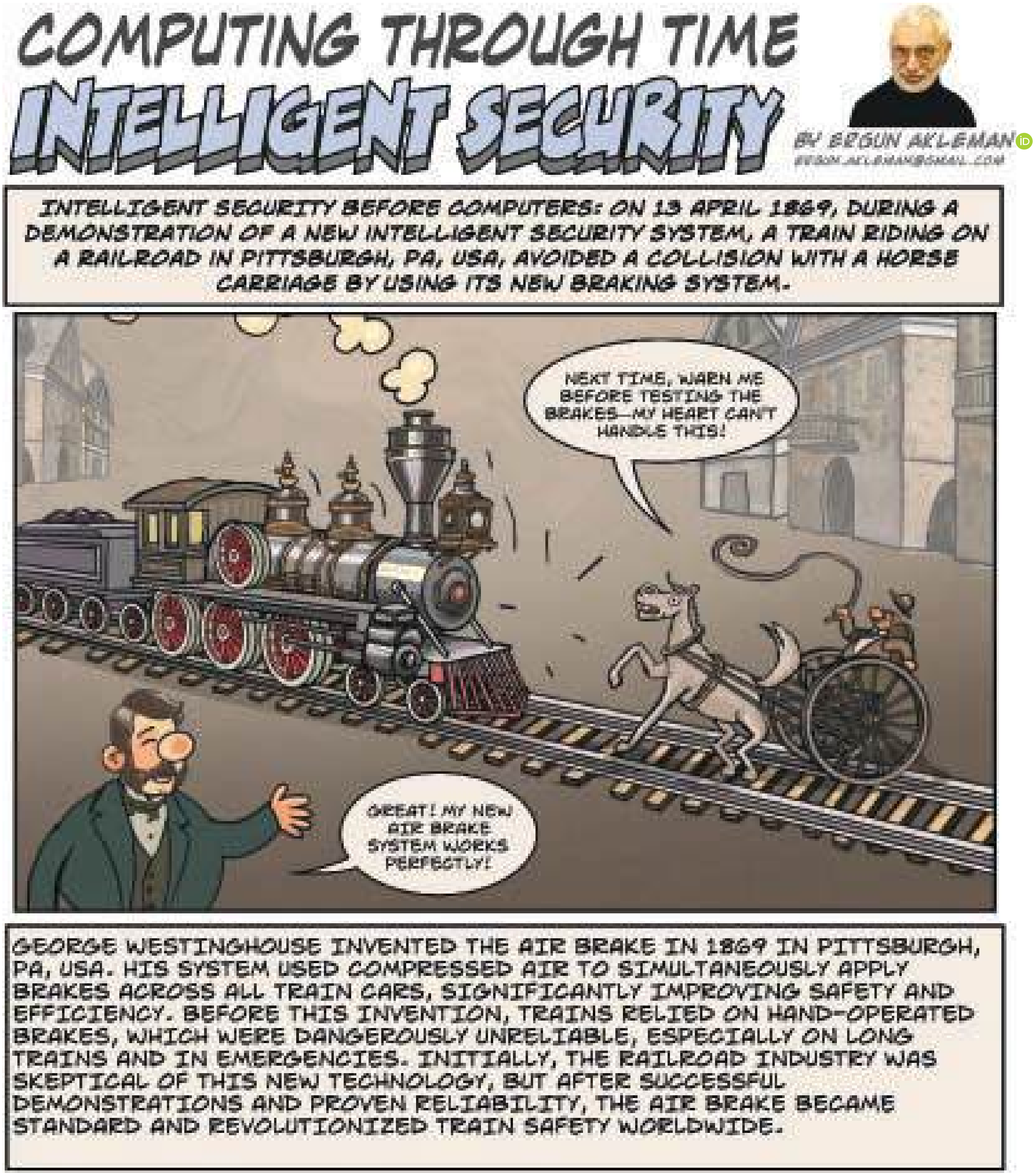
APRIL 2000

https://www.computer.org/csdl/magazine/co/2000/04

Denial-of-Service Attacks Rip the Internet; Lee Garber (p. 12): "The attacks, which observers say cost victims millions of dollars, sent shock waves through the industry because they crippled some of the world's premier ecommerce sites. … On Monday, 7 February, the first of the high-profile DDoS attacks hit Yahoo, the most popular site on the Web." (p.13) "Denial-of-service attacks have been around for years. The attacks have used several techniques to crash, hang up, or overwhelm servers with malformed packets or large volumes of traffic." (p. 17) "IP version 6 also offers authentication, so as more network equipment vendors support the protocol, users will have

Digital Object Identifier 10.1109/MC.2025.3527406
Date of current version: 27 March 2025

more protection against DDoS attacks." *[Editor's note: The article not only discusses distributed denial-of-service (DDoS) attacks but also zombies, trojans, and other viruses and offers some suggestions on how to protect against them. Of course, as we know the danger today is as large if not larger than 25 years ago. Especially as many governments seem to be developing and using more and more advanced techniques in the ongoing cyber war.]*



Digital Object Identifier 10.1109/MC.2025.3535914
Date of current version: 27 March 2025

News Briefs; Ed: Anne C. Lear (p. 18ff): UCITA: Misguided or Misunderstood?: "The Uniform Computer Information Transactions Act, now under consideration by individual states, specifies the rights of software vendors and buyers." *[Editor's note: The short, interesting article analyzes the pros and cons of the proposal but, as we now know, it was never broadly established and enforced.]* "**What's Next for the Next-Generation Net?** Next-Generation Internet, Internet2, and many other projects around the world are developing new technologies and capabilities that will soon begin to be deployed broadly across the Internet." *[Editor's note: Founded around that time the nonprofit organization Internet2 is still around, has many members and is still developing advanced concepts for the Internet.]*

Leveraging Inconsistency in Software Development; Bashar Nuseibeh et al. (p. 24): "Maintaining consistency at all times is counterproductive. In many cases, it may be desirable to tolerate or even encourage inconsistency to facilitate distributed teamwork and prevent premature commitment to design decisions." (p. 25) "At this point, you would choose from among several different inconsistency-handling strategies, including resolving the inconsistency immediately, ignoring it completely, or tolerating it for a while." *[Editor's note: This interesting article investigates different types of inconsistencies, how to detect them and how to handle them. This includes even the psychology of denials by different parts of the teams.]*

GUEST EDITORS: Configurable Computing: Technology and Applications; Ranga R. Vemuri et al. (p. 39): "The reconfigurable data path came in response to a call for more parallelism and higher performance in the now larger ICs. Reconfigurable data paths have coarser-grained reconfigurable units than their fine-grained FPGA *[Editor's note: field-programmable gate array]* predecessors; they can accommodate reconfigurable nibble, byte, or wider arithmetic logic units (ALUs)." (p. 40) "This special issue continues with five articles representing these historical and emerging trends. — First, André DeHon surveys the state of the art in FPGA based custom-computing machines. He then delves into their computational-density advantage and explains how to exploit this advantage. — In the next article, Simon Haynes and his colleagues describe Sonic. This flexible, scalable FPGA-based custom architecture handles real-time video image-processing applications such as fractal image generation, filtering, and transformation. — Recently, some researchers have proposed using configurable computers to solve combinatorial search problems. Configurable computers' high degree of concurrency makes efficient implementation of parallel algorithms possible. For example, the short article by Marco Platzner introduces a custom configurable architecture for solving the Boolean satisfiability problem. Although solution speedups of three to five orders of magnitude are not uncommon, this technique is still emerging. — The article by Tim Callahan and his colleagues introduces the Garp architecture and describes techniques for compiling standard C programs onto it. Garp includes a reconfigurable array coprocessor attached to a main MIPS processor. The Garp C compiler uses various techniques to identify and accelerate computationally intense kernels in a C program using the reconfigurable array. The architecture, along with these associated techniques, exploits inherent application-level, instruction-level parallelism, making it suitable for accelerating many applications. — In the last article of this special issue, Seth Goldstein and his colleagues describe PipeRench, an architecture that provides a scalable, configurable pipeline structure to accelerate stream-based computations. PipeRench allows graceful degradation of performance with application size. *[Editor's note: As these five summaries already identify very well the main issues discussed in the quite detailed articles, I will refrain from extensively summarizing them again and will only add my own comments. Not surprisingly, in most of these featured articles the investigated applications allow for a high degree of parallelization.]*

The Density Advantage of Configurable Computing; André DeHon (p. 47): "The array computes the application's performance-limiting portions (10 percent of the code, 90 percent of the computations) with high parallelism on densely packed spatial operators. The processor packs the computation's noncritical portions (90 percent of the code, 10 percent of the computation) into minimum space." *[Editor's note: Reading this article, one wonders why FPGAs have not taken over more of the world as has happened. Established market pressure has been cited.]*

Video Image Processing with the Sonic Architecture; Simon D. Haynes et al. (p. 51): "As Figure 1 shows, the design consists of plug-in processing elements (PIPEs) connected by the PIPE bus and PIPEflow buses. It also enables design reuse and supports the software plug-in methodology." *[Editor's note: As the authors themselves state, the design is oriented towards video and image processing and is limited by the performance of the components available 25 years ago.]*

Reconfigurable Accelerators for Combinatorial Problems; Marco Platzner (p. 58): "Tailoring hardware to a specific algorithm and a specific set of input data can boost execution several fold. One hardware circuit that solves Boolean satisfiability improved execution time by a factor of 140,000 over state-of-the- art software solvers." (p. 59) "A practical acceleration engine will have to combine an instance-specific accelerator with software SAT solvers." *[Editor's note: The article investigates only partially the switchover between software and hardware that is needed to get the claimed acceleration.]*

The Garp Architecture and C Compiler; Timothy J. Callahan et al. (p. 62): "Garp combines a single-issue MIPS processor core with reconfigurable hardware to be used as an accelerator. We designed both the reconfigurable hardware and the interfaces among the system components, tailoring them for general-purpose computing." *[Editor's note: The authors confess in this article that in production systems hand-coded libraries will have to be used together with the Garp architecture. Of course, with today's huge software libraries that is now a common programming technique.]*

PipeRench: A Reconfigurable Architecture and Compiler; Seth Copen Goldstein et al. (p. 70): "PipeRench is a reconfigurable fabric—an interconnected network of configurable logic and storage elements." (p. 76) "Compilation has two components: generating optimized multipliers and generating the rest of the pipeline." *[Editor's note: The article describes in detail those two aspects of the compilation but again concentrates on stream-based applications.]*

Communications: Recent Advances in Wired Networking; Upkar Varshney (p. 107): "This column — the inaugural installation of *Computer*'s Communications department—looks into the latest advances in wired networking. ... lot of research on using the enormous bandwidth of fiber has shifted to other frontiers, such as wavelength division multiplexing (WDM) and Solitons." *[Editor's note: Here we see that 25 years ago there were still issues when developing technology and protocols for multiple 100-Gigabit networks that are more or less standard today. However, recently 400 Terabits was achieved by a Japanese research lab.]*

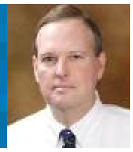
Using Technology and Innovation to Simulate Daily Life; Michael Macedonia (p. 110): "I knew something was up when I saw my daughter shouting at our computer, scolding one of her Sims—a simulated male who kept making mess of his house. ... Thus, the primary skill you need to play the game is the ability to plan and queue instructions for your Sims." (p. 111) "Despite all its technical bells and whistles, The Sims' ultimate beauty lies in its ability to immerse you in the Sim world and captivate you with each Sim's autonomy." *[Editor's note: Despite the enthusiasm conveyed in this article, the Sims World never became a public rage, but, in my mind unfortunately, it became essential in today's military simulations of strategies, tactics, and even guiding executions.]*

XML: An Interview with Peter Flynn; Ed: Charles Severance (p. 113): "As such, XML is becoming an important tool in application integration. Peter Flynn was a key voice in developing the standard and remains an active observer of how XML is finding its place in the world. ... If business can actually agree on what constitutes an invoice order, we might actually see some real XML-enabled e-commerce." *[Editor's note: This interesting article is mostly concerned with the need for universal standardization of application style sheets. Of course, that never happened, and still XML and its variants are practically behind every page on the Web.]*

Software Technologies: Fundamental Ideas and Change; Michael Lutz; (p. 115): "Welcome to the first installment of the new Software Technologies department. ... I'll give you some essential information and I'll present you with technologies that you otherwise might have missed. In Software Technologies, we will explore everything from the mainstream to the marginal." *[Editor's note: I am eager to see and extract for my readers the interesting issues that may be described.]*

When Atoms and Bits Collide: Detroit Goes Digital; Ted Lewis (p. 120): "Can old-economy manufacturing companies benefit from the Web? If so, how might they be different in 10 years? In general, how might the new economy transform traditional manufacturing businesses? ... Create business-to-business auctions and procurement systems on the Web and reduce the average order-processing cost from $100 to $10. ... To be competitive in the atoms business, manufacturers must assemble atoms without owning them." *[Editor's note: In essence, this article argues for an Internet-based integrated supply chain and for outsourcing of all "atom" components but also for franchising manufacturing and services. Much of this has actually happened but not necessarily for the benefit of the customers.]*

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



Big Tech, Big Energy, and AI

Nir Kshetri, The University of North Carolina at Greensboro
Jeffrey Voas, IEEE Fellow

This message considers the energy needs of data centers and rumors of big tech looking to acquire energy providers to satisfy their needs.

We continually hear about the energy needs of data centers, and we've heard rumors of big tech looking to acquire energy providers to "feed their needs." So, we thought it would be interesting to look at the numbers. Here's what we uncovered.

The rising adoption of generative artificial intelligence (AI) that began in 2022 is driving the demand for more data centers. Meta's electricity consumption grew by 33% in 2023, reaching more than 15 terawatt-hours (TWh), nearly tripling its power usage since 2019.[1] Generative AI demands immense power for both development and operation, prompting a race to secure future energy capacity.[2] Newmark's report projected that data center energy consumption will more than double by 2030 compared to 2022.[3] According to a Lawrence Berkeley National Laboratory analysis, data centers are projected to consume up to 12% of U.S. electricity by 2028, totaling 132 gigawatts (GW) annually, tripling their 2023 consumption levels.[4] In short, tech giants are significantly invested in energy infrastructure and server farms.[2] In 2024, Amazon Web Services, Google, and Microsoft collectively invested US$180 billion in expanding their data centers.[5]

Let's briefly look at big tech's eye toward nuclear, sustainable, and renewable energy.

NUCLEAR

Big tech firms, including Microsoft, Google, and Amazon, are proactively investing in their own energy sources. A key trend in meeting energy demands is the growing investment in nuclear power, with an emphasis on partnerships with existing facilities and the development

DISCLAIMER

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

Digital Object Identifier 10.1109/MC.2025.3529175
Date of current version: 27 March 2025

TABLE 1. Big tech's nuclear power investments.

Company	Energy consumption	Nuclear power investment
Amazon	Does not release energy consumption figures[9]	2024: Amazon announced a plan to invest US$500 million in nuclear energy projects in Washington State, Pennsylvania, and Virginia.[10]
Google	2.9 TWh in 2011, 12.8 TWh in 2019, 25.9 TWh in 2023[11]	2024: Google signed a deal with Kairos Power to purchase nuclear energy, marking the first such corporate agreement. Kairos Power plans to operate its first reactor within six years, contributing 500 MW to the U.S. grid by 2035.[12]
Meta	5.1 TWh in 2019 to 15.3 TWH in 2023[1]	Dec 2024: Meta was seeking nuclear power proposals to meet its AI and sustainability goals. The company plans to add 1–4 GW of nuclear capacity in the U.S. starting in the early 2030s.[8]
Microsoft	11.2 TWH in 2020 to 18.6 TWH in 2022[13]	2024: Microsoft secured a 20-year agreement to purchase energy from the Three Mile Island nuclear plant, which was closed in 2019. The plant is scheduled to reopen by 2028 following a US$1.6 billion investment for upgrades and permits. The energy from the plant will support Microsoft's data centers, crucial for its AI and cloud computing operations.[14]

of small modular reactors to provide sustainable and scalable energy solutions[6] (Table 1). Nuclear energy offers a clean and sustainable power solution that reduces harmful emissions, occupies less land, and produces minimal waste, making it a viable option for tech companies aiming to meet sustainability targets.[7]

Microsoft has committed to purchasing all electricity from the decommissioned Three Mile Island nuclear power plant, slated to reopen by 2028. This move secures a stable and emissions-free energy supply for Microsoft's data centers, addressing both the power-intensive requirements of AI operations and sustainability objectives.[6] (A typical U.S. nuclear plant has a capacity of approximately 1 GW.[8])

> Generative AI demands immense power for both development and operation, prompting a race to secure future energy capacity.

SUSTAINABLE AND RENEWABLE

Tech companies are increasingly securing renewable power purchase agreements (PPAs) to meet their sustainability goals by purchasing energy from wind and solar farms. Meta has funded solar projects in Denmark and Norway.[15] In 2018, Meta entered a 15-year PPA for the entire 294-MW output of a wind farm cluster in Norway.[16] Meta has also entered a PPA with Longroad Energy to secure 300 MW of solar energy for its Texas data center. The energy will be sourced from Longroad's "1000 Mile" solar project in Yoakum County, TX, USA, which has a capacity of 300 MWac (400 MWdc).[17]

From 2010 to 2021, Google secured more than 7 GW of renewable energy through various purchase agreements (https://sustainability.google/progress/energy/). In 2017, Google acquired 160 MW of wind power in Norway.[16] In early 2024, Google announced that it would add more than 700 MW of clean energy capacity, primarily from two new offshore wind farms in The Netherlands. This is expected to contribute to achieving 90% clean energy for its Dutch data centers in 2024.

In closing, we've offered a glimpse of budding relationships between big energy and big tech. And since "energy-thirsty" AI has huge national security ramifications, with nation-states competing to dominate "everything AI," cooperation between big energy with big tech is something to keep an eye on [since the winner(s) will likely need unlimited energy]. And for a more in-depth discussion of this topic, see C. Bash et al.[18]

REFERENCES

1. B. Alves, "Meta's electricity consumption FY 2019-2023," Statista, Hamburg, Germany, Nov. 11, 2024. Accessed: Jan. 3, 2025. [Online]. Available: https://www.statista.com/statistics/1500373/meta-electricity-consumption-intensity/#:~:text=Meta's%20electricty%20consumption%20increased%2033,company's%20power%20consumption%20nearly%20tripled
2. I. A. Hamilton, "2024: The year big tech got into the energy game," *The Daily Upside*, Dec. 30, 2024. Accessed: Jan. 2, 2025. [Online]. Available: https://www.thedailyupside.com/technology/big-tech/big-tech-got-into-the-energy-game

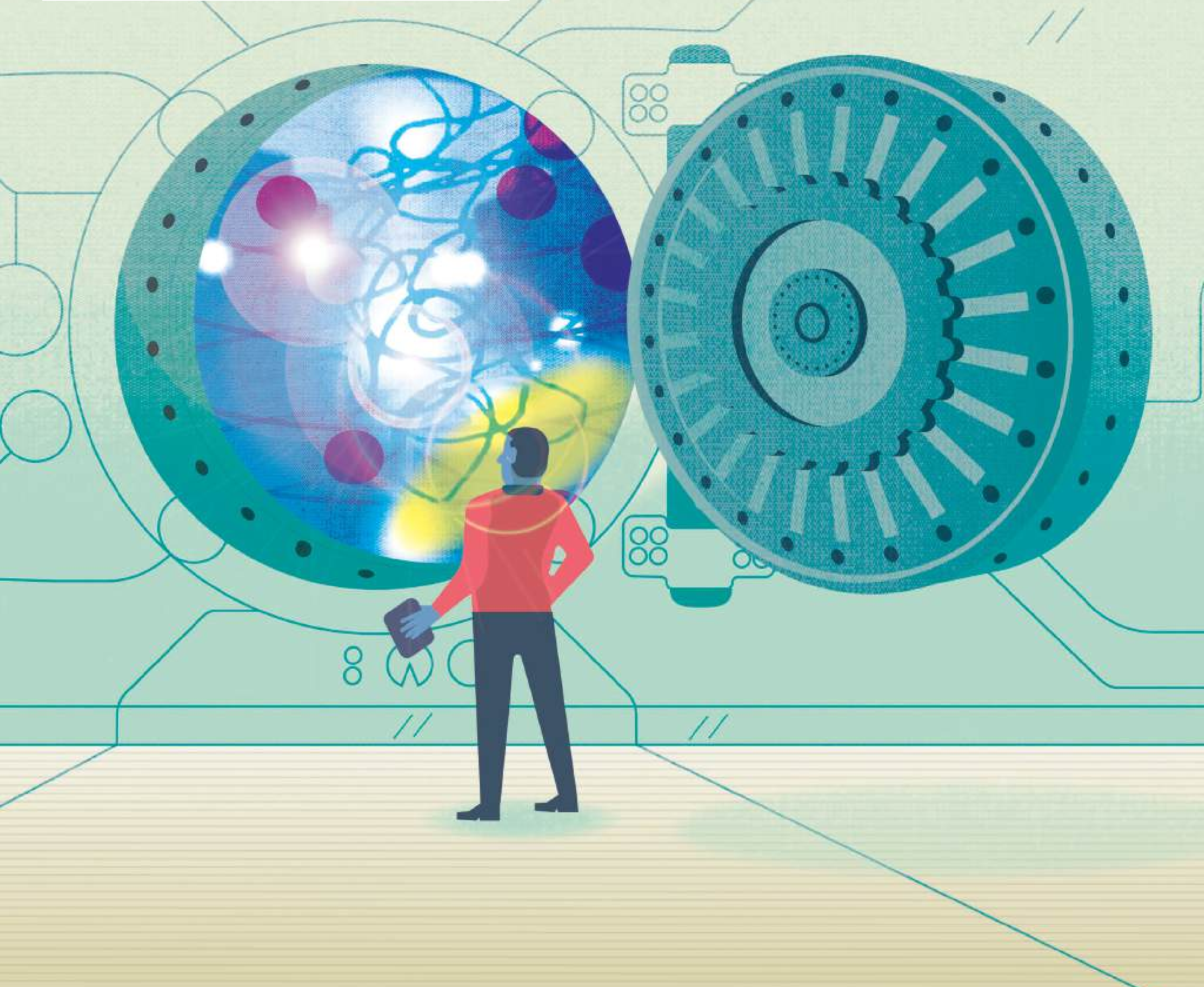
3. N. Rubio-Licht, "Big Tech Has an Energy Problem: How tech firms reconcile their desire for endless growth with bold climate goals," *The Daily Upside*, Apr. 22, 2024. Accessed: Jan. 2, 2025. [Online]. Available: https://www.thedailyupside.com/industries/energy/big-tech-has-an-energy-problem/
4. H. Clancy, "The data center boom threatens climate goals: How to fight back," *Trellis*, Jan. 2, 2025. Accessed: Jan. 2, 2025. [Online]. Available: https://trellis.net/article/data-center-boom-threatens-climate-goals-how-to-fight-back/
5. J. Laws, "Big tech increases spending on infrastructure amid AI boom," *Newsweek*, Dec. 5, 2024. Accessed: Jan. 3, 2025. [Online]. Available: https://www.newsweek.com/big-tech-increases-spending-infrastructure-amid-ai-boom-1995889
6. G. Scheig, "Big tech's strategic shift to nuclear power for AI energy demands," *Stout*, Nov. 7, 2024. Accessed: Jan. 2, 2025. [Online]. Available: https://www.stout.com/en/insights/commentary/big-techs-strategic-shift-nuclear-power-ai-energy-demands
7. Office of Nuclear Energy, "3 reasons why nuclear is clean and sustainable," *U.S. Department of Energy*, Mar. 31, 2021. Accessed: Jan. 3, 2025. [Online]. Available: https://www.energy.gov/ne/articles/3-reasons-why-nuclear-clean-and-sustainable
8. "Meta seeks nuclear power developers for reactors to start in early 2030s," *CNBC*, Dec. 3, 2024. Accessed: Jan. 3, 2025. [Online]. Available: https://www.cnbc.com/2024/12/03/meta-seeks-nuclear-power-developers-for-reactors-to-start-in-early-2030s-.html
9. A. Leonard, "The carbon footprint of Amazon, Google, and Facebook is growing: How cloud computing—and especially AI—threaten to make climate change worse," *Sierra*, Sep. 11, 2024. Accessed: Jan. 3, 2025. [Online]. Available: https://www.sierraclub.org/sierra/2024-3-fall/feature/carbon-footprint-amazon-google-and-facebook-growing
10. E. Davis, "Amazon is going nuclear, joining big tech in the race for more power: 'There's not going to be enough wind and solar to meet needs'," *Entrepreneur*, Oct. 16, 2024. Accessed: Jan. 3, 2025. [Online]. Available: https://www.entrepreneur.com/business-news/amazon-is-going-nuclear-joining-big-tech-in-energy-race/481398
11. L. Fernández, "Google energy consumption 2011-2023," Statista, Hamburg, Germany, Oct. 11, 2024. Accessed: Jan. 3, 2025. [Online]. Available: https://www.statista.com/statistics/788540/energy-consumption-of-google/
12. S. Shibu, "Google follows Microsoft's lead with nuclear power for AI," *Entrepreneur*, Oct. 15, 2024. Accessed: Jan. 3, 2025. [Online]. Available: https://www.entrepreneur.com/business-news/google-follows-microsoft-turns-to-nuclear-power-for-ai/481316
13. "2022 environmental sustainability report data fact sheet," Microsoft, Redmond, WA, USA. Accessed: Jan. 3, 2025. [Online]. Available: https://query.prod.cms.rt.microsoft.com/cms/api/am/binary/RW13PLE
14. S. Shibu, "Will it take nuclear power to sustain AI? Microsoft is betting on it," *Entrepreneur*, Sep. 23, 2024. Accessed: Jan. 3, 2025. [Online]. Available: https://www.entrepreneur.com/business-news/microsofts-next-power-source-for-ai-data-centers-is-nuclear/480257
15. M. Jones, "Big tech's energy needs are growing so fast that power grids can't keep up," *Mother Jones*, Jun. 17, 2024. Accessed: Jan. 3, 2025. [Online]. Available: https://www.motherjones.com/politics/2024/06/big-tech-ai-energy-electricity-use-growing-power-grid-restraints/
16. P. Judge, "Facebook signs for 294 MW of wind power in Norway," *Data Center Dynamics*, May 24, 2018. Accessed: Jan. 3, 2025. [Online]. Available: https://www.datacenterdynamics.com/en/news/facebook-signs-for-294mw-of-wind-power-in-norway/
17. A. Paleja, "300 MW solar power to meet Meta's data center demands, Longroad Energy signs deal," *Interesting Engineering*, Dec. 30, 2024. Accessed: Jan. 3, 2025. [Online]. Available: https://interestingengineering.com/energy/meta-data-center-300-mw-solar-power
18. C. Bash, J. McDonald, D. Milojicic, J. Mueller, T. Wilson, and Y. Zhang, "Convergence of energy and IT sectors," *Computer*, vol. 58, no. 1, pp. 151–158, Jan. 2025, doi: 10.1109/MC.2024.3498572.

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, and the "Computing's Economics" column editor for *Computer*. Contact him at nbkshetr@uncg.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.

SECURE AND INTELLIGENT SYSTEMS

D. Frank Hsu, Fordham University
Christina Schweikert, St. John's University



Digital Object Identifier 10.1109/MC.2025.3534447
Date of current version: 27 March 2025

Secure and intelligent systems are those that not only employ methods and technologies such as machine learning, informatics, and artificial intelligence (AI) in decision making but also incorporate robust security measures to protect sensitive data and prevent unauthorized access or adversarial attacks. In essence, a secure and intelligent system combines security and intelligence to function efficiently and effectively while increasing reliability and mitigating risks.

In this issue, we are very pleased to include four articles presenting innovative and timely contributions to secure and intelligent computing.

ABOUT THIS ISSUE

Kuhn et al.[A1] present an insightful method for measuring and visualizing the relative strength of a dataset in machine learning. The authors use a combinatorial method to effectively measure the convergence of interactions between features, an essential consideration in judging data adequacy for training, validation, and testing.

Next, Gavrilova[A2] provides a comprehensive overview of information fusion in the context of innovations in biomimetic multimodal systems. These include unimodal/multimodal system design, deep learning architectures, new behavioral traits based on online social media analytics, and paying attention to the user's data security and privacy.

Then, Nakao et al.[A3] provide a thorough analysis of the escalating complexity and sophistication of cyberattacks within Internet of Things (IoT) environments. It covers detailed case studies

APPENDIX: RELATED ARTICLES

A1. D. R. Kuhn, M S Raunak, and R. N. Kacker, "Measuring and visualizing dataset coverage for machine learning," *Computer*, vol. 58, no. 4, pp. 18–26, Apr. 2025, doi: 10.1109/MC.2025.3527374.

A2. M. L. Gavrilova, "Information fusion: A decade of innovations in biometric multimodal research," *Computer*, vol. 58, no. 4, pp. 27–36, Apr. 2025, doi: 10.1109/MC.2025.3526135.

A3. K. Nakao, D. Inoue, and K. Yoshioka, "Unveiling IoT threats: A case study on darknet and honeypot analysis," *Computer*, vol. 58, no. 4, pp. 37–45, Apr. 2025, doi: 10.1109/MC.2025.3531364.

A4. E. Owusu, M. Mapkar, M. Rahouti, and D. C. Verma, "Robust Intrusion detection with combinatorial fusion and generative artificial intelligence," *Computer*, vol. 58, no. 4, pp. 46–57, Apr. 2025, doi: 10.1109/MC.2024.3524302.

that highlight the vulnerabilities of IoT devices, advanced threat detection, and an analysis of IoT malware, all through the lens of darknet observation and honeypot technologies.

Finally, Owusu et al.[A4] present a novel intrusion detection system that combines combinatorial fusion analysis with generative AI to enhance anomaly detection in intelligent systems. The system addresses challenges in detecting low-profile and evolving threats, in particular for imbalanced datasets. ■

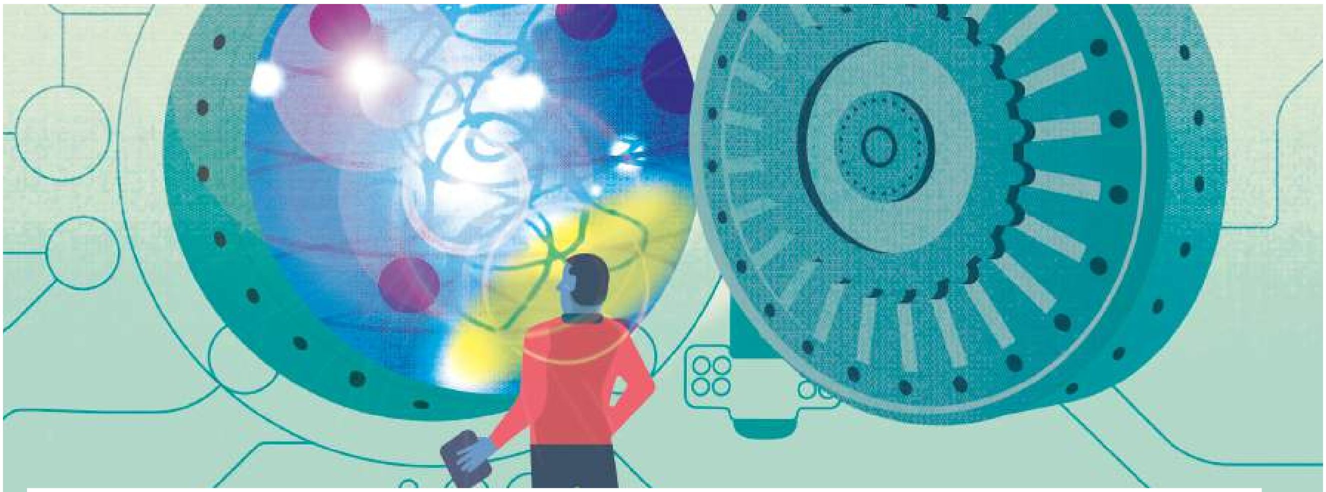
ABOUT THE AUTHORS

D. FRANK HSU is the Clavius Distinguished Professor of Science, a professor of computer and information science, and director of the Laboratory of Informatics and Data Mining, Fordham University, New York, NY 10023 USA. Contact him at hsu@fordham.edu.

CHRISTINA SCHWEIKERT is an associate professor of computer science and the program director for the Master of Science in Data Science program, St. John's University, Queens, NY 11439 USA. Contact her at schweikc@stjohns.edu.

Digital object identifier 10.1109/Mc.2025.3546129

Digital Object Identifier 10.1109/MC.2025.3533814



Measuring and Visualizing Dataset Coverage for Machine Learning

D. Richard Kuhn, M S Raunak, and Raghu N. Kacker, National Institute of Standards and Technology

Measuring and visualizing the relative strength of datasets in machine learning can be highly useful. Combinatorial methods effectively measure coverage of interactions between features, an essential consideration in judging data adequacy for training and testing.

When training a machine learning (ML) model, we need to ask, do we have enough training data, and is it representative of the environment where the ML model will be used? This decision is akin to software engineers' selection of a set of values with which to test a software system, and how many tests are needed. Characterizing the training dataset requires measuring the data using some metric and against some criteria. These criteria can also be used to compare relative strengths or usefulness of two datasets, be it for regular software testing or for the purpose of training an ML model. Common approaches used for designing test suites or choosing training datasets vary from random selection to using different types of source code-based strategy such as branch coverage or stronger structural coverage criteria. In the case of ML modeling, various forms of neuron coverage are often advocated. Some research, however, indicates that neuron coverage is not sufficiently effective in properly characterizing the strength of a training dataset of an ML model.[1,2,3] Measures of combinatorial coverage,[4] combinatorial coverage differences,[5,6] and combination frequency differences[7] may be useful as metrics and measurement techniques in this regard. These measures have uses in different phases of the ML

Digital Object Identifier 10.1109/MC.2025.3527374
Date of current version: 27 March 2025

lifecycle, including data gathering, model development, and eventual system monitoring and maintenance.[8,9] In this article we explain these coverage metrics, their measurement mechanisms, and how they can be used to assess the relative strength or weakness of a training or testing dataset.

BASIC COMBINATORIAL COVERAGE MEASURES

The concepts described here originate from combinatorial methods[10] used in traditional software testing. Combinatorial coverage can measure the level of feature value combinations present in a dataset. This dataset can represent a test suite, where each row is a test case or it can represent a training set for a supervised ML model that accepts tabular data, where each row is a training sample. These coverage-related concepts can be illustrated using the example shown in Table 1. It shows a table of specific values of four binary variables. Let us consider that each row defines a set of feature values that represent a specific sample, that is, instance used for training an ML model. We then define the following two terms:

- *t-way value combination*: A combination of (feature) values present in a training dataset. For example, ($a = 0, b = 0$) is one two-way value combination, and ($a = 1, b = 0$) is a different two-way value combination for the same feature combination. Similarly, ($a = 1, b = 0, c = 1$) is a three-way value combination.
- *t-way coverage*: The proportion of total $C(n, t) \times v^t$ t-way value combinations present in at least one sample (that is, row) in the training dataset. Here, n is the number of training features or variables, t is the number of interactions within those input features being covered, and v represents the number of possible values a feature can take.

The convention for describing an array configuration for combinatorial testing or coverage is $v_1^{n1} v_2^{n2}$..., where the v are the number of feature values and n are the number of occurrences of features that have this number of values. So, the simple array in Table 1 is a 2^4 configuration (four features of two values each) and the example in Table 2 is a $2^4 3^1 3^1$ configuration.

Example

Table 1 represents a training set with four binary features: a, b, c, and d. There are $C(4, 2) = 6$ possible feature combinations: [{a, b}, {a, c}, {a, d}, {b, c}, {b, d}, {c, d}] and $C(4, 2) \times 2^2 = 24$ possible two-way feature value combinations. For notational shorthand, let us refer to these feature combinations as {ab, ac, ad, bc, bd, and cd}. Nineteen of these 24 value combinations are present in the training dataset, that is, they are covered. The missing combinations are $ab = 11$, $ac = 11$, $ad = 10$, $bc =$ 01, $bc = 10$. Note that only two, bd and cd, are covered with all four possible feature value pairs. So, we have 79% (19/24) for the value combination coverage metric. For a better understanding of this test set, we can compute the configuration coverage for each of the six feature combinations, as shown in Figure 1. For this dataset, one of the combinations (bc) is covered at the 50% level, three (ab, ac, ad) are covered at the 75% level, and two (bd, cd) are covered at the 100% level. And, as noted previously, for the whole set of tests, 79% of variable-value configurations are covered. Although the

TABLE 1. Example training dataset with four binary features.

a	b	c	d
0	0	0	0
0	1	1	0
1	0	0	1
0	1	1	1

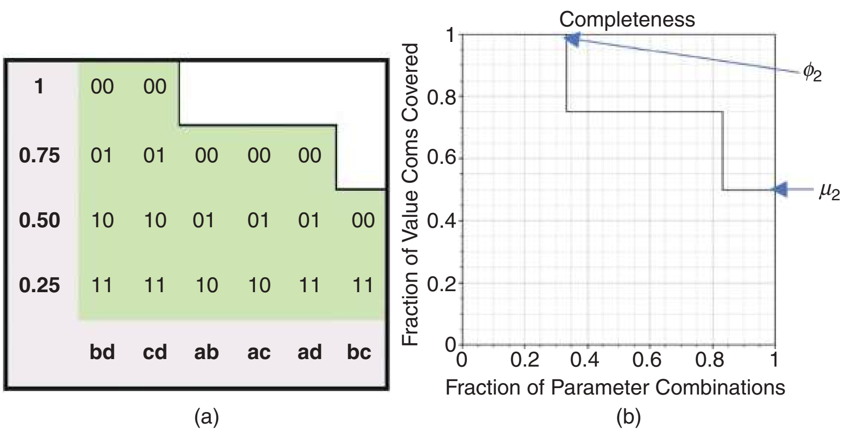


FIGURE 1. Graphical representation of two-way input space covering. (a) Table mapped to graph. (b) The concept of μ_t and ϕ_t.

example in Table 1 uses features with the same number of values, this is not essential for the measurement, and the same approach can be used to compute coverage for datasets in which features have differing numbers of values.[11]

easy to see that the (feature) value combinations in Figure 1(a) cover 19/24 = 79% of all possible two-way value combinations. Based on the ideas just presented, we define three measures S_t, μ_t, and ϕ_t. The measures are illustrated for two-way covering in Figure 1(b).

THE AREA UNDER THE CURVE REPRESENTS THE (FEATURE) VALUE COMBINATIONS COVERED, AND THE AREA ABOVE THE CURVE REPRESENTS COMBINATIONS THAT HAVE *NOT BEEN COVERED* IN THE TRAINING SET.

To make the data in Figure 2 more understandable, it will help to produce a graph whose components are easily tied to the data. One way to do this is shown in Figure 2. If the feature value combinations are listed as columns and sorted by the level of coverage for each combination, a graph as seen in Figure 1(a) results, called a combinatorial *completeness graph*.

The area under the curve represents the (feature) value combinations covered, and the area above the curve represents combinations that have *not been covered* in the training set. It is also

Vars	Configurations	Coverage
a b	00, 01, 10	0.75
a c	00, 01, 10	0.75
a d	00, 01, 11	0.75
b c	00, 11	0.50
b d	00, 01, 10, 11	1
c d	00, 01, 10, 11	1

FIGURE 2. A test array covering two-way combinations of a, b, c, and d to different levels.

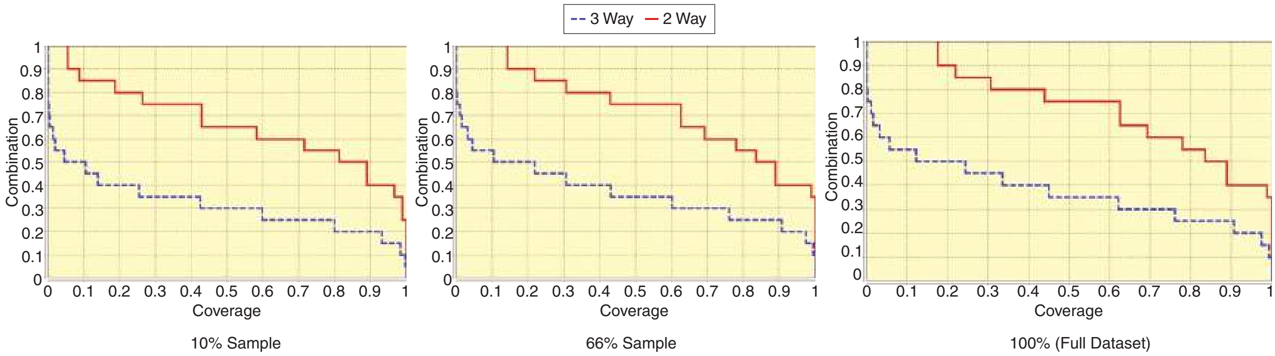
- S_t: Fraction of valid possible value combinations covered, that is, the area under the curve in a completeness graph, as shown in Figure 1(b).
- μ_t: The minimum *t*-way coverage present in the training space.
- ϕ_t: The proportion of *t*-way combinations for which all *t*-way value combinations are included.

ML significance

If μ_t is low, then there is at least one *t*-way (feature) value combination that is not covered well by the training dataset, so there may be insufficient evidence that the ML model will produce the intended output for that feature value combination. For example, a simple autonomous vehicle dataset might include features with two possible values each: *pavement (dry, wet)* and *lighting (light, dark)*. If the training includes combinations (*wet*, *light*), (*dry*, *light*) and (*dry*, *dark*), but not the other possible pair, then coverage for the feature pair *{pavement, lighting}* is 0.75. If this is the lowest coverage among all feature pairs, then $\mu_2 = 0.75$ and the *Y* intercept on the right side of the coverage graph would be at 0.75. In this case of $\mu_2 = 0.75$, we would know that we have at least 75% coverage for all pairs of features.

The value of S_t, area under the curve, is even more important. The feature value combinations covered are represented by the area S_t below the curve [see Figure 1 (b)]. Combinations that have not been included in any of the training samples are those above the curve. In the case of the example, these are *ab* = 11, *ac* = 11, *ad* = 10, and *bc* = 01, *bc* = 10. In most cases, there are two potential explanations for the missing combinations: 1) the training set is too small, or 2) there are constraints, known or unknown, among feature values. In the first case, it is difficult to say with confidence what will happen if these feature value combinations are encountered by the ML model, raising the possibility of classification error or unintended output. Thus, there may be a need to increase the training set size. However, simply increasing the training set randomly will not guarantee inclusion of different feature interactions.

The second situation is not uncommon in many supervised ML modeling scenarios. For example, if the problem is to identify dog breeds, there may be some dogs with a height of 8 in and some with a weight of 150 lb, but there will be none with the combination (height: 8 in, weight: 150 lb). Thus, if we measure the combinatorial coverage of the entire dataset, we would not find 100% coverage of possible two-way combinations. To help



pct Sample	2-Way Coverage	3-Way Coverage	2-Way Fraction of Full Set	3-Way Fraction of Full Set
10	0.665	0.319	0.907	0.826
33	0.711	0.359	0.970	0.930
66	0.729	0.379	0.995	0.982
80	0.732	0.382	0.999	0.990
100	0.733	0.386	1	1

FIGURE 3. Coverage of samples of full dataset, smoke detection.

understand the coverage obtained in training datasets, we can compute the combinatorial coverage of present in the training set as a proportion of the coverage in the full dataset. An example is shown in Figure 3, which shows coverage of a smoke detection dataset[12] that contains 35,746 observations. Even with this large set, many combinations are not seen, most likely due to constraints among the feature values. The full set contains 73.3% of the possible two-way combinations.

As noted, it is common in ML to have some value combinations that do not occur in the real world, or the environment where the model will be used. In this case, the best information on the possible values of features is represented by the complete dataset, so it is important that the sample used for training match the feature value combination in the full dataset. From the graphs and coverage figures, we can see that the 66% sample includes 72.9% of possible two-way combinations, or 99.5% of the full dataset.

COMBINATORIAL COVERAGE DIFFERENCE MEASURES

The previous section discussed the coverage defined by sets of value combinations of features, but after these sets have been computed it may be useful to do comparisons and differences between sets. For classification problems, where the goal is to determine the class (for example, type of animal) of an item, we refer to sets being distinguished as *Class* or *Non-class* sets (see Figure 4). A *Class set* is the one where certain value combinations of features, inputs, or configuration are present while in a *Non-class set*, these value combinations are absent or rare. Thus, there are two sets of interest as follows:

- C_t = set of t-way value combinations in Class file.
- N_t = set of t-way value combinations in Nonclass file.

We define a t-way combination c_t as a *distinguishing combination* for the class C if it is present in class instances C, and absent in nonclass instances N, or if it is more common in C than N as determined by a threshold value. A threshold T determines if a particular t-way value combination c_t is common in set C_t and rare in set N_t, and thus distinguishes one set from the other. That is, combination x_t for a class C is *distinguishing* iff occurrences $(x_t, C_t) > T \times$ occurrences (c_t, N_t), where occurrences $(x, Y) =$

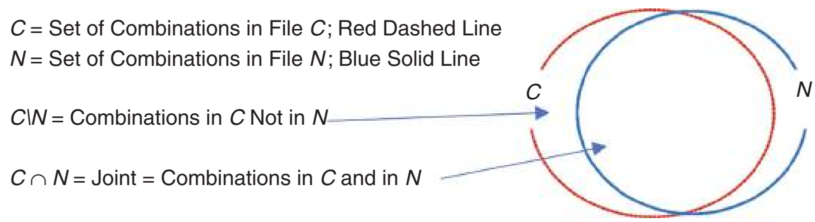


FIGURE 4. Class and nonclass file combinations.

frequency of value combination x in set of value combinations Y. If occurrences $(x_t, N_t) = 0$, then the combination x_t is unique to the class file. If T is some multiple, then the combination is rare in N and common in C; therefore, it is more strongly associated with C. The higher the value of T, the more strongly distinguishing combinations are associated with C. This threshold value is chosen based on the application.

ML significance

To illustrate these measures and their potential use, we can consider a small example. Suppose we have a database of animals as shown next (a $2^4 3^1 4^1$ configuration), and the task is to determine levels of similarity and differences. Consider differences between birds and nonbirds in the database, using the following attributes: fur: y, n; eggs: y, n; legs: 2, 4; wings: y, n; size: s, m, l; color: b(lack), w(hite), g(ray), r(ed). What t-way combinations are associated with these two classes?

Measuring one-way through four-way combinations, there are five single values that are common between the bird and nonbird classes, or 33.3% of the total number of single values. To make this easier to visualize, Figure 5 shows Venn diagrams of the combinations of features in the two classes. Only three two-way combinations are common between the two, and only a single three-way combination is shared. Because there is little overlap between the two classes, even for two-way combinations, we see immediately that this will be an “easy” problem for ML.

TABLE 2. Example animal identification task.

	fur	eggs	legs	wings	size	color
dog1	y	n	4	n	m	b
dog2	y	n	4	n	s	w
bat1	y	n	2	y	s	g
bat2	y	n	2	y	s	b
bird1	n	y	2	y	s	r
bird2	n	y	2	y	m	w
bear1	y	n	4	n	l	b
bear2	y	n	4	n	l	w

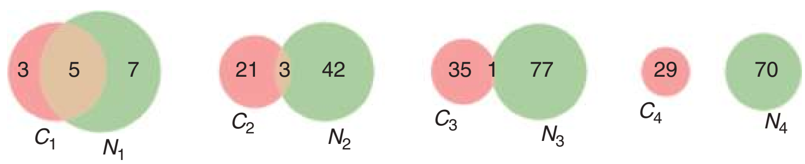


FIGURE 5. Distinguishing combinations for bird identification task.

SET DIFFERENCE COMBINATORIAL COVERAGE

The approach described previously can be refined into a measure of set difference combinatorial coverage (SDCC).[6] Given two sets A and B, the combinations in A but not in B will be A, and the SDCC measure is $|A \setminus B| \,/\, |A|$, the proportion of t-way combinations in A but not in B.

ML significance

SDCC has been used to evaluate the adequacy of training sets in a *transfer learning* context,[6] where a model is used in a new or changed environment that differs from the one used in training. Let us consider two datasets, a source dataset S, and a target dataset T. The ML model will be trained with S and tested on T to see how well the learning done on the source apply or transfer to the target. Let us compute SDCC $|T_t \setminus S_t| / |T_t|$ for some specified level of t-way combinations. This metric measures the level of value combinations in the target environment for which the model (derived from source S) has not been trained. This measure has been used in explaining an example where a model trained on a small dataset retained accuracy when applied to a dataset nearly twice as large, but with less diversity in feature value interactions than the smaller dataset. The SDCC measure showed that the small set had few combinations on which it had not been trained.

PER-COMBINATION COVERAGE

In some scenarios, measuring the per combination coverage present in a dataset can be useful. Consider the toy dataset shown previously. Here, Figure 6(b) shows all known value combinations for six features, and

Figure 6(a) shows two samples used for testing the model. Figure 7 shows two-way and three-way coverage of per-combination view. Each block in a coverage chart shows the level of coverage for a specific combination. Thus, there are 15 blocks in the two-way coverage chart because there are $C(6,2) = 15$ two-way combinations, and similarly 20 blocks for the three-way combinations. The 15 blocks for two-way heatmap shows (reading from left to right, top to bottom) coverage for (a, b), (a, c), (a, d), (a, e), (a, f), (b, c), ... (e, f). It can be seen that Figure 7 (left) contains all two-way value combinations for (a, b), (a, d), (a, f), (b, d), and so on. Here, the black square represents complete coverage. Similarly, lower levels of coverage for other combinations can be seen in other blocks such as (a, c), (a, e), (e, f) and so on. Figure 7 (right) shows per combination coverage of the $C(6,3) = 20$ three-way combinations for (a, b, c), (a, b, d), (a, b, e), and so on. For example, a black square indicates complete coverage for (a, b, d), while (a, b, c) has lower coverage indicated by the lighter (brown) color.

ML significance

For a real-world example of this visualization approach, see Figure 8, which shows graphs of two-way and three-way combinations in a 22-attribute mushroom identification ML problem.[13] The objective of this task is to identify combinations that are strongly or weakly associated with edible or poisonous mushrooms. The left chart shows summary coverage for $C(22,2) = 231$ two-way combinations, and the right chart shows $C(22,3) = 1{,}540$ three-way combinations. If all attribute combinations are covered to about the same level, then the matrices in Figure 8 will have similar colors, with coverage level indicated by the heatmap color bar on the right. In an ML context, the attribute or feature interaction coverage heatmaps could be used in comparing model training and test sets, to check that

a	b	c	d	e	f
1	0	1	0	1	1
1	0	1	1	1	1

(a)

a	b	c	d	e	f
1	0	1	0	1	1
1	0	1	1	1	1
1	0	0	1	0	1
1	0	1	0	1	1

(b)

FIGURE 6. Six-feature test set (a) Test set. (b) All known observations.

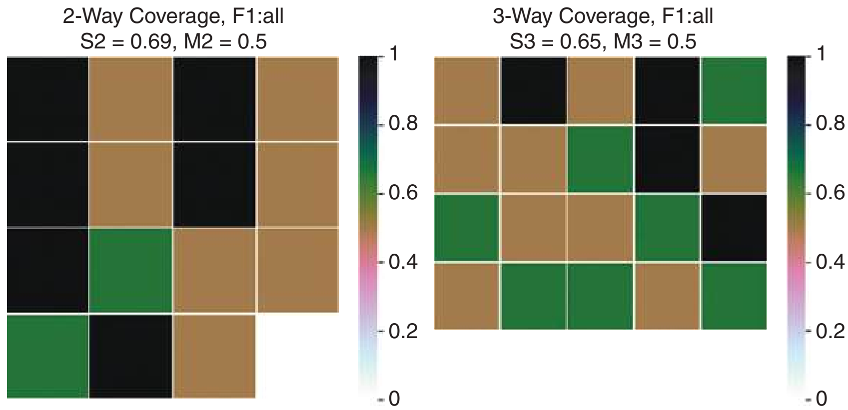


FIGURE 7. Per-combination coverage.

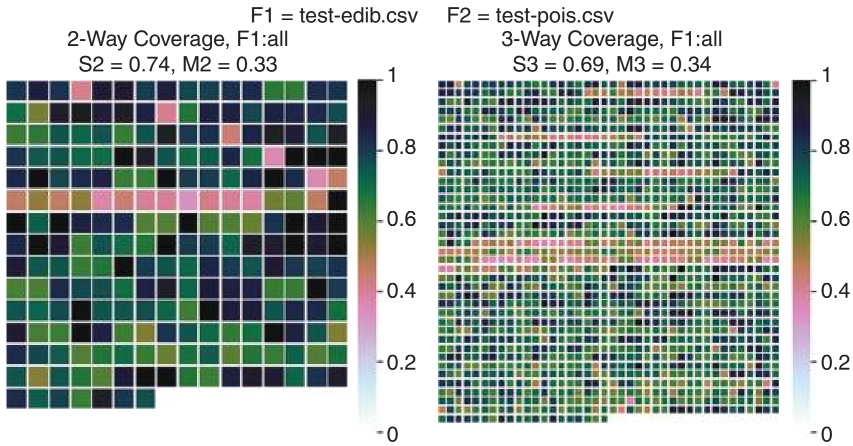


FIGURE 8. Summary coverage, two-way and three-way combinations for mushroom problem.

distribution of combination coverage is nearly the same for both sets. Note that this coverage may vary with specific attributes or features, and they may not have the same level of value coverage. But if training and test sets both have similar feature interaction distribution, then the heatmaps should be very similar. This approach thus gives us a quick, visual way to identify weak spots in a dataset, or confirm similarity between datasets, especially for training and testing datasets, for ML or other applications.

> **THE OBJECTIVE OF THIS TASK IS TO IDENTIFY COMBINATIONS THAT ARE STRONGLY OR WEAKLY ASSOCIATED WITH EDIBLE OR POISONOUS MUSHROOMS.**

COMBINATORIAL FREQUENCY DIFFERENCES

As noted previously, another useful combinatorial measure is the degree to which a particular combination is associated with a class. If a feature combination is seen in a high proportion in one class but is rare in other classes, it may help to identify that class. Informally, we can define this frequency difference measure as the difference between occurrence rate for a combination in one class compared with another.[7,14] For example, if a combination of *long tail* and *curly fur* occurs in 0.90 of a particular dog breed and in 0.15 of another, then the two-way combinatorial frequency difference (CFD) is 0.75. We should note that we use the terms combinatorial frequency difference and combination frequency difference interchangeably in our publications. Measuring these differences across all *t*-way combinations in a dataset can be useful in visualizing ML model effectiveness, as shown in the following example.

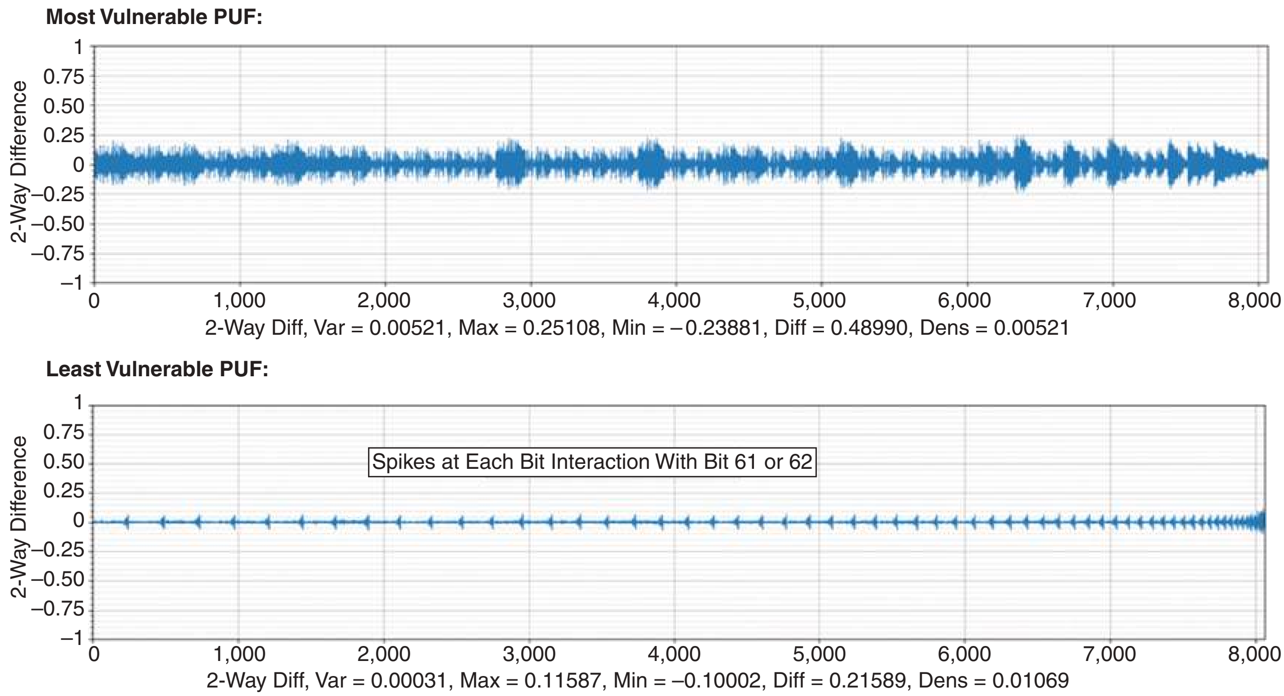


FIGURE 9. Frequency differences for PUFs.

A physically unclonable function (PUF) is a circuit implementing a black box function that produces a response *r* for a given challenge string of bits *c*, that is, $r = f(c)$, with a response of 0 or 1. PUFs utilize unique properties of the physical elements within the hardware such as the small variation in propagation delays between identical circuit gates or small threshold mismatches in transistor feedback loop due to process variation. These physical characteristics are very difficult to reproduce in the hardware, which is what makes them physically unclonable and useful for applications such as secure storage, detection of hardware counterfeiting, and so on.

ML significance

Many authors have shown that ML models can be constructed to predict the output of PUFs for a given input string, thus "breaking" the PUF by defeating its authentication function. Vulnerability to breaking through ML attacks can vary significantly with PUF design, and a challenge in developing PUFs is to identify potential weaknesses before constructing the PUF. Figure 9 is a CFD graph.[14] It shows differences in the frequency of response "1" (above the center line) or "0" (below the center line) for two different classes of PUF. Note that this is the output of the PUF and not the "*y*-axis" value of the graph. The more vulnerable PUF with an easier-to-predict output will have a dense set of input value combinations associated with a particular response of "0" or "1," while the least-vulnerable PUF will have weak associations, barely above the zero centerline. Thus, the comparative resistance of a PUF design to an ML attack is immediately obvious from the graphs.

ABOUT THE AUTHORS

D. RICHARD KUHN is a computer scientist in the Computer Security Division, National Institute of Standards and Technology, Gaithersburg, MD 20899 USA. His research interests include combinatorial methods in software verification and testing, including assurance and explainability in artificial intelligence and machine learning. Kuhn received his M.S. in computer science from the University of Maryland, College Park. He is a Fellow of IEEE. Contact him at kuhn@nist.gov.

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

RAGHU N. KACKER is a scientist in the National Institute of Standards and Technology, Gaithersburg, MD 20899 USA. His research interests include testing of software-based systems for trust and security, and mathematics of measurement. Kacker received his Ph.D. from the Iowa State University. Contact him at raghu.kacker@ nist.gov.

We have introduced some measures and associated visualizations for evaluating training and testing data used in ML. Three classes of measures include combinatorial coverage, combinatorial coverage differencing, and combinatorial frequency differencing. Primary applications include the following:

- *Training sample evaluation*: Computing *t*-way combinatorial coverage (CC) of a training or testing sample as a fraction of coverage of the full dataset (see Figure 3) provides a view of sample adequacy.
- *Transfer learning*: SDCC can help determine if a model trained in a source environment will be adequate in a target environment.[6]
- *Model understanding*: CFD can provide an initial idea about how easy or difficult a model will be to learn.
- *ML explainability*: CC, SDCC, and CFD can work together to provide insight into what an ML model has learned or what factors or features interactions have been most influential in an ML model.

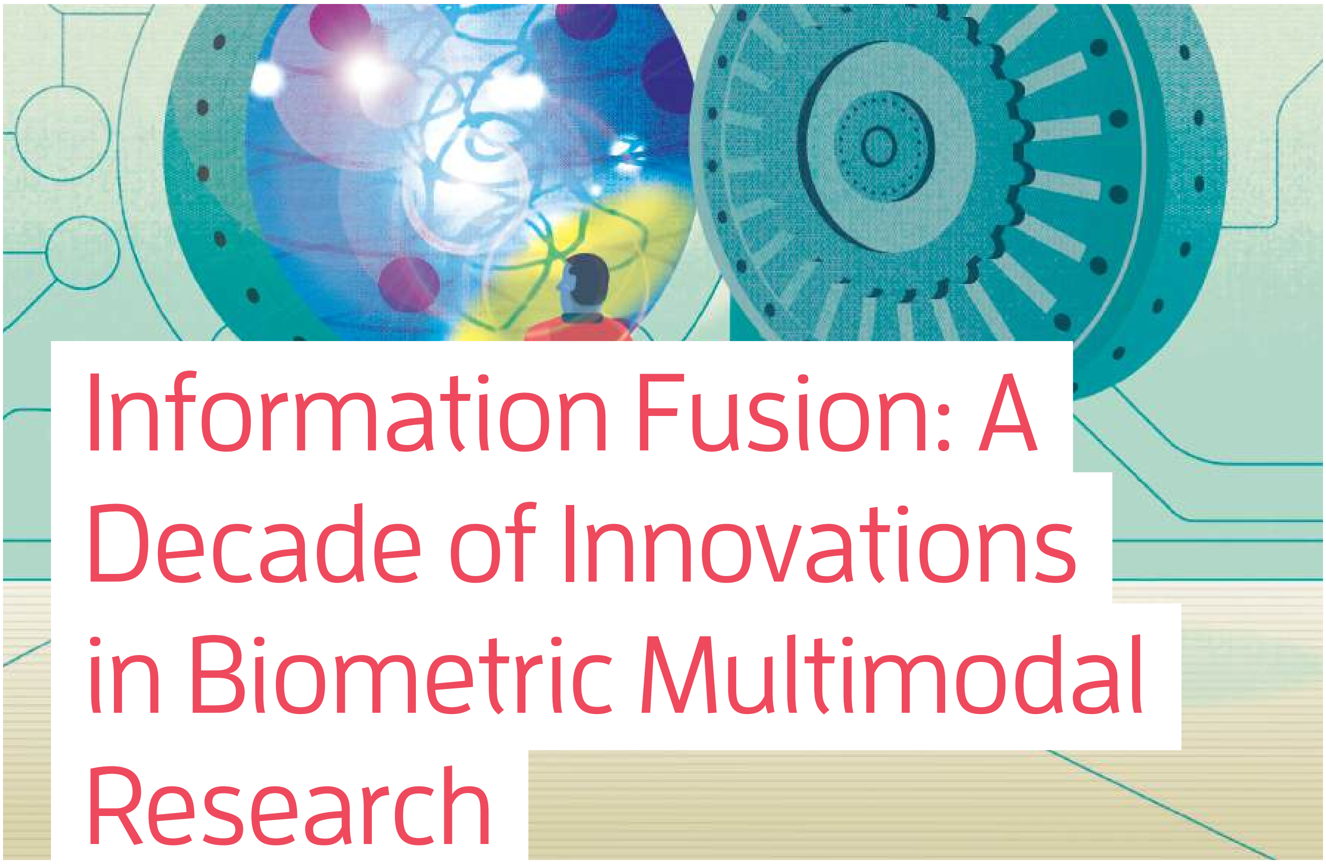
Using *t*-way combinations of feature or attribute values provides a new approach for measuring similarity and difference between datasets. Additionally, *t*-way combinations may be used to distinguish between datasets that may appear similar based only on single-attribute values. These measures are useful in providing quantitative measures of data completeness, transfer learning, training dataset evaluation, model understanding, and explainable AI. As one of the few quantitative approaches for evaluating ML training data adequacy, combinatorial coverage methods have significant potential in domains with high assurance needs, such as autonomous vehicles or health care. More information and publications can be found at the project site.[15] C

ACKNOWLEDGMENT

Reference to commercial products or organizations is for information only; it does not imply recommendation or endorsement by NIST, nor that the products mentioned are necessarily the best available for the purpose.

REFERENCES

1. M. Trujillo, M. Linares-Vásquez, C. Escobar-Velásquez, I. Dusparic, and N. Cardozo, "Does neuron coverage matter for deep reinforcement learning? A preliminary study," in *Proc. IEEE/ACM 42nd Int. Conf. Softw. Eng. Workshops*, pp. 215–220, 2020, doi: 10.1145/3387940.3391462.
2. F. Harel-Canada, L. Wang, M. A. Gulzar, Q. Gu, and M. Kim, "Is neuron coverage a meaningful measure for testing deep neural networks?" in *Proc. 28th ACM Joint Meeting Eur. Softw. Eng. Conf. Symp. Found. Softw. Eng.*, pp. 851–862, 2020, doi: 10.1145/3368089.3409754.
3. Z. Yang, J. Shi, M. Hilmi Asyrofi, and D. Lo, "Revisiting neuron coverage metrics and quality of deep neural networks," in *Proc. IEEE Int. Conf. Softw. Anal., Evol. Reengineering (SANER)*, Piscataway, NJ, USA: IEEE Press, pp. 408–419, 2022, doi: 10.1109/SANER53432.2022.00056.
4. D. R. Kuhn, I. Dominguez Mendoza, R. N. Kacker, and Y. Lei, "Combinatorial coverage measurement concepts and applications," in *Proc. IEEE 6th Int. Conf. Softw. Testing, Verification Validation Workshops*, Piscataway, NJ, USA: IEEE Press, 2013, pp. 352–361, doi: 10.1109/ICSTW.2013.77.
5. D. R. Kuhn, M. S. Raunak, and R. Kacker, "Combinatorial coverage difference measurement," National Institute of Standards and Technology, Gaithersburg, MD, USA, Tech. Rep., 2021. [Online]. Available at: https://nvlpubs.nist.gov/nistpubs/CSWP/NIST.CSWP.06222021-draft.pdf
6. E. Lanus, L. Freeman, D. R. Kuhn, Raghu, N. Kacker, and Y. Lei, "Combinatorial testing metrics for machine learning," in *Proc. IEEE Int. Conf. Softw. Testing, Verification Validation*, Piscataway, NJ, USA: IEEE Press, 2021, pp. 81–84, doi: 10.1109/ICSTW52544.2021.00025.
7. R. Kuhn, M. S. Raunak, and R. Kacker, "Combination frequency differencing," US Department of Commerce, Tech. Rep., Washington, DC, USA, 2021. [Online]. Available at: https://nvlpubs.nist.gov/nistpubs/CSWP/NIST.CSWP.12062021-draft.pdf
8. J. Chandrasekaran et al., "Leveraging combinatorial coverage in the machine learning product lifecycle," *Computer*, vol. 57, no. 7, pp. 16–26, Jul. 2024, doi: 10.1109/MC.2024.3366142.
9. T. Cody, E. Lanus, D. D. Doyle, and L. Freeman, "Systematic training and testing for machine learning using combinatorial interaction testing," in *Proc. IEEE Int. Conf. Softw. Testing, Verification Validation Workshops (ICSTW)*, Piscataway, NJ, USA: IEEE Press, 2022, pp. 102–109, doi: 10.1109/ICSTW55395.2022.00031.
10. M. S. Raunak, D. R. Kuhn, R. N. Kacker, and Y. Lei, "Ensuring reliability through combinatorial coverage measures," *IEEE Rel. Mag.*, vol. 1, no. 2, pp. 20–26, Jun. 2024, doi: 10.1109/MRL.2024.3389629.
11. D. Richard Kuhn, D. R. Wallace, and A. M. Gallo, "Software fault interactions and implications for software testing," *IEEE Trans. Softw. Eng.*, vol. 30, no. 6, pp. 418–421, Jun. 2004, doi: 10.1109/TSE.2004.24.
12. "Smoke detection dataset." Kaggle. Accessed: Oct. 1, 2024. [Online]. Available: https://www.kaggle.com/datasets/deepcontractor/smoke-detection-dataset
13. "Edible mushroom classification dataset." UCI Machine Learning Repository. Accessed: Oct. 1, 2024. [Online]. Available: https://archive.ics.uci.edu/dataset/73/mushroom
14. D. R. Kuhn, M. S. Raunak, C. Prado, V. C. Patil, and R. N. Kacker, "Combination frequency differencing for identifying design weaknesses in physical unclonable functions," in *Proc. IEEE Int. Conf. Softw. Testing, Verification Validation Workshops (ICSTW)*, Piscataway, NJ, USA: IEEE Press, 2022, pp. 110–117, doi: 10.1109/ICSTW55395.2022.00032.
15. "Combinatorial methods for trust and assurance." National Institute of Standards and Technology (.gov). Accessed: Oct. 1, 2024. [Online]. Available: https://csrc.nist.gov/acts



Information Fusion: A Decade of Innovations in Biometric Multimodal Research

Marina L. Gavrilova, University of Calgary

This article presents an overview of the decade of exploration into the emerging traits related to the development of multimodal biometric systems. It showcases advantages of deep learning architectures, surveys new behavioral traits based on online social media, and identifies importance of keeping user's data privacy in mind when developing new generation of trustworthy and explainable biometric systems.

Automatic decision making, especially when dealing with crucial cybersecurity and privacy issues, has become one of the key directions in digital identity research. Trustworthy human–computer collaboration highlights the multifaceted role of artificial intelligence (AI) and deep learning in biometric online security. Intricate relationships among biometric security, privacy, and human-driven decision making, lead to new insights on how to ensure safe and secure digital space. Within this framework, information fusion is increasingly important to ensure accurate, reliable, and trustworthy decision making,[1,23] Various scoring systems—including feature, rank, and decision-level fusion—have been broadly utilized in intelligent biometric systems, cognitive systems, artificial neural networks, pattern recognition, and AI. Examples of truly remarkable innovations in this space include a new method to evaluate the predictive behavior of a scoring system through combinatorial fusion

Digital Object Identifier 10.1109/MC.2025.3526135
Date of current version: 27 March 2025

analysis.[2] The approach is based on combining the rank-score characteristic function and cognitive diversity.[2] A seminal work recently published in *Computer* summarizes the paradigm of combinatorial fusion through combining various scoring systems across computational modeling, informatics, and intelligent systems.[3]

In addition to multimodal systems, biometric online user recognition has been further investigated from the point of view of security and privacy. Insights from those investigations are crucial in assisting policy makers and governments to pave the way toward a safer society. Biometric security is no longer limited to the physiological features of human beings, especially in this era of rapidly growing online activities. Human behavior, which is expressed through social interaction and communication, is an important source of biometric data. These so-called *digital footprints* reveal character traits and habitual patterns, which can be utilized for person authentication or granting access to a resource. The applications of this research include a wide variety of domains, such as person authentication, anomaly detection, behavior analysis, customer profiling, e-commerce, and situation awareness.[1]

AUTOMATIC DECISION MAKING, ESPECIALLY WHEN DEALING WITH CRUCIAL CYBERSECURITY AND PRIVACY ISSUES, HAS BECOME ONE OF THE KEY DIRECTIONS IN DIGITAL IDENTITY RESEARCH.

Over the past decade, research has explored human behavior in a new light, as a rich source of biometric information. In 2014, an idea was born that an individual's social communications contain valuable sources of unique human behavior that can be explored for person authentication. Thus, the social behavioral biometrics (SBB) has been introduced to explore unique behavioral traits expressed through users' communication.[2] In the era of online social networks, where peoples' identity has expanded beyond the real world to the cyberspaces and virtual reality, SBB has demonstrated its ability to capture inherent habits and sentiments of humans expressed through interactions in virtual worlds. This article looks at the core of the problem from multiple perspectives by examining biometrics and information fusion research over the past decade.

The subsequent sections summarize a decade of multimodal biometric system design, conducted at the Biometric Technologies Laboratory (BTLab) at the University of Calgary in Canada. The material is based on the Fordham Clavius Distinguished lecture, delivered by the author in November 2023, as well as research conducted at the BTLab over past decade and disseminated through book chapters.[1,24] Some of the presented material follows closely a book chapter in *Breakthroughs in Digital Biometrics and Forensics*,[1] presenting a concise description of original research conducted at the BTLab at the University of Calgary.

MULTIMODAL BIOMETRIC SYSTEMS

Over the past decade, processing, visualizing, and interpreting vast amounts of information, coming from a variety of sources (sensors, cameras, social media, etc.) and including various applications (biometric data, medical data, geospatial data, etc.), have radically changed. Innovations in data analytics and AI domains are introduced with accelerating pace, thus creating very real challenges to the way we as humans make informed decisions. At the same time, standards on the development of ethical and trustworthy AI-powered decision-making systems remain at their infancy.

This article looks at the core of the problem from multiple angles by examining the past decade of biometric and information fusion research. From the early stages of multimodal biometric system development and information fusion to social behavioral authentication and emotion-aware robots, it examines societal changes, which the development of new transformative technologies brings. Recent state-of-the-art methods in the information security domain are based on deep learning systems integration with multimodal biometric research. This article discusses the efficacy of hybrid deep learning architectures, leveraging handcrafted features based on domain knowledge with powerful deep-network architectures. Examples of concrete projects from public health, communication, and cybersecurity domains support the methodological development.

Currently, the potential for new innovations is contrasted with an increased risk to an individual's privacy and ethical decision making. Solutions based on the notion of cancellability and adaptive de-identification are presented as powerful mechanisms to address the privacy concerns. The article is concluded with discussion of open problems and challenges of mitigating the impact of security-enhancing technologies—such as video surveillance, location tracking, and smart sensors—on user's privacy and trust.

TRUSTWORTHY BIOMETRIC SYSTEMS

The need for trustworthy, reliable, and explainable models is especially high in the biometric authentication domain, where real-time identification of an individual or detection of a high-risk situation require both high precision and trust. While a significant progress was made in the development of powerful deep learning architectures, key unresolved challenges relate to their black box nature, difficulty in generalization to other datasets, and inability to provide confidence in their decisions. This research program will address the aforementioned deficiencies by focusing on the issues of trust, privacy, confidence, reliability, authenticity, and explainability of biometric decision making. One of the key research challenges is in how a multimodal biometric-based authentication system can be developed in a manner that ensures confidence in its decisions, guarantees result accuracy and reliability, provides decision-making transparency, and warrants biometric data privacy.

There is a pressing need to inform the development of new-generation biometric-based identity management systems, incorporating principles of trustworthiness, explainability, reliability, and data privacy. According to the Computer Security Resource Center definition: "Trustworthiness is a concept that includes privacy, reliability, resilience, safety and security." Trustworthy decision making is defined as the ability of an intelligent computer system to perform a real-time task repeatedly, reliably, and dependably in complex real-world conditions, combined with the response of a human user who needs to accept and trust the system's recommendation. However, very few biometric systems include metrics to report how certain they are in the accuracy and correctness of their decisions, and only a fraction possess self-awareness to admit they are uncertain in their recommendations.[1,25,28]

EXPLAINABILITY AND INTERPRETABILITY

Explainable and interpretable decision making powered by the new generation of deep learning architectures recently emerged as one of key societal challenges, requiring immediate attention. The need for explainable models has been documented in medical diagnostics, recommender systems, emergency response, and risk management, to name a few.[1,24,28] In the biometric security domain, the amount of research devoted to the development of deep learning architectures has skyrocketed.[4,22,24] The recently proposed state-of-the-art methods include a graph convolutional neural network (CNN) for gait authentication, a feed-forward neural network with custom loss function for masked face recognition, and a bimodal sequential neural network with handcrafted features injection for gait emotion recognition.[4,5,25] These methods focus predominantly on achieving the highest possible identification rate while ensuring that the model does not overfit. While performing ablation studies and extensive hyperparameter tuning have quickly became the domain norm, the same cannot be said about deep learning feature engineering, algorithmic transparency, and explainability of results.

Generally, a deep neural network is considered a black box that does not provide precise information about its internal processes.[27] Here, by "explainable biometric system," we mean the system where the reasoning behind the decisions and predictions is understandable and transparent.[26] To achieve explainability of AI-based systems, prediction and classification outcomes of AI models must be transparent and understood by a human.[26] Current approaches to AI model explainability are usually restricted to annotating and labeling the datasets, and to stepping through algorithms during their execution. The approaches we believe would be most beneficial to drive innovations include incorporating a domain-specific feature learning into personable explainable models and developing a visual framework for interpretability of the biometric system decision. This will lead to building inherently explainable models, with a clearly visible decision-making transparency.

BIOMETRICS PRIVACY

A recent discourse in biometric literature highlighted the pressing need for privacy preservation. Biometric data collection, storage, preprocessing, and subsequent use for authentication or verification purposes should be governed by strict regulations. It is also of significant importance not to compromise individuals or allow for subsequent data mining for all actors to instill proper procedures for access to the data, understanding how the technology is used, in what form it can be stored, and who is allowed to access this data. Furthermore, the recent rise of deep neural networks now provides unparalleled opportunities to mine information not only pertaining to a single individual, but to the whole population.

While biometric template protection schemes, such as cancellable biometrics and biohashing, have been extensively studied, exciting new approaches to biometric privacy based on information concealment through the process of de-identification are only starting to gain momentum.[24] Existing de-identification research is predominantly restricted to visual biometric data (facial images and videos). Original methods, based on blurring, pixelation, or clustering, are inadequate in producing realistic-looking images. The recent development of deep learning architectures assisted with this task; however, the inability of such approaches to adaptively control the parameters of de-identified biometrics, as well as significant loss in system identification ability, are problematic. Current state-of-the-art research addresses both deficiencies by developing the personably controllable multimodal de-identification system,[24] which enables adaptive de-identification while retaining accurate and reliable biometric recognition.

TO ACHIEVE EXPLAINABILITY OF AI-BASED SYSTEMS, PREDICTION AND CLASSIFICATION OUTCOMES OF AI MODELS MUST BE TRANSPARENT AND UNDERSTOOD BY A HUMAN.

Investigating de-identification approaches also supports the development of a trustworthy authentication system: The original biometric can be changed or transformed in a controllable manner. In case it is compromised, the overall system integrity and recognition efficacy remain intact. Thus, we believe that controllable biometric de-identification through selective biometric concealment and cancelable multimodal template protection mechanisms are the ways of future privacy-preserving approaches.

BIOMETRIC SYSTEM CLASSIFICATION

A biometric system is defined as "a pattern-recognition system that is designed to verify or identify a person based on their intrinsic physiological or behavioral characteristics."[3] A general taxonomy of biometrics is depicted in Figure 1.[1] Over the past decade, the use of biometric technologies and systems has significantly expanded within the public and private sectors by offering very compelling person identification and verification solutions. In contrast to the traditional identification and verification methods that use passwords or personal identification numbers, biometric systems are deemed to be exceptional in terms of nontransferability, portability, and acceptance.[3] Multiple issues of their integration to new technologies were examined, including privacy[3] and data quality.[28] One of the most fascinating and fastest growing domains of biometric research is related to human aspects of decision making.

A new avenue for next-generation biometric systems development was opened decades ago, with the transformational shift toward multimodal systems relying on information fusion. A key goal of a multimodal system is the same as of a unimodal one, using a single biometric trait or multiple traits for verification or identification.[3] Multimodal biometrics have been proven to be more reliable, accessible, accurate, and versatile, being used in numerous applied settings, from customs offices to amusement park entries. Information fusion based on probabilistic models ensures not only an accurate identification, but also sets a confidence in such identification.

AI AND DEEP LEARNING METHODS FOR BIOMETRIC

The past decade has witnessed significant developments in the domain of physiological and behavioral biometric recognition. Numerous examples of successful integration of the latest biometric technologies in practice can be found in the domains of information security, surveillance, cybersecurity, medicine, finance, education, retail, and others. Recently, physiological and behavioral biometrics harnessed unprecedented opportunities for the integration of the deep learning framework, with physiological and behavioral biometric data processing. Research studying both physiological biometrics—such as face, fingerprint, ear, and iris—and behavioral biometrics—such as gait, voice, and signature—have embraced the successful deployment of deep learning. One of such examples can be found in the work by Grm et al.,[4] exploring deep learning architecture for face recognition in the presence of image degradation.

False positives and false negatives in traditional biometric methods present continuing problems for the abovementioned applications. The introduction of multimodal systems helped to minimize their negative impact on accuracy and precision. It is anticipated that further advancements stemming from incorporating deep learning in the data preprocessing and decision making would be a game changer in this respect.

The recent evolution in AI is due to the deep learning advancements. It is remarkably different from the traditional machine-learning approaches that used to heavily rely on handcrafted features. Deep learning thus has improved the ability of computers to detect, recognize, classify, and describe patterns, including perceived cybersecurity risks, medical diagnostics, or biometric identification. CNN was the first to be integrated in biometric research.[5] They are typically used for classification and detection tasks.[5]

To effectively deal with sequential data, and to incorporate previous computational steps into the architecture, the recurrent neural network was designed.[5] They are particularly useful for natural language processing, translation from other languages, and speech diarization tasks. They have also been successfully used in biometric gait recognition.[5]

For the task of generating high-quality synthetic images and signals, generative adversarial networks are highly useful.[7] In the biometric security domain, such architectures are used for image and video synthesis, voice and sound generation, and text-to-image conversion. The architecture consists of a generator and a discriminator.[7] The generator attempts to create data, and the discriminator tries to distinguish original and synthetic data. The two architectures are in constant competition and thus improvement.

Deep learning algorithms require a large amount of training data to reduce loss and prevent overfitting. However, sometimes it is difficult to gather a large amount of data due to several limitations. Transfer learning methods were proposed to aid

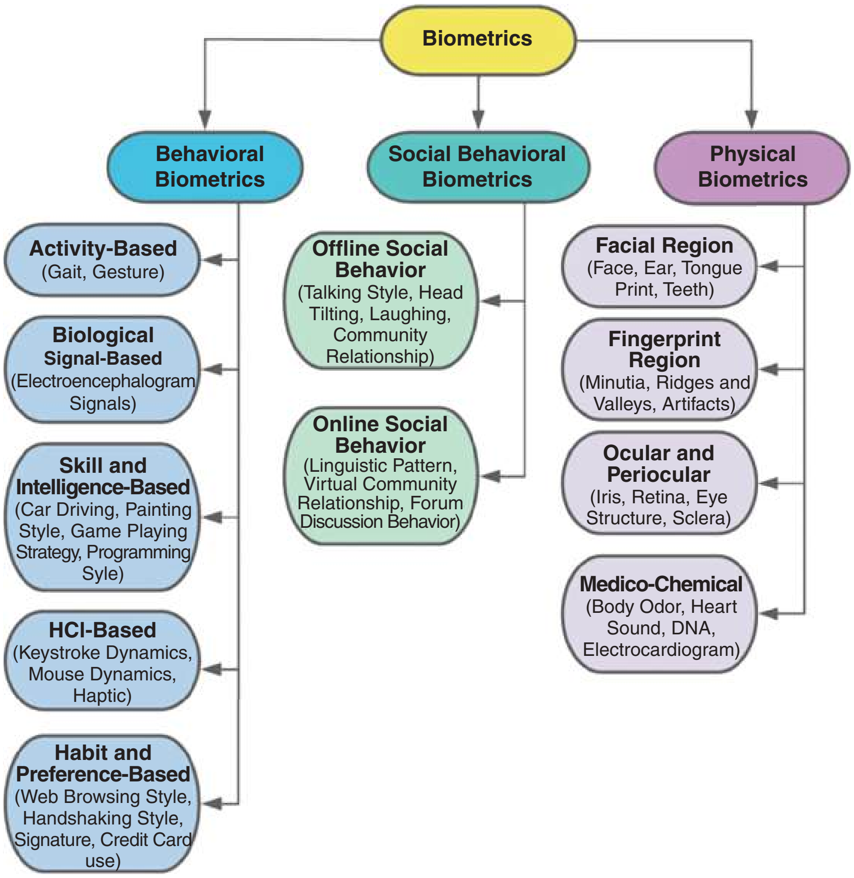


FIGURE 1. A taxonomy of biometric traits.[1]

this problem.[6] In transfer learning, a model that was trained for a specific problem can be reused for the purpose of solving a different problem.

Transformers became the newest addition to the fastest-growing deep learning architectures. Transformers make full use of self-attention by replacing earlier neural network models that used attention between encoder and decoder.[6]

Based on the aforementioned survey, the vast variety, constant innovation, and advanced capabilities of deep learning architectures make them highly useful for multimodal biometric architectures in a variety of heterogeneous modalities. Over the recent decade, deep learning architectures have been successfully integrated into facial recognition, fingerprint identification, speaker diarization, gait and activity recognition, and other biometric authentication systems.[6] In 2018, the first multimodal biometric data-driven CNN architecture based on face, iris, and fingerprint features was introduced.[7]

A deep learning–based biometric system resembles a classical architecture and consists of two stages: enrollment and authentication.[1] During the enrollment phase, the biometric data are obtained using sensors (camera, video, sound recording, etc.). The preprocessing is then applied, including noise reduction, data anonymization, etc. After training and hyperparameter tuning, the model is ready to be used for identification or verification. First, the enrolled user data are sent into the deep learning model to obtain two intermediate embedding feature vectors. Subsequently, a proper distance function is utilized to compute the distance between these two feature vectors. If the result is higher than a particular threshold, the authentication is successful. The generic architecture of the biometric system based on deep learning is presented in Figure 2.[1]

EMERGING BEHAVIORAL BIOMETRIC TRAITS

Recent advancements in the biometric domain have led to new concepts of behavioral biometric, encompassing contextual, soft, and communication traits. These are referred to as *SBB*.[1] Typically, those traits include building relationship networks to represent relationships of friendship, user hobbies, temporal, emotional, psychological, linguistic, and esthetic patterns.[2] Such relationships play an increasingly important role in the domains of online security and cybersecurity.

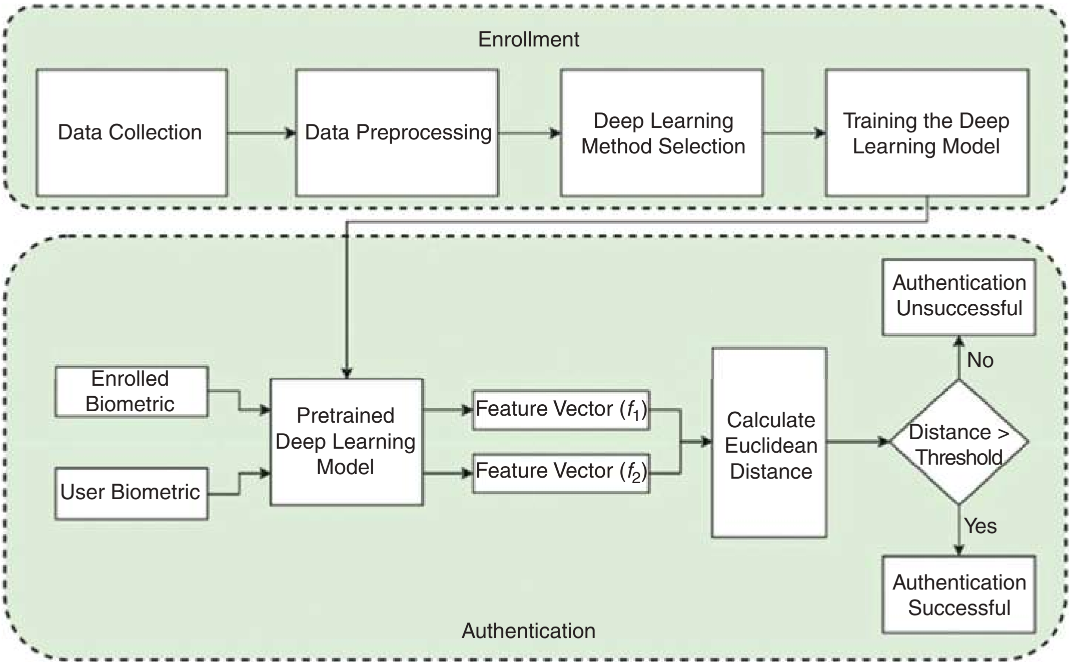


FIGURE 2. A schematic representation of a generic architecture of biometric authentication system based on deep learning.[1]

Figure 3 presents an overview of all of the emerging research domain areas currently being investigated at the Biometric Technologies Lab at the University of Calgary, Canada.

HUMAN FACTORS FOR TRUST AND EXPLAINABILITY

Humans experience emotions in response to life events, such as career advancements, interpersonal relationships, and professional careers. This success is also driven by human personality traits.[8]

A recent study[9] designed a personality traits classification metric based on communication patterns of online social platform (Twitter) users. It established that language-based models can consistently predict certain personality traits and introduced a Linguistic Personality Traits Assessment system for personality identification of users online from their written texts.[9] The emotional states expressed in online communications across social media is another important digital footprint. Recently, BTLab research introduced a new behavioral modality based on human esthetic preferences.[1] Additional discoveries were made in understanding how human emotional states can be used to perform behavioral pattern analysis of online users.[10,11,12] Researchers also discovered that human esthetic preferences can be utilized for gender and user authentication.[1,14]

In the BTLab, we have introduced new biometric esthetic systems (visual esthetic, audio esthetic, and video esthetic) and combined them into a multimodal esthetic system.[24] Applications of biometric esthetic systems include continuous authentication, recommendation systems, and consumer behavioral analysis.

Human factors play a crucial role in security systems design. Prior biometric security research mainly focused on physiological and behavioral biometrics. Current research we conduct is focused on social behavioral traits.[8] The current research

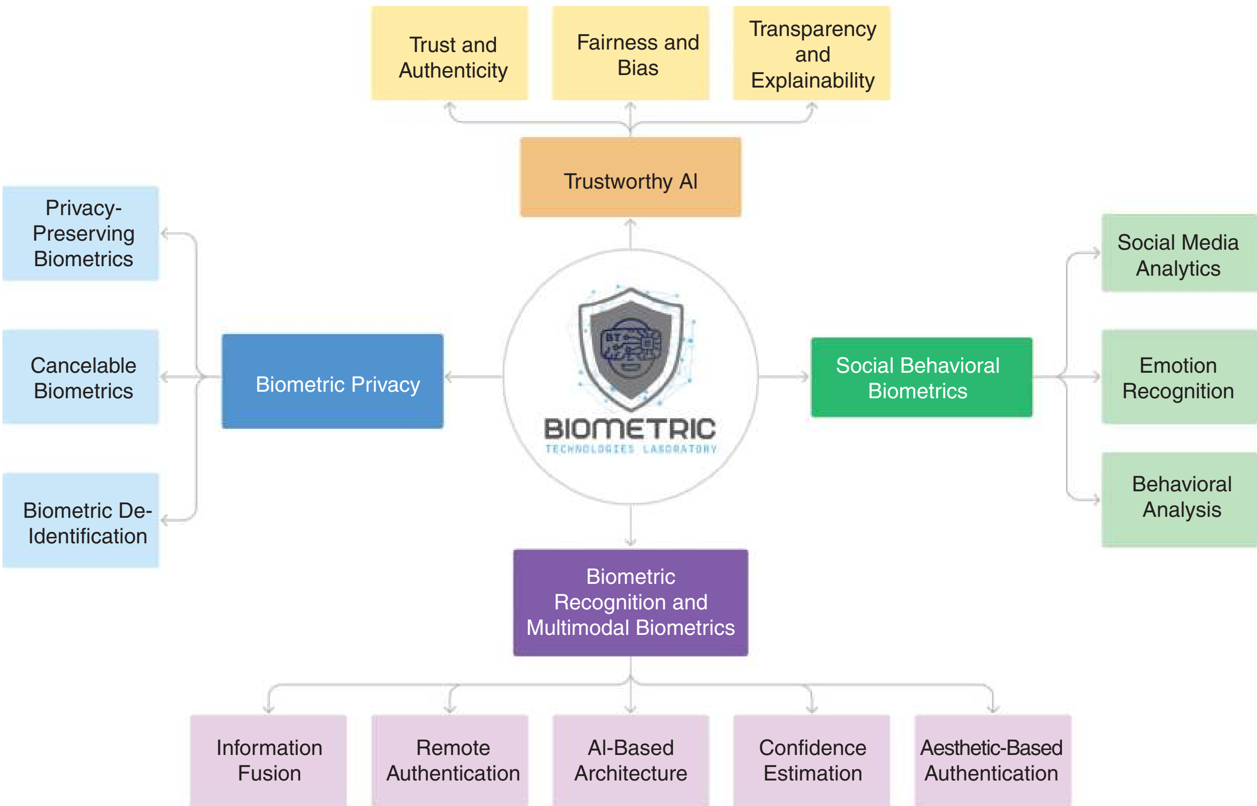


FIGURE 3. Emerging research in multimodal biometric domain conducted at the BTLab.

leverages these findings to create explainable and transparent multimodal systems by considering behavioral factors in their design. Incorporating features obtained from emotional, esthetics, and nonverbal expressions into the biometric system will allow creation of personable and explainable models. Finally, a potent direction for future research is to study the role of online human communication in trust and explainability of decision making.

FUTURE RESEARCH DIRECTIONS

Introduction of new deep learning architectures, capable of targeted injection of domain-specific knowledge and feature learning at deep layers of the network, is one of the emerging research domains. This research will uncover deep learning systems reasoning by incorporating insights from iterative graph-like cognitive system structure with nonmonotonic logical reasoning and deep inference. The main advantage of this approach is that it allows incremental discovery of previously unknown deep features, while merging them with a traditional deep learning architecture for subsequent biometric matching. It is anticipated that such novel hybrid biometric architectures will not only increase reliability of biometric identification, but also provide added benefits of explainability and interoperability of decisions.

Discovery of new multimodal systems and information fusion methods can serve as a further catalyst for advanced research in the domain of cybersecurity, medical forecasting, virtual reality, online games, and consumer electronics. There are also many open problems related to ensuring privacy in social behavioral biometrics data acquisition, training AI-based systems without bias, and using the data ethically for decision-making, while ensuring accuracy, reliability, and trustworthiness. The next decade will bring forward solutions to pressing societal challenges related to fairness, explainability, and trustworthiness of decision making, balancing the user's data privacy with fast and convenient access to systems and resources and increasing reliability of rea-time user authentication based on biometric data.

Developing hybrid approaches based on a combination of handcrafted features with deep learning architectures, integrating concepts of data de-identification, and cancellability in real-time safety crucial applications, and ensuring seamless translation of research to our society will drive forward innovation in the biometric security research domain.

Another important future direction of research is related to data interchange. Different decision-making systems produce different data items, often stored in inconsistent ways. This is particularly relevant to medical and biometric data. It will be important to consider how the fusion approaches could help with data interchange problems.

Development of trustworthy and explainable decision-making architectures became a key focus of most recent research discourse. In addition, bias mitigation when working with data, as well as its ethical use, are of paramount importance. Incorporation of personality features, extrapolated from the online communication, in combination with emotional and psychological traits, could be one of the approaches to address the issue.

Finally, examining the concept of trustworthy and explainable multimodal systems is brand-new emerging research. Overall, active research focused on incorporation of information fusion with the development of a new generation of trustworthy AI-based systems will be crucial for ensuring safe and secure societies. ■

ACKNOWLEDGMENT

This research was partially funded by a Natural Sciences and Engineering Research Council of Canada (NSERC) Discovery grant, a NSERC Strategic Partnership grant, the NSERC Alliance, the University of Calgary Institute for Transdisciplinary Studies, and the University of Calgary Research Excellence Chair in Trustworthy and Explainable Artificial Intelligence Program.

REFERENCES

1. M. Gavrilova et al., "A multifaceted role of biometrics in online security, privacy, and trustworthy decision making," in *Breakthroughs in Digital Biometrics and Forensics*, K. Daimi, G. Francia III, L. H. Encinas, Eds., Cham, Switzerland: Springer-Verlag, 2022, ch. 14, pp. 303–324.
2. L. Hurley, B. Kristal, S. Sirimulla, C. Schweikert, and D. F. Hsu, "Multi-layer combinatorial fusion using cognitive diversity," *IEEE Access*, vol. 9, pp. 3919–3935, 2021, doi: 10.1109/ACCESS.2020.3047057.
3. D. F. Hsu, B. Kristal, and C. Schweikert, "Combinatorial fusion analysis," *Computer*, vol. 57, no. 9, pp. 96–100, Sep. 2024, doi: 10.1109/MC.2024.3406058.
4. K. Grm, V. Štruc, A. Artiges, M. Caron, and H. K. Ekenel, "Strengths and weaknesses of deep learning

ABOUT THE AUTHOR

MARINA L. GAVRILOVA is a Research Excellence Chair in Trustworthy and Explainable Artificial Intelligence, and a director of the Biometric Technologies Laboratory, University of Calgary, AB T2N 1N4, Canada. She published over 300 articles in the domains of pattern recognition, information fusion, machine learning, biometric systems, and cybersecurity. Gavrilova received a Ph.D. from the University of Calgary. She is an Editor-in-Chief of *Transactions on Computational Sciences* (Springer) and the *International Journal of the Digital Human* (Interscience). She is an associate editor for *IEEE Access* and *IEEE Transactions on Computational Social Systems*. She is a Senior Member of IEEE. Contact her at mgavrilo@ucalgary.ca.

models for face recognition against image degradations," *IET Biom.*, vol. 7, no. 1, pp. 81–89, 2018, doi: 10.1049/iet-bmt.2017.0083.

5. W. Jiang, L. Ma, Y. G. Jiang, W. Liu, and T. Zhang, "Recurrent fusion network for image captioning," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 510–526.
6. A. B. López, "Deep learning in biometrics: A survey," *Adv. Distrib. Comput. Artif. Intell.*, vol. 8, no. 4, pp. 19–32, 2019.
7. S. Soleymani, A. Dabouei, H. Kazemi, J. Dawson, and N. M. Nasrabadi, "Multi-level feature abstraction from convolutional neural networks for multimodal biometric identification," in *Proc. 24th Int. Conf. Pattern Recognit.*, Piscataway, NJ, USA: IEEE Press, 2018, pp. 3469–3476, doi: 10.1109/ICPR.2018.8545061.
8. A. Terracciano, C. E. Lockenhoff, A. B. Zonderman, L. Ferrucci, and P. T. Costa Jr., "Personality predictors of longevity: Activity, emotional stability, and conscientiousness," *Psychosom Med.*, vol. 70, no. 6, pp. 621–627, 2008, doi: 10.1097/PSY.0b013e31817b9371.
9. K. P. Kumar, and M. L. Gavrilova, "Personality traits classification on Twitter," in *Proc. 16th IEEE Int. Conf. Adv. Video Signal Based Surveillance (AVSS)*, Piscataway, NJ, USA: IEEE Press, 2019, pp. 1–8, doi: 10.1109/AVSS.2019.8909839.
10. M. De Marsico, and G. Dionisi, "Your face may say the truth when you lie," in *Proc. Int. Conf. Adv. Vis. Interfaces*, pp. 1–3, 2022, doi: 10.1145/3531073.3534486.
11. C. S. Montero, M. Munezero, and T. Kakkonen, "Investigating the role of emotion-based features in author gender classification of text," in *Proc. Int. Conf. Intell. Text Process. Comput. Linguistics*, 2014, pp. 98–114.
12. J. K. Rout, K. K. R. Choo, A. K. Dash, S. Bakshi, S. K. Jena, and K. L. Williams, "A model for sentiment and emotion analysis of unstructured social media text," *Electron. Commerce Res.*, vol. 18, no. 1, pp. 181–199, 2018, doi: 10.1007/s10660-017-9257-8.
13. N. Rodriguez-Fernandez, S. Alvarez-Gonzalez, I. Santos, A. Torrente-Patiño, A. Carballal, and J. Romero, "Validation of an aesthetic assessment system for commercial tasks," *Entropy*, vol. 24, no. 1, pp. 103–121, 2022, doi: 10.3390/e24010103.
14. C. Segalin, A. Perina, and M. Cristani, "Personal aesthetics for soft biometrics: A generative multiresolution approach," in *Proc. Int. Conf. Multimodal Interact.*, New York, NY, USA: Association for Computing Machinery, 2014, pp. 180–187.
15. P. Lovato et al., "Tell me what you like and I'll tell you what you are: Discriminating visual preferences on Flickr data," in *Proc. Asian Conf. Comput. Vis.*, 2012, pp. 45–56.
16. J. Veselsky et al., "Establishing trust in vehicle-to-vehicle coordination: A sensor fusion approach," in *Proc. 2nd Workshop Data-Driven Intell. Cyber-Phys. Syst. Smart Cities Workshop (DI-CPS)*, Piscataway, NJ, USA: IEEE Press, 2022, 7–13, doi: 10.1109/DI-CPS56137.2022.00008.
17. S. K. Sowe, E. Simmon, K. Zettsu, F. De Vaulx, and I. Bojanova, "Cyber-physical-human systems: Putting people in the loop," *IT Prof.*, vol. 18, no. 1, pp. 10–13, Jan./Feb. 2016, doi: 10.1109/MITP.2016.14.
18. J. Granatyr et al., *Inferring Trust Using Personality Aspects Extracted from Texts*, in *Proc. IEEE Int. Conf. Syst., Man Cybern. (SMC)*, 2019, pp. 3840–3846, doi: 10.1109/SMC.2019.8914641.
19. N. Voloch, N. Gal-Oz, and E. Gudes, "A trust based privacy providing model for online social networks," *Online Soc. Netw. Media*, vol. 24, Jul. 2021, Art. no. 100138, doi: 10.1016/j.osnem.2021.100138.
20. M. Alkhamees, S. Alsaleem, M. Al-Qurishi, M. Al-Rubaian, and A. Hussain, "User trustworthiness

in online social networks: A systematic review," *Appl. Soft Comput.*, vol. 103, May 2021, Art. no. 107159, doi: 10.1016/j.asoc.2021.107159.

21. M. Sultana, P. P. Paul, and M. Gavrilova, "A concept of social behavioral biometrics: Motivation, current developments, and future trends," in *Proc. Int. Conf. Cyberworlds*, Santander, Spain, 2014, pp. 271–278, doi: 10.1109/CW.2014.44.
22. S. Prabhakar, S. Pankanti, and A. K. Jain, "Biometric recognition: Security and privacy concerns," *IEEE Security Privacy*, vol. 1, no. 2, pp. 33–42, Mar./Apr. 2003, doi: 10.1109/MSECP.2003.1193209.
23. Y. Wang et al., "Abstract intelligence: Embodying and enabling cognitive systems by mathematical engineering," in *Cognitive Analytics: Concepts, Methodologies, Tools, and Applications*, Hershey, PA, USA: IGI Global, 2020, pp. 52–69.
24. F. Anzum et al., "A comprehensive review of trustworthy, ethical, and explainable computer vision advancements in online social media," in *Global Perspectives on the Applications of Computer Vision in Cybersecurity*, Hershey, PA, USA: IGI Global, 2024, ch. 1, pp. 1–46.
25. G.-S. J. Hsu, H.-Y. Wu, C.-H. Tsai, S. Yanushkevich, and M. Gavrilova, "Masked face recognition from synthesis to reality," *IEEE Access*, vol. 10, pp. 37,938–37,952, 2022, doi: 10.1109/ACCESS.2022.3160828.
26. S. Thiebes, S. Lins, and A. Sunyaev, "Trustworthy artificial intelligence," *Electron. Markets*, vol. 31, no. 2, pp. 447–464, 2021, doi: 10.1007/s12525-020-00441-4.
27. D. Minh, H. X. Wang, Y. F. Li, and T. Ngyen, "Explainable artificial intelligence: A comprehensive review," *Artif. Intell. Rev.*, vol. 55, no. 5, pp. 3503–3568, 2022, doi: 10.1007/s10462-021-10088-y.
28. S. Tumpa, A. Gavrilov, O. Z. Duran, F. Tuz Zohra, and M. Gavrilova, "Quality estimation for facial biometrics, innovations, algorithms, and applications," in *Cognitive Informatics and Natural Intelligence*, Hershey, PA, USA: IGI Global, 2020, pp. 298–320.

Digital Object Identifier 10.1109/MC.2025.3546133

Unveiling IoT Threats: A Case Study on Darknet and Honeypot Analysis

Koji Nakao and **Daisuke Inoue**, National Institute of Information and Communications Technology
Katsunari Yoshioka, Yokohama National University

This article delves into the escalating complexity and sophistication of cyberattacks within the Internet of Things (IoT) landscape. It presents detailed case studies that highlight the vulnerabilities of IoT devices, advanced threat detection, and analysis of IoT malware, all examined through the lens of darknet observation and honeypot technologies.

Recent cyberattacks have become more sophisticated and diverse, making analysis increasingly difficult. The rise in vulnerabilities has also heightened the threats to organizations. While measures against known threats have been strengthened, responses to new attacks are lagging. In this ever-changing cyber landscape, various applications and services are being developed, and many Internet of Things (IoT) devices are being utilized. Especially in 5G environments, IoT devices play a crucial role in edge systems such as automotive and health care, as well as in cyberphysical systems, digital twins, and the metaverse.

When looking at IoT security measures from a global perspective, in the United States, United Kingdom, Europe, and Singapore, measures are being promoted based on requirements and certification for IoT devices that will be manufactured in the future. In Japan, in addition to IoT device certification, activities such as searching for vulnerabilities in existing IoT devices and issuing alerts to

Digital Object Identifier 10.1109/MC.2025.3531364
Date of current version: 27 March 2025

owners of vulnerable IoT devices are also being carried out through projects such as NOTICE.[1] While there are various IoT security measures in place, this article focuses on the threat of IoT malware. It introduces specific research and development methods and analysis cases based on two technologies: 1) darknet observation (sinkhole monitoring technology) and 2) honeypot technology. By utilizing these observation and analysis technologies, we can identify malicious behavior that has not been detected by conventional intrusion detection system, and so on, promote the sharing of threat intelligence information, and lead to the development of future security measures.

In the following sections, we will introduce our darknet observation and honeypot technologies, provide specific examples of our analyses. We will further showcase artificial intelligence (AI)-driven analysis and discuss current challenges and future directions.

DARKNET OBSERVATION SYSTEM

A darknet is a collection of IP addresses that are not being used by any organization. Although the IP addresses of darknets are registered and openly available in public, they are not assigned to any legitimate hosts/PCs, so it is generally assumed that they do not receive any incoming traffic. However, it is possible to observe a large amount of incoming traffic and is possible to observe new types of network attacks, such as network scans by malware, the behavior of malware infection, and distributed denial-of-service (DDoS) backscatter. In this type of observation using a darknet because unused IP addresses are used, there are no privacy issues that are discussed in actual network monitoring. Another feature is the ability to observe malicious behavior in real time through the analysis of incoming traffic. While it may not be possible to detect pinpoint attacks, such as targeted attacks, within the darknet, it is feasible to monitor global trends in attack behavior. This method of darknet observation is also known as "sinkhole monitoring."

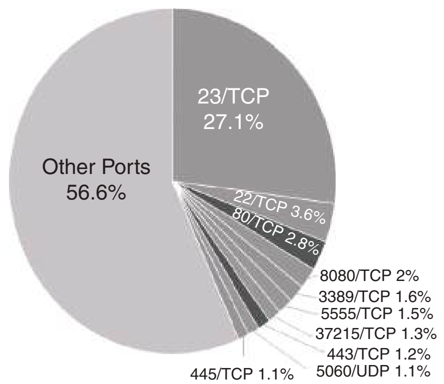
NICTER system

We have developed a darknet observation system called NICTER,[2] which includes multiple/16 and/24 darknets and currently monitors more than 300,000 darknet addresses. System operators can visualize the received traffic to comprehensively and in real time understand the trends in attack behaviors (scans).

Since 2005, NICTER has been conducting observations, and in 2023, it recorded 6.197 billion packets and 289,686 unique Internet Protocol (IP) addresses, equating to more than 2.26 million packets per IP address annually. However, in recent years, there has been an increase in research scans by organizations like Censys and Shodan, with 90 organizations conducting research that reached NICTER in 2023. Although research scans differ from attack-related scans, more than 70% of all scans in the third quarter of 2023 were research-related.

As a specific observation case, an investigation into the destination port numbers used by scan packets was conducted with 2023 statistics. The pie chart and table in Figure 1 show the statistics of the destination port numbers used. These statistics reveal scans targeting IoT devices such as webcams and routers on port 23/TCP (Telnet), as well as scans targeting web servers using ports 80/TCP and 443/TCP.

To facilitate future functional enhancements for NICTER, it is imperative to expand the current pool of 300,000 darknet addresses. This expansion is crucial for the effective management of extensive and heterogeneous IoT networks and for addressing evolving attack vectors. The global deployment of NICTER's darknet observation agents and the subsequent analysis of aggregated data will significantly enhance the comprehensiveness of observations on a worldwide scale. This approach will enable the global monitoring of IoT device scanning activities and the identification of sophisticated



Port	Target Service
23/TCP	Telnet (Router, Web Camera)
22/TCP	SSH (Server, Router)
80/TCP	HTTP (Web Server)
8080/TCP	HTTP (Web Console)
3389/TCP	Remote Desktop
5555/TCP	ADB (Android)
37215/TCP	Huawei Router
443/TCP	HTTPS (Web Server)
5060/UDP	SIP
445/TCP	Windows SMB

FIGURE 1. Top 10 destination ports observed by NICTER (2023).

attack behaviors, thereby augmenting analytical capabilities through the use of artificial intelligence.

Examples of analysis utilization from NICTER

The observation results from NICTER are shared with many attack analysts and organizations aiming to ensure cybersecurity. Through NICTER Web (https://www.nicter.jp/en/), NICTER Report (https://www.nict.go.jp/en/index.html), and NICTER Blog, information sharing and collaboration with numerous stakeholders are promoted. Below are specific examples of NICTER utilization:

Destination ports used in attacks. The destination ports in scans are useful for identifying the devices and systems targeted by attacks. NICTER recognizes that not only 23/tcp but many other destination ports are targeted in scans. As shown in Table 1, multiple devices such as routers are targeted, including 8291/tcp (MikroTik Router OS Winbox), 34567/tcp [Xiongmai DVR application programming interface (API)], and 37215/tcp (Huawei HG532 router). It is crucial to pay attention to suddenly emerging port numbers, as illustrated by the identification of specific devices next.

Identification of IoT devices. In 2022 May, NICTER detected a surge in scans targeting port 23/tcp. A scan-back on port 80/tcp revealed that the source was a router without Mirai characteristics. Upon connecting to the Internet, attack behaviors leading to infection by the InfectedSlurs bot were observed, including successful authentication using factory default credentials and command injection via network time protocol server settings. The analysis results were reported to the router vendor, who resolved the vulnerability through updates. We have established a systematic process from NICTER detection to IoT device identification, collaborating with device vendors. Additionally, NOTICE[1] leverages NICTER's detection results of Mirai-infected IoT devices to reduce the number of infected devices nationwide.

DDoS detection. By observing Syn-ack packets with NICTER, we can detect backscatter packets from organizations targeted by DDoS attacks, as

TABLE 1. Ports related to IoT vulnerability.

2022	Jan-23	Feb-23	Mar-23	Apr-23	May-23	Jun-23	Jul-23	Aug-23	Sep-23
23/tcp	23/tcp	23/tcp	23/tcp	23/tcp	23/tcp	23/tcp	23/tcp	23/tcp	23/tcp
22/tcp	22/tcp	22/tcp	80/tcp	22/tcp	22/tcp				
80/tcp	80/tcp	80/tcp	22/tcp	80/tcp	80/tcp				
5555/tcp	3389/tcp	5555/tcp	3389/tcp	3389/tcp	5555/tcp				
6379/tcp	8080/tcp	3389/tcp	**37215/tcp**	443/tcp	3389/tcp				
2375/tcp	5555/tcp	**37215/tcp**	443/tcp	6379/tcp	443/tcp				
443/tcp	443/tcp	8080/tcp	8080/tcp	8080/tcp	8080/tcp				
445/tcp	81/tcp	443/tcp	5555/tcp	5060/udp	**37215/tcp**				
2376/tcp	6379/tcp	81/tcp	6379/tcp	60023/tcp	445/tcp	5060/udp	81/tcp	5555/tcp	27610/tcp
81/tcp	445/tcp	6379/tcp	5060/tcp	**37215/tcp**	81/tcp	8081/tcp	6379/tcp	6379/tcp	15734/udp
8080/tcp	**37125/tcp**	445/tcp	445/tcp	5555/tcp	6379/tcp	81/tcp	445/tcp	445/tcp	6379/tcp
5060/udp	2375/tcp	5060/udp	2375/tcp	445/tcp	5060/udp	2375/tcp	8443/tcp	81/tcp	445/tcp
3389/tcp	5060/udp	60023/tcp	81/tcp	81/tcp	2375/tcp	3128/tcp	3128/tcp	3128/tcp	8088/tcp
2323/tcp	8081/tcp	2376/tcp	60023/tcp	2375/tcp	123/udp	6379/tcp	8081/tcp	53/udp	5555/tcp
123/udp	2222/tcp	123/udp	123/udp	53/udp	8081/tcp	**52869/tcp**	2375/tcp	2222/tcp	53/udp
37215/tcp	123/udp	**34567/tcp**	2376/tcp	2376/tcp	2376/tcp	2323/tcp	**52869/tcp**	2375/tcp	2222/tcp
1433/tcp	60023/tcp	8081/tcp	**34567/tcp**	123/udp	8443/tcp	8443/tcp	53/udp	**8291/tcp**	**8728/tcp**
4200/tcp	2323/tcp	2222/tcp	53/udp	1433/tcp	2323/tcp	2376/tcp	2222/tcp	8081/tcp	2375/tcp
111/tcp	2376/tcp	8443/tcp	8443/tcp	21/tcp	53/udp	123/udp	8088/tcp	8443/tdp	123/udp
8443/tcp	1433/tcp	2376/tcp	1433/tcp	8443/tcp	60023/tcp	21/tcp	21/tcp	8088/tcp	3128/tcp

8291/ tcp	MikroTik Router OS Winbox
8728/ tcp	MikroTik Router OS API
34567/ tcp	Xiongmai DVR API
37215/ tcp	Huwaei HG532 Router
52869/ tcp	Realtek SDK (Software Development Kit)

attackers often randomize source IP addresses. Our analysis shows that full-scale DDoS attacks typically start a few hours to days after Syn-ack packets are first observed. Therefore, we provide immediate DDoS detection alerts to relevant organizations upon observing Syn-ack packets. For instance, this alert system helped ensure the stable operation of the 2021 Tokyo Olympics management system.

X-POT: A NEW GENERATION IOT HONEYPOT

We introduce X-POT,[3] an adaptive honeypot framework designed to emulate various IoT devices using responses gathered from Internet-wide scans. As mentioned previously, attackers frequently conduct broad IP address scans to propagate infections. These random scans are also detectable by honeypot sensors. However, the increasing diversity of ports used by Internet-connected devices necessitates enhanced observation capabilities for honeypots. Attackers target such devices, making it challenging to discern the type of attack and its associated port. To address this issue, we construct X-POT, which updates honeypot responses based on Internet-wide scan findings to emulate Internet device behaviors.

Figure 2 shows the basic concept of X-POT. It comprises two main units: the Inner Unit (Attack Monitor and Response Collector) and the Outer Unit (Service Emulator). The Inner Unit comprehensively monitors all TCP ports, collecting responses from network scans. These responses inform the Service Emulator, which then emulates attack targets using the collected data. The primary objective of X-POT is to enhance honeypot observation capabilities.

Overview of observations using X-POT

During our study, we deployed HTTP X-POT across multiple virtual private servers (AWS, DigitalOcean, Vultr, Conoha, Sakura VPS) from 15 March to 20 May 2020. Alongside an analysis server hosted by a commercial ISP, we observed attacks on 23 IP addresses. Initially, we deployed a basic honeypot on AWS, conducting Internet-wide scans and collecting responses from 22 July 2019 to 15 March 2020. Subsequently, we deployed HTTP X-POT with collected responses from 15 March to 20 May 2020, continuously modifying responses based on access from approximately five IP addresses daily. HTTP X-POT processed 4,729,097 HTTP requests across 64,912 ports, originating from 85,849 unique IP addresses, and identified 1,276 unique malware samples. Our findings included attacks on various IoT device web interfaces (for example, routers, IP cameras) and samples of IoT-specific malware such as Mirai and Gafgyt, with several responses containing device-specific data indicative of characteristic attacks and vulnerabilities.

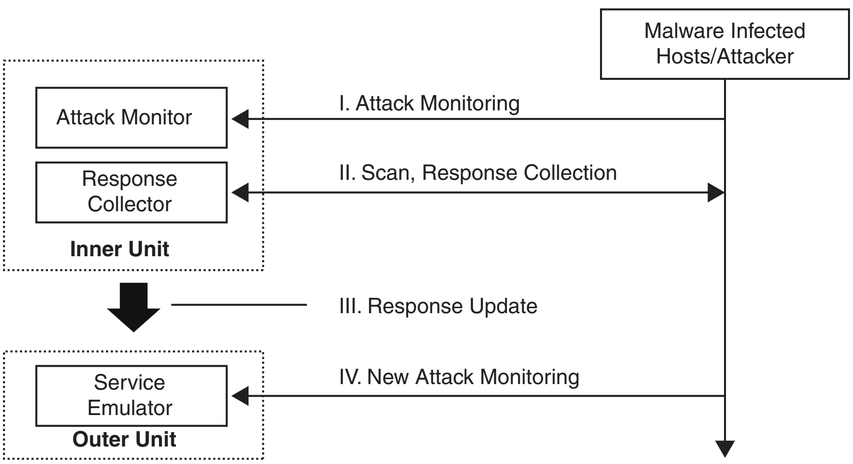


FIGURE 2. X-POT concept.

Examples of attack analysis using X-POT

We successfully collected a significant number of responses. The responses from these Access-uniform resource locators (URLs) changed over the observation period. The scan primarily targeted ports that received numerous requests from multiple hosts, such as port 8000/TCP (commonly associated with web servers) and the tmUnblock.cgi script, which is used in Cisco/Linksys router firmware and other targeted attacks.

The request and response data proved valuable in identifying the devices that attackers were targeting. For instance, HTTP requests like GET/v1.16/version on port 2375 and GET/ on port 200 suggest that the targeted devices were likely Docker APIs or Elasticsearch services (Figure 3). Additionally, collected responses from GET/ on port 1200 indicate that the Access-URL was related to RSShub. Although we did not observe any remote code execution (RCE) exploits for RSShub during our experiment, the information is beneficial for developing high-interaction honeypots. It helps us identify potential attack targets based on the content of the requests or the responses obtained. Figure 4 shows that the attack observed on 5984/TCP port. When the HTTP X-POT returned a response containing "CouchDB" in the server header, the client initiated a request to create a user account using the PUT method. It was observed that this account creation request occurred only when the server responded with CouchDB versions 2.1.1 or 2.2.0. Versions 2.3.0 and 2.3.1 did not appear to be targeted in the attack. Subsequently, the username and password were utilized for basic authentication, after which the attacker attempted to execute commands. Using the wget command, the attacker downloaded a file named "2start.jpg" and attempted to execute it.

XPOT is limited in the number of IP addresses it can utilize and the variety of devices it can emulate. To mitigate these constraints, it is essential to harness both the extensive scale of the darknet and the granular observability provided by honeypots. One potential improvement involves collecting responses from compromised devices detected in large scale darknet (NICTER) observations and integrating these responses into honeypot (XPOT) operations to accurately emulate devices targeted by emerging attacks.

More recent observation of XPOT is described at https://sec.ynu.codes/iot. Interested researchers could contact us for the dataset.

ANALYSIS STUDIES USING NICTER AND X-POT

Mirai: IoT malware

The malware known as "Mirai," which is said to be the origin of malware targeting IoT devices, first appeared in 2016. At that time, a surge in scans targeting 23/TCP and 2323/TCP was observed on the darknet (by NICTER). For 23/TCP, the number of scans rapidly increased to 2 million per day after 13 September 2016. As for 2323/TCP, no scans were observed until 1 September 2016, but the number of scans gradually increased from that point, reaching up to 150,000 scans per day from 13 September, similar to 23/TCP.

In response to this behavior, the observation results of X-POT confirmed infections due to scans targeting 23/TCP and 2323/TCP. Analysis of the captured malware revealed that its behavior completely matched that of "Mirai," which was a hot topic worldwide. The analysis showed that "Mirai" scans 23/TCP and 2323/TCP, performs dictionary attacks, and has the characteristic of having the same destination IP address and TCP sequence number.

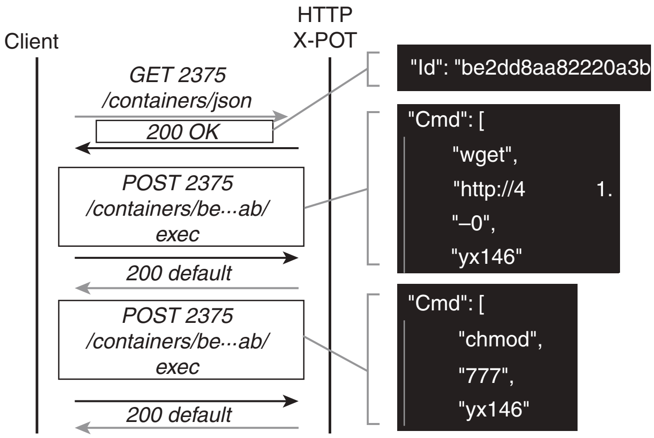


FIGURE 3. Docker API attack observed by XPOT.[2]

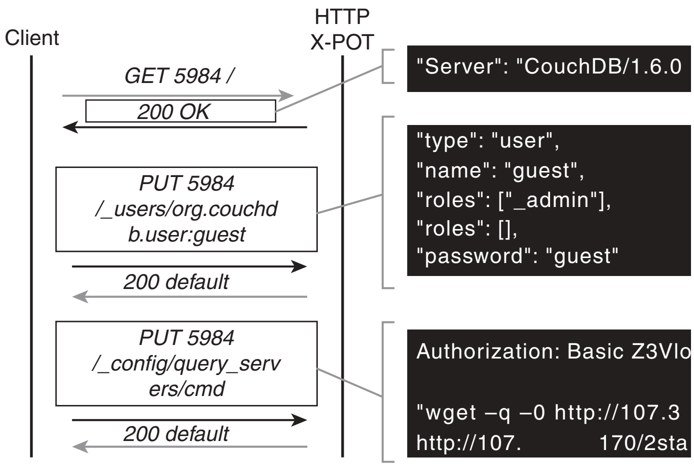


FIGURE 4. CouchDB attack observed by XPOT.[2]

Particularly, using the characteristic of "Mirai" that the destination IP address and TCP sequence number are the same, NICTER confirmed that 40 million packets were received from hosts infected with Mirai in late September 2016. This result was almost the same as the previously mentioned observation results for 23/TCP and 2323/TCP, and the number of hosts infected with Mirai peaked at 14 million.

Furthermore, checking the current status of Mirai infections with NICTER, it was found that the Mirai malware continues to spread, with many infected devices still existing in countries such as China and India. It was also found that there are still between 40,000 and 100,000 Mirai-infected devices worldwide.

IoT gateway (mobile router)

On 11 July 2023, NICTER conducted a scan-back on Mirai-infected hosts observed, identifying 70 hosts returning the server header thttpd (Server: thttpd/2.25b 29dec2003). To verify the characteristics of these hosts, we checked the Censys scan results and found that all hosts had port 6666/TCP open. To analyze IoT devices in more detail, we used a group of vulnerable IoT devices to check the banners of these devices, confirming that the Mirai-infected hosts was a router. Consequently, an X-PoT honeypot, simulating router behavior, was deployed for attack analysis.

This honeypot observed a large number of login attempts from The Netherlands around 12:38 p.m. on 29 July 2023. The specific attack process was as follows:

1. The attacker first checks the server header information to confirm if the target router is running thttpd.
2. They verify if port 6666/TCP (Telnet) is accessible (if accessible, proceed to step 5).
3. Log in to port 80/TCP with admin:1234 credentials and enable Telnet (6666/TCP).
4. Access port 80/TCP without entering BASIC authentication information (repeat the same procedure as step 1).
5. Access port 6666/TCP (Telnet) and execute commands via ping.

In this case, various investigations and analyses were conducted using NICTER, Censys, and honeypots, but no vulnerabilities were found in the router itself. This indicates that the person who set up the router (user: system integrator) intentionally changed the settings to allow network access for the maintenance. Furthermore, the cause of this incident can be attributed to issues such as not changing the default password. The aforementioned causes are thought to have led to the attack, but since the latest firmware for the router in question forces the password to be changed, it is crucial to enhance the security awareness of users (installers).

SPOT: IN-DEPTH ANALYSIS OF IOT RANSOMWARE ATTACKS USING BARE METAL NAS DEVICES

In recent years, IoT devices, such as network-attached storage (NAS) systems, have increasingly become targets for ransomware attacks. These attacks often exploit remote vulnerabilities in the devices, bypassing traditional ransomware infection methods such as phishing. To better understand the nature of these attacks and the behavior of ransomware targeting NAS devices, we developed an observation system called SPOT, as an extension of X-POT using bare-metal QNAP NAS devices. This system allowed us to collect detailed data on the exploitation attempts, which can be difficult to achieve through virtualized environments alone.

SPOT system overview

For our experiment, we implemented SPOT[4] using two physical QNAP devices. One device was configured as a honeypot, while the other served as a malware sandbox for deeper analysis of observed attacks. By using bare-metal devices rather than virtualized environments, we ensured that the full functionality of QNAP devices, including dynamically changing file paths and interrelated application files, was maintained. This allowed us to accurately capture and analyze ransomware attacks that exploit vulnerabilities in these specific NAS devices.

The SPOT system was composed of several key components, including Web and Secure Socket Shell (SSH) reverse proxies, a packet monitor, and a QNAP monitor. These components worked together to decrypt, record, and analyze traffic targeting the NAS device, as well as to monitor internal device behaviors.

Components of SPOT

Figure 5 shows the architecture of SPOT. It relies on five critical components (see next) to monitor and analyze the incoming traffic targeting the QNAP honeypot. These components ensure that we captured a wide range of data on potential ransomware attacks, as well as other types of exploitation attempts.

Web reverse proxy. This handles incoming traffic on HTTP/TLS ports (80/tcp, 443/tcp, and 8080/tcp). This proxy decrypts TLS payloads and records all incoming HTTP requests, allowing potential attacks against the NAS web interface to be identified.

SSH reverse proxy. This is a Python-based proxy to monitor attackers attempting to log in via SSH.

Packet monitor. This uses signature matching to detect and analyze HTTP requests targeting QNAP-specific paths. It also monitors the resolution of official domains by QNAP devices.

QNAP monitor. This tracks internal activity on the NAS device, such as system processes, open ports, and changes to specific files.

Access controller. Using iptables to control the flow of traffic through the system, this controller ensures that the QNAP device could only communicate with official domains, preventing malicious files from entering the device.

Observation results

During the observation period, SPOT captured a wide range of malicious activities targeting the QNAP honeypot. These activities included brute-force login attempts, web-based exploitation attempts, and attempts to download malware.

SSH reverse proxy results. On average, SPOT observed 1,700 SSH login attempts per day, primarily consisting of brute-force attacks. Out of these attempts, 177 were successful logins where attackers executed commands. The most common command was "/ip cloud print," which is specific to Mikrotik's Router OS and therefore not functional on the QNAP device. This suggests that many of the attacks were not specifically targeting QNAP but rather IoT devices in general.

Web reverse proxy results. The Web reverse proxy observed an average of 130 HTTP sessions per day, including port scans and attempts to exploit web-based vulnerabilities. Of the 149 hosts that accessed the login authentication page of the QNAP device, 46 attempted to extract sensitive files, such as password files and SSH keys, through vulnerabilities in the Photo Station application.

Compared to other IoT honeypots, SPOT demonstrated a higher capture rate of QNAP-related access. For example, during the same observation period, a general-purpose IoT honeypot only observed four hosts accessing the QNAP login page, while SPOT observed 149. This highlights the effectiveness of using a dedicated, bare-metal NAS honeypot for observing ransomware attacks targeting specific devices like QNAP. Figure 6 is the overview of eCh0raix ransomware attack we observed.

MALWARE ACTIVITY DETECTION TECHNOLOGY USING AI

Malware is pivotal in cyberattacks. X-POT analysis indicates that numerous IoT malwares commence spontaneous scanning post-infection to identify vulnerable hosts. Upon detecting an open vulnerability-related port on an IoT device, attackers typically infect the device with malware, facilitating further malicious activities.

We focused on the fact that devices infected with the same IoT malware use the same scanning transmission module and used AI to detect the synchrony of scanning behavior in terms of both method and timing. As a specific AI method for synchrony

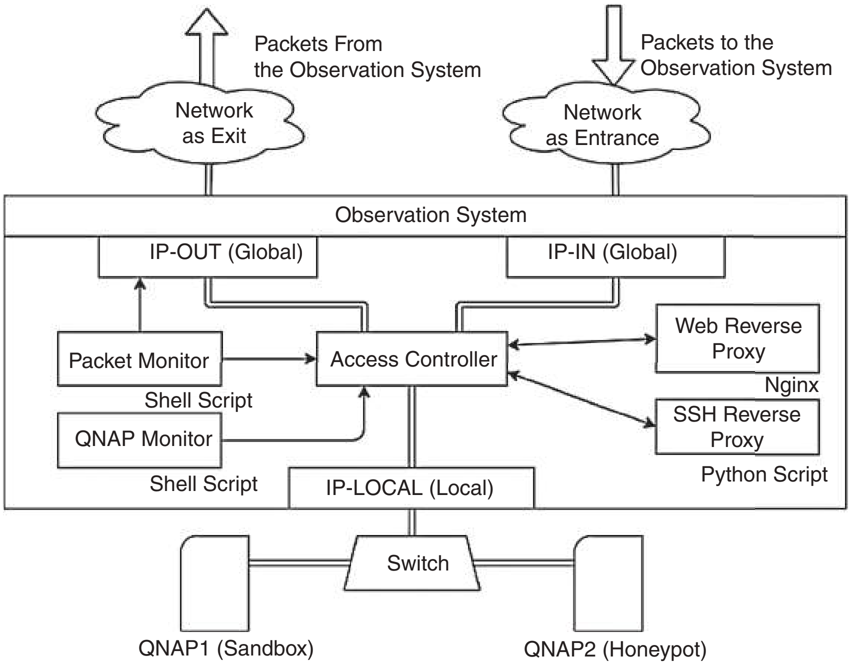


FIGURE 5. Architecture of SPOT.

detection,[5,6] we used the "GLASSO engine," a system that uses a sparse structure learning algorithm called *graphical lasso* to estimate the synchrony of infected hosts in real time. The main feature of this method is that it can detect related attack campaigns by estimating the synchronization between many infected hosts in any time slot and detecting abnormal values in synchronization between different time slots. Furthermore, this method is robust to noise, and it has the feature that it can ignore events with weak synchronization due to the effect of the ℓ1 regularization penalty term of the graphical lasso algorithm.

In the evaluation experiment, a list of cyberthreats, including relatively large-scale infection events, was collected and created for each destination port of scans sent from infected hosts over a certain period of time in the darknet (NICTER), and this threat list was used as "correct data." As a result, it was confirmed that the GLASSO engine could detect cyberthreats involving host-scale infections of a certain scale or larger with high accuracy (97.14% accuracy), demonstrating its effectiveness and usefulness.[6]

Therefore, the derivation of synchronized scanning behavior in NICTER using the GLASSO engine means that the group of hosts (IoT devices) that are sending out those scans are infected with the same malware, and it becomes possible to quickly grasp the spread of infection of IoT devices by new IoT malware bots. Furthermore, by correlating malware samples using X-POT, it is possible to analyze related malware, identify infected devices, and investigate them, minimizing damage to users of vulnerable IoT devices and enabling effective incident response.

AI technology enhances cybersecurity by automating threat detection, improving response times, and strengthening defenses against evolving risks. AI can analyze vast amounts of data to identify patterns, anomalies, and correlations, enabling real-time threat detection and risk mitigation. Additionally, AI automates routine tasks such as log analysis and vulnerability scanning, allowing human analysts to focus on more complex issues.

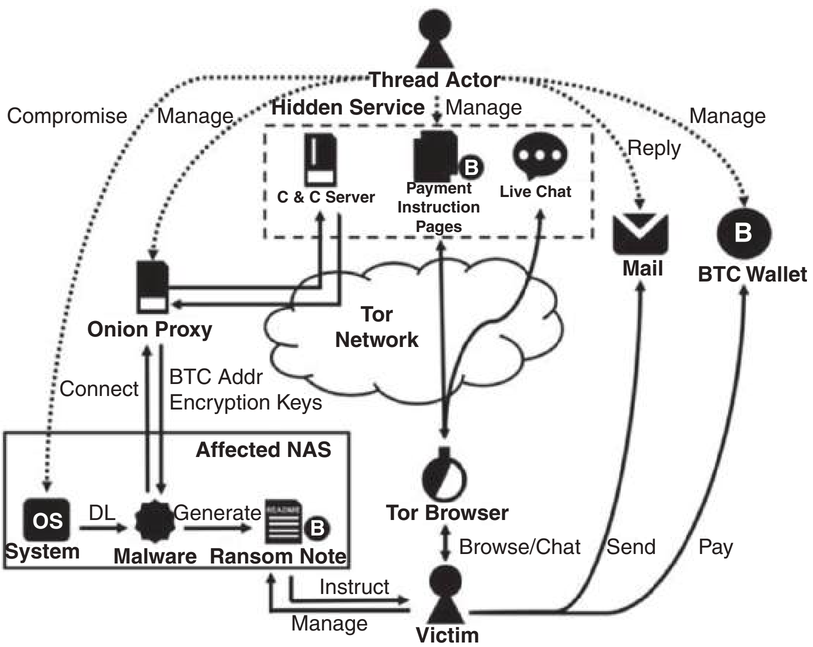


FIGURE 6. Overview of eChOraix ransomware attack observed by SPOT.[4]

Considering the future development of IoT, as discussed, utilizing darknets and honeypots introduced in this article is effective. Specifically, building honeypots specialized for certain IoT devices and systems is crucial. The malware capture module in the honeypot must be dynamically configurable to mimic attack targets flexibly. The research results using darknets and honeypots will be effective in detecting and analyzing diverse and sophisticated threats. However, it is also important to continue discussing privacy and safety issues related to these threats. Future research and development should consider the following:

1. The increasing complexity and range of attacks, now targeting IoT devices like automobiles and health care, make it hard to fully understand threats. Effective observation and analysis techniques, such as using darknet and honeypots, along with open source intelligence and threat intelligence, are essential to grasp the attack's full scope, intent, and background. Identifying threats requires detailed, broad-perspective analysis, including AI.

2. While analyzing current attack behaviors is important, it is also crucial to identify and analyze signs of future attacks. Research on such attack signs includes abnormal activities on the darknet, the evolution of malware functions, and AI-based attack predictions. These efforts and studies are expected to technically support "active cyber defense."
3. AI can be used for both defense and attack, and AI-driven phishing and deepfake fraud are already increasing. When using AI in cybersecurity, it is essential to recognize AI threats such as backdoors, poisoning, and evasion attacks. In particular, thorough verification of AI-derived results is necessary, and privacy risks must be considered for AI monitoring systems in IoT devices.
4. In cybersecurity research, it is necessary to comprehensively consider safety, privacy, resilience, and reliability in addition to focusing on security properties (such as CIA) tailored to specific models and scenarios. The ultimate goal is to ensure overall "trustworthiness" by encompassing all these elements.

This article focused on IoT threats and introduced our research findings, including NICTER, X-POT, and several analysis cases. It is no exaggeration to say that the observation, analysis, and application of cyberthreat data will continue to play an important role in cybersecurity efforts. We hope you find this article valuable and that it contributes to your future research endeavors. C

ABOUT THE AUTHORS

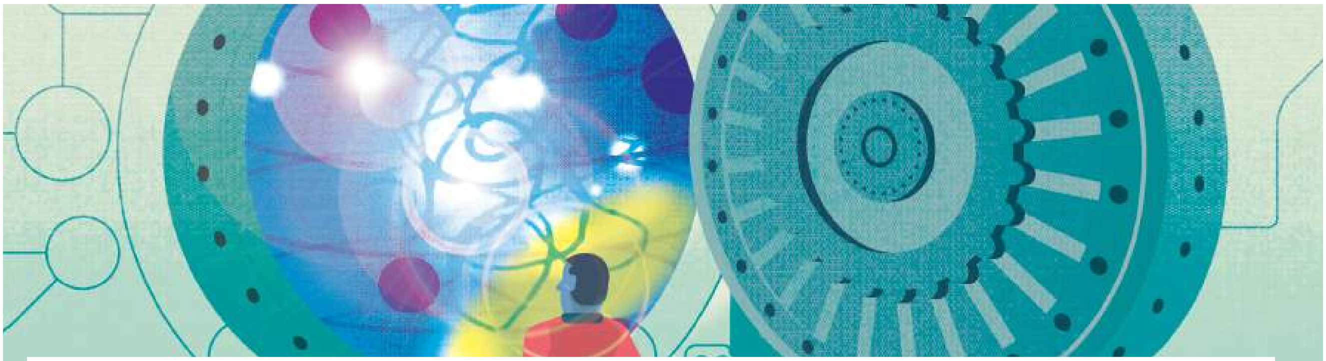
KOJI NAKAO is a distinguished researcher at the Cybersecurity Research Institute, National Institute of Information and Communications Technology, Tokyo, Japan, and a visiting professor at Yokohama National University, Yokohama, 240-0067, Japan. His research interests include Internet of Things security, 5G security, and supply chain security. Koji received his B.E. in mathematics from Waseda University. Contact him at ko-nakao@nict.go.jp.

DAISUKE INOUE is the director of the Cybersecurity Research Institute, National Institute of Information and Communications Technology, Tokyo, Japan. His research interests include practical cybersecurity research such as Internet of Things security, usability security, machine learning for security, and cybersecurity human resource development. Daisuke received his Ph.D. in engineering from Yokohama National University. Contact him at dai@nict.go.jp.

KATSUNARI YOSHIOKA is a professor at Yokohama National University, Yokohama, Japan. His research interests include Internet of Things security based on system security and network security, and monitoring of threat actors. Katsunari received his Ph.D. in engineering from Yokohama National University. Contact him at yoshioka@ynu.ac.jp.

REFERENCES

1. "Introduction to the NOTICE project." NOTiCE. Accessed: Jan. 25, 2025. [Online]. Available: https://notice.go.jp/en
2. K. Nakao, D. Inoue, M. Eto, and K. Yoshioka, "Practical correlation analysis between scan and malware profiles against zero-day attacks based on darknet monitoring," *IEICE Trans. Inf. Syst.*, vol. E92-D, no. 5, pp. 787–798, May 2009, doi: 10.1587/transinf.E92.D.787.
3. S. Kato, R. Tanabe, K. Yoshioka, and T. Matsumoto, "Adaptive observation of emerging cyber attacks targeting various IoT devices," in *Proc. IFIP/IEEE Int. Symp. Integr. Netw. Manage. (IM)*, 2021, pp. 143–151.
4. H. Yasui, T. Inoue, T. Sasaki, R. Tanabe, K. Yoshioka, and T. Matsumoto, "SPOT: Analyzing IoT ransomware attacks using bare metal NAS devices," in *Proc. 17th Asia Joint Conf. Inf. Secur. (AsiaJCIS2022)*, 2022, pp. 16–23, doi: 10.1109/AsiaJCIS57030.2022.00013.
5. C. Han, A. Tanaka, and T. Takahashi, "Darknet Analysis-based early detection framework for malware activity: Issue and potential extension," in *Proc. IEEE Workshop Big Data Cybersecurity*, 2022. pp. 2980–2988, doi: 10.1109/BigData55660.2022.10021021.
6. C. Han, J. Takeuchi, T. Takahashi, and D. Inoue, "Dark-TRACER: Early detection framework for malware activity based on anomalous spatiotemporal patterns," *IEEE Access*, vol. 10, pp. 13,038–13,058, 2022, doi: 10.1109/ACCESS.2022.3145966.



Robust Intrusion Detection With Combinatorial Fusion and Generative Artificial Intelligence

Evans Owusu, **Mariyam Mapkar**, and **Mohamed Rahouti**, Fordham University
Dinesh C. Verma, IBM Thomas J. Watson Research Center

This article proposes an advanced intrusion detection system that combines combinatorial fusion analysis with generative artificial intelligence to improve anomaly detection in intelligent systems. It addresses challenges in detecting low-profile and evolving threats, especially in imbalanced datasets.

The rapid evolution of cyberthreats, particularly sophisticated denial-of-service (DoS) attacks, presents major challenges to the security of intelligent systems across sectors like critical infrastructure, autonomous vehicles, and smart cities.[1] As these attacks become more adaptive and distributed, traditional machine learning (ML) models in intrusion detection systems (IDSs) struggle to detect low-profile threats that mimic normal behavior, despite excelling at identifying well-known attacks.[2] Furthermore, their closed-box nature limits transparency and interpretability, making it difficult for security professionals to trust and understand decisions made by IDSs, especially in critical environments where explainability is crucial.

Further, the adaptability of existing IDSs is another critical shortcoming.[3] In today's network environments, characterized by heterogeneous, distributed, and dynamic architectures such as multicloud platforms, Internet of Things (IoT) ecosystems, and edge computing, traditional IDS approaches that rely on static models and fixed thresholds become insufficient. These systems

Digital Object Identifier 10.1109/MC.2024.3524302
Date of current version: 27 March 2025

require constant updates and manual tuning to handle new and unknown threats, which are not feasible in the real-time operational contexts of modern intelligent systems. Consequently, there is a need for a more accurate and robust IDS, which is also highly adaptable and capable of real-time learning and decision making.[4]

To address these challenges, this article proposes leveraging combinatorial fusion analysis (CFA)[5,6,7] as a groundbreaking approach that aligns with the goals of building secure and intelligent systems. CFA offers a robust framework integrating multiple ML models through advanced score combinations (SCs) and rank combinations (RCs) to leverage their diverse strengths while mitigating individual weaknesses. Unlike conventional ensemble methods that may rely on simple averaging or voting, CFA utilizes rank-score characteristic (RSC) functions and cognitive diversity (CD) metrics to assess and combine models based on their complementarity and diversity. This approach enables the development of IDSs that are more intelligent and context-aware, capable of detecting a broader range of attacks, from volumetric to low-and-slow attacks, and even novel, adaptive threats.

The goal of this article is to demonstrate how CFA can serve as a cornerstone for next-generation IDSs, offering enhanced detection accuracy, resilience, and adaptability. By fusing multiple models capable of recognizing diverse threat patterns, the IDS can dynamically adjust to new attacks and environments without needing extensive retraining. This approach also improves interpretability, fostering trust and confidence in security operations. Key contributions include presenting CFA's use of CD and weighted fusion techniques to enhance performance through advanced SCs and RCs and demonstrating the effectiveness of generative artificial intelligence (GAI)-integrated CFA in a case study on DoS attack detection, with significant improvements in performance and adaptability for both high-frequency and low-profile attacks.

BY FUSING MULTIPLE MODELS CAPABLE OF RECOGNIZING DIVERSE THREAT PATTERNS, THE IDS CAN DYNAMICALLY ADJUST TO NEW ATTACKS AND ENVIRONMENTS WITHOUT NEEDING EXTENSIVE RETRAINING.

THE NEED FOR ADVANCED INTRUSION DETECTION IN SECURE AND INTELLIGENT SYSTEMS

Traditional ML-based IDSs face challenges in interpretability, trust, and detecting low-profile, adaptive attacks, which are crucial as intelligent systems increasingly support critical infrastructures like smart grids and healthcare. Many ML models, especially deep learning, lack transparency, making it difficult for cybersecurity professionals to understand malicious packet detections, undermining trust and compliance.[8] Modern DoS attacks mimic legitimate traffic to evade detection, and traditional models often fail to catch these subtle threats or produce false positives.[9] Additionally, traditional IDSs struggle with scalability and adaptability as intelligent systems expand, integrating cloud services, IoT, and edge computing.[10] The need for constant retraining to address evolving threats is impractical, requiring more autonomous and adaptable solutions.

The evolving cyberthreat landscape demands IDSs that are both secure and intelligent, capable of learning, adapting, and scaling autonomously in response to new challenges. Today's threats are not only more frequent but also more varied and unpredictable. Attackers increasingly leverage AI and ML to develop malware that can adapt to different environments and employ multivector strategies to overwhelm traditional defense mechanisms. This requires IDSs to be intelligent enough to anticipate, learn from, and counteract new types of attacks dynamically.

Further, the integration of secure and intelligent systems in sectors like critical infrastructure, autonomous vehicles, smart cities, and healthcare makes them prime targets for advanced persistent threats, zero-day attacks, and insider threats.[11] Such attacks require more than static, rule-based defenses; they require systems that can continuously learn from diverse data points, detect anomalies in real time, and provide actionable insights for preemptive measures.

As these systems become more heterogeneous and distributed, the need for adaptive and scalable detection mechanisms grows. For instance, in a smart city environment, multiple interconnected subsystems (for example, traffic control, public safety, and utilities) must work together to detect and respond to potential threats. In this scenario, a robust IDS must not only secure each subsystem but also intelligently correlate information across the network to detect coordinated attacks.[12] Traditional IDS approaches fall short in providing this level of intelligence and integration, highlighting the need for advanced methods.

CFA: A GAME CHANGER

To build secure and intelligent systems capable of meeting the above challenges, we propose leveraging CFA,[5,6,7] which provides a framework that integrates multiple ML models through a combination of score and rank functions, enabling a more comprehensive, adaptive, and secure detection mechanism. Unlike traditional methods, CFA emphasizes the diversity and complementarity of multiple models, resulting in improved detection capabilities, lower false positives, and higher interpretability, key attributes for intelligent and secure systems.

Unlike traditional ensemble methods, which often rely solely on averaging or majority voting, CFA leverages both SC in Euclidean space and RC in Kemeny space to integrate diverse models' outputs effectively. The essence of CFA lies in its use of CD, a measure of dissimilarity among scoring systems, to determine the optimal fusion of models.[13]

THE EVOLVING CYBERTHREAT LANDSCAPE DEMANDS IDSs THAT ARE BOTH SECURE AND INTELLIGENT, CAPABLE OF LEARNING, ADAPTING, AND SCALING AUTONOMOUSLY IN RESPONSE TO NEW CHALLENGES.

A scoring system A on a dataset $D=\{d_1, d_2, \ldots, d_n\}$, defined by a score function $s_A: D \mapsto \mathbb{R}$ and a rank function $r_A: D \mapsto \mathbb{N}$, was introduced in Hsu et al.[5,7] Sorting s_A values in descending order gives r_A, where $\mathbb{N}=\{1, 2, \ldots, n\}$. The RSC function $f_A: \mathbb{N} \mapsto \mathbb{R}$ for A is defined as

$$f_A(i) = s_A(r_A^{-1}(i)) = (s_A \circ r_A^{-1})(i). \quad (1)$$

The CD among scoring systems is defined as the difference among their RSC functions.[13] The diversity of RSC functions can be computed for each pair of scoring systems. The CD between two systems A_i and A_j, denoted as $CD(A_i, A_j)$, is based on their RSC functions f_{A_i} and f_{A_j}. For a rank $k \in \{1, 2, \ldots, n\}$, $CD(A_i, A_j)$ is defined as $\sqrt{(1/(n^2-n))\Sigma_{k=1}^{n}(f_{A_i}(k) - f_{A_j}(k))^2}$.

The diversity strength (DS) of system A is the average CD between A and all other systems. To compute this, we first need to calculate the RSC functions for each cross-validation split $D_j = \{d_1, \ldots, d_n\}$ and scores $j \in \{1, \ldots, P\}$. The score function $s_{kj}(d)$ assigns a real number to each $d \in D_j$, representing the score given by model M_k for the jth split. By sorting these scores in descending order and ranking the candidates in D_j, we obtain the rank function $r_{kj}(d)$. To facilitate comparison across multiple scoring systems, we apply linear normalization, transforming $s_{kj}(d): D \to \mathbb{R}$ to $s_{kj}^*(d): D \to [0, 1]$ using $s_{kj}^*(d) = (s_{kj}(d) - s_{\min})/(s_{\max} - s_{\min})$, where $s_{\max} = \max\{s_{kj}(d) | d \in D\}$ and $s_{\min} = \min\{s_{kj}(d) | d \in D\}$. The RSC functions are then derived by sorting these normalized scores for each system in descending order, using the rank values as keys. This process allows for a robust calculation of DS by measuring the differences among the RSC functions of various systems.

The superiority of RC over SC can be evaluated under specific conditions, where RCs may yield better performance when larger CD values exist between pairs of the five ML models.[14] We can integrate results from m scoring systems, each defined by its score function $s_{kj}(d)$ and rank function $r_{kj}(d)$ for label j, with k as the system index. Techniques include SC, RC, voting, average combination (AC), and weighted combination (WC), which can be computed using the following weighting metrics. The AC computes the average score and rank as $s_s(d) = \Sigma_{i=1}^{m}[w_i s_{ij}(d)]$ and $s_R(d) = \Sigma_{i=1}^{m}[w_i r_{ij}(d)]$, where $w_i = 1/m$, and s_s and s_R represent the score and rank functions of SC and RC, respectively. The WC by DS (WCDS) includes weighted SCs (WSCDS) and RCs (WRCDS), with weights defined as $W_i = (1/N)$

for AC and $W_i = (ds(A_i)) / (\Sigma_{i=1}^{N} ds(A_i))$ for WCDS. The WSCDS is calculated as $\mathrm{WSCDS}_{ij}(d) = ((\text{weight of model } i) * s_{ij}(d)) / (\text{sum of weights})$, while WRCDS uses $r_{ij}(d)$ instead of $s_{ij}(d)$ and inverts w_i to $1/w_i$.

By optimizing SCs and RCs with weighted DS, CFA offers a comprehensive fusion method that accounts for both model performance and diversity. This approach is particularly effective in intelligent systems, like cybersecurity, where complex decision making is required across diverse data landscapes. CFA's ability to integrate models based on diversity and performance makes it ideal for IDSs, enhancing precision and adaptability in detecting a range of attack profiles. Traditional IDSs often struggle with missed detections or high false-positive rates due to reliance on single models. CFA, by aggregating confidence levels through SCs and distinguishing prediction outputs with RCs, creates an intelligent, adaptive mechanism that identifies both high- and low-profile threats. The WCDS further enhances robustness by prioritizing diverse models, fully leveraging CD to secure intelligent systems.

CASE STUDY: APPLYING CFA TO DoS ATTACK DETECTION

Dataset and methodology

To demonstrate the effectiveness of CFA in intrusion detection, a case study was conducted using the LYCOS-IDS2017 dataset,[15] which includes a wide range of network traffic types representing various attack profiles, such as benign traffic, DoS attacks (for example, GoldenEye, Hulk, and Slowloris), distributed DoS (DDoS), and web-based attacks (for example, SQL Injection and XSS), as shown in Table 1. CFA combines multiple ML models' outputs using the WSCDS method, where individual models' DSs serve as weights to improve final predictions. The models used include support vector machine (SVM), AdaBoost (ADB), multilayer perceptron (MLP), Gaussian naive Bayes (GNB), and linear discriminant analysis (LDA), representing five scoring systems (A, B, C, D, and E). The diversity strengths were computed for each data item, and combinations of two to five models were explored to enhance accuracy. After training, the model performance was evaluated on unseen test data across 14 traffic classes (for example, benign, DoS/DDoS, and web attacks), where the class with the highest prediction probability, reflecting each model's confidence, was chosen, and these probabilities formed a new dataset for further insights through CFA's fusion approaches.

Key improvements through WSCDS

The WSCDS method demonstrates significant performance improvements by combining the diversity strengths of different models. For instance, data item d_{10} leads to a prediction of class C_6 using the combined models A, B, C,

TABLE 1. Network traffic types in the LYCOS-IDS dataset[15] with a 75%–25% split: 660,944 training items (X) and 220,312 test items.

Traffic	Encoding	X	Y	Test set
Benign	0	330,474	405,441	110,158
Bot	1	550	37,672	183
DDoS	2	71,761	137,764	23,920
DoS Goldeneye	3	5,073	55,647	1,691
DoS Hulk	4	119,241	188,617	39,747
DoS Slowhttptest	5	3,649	52,135	1,216
DoS Slowloris	6	4,255	53,548	1,418
FTP Patator	7	3,001	50,396	1,000
Heartbleed	8	7	12,317	2
Portscan	9	119,197	188,290	39,732
SSH Patator	10	2,218	47,388	739
Webattack Bruteforce	11	1,020	42,098	340
Webattack Sql Injection	12	9	13,558	3
Webattack XSS	13	489	37,017	163
Total		660,944	1,321,888	220,312

Using CTGAN, 660,944 synthetic entries were generated, doubling the training set to 1,321,888 items (Y).

D, and E, where the DSs of the models are calculated as 0.359996, 0.523272, 0.337381, 0.472812, and 0.598646, respectively. This calculation is systematically applied across all combinations of models, resulting in outstanding detection accuracy. The WSCDS method consistently enhances the accuracy of predictions compared with single-model performance. The method utilizes the diversity strengths of models across multiple classes of DoS attacks, achieving high precision in detecting attacks that individual models fail to classify accurately.

Model diversity and RSC functions

One of the critical aspects of CFA is the role that model diversity plays in improving predictions. The diversity among the RSC functions of each model is visualized in several figures. For example, Figure 1 presents the diversity among the RSC functions of five models for data item d_{10}. The area between any two RSC functions represents the diversity between the corresponding models, and larger areas indicate greater CD. In this case, models B and E demonstrate the highest diversity, which contributes to the effectiveness of the WSCDS method when combining these models.

This diversity is critical in improving detection accuracy as models with higher diversity provide more complementary predictions, reducing errors.

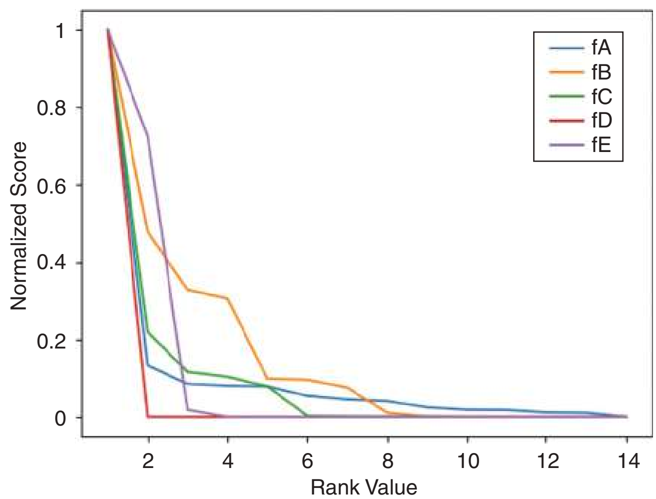


FIGURE 1. The CD among the five scoring systems based on their RSC functions for data item d_{10} from the test dataset. The models were trained on dataset X. The area between RSC functions represents diversity. Diversity strengths: A, 0.359996; B, 0.523272; C, 0.337381; D, 0.472812; E, 0.598646.

Single and combined model performance

Table 2 presents the best-performing models for each attack class based on the ratio of correct predictions. Instances where multiple models are listed for a given class indicate a tie in performance. The results demonstrate that the CFA improved the performance of the individual models in six classes and achieved performance equivalent to the highest-ranking individual models in five other classes. Notably, in four of these five ties, both the individual and combined models achieved perfect scores, detecting all attacks. This suggests that for these classes, the CFA reached the peak possible performance, indicating that further improvements through CFA were not possible.

For class 7, multiple CFA models matched the performance of individual model D, including combinations such as AC, AD, BD, CD, DE, ABC, ABD, ACD, ADE, BCD, BDE, CDE, ABDE, ACDE, BCDE, and ABCDE. In class 8, several CFA models tied with individual models A, D, and E, achieving perfect scores with combinations like AB, AD, AE, BD, BE, CD, CE, DE, ABC, ABD, ABE, ACD, ACE, ADE, BCD, BCE, BDE, CDE, ABCD, ABCE, ABDE, ACDE, BCDE, and ABCDE. Similar ties were seen in classes 12 and 13, where CFA models matched the perfect performance of the top individual models. While some improvements, such as the 0.0032% increase for class 2 in Table 2, are marginal and statistically measurable, they hold limited practical significance. More substantial improvements, such as the 3.31% increase in detection accuracy for class 6, demonstrate the method's effectiveness in addressing challenging attack scenarios.

CFA outperforms individual models. CFA outperformed individual models as follows:

- *Class 2*: The CFA model ABC achieved a slightly higher ratio (0.999875) compared with the individual models C and E (0.999833). This means one additional correct prediction, showcasing the CFA's ability to enhance performance even when individual models are already performing at a high level.

- *Class* 3: CFA model BC improved the correct predictions by 10, increasing the ratio from 0.984624 (individual model C) to 0.990538, thus enhancing the detection rate by approximately 0.6%. This improvement is significant in critical security applications where every detected attack counts.
- *Class* 4: The CFA model BC increased the number of correct predictions by 27 compared with the individual model C, raising the ratio from 0.998163 to 0.998843. While the percentage increase is small, it reflects the detection of additional attacks that might have been missed otherwise.
- *Class* 6: The most substantial improvement is observed here, where the CFA model BC increased the correct predictions by 47, raising the ratio from 0.957687 (95.77%) to 0.990832 (99.08%). This enhancement demonstrates the

TABLE 2. The best-performing models for each class based on the correct prediction ratio for training dataset X.

Class	Individual	Correct predictions (I)	Percentage	CFA	Correct predictions (Z)	Percentage	Z − I	Winner
0	B	109,844	99.714955	BCE	107,536	97.619782	−2,308	Individual
1	C	183	100.0	BCD	183	100.0	0	Tie
2	C, E	23,916	99.983278	ABC	23,917	99.987458	1	CFA
3	C	1,665	98.462448	BC	1,675	99.053814	10	CFA
4	C	39,674	99.816338	BC	39,701	99.884268	27	CFA
5	C	1,191	97.944079	BC	1,180	97.039474	−10	Individual
6	C	1,358	95.768688	BC	1,405	99.083216	47	CFA
7	D	998	99.8	AC, AD, BD, CD, DE, ABC, ABD, ACD, ADE, BCD, BDE, CDE, ABDE, ACDE, BCDE, ABCDE	998	99.8	0	Tie
8	A, D, E	2	100.0	AB, AD, AE, BD, BE, CD, CE, DE, ABC, ABD, ABE, ACD, ACE, ADE, BCD, BCE, BDE, CDE, ABCD, ABCE, ABDE, ACDE, BCDE, ABCDE	2	100.0	0	Tie
9	D	39,632	99.748314	BCD, ADE, ACD	39,702	99.924494	70	CFA
10	A	732	99.052774	ABC	730	98.782138	−2	Individual
11	A	319	93.823529	ABC, ACE	340	100.0	21	CFA
12	D	3	100.0	AD, BD, CD, DE, ABD, CDE	3	100.0	0	Tie
13	C, D	163	100.0	AD, BD, CD, ABD, ACD, BCD, BDE, CDE, ABDE, ACDE, BCDE, ABCDE	163	100.0	0	Tie

Ties indicate equal performance. CFA improved six classes and matched the best individual models in five. In four ties, both CFA and individual models achieved perfect scores, indicating peak performance.

CFA's effectiveness in detecting attacks that individual models may struggle with.

CFA achieves perfect scores (ties with individual models). In classes 1, 7, 8, 12, and 13, the CFA models achieved the same perfect scores as the best individual models. This indicates that the CFA approach reached the maximum possible performance for these classes. Since both the individual and combined models detected all attacks correctly, further improvement was not possible.

Consistency across multiple classes. The CFA models show consistent performance improvements or equivalence across multiple classes. This consistency is crucial in IDSs, where the ability to reliably detect a wide range of attack types is vital for maintaining security.

Individual models outperform CFA. In classes 0, 5, and 10, individual models performed better than the CFA models:

- *Class 0*: Individual model B achieved a higher ratio (0.99715) compared with the CFA model BCE (0.976198), with the CFA model having 2,308 fewer correct predictions.
- *Class 5*: Individual model C outperformed the CFA model BC, with a ratio of 0.979441 (97.94%) versus 0.970395 (97.04%).
- *Class 10*: Individual model A slightly outperformed the CFA model ABC, although the difference was minimal (0.990528 versus 0.987821).

These instances suggest that in some cases, the CFA may not always enhance performance and that individual models can be more effective, possibly due to overfitting in the combined models or the strengths of individual models in specific attack types.

Significance of CFA in improving detection rates. The CFA approach demonstrates its strength in enhancing detection rates for several classes, particularly where individual models have limitations. By combining models, CFA leverages the diverse strengths of different algorithms, leading to improved overall performance. The improvements in classes 2, 3, 4, 6, 9, and 11 are particularly noteworthy as they involve critical attack types where enhanced detection is essential.

Figure 2 illustrates the performance of all individual classes and combined models, where the CFA model consistently outperforms the individual models. Figure 2 shows the number of correct predictions for each of the

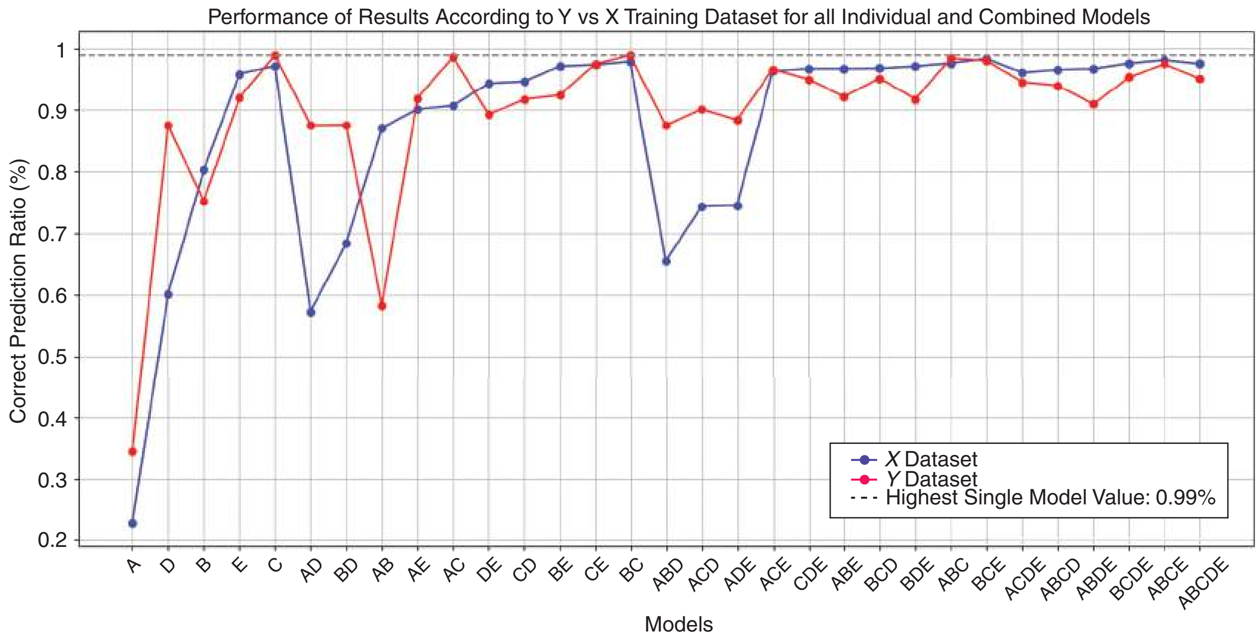


FIGURE 2. The percentage of correct predictions for each of the single and combined models using the WSCDS for all 14 classes combined.

single and combined models using WSCDS for all 14 classes combined. For example, in class 0, the combined CFA model achieves a prediction accuracy of 97.6%, compared with 99.7% for model B. In class 1, the CFA model reaches a perfect prediction accuracy of 100%, demonstrating its effectiveness in handling high-profile attacks.

The evolution of cyberthreats, like DoS/DDoS attacks, challenges the security of modern systems. Traditional ML-based IDSs often struggle with low-profile or evolving threats that mimic legitimate behavior. Integrating GAI and CFA provides a novel framework to enhance IDSs in secure, intelligent systems.

LEVERAGING GAI FOR ENHANCED DATA REPRESENTATION

In secure and intelligent systems, accurate anomaly detection relies on extensive training data that capture both normal and malicious behaviors. The integration of GAI models, such as generative adversarial networks (GANs), enhances the detection capabilities of IDSs by generating synthetic "normal" network traffic to augment training datasets. This helps address challenges like data imbalance and evolving threat landscapes, enabling IDSs to better distinguish among legitimate and malicious activities. Models like conditional GANs (CTGANs) can create balanced, diverse datasets. Next, we explore how GAI, when combined with CFA, can enhance intrusion detection by providing better data representation and facilitating more effective model fusion.

Synergy between GAI and CFA for intelligent decision making

By applying the WSCDS method, CFA enhances model accuracy and interpretability, with GAI further improving performance. CTGAN-generated synthetic data add diversity, helping models detect a wider range of attack behaviors. Combined with CFA's fusion techniques, this creates a more adaptable detection system for various attacks. The original dataset X had 660,944 entries, while the augmented dataset Y, with 1,321,888 data points, doubled its size. GAI models like CTGAN address the challenge of limited training data by generating synthetic data that balance imbalanced datasets, improving detection of underrepresented attacks such as DoS and DDoS.

Algorithm 1 presents our proposed method to enhance intelligent decision making by integrating CFA with GAI. It begins by training a CTGAN model on real data to generate synthetic data, which are then processed by multiple ML models. The outputs of these models are fused using CFA, and the CD is computed to assess model heterogeneity. If the CD meets a predefined threshold, the fusion is optimized; otherwise, additional synthetic data are generated, and the process repeats until optimized fusion is achieved. The final decision output is the result of the optimized fusion.

Further, CTGAN was used to generate synthetic data for augmenting the original training set of the LYCOS-IDS dataset.[15] This augmentation led to a twofold increase in the size of the training set, providing more robust data coverage and contributing to improved detection performance, particularly for attacks that individual models found challenging to detect.

Performance analysis

To evaluate the impact of GAI-integrated CFA, we used the same ML models employed in our baseline case study: SVM, ADB, MLP, GNB, and LDA. We compared the performance of individual models and their combinations under the WSCDS metric, both with and without synthetic data augmentation.

Performance across models and combinations. The summary of results in Table 3 further validates the

ALGORITHM 1:
GAI-integrated CFA: implementation use case.

1: **Input:** Real Data D_{real}, ML Models $\{M_1, M_2, \ldots, M_n\}$, CTGAN Model
2: **Output:** Optimized Model Fusion and Final Decision Output
3: Initialize Generative Model: CTGAN $\leftarrow$ train on D_{real}
4: Generate Synthetic Data: $D_{syn} \leftarrow$ CTGAN.generate()
5: Apply ML Models: $\{O_1, O_2, \ldots, O_n\} \leftarrow \{M_1(D_{syn}), M_2(D_{syn}), \ldots, M_n(D_{syn})\}$
6: Fuse Model Outputs: $O_{fused} \leftarrow$ CFA$(\{O_1, O_2, \ldots, O_n\})$
7: Compute Cognitive Diversity: CD $\leftarrow$ diversity$(\{O_1, O_2, \ldots, O_n\})$
8: **if** CD $\geq$ threshold **then**
9: Optimize Model Fusion: $O_{optimized} \leftarrow$ CFA.optimize(O_{fuses}, CD)
10: **else**
11: Generate More Synthetic Data: $D_{syn} \leftarrow$ CTGAN.generate()
12: **Repeat** Steps 5-7
13: **end if**
14: **Final Decision Output:** $O_{optimized}$

efficiency and superiority of CFA but this time integrated with GAI. Specifically, the superiority of model combinations over individual models for several classes is observed. The number of correct predictions across different attack types for both single and multimodel combinations showed notable improvements. For instance, class 9 saw a performance boost from 99.861573% to 99.879191% with the CFA approach.

GAI-integrated CFA versus baseline CFA: Comparative analysis. Table 4 further provides a class-level performance breakdown, demonstrating that GAI integration benefits some classes more than others. For example,

TABLE 3. The top-performing models for each class based on correct predictions for training dataset Y.

Class	Individual	Correct predictions (I)	Percentage	CFA	Correct predictions (Z)	Percentage	Z − I	Winner
0	C	108,184	98.208028	BC	108,046	98.082754	−138	Individual
1	D	183	100.0	AD, BD, CD, ABD, ACD, ADE, ABCD, ABDE	183	100.0	0	Tie
2	A, B, C	23,916	99.983278	AB, AC, BC, BD, BE, CE, ABC, ABD, ABE, ACD, ACE, ADE, BCD, BCE, BDE, CDE, ABCD, ABCE, ABDE, ACDE, BCDE, ABCDE	23,916	99.983278	0	Tie
3	C	1,649	97.516263	BC	1,649	97.516263	0	Tie
4	C	39,727	99.949682	AC, BC	39,727	99.949682	0	Tie
5	C	1,188	97.697368	BC	1,198	98.519737	10	CFA
6	D	1,343	94.71086	CDE	1,356	95.627645	13	CFA
7	B, C, D, E	996	99.6	AC, AD, AE, BC, BD, BE, CD, CE, DE, ABC, ABD, ABE, ACD, ACE, ADE, BCD, BCE, BDE, CDE, ABCD, ABCE, ABDE, ACDE, BCDE, ABCDE	996	99.6	0	Tie
8	A, B, C	2	100.0	AB, AC, AE, BC, BE, CE, ABC, ABD, ABE, ACD, ACE, BCD, BCE, ABCD, ABCE, ABDE, ACDE, BCDE, ABCDE	2	100.0	0	Tie
9	C	39,677	99.861573	ABC	39,684	99.879191	7	CFA
10	C, E	728	99.052774	AB	730	98.782138	2	CFA
11	C	338	99.411765	ABC	339	99.705882	1	CFA
12	A, D	3	100.0	AB, AD, BD, CD, DE, ABD, ACD, ADE, BCD, BDE, ABCD, ABDE, ACDE, ABCDE	3	100.0	0	Tie
13	D	159	97.546012	AD, BD, CD, DE, ABD, ACD, ADE, BCD, BDE, CDE, ABCD, ABDE, ACDE, BCDE, ABCDE	159	97.546012	0	Tie

Ties indicate equal performance. GAI-integrated CFA improved five classes and matched the highest-ranking models in eight others. Exploring alternative CFA metrics may resolve ties and enhance performance.

in class 5 (DoS/DDoS), the GAI-integrated CFA model BC improved performance from 97.94% (baseline) to 98.52%, showing a modest improvement of 0.58%. However, in class 0 (Bot), the baseline model B outperformed the GAI-integrated model, achieving 99.72% compared with the GAI model C's 98.21%, resulting in a decrease of 1.51%.

In more challenging classes, such as class 6 (DoS), the baseline CFA model BC performed significantly better, achieving 99.08%, while the GAI-integrated model CDE reached only 95.63%, marking a drop of 3.45%. Similarly, for class 3, the baseline model BC attained a higher accuracy of 99.05%, while the GAI-integrated version experienced a decrease to 97.52%.

TABLE 4. A class-level performance comparison between the GAI-integrated CFA (Y dataset) and baseline CFA (X dataset), highlighting the differences in correct prediction ratios for each class.

Class	Performance (X dataset)	Top model (X dataset)	Performance (Y dataset)	Top model (Y dataset)	Difference (%)
0	99.72%	B	98.21%	C	−1.51%
1	100%	C	100%	D, AD, BD, CD, ABD, ACD, ADE, ABCD, ABDE	0.0%
2	99.99%	ABC	99.98%	A, B, C, AB, AC, BC, BD, BE, CE, ABC, ABD, ABE, ACD, ACE, ADE, BCD, BCE, BDE, CDE, ABCD, ABCE, ABDE, ACDE, BCDE, ABCDE	−0.01%
3	99.05%	BC	97.52%	C, BC	−1.53%
4	99.88%	BC	91.37%	C, AC, BC	−8.51%
5	97.94%	C	98.52%	BC	0.58%
6	99.08%	BC	95.63%	CDE	−3.45%
7	99.80%	D, AC, AD, BD, CD, DE, ABC, ABD, ACD, ADE, BCD, BDE, CDE, ABDE, ACDE, BCDE, ABCDE	99.60%	B, C, D, E, AC, AD, AE, BC, BD, BE, CD, CE, DE, ABC, ABD, ABE, ACD, ACE, ADE, BCD, BCE, BDE, CDE, ABCD, ABCE, ABDE, ACDE, BCDE, ABCDE	−0.20%
8	100.0%	A, D, E, AB, AD, AE, BD, BE, CD, CE, DE, ABC, ABD, ABE, ACD, ACE, ADE, BCD, BCE, BDE, CDE, ABCD, ABCE, ABDE, ACDE, BCDE, ABCDE	99.60%	A, B, C, AB, AC, AE, BC, BE, CE, ABC, ABD, ABE, ACD, ACE, BCD, BCE, ABCD, ABCE, ABDE, ACDE, BCDE, ABCDE	−0.4%
9	99.92%	BCD, ADE, ACD	99.88%	ABC	−0.04%
10	99.05%	ABC, ACE	98.78%	AB	−0.27%
11	100.0%	A	99.70%	ABC	−0.3%
12	100.0%	D, AD, BD, CD, DE, ABD, CDE	100.0%	A, D, AB, AD, BD, CD, DE, ABD, ACD, ADE, BCD, BDE, ABCD, ABDE, ACDE, ABCDE	0.0%
13	100.0%	C, D, AD, BD, CD, ABD, ACD, BCD, BDE, CDE, ABDE, ACDE, BCDE, ABCDE	100.0%	D, AD, BD, CD, DE, ABD, ACD, ADE, BCD, BDE, CDE, ABCD, ABDE, ACDE, BCDE, ABCDE	0.0%

Overall, the integration of synthetic data via GAI typically enhances the performance of CFA models, especially in situations where increased data diversity is critical. However, the results are mixed at the class level, with some classes showing declines in performance. While GAI integration generally improves model robustness and detection accuracy, there are certain instances, such as in class 0 and class 6, where the baseline CFA models outperform their GAI-integrated counterparts.

GAI-integrated CFA: Learned lessons

Integrating synthetic data with CFA in IDS improves detection performance by enhancing training data diversity compared with baseline CFA (without GAI). While the baseline CFA (see the section "Case Study: Applying CFA to DoS Attack Detection") effectively combined ML models for attack detection, it lacked data diversity, limiting adaptability to evolving threats. Synthetic data overcome these limitations, boosting generalization and detection across a wider range of attacks.

Enhanced detection accuracy and recall. Baseline CFA models like BE and CE showed high recall but struggled with low-profile attacks (for example, class 1, Bot, and class 11, Webattack Bruteforce). GAI-enhanced models improved detection in difficult cases (for example, model DF increased recall for class 1). Synthetic data expanded decision boundaries, allowing better traffic differentiation (for example, model AF achieved higher recall for class 3 compared with baseline CFA).

Impact of CD on model fusion. CFA prioritizes CD to optimize model fusion, and GAI increased CD in many cases, enhancing overall fusion results. Models with higher CD outperformed others in recall and precision, with GAI-integrated CD enabling advanced WCs for better outcomes.

Broader generalization and adaptability. Synthetic data improve generalization in real-time environments, helping models adapt to emerging threats with less retraining. A comparison between synthetic and real data showed close alignment, boosting performance in real-world deployments. GAI-enhanced models enable scalable, continuously learning IDSs, essential for intelligent systems facing evolving threats.

Deployment considerations. Deploying the proposed system in large-scale real-world networks poses challenges such as increased processing time $(O(m^2 \cdot n))$ and memory demands for storing synthetic datasets and intermediate outputs. Addressing these challenges might require distributed computing, efficient memory management, and model optimization to ensure scalability and real-time performance.

Last, the proposed CFA-GAI framework has the potential to enhance scalability through CD and weighted fusion, enabling efficient model integration without extensive retraining. CTGAN-generated synthetic data further mitigate data imbalance, improving performance across diverse scenarios.

This article highlighted the potential of CFA to enhance intrusion detection in secure and intelligent systems. By combining multiple ML models using advanced SCs and RCs and leveraging CD, CFA improves detection accuracy, reduces false positives, and strengthens the accuracy of IDS against evolving cyberthreats. As cyberenvironments grow more complex, integrating CFA with GAI for real-time monitoring offers a path toward intelligent and adaptive cybersecurity solutions. C

ACKNOWLEDGMENT

Mohamed Rahouti's research was supported by the National Science Foundation, the Florida Center for Cybersecurity, and the Qatar Research Development and Innovation Council.

REFERENCES

1. M. A. O. Rabah, H. Drid, Y. Medjadba, and M. Rahouti, "Detection and mitigation of distributed denial of service attacks using ensemble learning and honeypots in a novel SDN-UAV network architecture," *IEEE Access*, vol. 12, pp. 128,929–128,940, 2024, doi: 10.1109/ACCESS.2024.3443142.
2. S. Muneer, U. Farooq, A. Athar, M. A. Raza, T. M. Ghazal, and S. Sakib, "A critical review of artificial intelligence based approaches in intrusion detection: A comprehensive analysis," *J. Eng.*, vol. 2024, no. 1, 2024, Art. no. 3909173, doi: 10.1155/2024/3909173.
3. A. Heidari and M. A. Jabraeil Jamali, "Internet of things intrusion detection systems: A comprehensive review and future directions," *Cluster Comput.*, vol. 26, no. 6, pp. 3753–3780, 2023, doi: 10.1007/s10586-022-03776-z.
4. T. Saranya, S. Sridevi, C. Deisy, T. D. Chung, and M. K. A. A. Khan, "Performance analysis of machine learning algorithms in intrusion detection system: A review," *Procedia Comput. Sci.*,

vol. 171, pp. 1251–1260, Dec. 2020, doi: 10.1016/j.procs.2020.04.133.
5. D. F. Hsu, Y.-S. Chung, and B. S. Kristal, "Combinatorial fusion analysis: Methods and practices of combining multiple scoring systems," in *Advanced Data Mining Technologies in Bioinformatics*, H.-H. Hsu, Ed., IGI Global, 2006, pp. 32–62.
6. L. Hurley, B. S. Kristal, S. Sirimulla, C. Schweikert, and D. F. Hsu, "Multi-layer combinatorial fusion using cognitive diversity," *IEEE Access*, vol. 9, pp. 3919–3935, 2020, doi: 10.1109/ACCESS.2020.3047057.
7. D. F. Hsu, B. S. Kristal, and C. Schweikert, "Combinatorial fusion analysis," *Computer*, vol. 57, no. 9, pp. 96–100, Sep. 2024, doi: 10.1109/MC.2024.3406058.
8. X. Cao and R. Yousefzadeh, "Extrapolation and AI transparency: Why machine learning models should reveal when they make decisions beyond their training," *Big Data Soc.*, vol. 10, no. 1, 2023, Art. no. 20539517231169731, doi: 10.1177/20539517231169731.
9. R. Uddin, S. A. P. Kumar, and V. Chamola, "Denial of service attacks in edge computing layers: Taxonomy, vulnerabilities, threats and solutions," *Ad Hoc Netw.*, vol. 152, 2024, Art. no. 103322, doi: 10.1016/j.adhoc.2023.103322.
10. N. Sahani, R. Zhu, J.-H. Cho, and C.-C. Liu, "Machine learning-based intrusion detection for smart grid computing: A survey," *ACM Trans. Cyber-Physical Syst.*, vol. 7, no. 2, pp. 1–31, 2023, doi: 10.1145/3578366.
11. P. M. Rao and B. D. Deebak, "Security and privacy issues in smart cities/industries: Technologies, applications, and challenges," *J. Ambient Intell. Humanized Comput.*, vol. 14, no. 8, pp. 10,517–10,553, 2023, doi: 10.1007/s12652-022-03707-1.
12. J. M. Kizza, "System intrusion detection and prevention," in *Guide to Computer Network Security*, Cham, Switzerland: Springer-Verlag, 2024, pp. 295–323.
13. D. F. Hsu, B. S. Kristal, Y. Hao, and C. Schweikert, "Cognitive diversity: A measurement of dissimilarity between multiple scoring systems," *J. Interconnection Netw.*, vol. 19, no. 1, 2019, Art. no. 1940001, doi: 10.1142/S0219265919400012.
14. D. F. Hsu, B. S. Kristal, and C. Schweikert, "Rank-score characteristics (RSC) function and cognitive diversity," in *Proc. Int. Conf. Brain Informat.*, 2010, pp. 42–54.
15. A. Rosay, F. Carlier, E. Cheval, and P. Leroux, "From CIC-IDS2017 to LYCOS-IDS2017: A corrected dataset for better performance," in *Proc. IEEE/WIC/ACM Int. Conf. Web Intell. Intell. Agent Technol. (WI-IAT)*, New York, NY, USA: Association for Computing Machinery, 2021, p. 6, doi: 10.1145/3486622.3493973.

ABOUT THE AUTHORS

EVANS OWUSU is pursuing a Ph.D. in computer science at the Department of Computer and Information Science, Fordham University, New York, NY 10458 USA. His research interests include network intrusion detection, ML, and blockchain technology. Owusu received an M.S. in applied mathematics from Gaziantep University and a second M.S. in computer science from Fordham University. Contact him at eowusu3@fordham.edu.

MARIYAM MAPKAR received a master's degree in data science at the Department of Computer and Information Science, Fordham University, New York, NY 10458 USA. Her research interests include network intrusion detection and ML. Contact her at mmapkar@fordham.edu.

MOHAMED RAHOUTI is an assistant professor at the Department of Computer and Information Science, Fordham University, New York, NY 10458 USA. His research interests include computer networking and security, blockchain technology, and AI/ML. Rahouti received an M.S. in statistics and a Ph.D. in electrical engineering from the University of South Florida. He is a Member of IEEE. Contact him at mrahouti@fordham.edu.

DINESH C. VERMA is at IBM Thomas J. Watson Research Center, Yorktown Heights, NY 10598 USA. His research interests include exploratory science projects applying generative AI models to networks and distributed systems. Verma received a Ph.D. from the University of California, Berkeley. He is a Fellow of IEEE, an IBM Fellow, an AAIA Fellow, and a Fellow of the U.K. Royal Academy of Engineering. Contact him at dverma@us.ibm.com.

The Emerging Metaverse: IEEE ISEMV 2024 Retrospective and Future Directions

Louis Nisiotis, University of Central Lancashire, Cyprus
Cecilia Metra, University of Bologna
Jyotika Athavale, Synopsys
Dejan Milojcic, Hewlett Packard Enterprise
Stephen Dukes, Imaginary Universes, LLC
Jeewika Ranaweera and **Sin-Kuen Hawkins**, IEEE

At the intersection of the physical and digital worlds, the metaverse is driving a new technological revolution across industrial applications, entertainment and gaming, transportation, and numerous other sectors.

The potential and applicability of the metaverse span a wide range of domains, fueled by recent technological advancements and innovations and the growing shift toward digitally enhanced lifestyles. Realizing this vision requires collaboration across multiple levels, bringing innovators, researchers, and professionals together to address critical technological, ethical, and social challenges. The IEEE Metaverse community and the inaugural 2024 IEEE International Symposium on Emerging Metaverse (ISEMV 2024) highlight the importance of this collaboration, encouraging the discussions and innovations aimed at building an interoperable and sustainable metaverse.

This article reflects on the symposium's pivotal role in bridging academic research with industry practices to address challenges and shape the future of the metaverse. Drawing on these insights, it proposes future directions for metaverse R&D, emphasizing the need for collective efforts to create a disruptive, cohesive, and inclusive metaverse ecosystem.

THE EMERGING METAVERSE

Although there is no universally accepted definition of the metaverse,[1,2] the concept is rapidly materializing due to

Digital Object Identifier 10.1109/MC.2025.3534445
Date of current version: 27 March 2025

significant advancements in software, hardware, systems, networking, and supporting technologies. Often described as the next technological revolution, the metaverse is envisioned as the convergence of virtual and physical realities within a digital ecosystem where individuals can interact, work, and coexist in immersive interconnected spaces. Recognizing its transformative potential, major industrial, governmental, and academic organizations are driving research, and standards, influencing corporate strategies, and prompting policy adaptations at both national and international levels. For instance, the European Commission[3] has outlined strategic frameworks for its development, while the World Economic Forum has highlighted its far-reaching societal and economic implications.[4] Companies across diverse industries have redefined their business strategies to align with the metaverse,[5] and technology developers are investing heavily in building the tools, environments, and infrastructures that underpin this ecosystem.[6] Recent technological advancements have accelerated progress, bringing the metaverse closer to realization.[7] In particular, artificial intelligence (AI) has played a pivotal role in shaping interactive metaverse environments, enabling content generation, natural language processing, adaptive experiences, intelligent behaviors, and complex computations.[8] These advancements enhance virtual interactions, making them more engaging and dynamic. Meanwhile, increasingly powerful computing hardware enables sophisticated real-time rendering and complex computations, essential for creating immersive and richly detailed systems.

Additionally, a wide array of accessible software development tools, rendering engines, open standards, and frameworks empowers a broad community of developers to contribute to the metaverse.

These innovations drive the development of immersive and accessible metaverse applications, emphasizing the importance of collaborative communities such as the IEEE Metaverse Initiative of IEEE Future Directions. This initiative promotes interdisciplinary research and facilitates the exchange of innovative ideas and breakthroughs. Aligned with this vision, the inaugural ISEMV 2024 was conceived as a premier forum by the IEEE Metaverse Initiative, with support from IEEE Future Directions and the IEEE Computer Society. The symposium brings together leaders, researchers, and practitioners to explore cutting-edge technological advancements and address societal implications of the metaverse. By fostering collaboration and innovation, ISEMV 2024 positions the metaverse at the forefront of technological progress, shaping the future of our interconnected digital lives.

ISEMV 2024

The theme of ISEMV 2024, "Navigating the Virtual Frontier," invited students, researchers, practitioners, and innovators to engage in a dynamic forum designed to foster collaboration, address technical challenges, and advance the convergence of the physical and digital worlds. The symposium aimed to shape and influence the future of a sustainable and resilient metaverse.

As a full-day, in-person event on 21 October 2024, ISEMV 2024 was co-located with the 23rd IEEE International Symposium on Mixed and Augmented Reality in Bellevue, WA, USA. As the inaugural symposium, it featured a distinguished lineup of keynotes, visionary talks, panel discussions, peer-reviewed paper presentations, and engaging poster presentations, all showcasing innovative applications of metaverse technologies. The contributions spanned a broad spectrum of topics, from foundational technologies to industry-specific innovations, demonstrating the disruptive potential of the metaverse.

The areas of interest at ISEMV 2024 encompassed a broad range of metaverse-related technologies and innovations, including AI, extended reality (XR), digital twins, cyber-physical systems, data management, human–AI collaboration, interoperability, interaction design, ethical AI, blockchain, gamification, edge/fog and accelerated computing, dependability, and open standards. The applications spanned diverse domains from the CityVerse to the EduVerse, all with a shared focus on fostering a resilient and universally accessible metaverse ecosystem.

The ISEMV 2024 symposium began with opening remarks from Cecilia Metra, general chair of ISEMV 2024 and IEEE Metaverse Initiative cofounder and cochair (together with Jeewika Ranaweera and Stephen Dukes), and Louis Nisiotis, technical program chair of ISEMV 2024. They were followed by welcome messages from Jyotika Athavale, 2024 IEEE Computer Society president, and Kathy Hayashi, IEEE director Region 6.

The following keynote sessions featured esteemed leaders in the metaverse field, providing valuable insights into cutting-edge industrial applications and visionary concepts, setting the stage for meaningful discussions on the potential of the metaverse across various sectors:

- "The Industrial Metaverse: Supercharging Industrial Efficiencies Towards Sustainability," presented by Gerhard Kress (Siemens), highlighted how the industrial metaverse can drive competitiveness and sustainability through the effective use of digital technologies. Using

real-world examples, Kress illustrated how companies leverage the transformation to achieve measurable benefits. He also introduced Siemens Xcelerator, an open business platform designed to accelerate the development of the industrial metaverse. He emphasized the need for standardization to bridge diverse ecosystems and promote interoperability.
- "We Need a Super Cockpit for the Mind," presented by Tom Furness (University of Washington; Virtual World Society), revisited key insights from his pioneering work while developing the "super cockpit" for the U.S. Air Force and explored its relevance to today's metaverse and AI landscapes. Furness proposed the creation of a "super cockpit for the mind" as a new interface designed to navigate the complexities of the digital future. He emphasized the need for virtual environments that empower and enhance human capabilities, ensuring technology serves as a tool for enhancement rather than an obstacle.
- "Challenges of Metaverse Silicon Health," presented by Yervant Zorian (Synopsys), addressed the critical role of silicon reliability in enabling the future of the metaverse. He discussed the increasing demands on emerging silicon technologies that are driven by AI, edge computing, and 5G integration. He explored solutions such as prognostics and analytics to improve quality, reliability, safety, and resilience. Zorian also highlighted strategies for addressing aging, degradation, and security challenges to optimize the health and performance of these foundational technologies to support the metaverse growth.

Throughout the day, four invited panels facilitated in-depth discussions on a wide range of domains that are relevant to the metaverse, featuring expert speakers from academia, industry, and research organizations.

The symposium brings together leaders, researchers, and practitioners to explore cutting-edge technological advancements and address societal implications of the metaverse.

The panel on "Metaverse for Education," organized and moderated by Arnold Pears (KTH Royal Institute of Technology), featured Stephen T. Frezza (Franciscan University of Steubenville), Tomi Kauppinen (Aalto University), and May Dongmei Wang (Georgia Institute of Technology; Emory University). The panel discussed the potential of generative AI and advanced large language models for enhancing virtual- and augmented-reality learning environments. The discussion highlighted how these technologies can create immersive educational experiences, support collaborative and project-based learning, and enable expert-driven interactions, redefining the future of education in the metaverse.

The "Digital Twins and the Metaverse" panel, organized and moderated by Dejan Milojicic (Hewlett Packard Labs), brought together insights from Carrie Dossick (University of Washington), Amit Dubey (Agilent Technologies), Larry Kaplan (Hewlett Packard Enterprise), and Ruby Leung (Pacific Northwest National Laboratory). The panel explored the pivotal role of digital twins in bridging real and virtual spaces, highlighting their diverse applications across industries. The panel covered wide areas of digital twin adoption for bridges, data centers, airplanes, and even the entire Earth. The panelists examined the commonalities and distinctions in addressing physical artifacts of vastly different time horizons, offering valuable insights into the challenges and opportunities in this domain.

The "Technological Challenges to Enable the Emerging Metaverse" panel, organized and moderated by 2024 IEEE Computer Society President Athavale (Synopsys), featured esteemed panelists Bala Kumaravel (Microsoft), Anna Mary Mathew (Microsoft), and Milojicic. The discussion covered the infrastructure requirements for the metaverse, emphasizing real-time multimedia processing, robust security measures, and lifecycle reliability. The discussions also highlighted the integration of digital twins and teleportation systems as critical components for delivering seamless and immersive user experiences.

The "Standards for the Metaverse" panel, co-led by Dukes (Imaginary Universes, LLC) and Nikolai Leung (Qualcomm), brought together prominent experts comprising Leonardo Chiariglione (Cedeo.net) and key leaders from the Metaverse Standards Forum, including Neil Trevett (Nvidia; The Khronos Group) and Christine Perey (PEREY Research & Consulting). The panelists discussed the need for interoperability among content providers, communication networks, and devices. They also highlighted the significance of coordinated efforts across organizations such as The Khronos Group, IEEE, MPEG, and the Metaverse Standards Forum to develop comprehensive and open standards, fostering a unified and interoperable metaverse ecosystem.

The keynotes and panels were further enriched by an invited talk from Mar Gonzalez-Franco (Google) titled "The Metaverse as a Hypothetical Exercise." Dr. Gonzalez-Franco explored the concept of the metaverse as a gradient, reflecting on hypothetical scenarios in which users might fully immerse themselves in virtual environments, and highlighting the importance of understanding user needs and identifying gaps in current development pipelines. Additionally, Leonardo Chiariglione (Cedeo.net) delivered a talk titled "A Path for a Metaverse Standard," emphasizing the complexities of developing unified standards due to diverse user requirements, rapid technological evolution, and the lack of a universal metaverse definition.

Each one of these sessions showcased the technological advancements that are driving the metaverse and offered a platform to explore its industrial, academic, educational, and social implications.

The symposium also featured technical paper presentations, with peer-reviewed articles from the ISEMV 2024 proceedings presented across three dedicated sessions. The technical papers collectively advanced the understanding and development of key metaverse technologies, frameworks, and applications. Some of the contributions focused on digital twin systems for sectors such as education, urban planning, and transportation, showcasing their potential for real-time data integration, simulation, and decision-making support. Interoperability emerged as a central theme, with articles addressing the need for unified standards, cross-platform connectivity, and scalable frameworks to create an open and inclusive metaverse. Research on AI-driven architectures demonstrated novel ways to enhance interactivity and user experiences within cyber-physical social systems. Studies on human–computer interaction explored innovative XR input methods and rapid prototyping approaches to improve user engagement and accelerate development cycles. Additionally, the potential of the metaverse for remote work and collaborative environments was examined, alongside sustainability-focused discussions on energy-efficient XR systems and the design of mixed reality workspaces. The contributions addressed technical challenges and highlighted practical solutions that promote accessibility, security, and cross-disciplinary collaboration, reinforcing the metaverse's role as a disruptive concept applicable to various real-world applications. The proceedings of ISEMV 2024 can be found in the following link: https://ieeexplore.ieee.org/xpl/conhome/10763946/proceeding

THE VALUE OF IEEE ISEMV FOR IEEE AND THE METAVERSE COMMUNITY

ISEMV 2024 highlighted the dynamic evolution of the metaverse, presenting some of the key trends and opportunities that are driving the field. One of the most significant trends discussed was the integration of XR and AI with other emerging technologies to create interconnected systems and metaverse applications that are capable of solving complex industrial, educational, and societal problems. However, the symposium also discussed challenges the metaverse is facing. One of the key issues identified was the need for interoperability across platforms and systems. The majority of current metaverse applications exist in isolated ecosystems, hindering the creation of seamless, user-centric experiences. The lack of universal standards and protocols results in fragmented digital environments in which users, assets, and content cannot move freely between different platforms, affecting the general consumer and enterprise experience, acceptance, and adoption.

IEEE ISEMV advances IEEE's mission of enhancing technology for the benefit of humanity. By fostering interdisciplinary dialogue, the symposium focused on bridging the gap between technological innovation and real-world applications. The keynotes and panels illustrated how the metaverse could address key challenges, from sustainability in the industry to equitable access to education, and the technical papers presented forward-thinking and innovative research advancing the current state of the art. The symposium's discussions and presentations on interoperability and standards reinforce IEEE's commitment to creating universally accessible and ethically sound technologies, reducing barriers to adoption and fostering global collaboration. Furthermore, we continue discussions on ethical considerations, including security and inclusivity.

THE FUTURE OF THE METAVERSE

Society and the industry are now getting beyond the phase of digital transformation, entering into the era of digital adaptation and augmented intelligence, with AI and advanced technologies being a part of our everyday lives. This will help with the metaverse's growing importance and future potential and the growing need for standardization, collaboration, and policies. This way, the future IEEE ISEMV will become part of a collective effort toward developing sustainable metaverse/metaverse applications that are scalable and accessible to everyone.

Metaverse applications provide the potential for seamlessly blending virtual and physical worlds to support meaningful interactions and addressing challenges, but they require collective efforts at multiple levels. The metaverse of the future has the potential to be an "eco-society of systems" that brings together virtual environments, the real world, artificial agents and elements, and humans, creating an ecosystem that complements reality rather than supplementing it. The metaverse's complexity requires a multidisciplinary approach that combines expertise from various fields working together to innovate and address key challenges such as accessibility, interoperability, privacy, security, and accessibility. Open access to the metaverse through accessible software development kits (SDKs) and applications for creators, developers, and researchers would enable them to contribute their expertise and creativity. It should be affordable, user friendly, and inclusive to do so.

The ability for applications to work together seamlessly through the establishment of universal standards and protocols that allow communication among different metaverse platforms is a key area that requires collaboration. The metaverse requires standardized frameworks to ensure interoperability. Industry consortia and standardization bodies play a key role in developing these protocols; for example, organizations like the IEEE Standards Association lead the efforts to develop these standards and address challenges. Establishing open standards and protocols that allow technology interoperability, and the exchange of digital assets, data, and interactions across platforms can help users to be part of an immersive metaverse experience through a unified digital presence.

Additionally, the environmental impact of metaverse technologies, such as the energy consumption of servers. The carbon footprint of the hardware powering the metaverse further highlights the need for considering green computing solutions and eco-friendly protocols.

A practical implementation strategy (road map) for advancing the metaverse requires short-term strategic goals that are focused on identifying and defining the necessary core protocols, agreed-upon standards, and efforts toward ensuring the accessibility of development tools and services. This involves collaborating with standards organizations to establish foundational interoperability protocols, while simultaneously providing open access SDKs to creators, developers, and researchers. Medium-term goals would prioritize achieving seamless cross-platform interactions through the adoption of universal standards for digital assets, and the ability to integrate diverse technologies for interoperable user experiences. Emphasis should be placed on the use cases that meaningfully involve users. Finally, the long-term goals aim for full integration of physical and virtual worlds into eco-societies of unified systems where the real and virtual worlds, humans, and elements coexist together in a digitally augmented reality.

Figure 1 illustrates the direction of the emerging metaverse evolution over the next five years, emphasizing vertical domains, horizontal technologies, and the transition from consumer to enterprise-focused applications. By focusing on short-term goals for interoperability and accessibility, medium-term goals for cross-platform integration, and long-term goals for intelligent systems and inclusivity, the

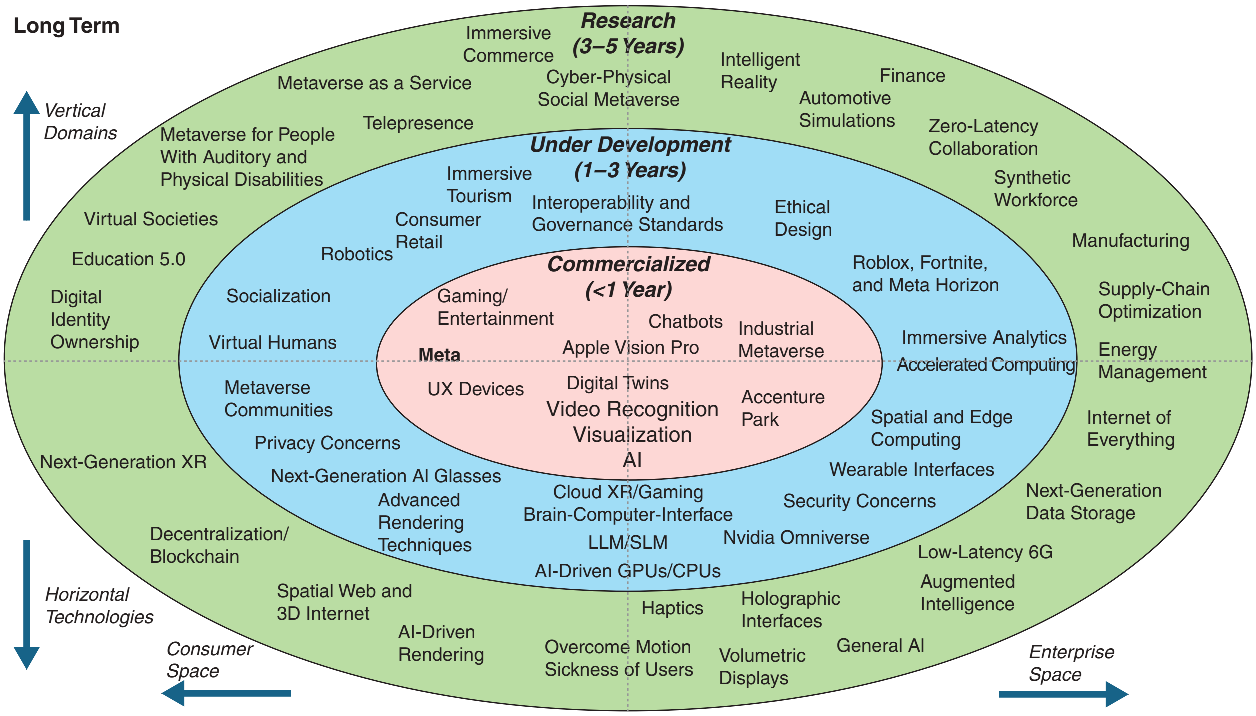


FIGURE 1. The direction of emerging metaverse evolution, from now to five years out, is categorized into three dimensions. 1) Horizontals [commercialized (less than one year)], under development (one to three years) and research (three to five years); 2) consumer versus enterprise space; and 3) vertical domains versus horizontal technologies. LLM/SLM: large language model/small language model; UX: user experience.

metaverse can evolve into an interconnected digital ecosystem that benefits individuals, industries, education, academia, and society in general.

The future of the metaverse focuses on how to develop applications that blend the real with the digital world through the seamless interplay of technology, humans, artificial agents, and elements. The emphasis will be on the development of metaverse applications that are not replacing or virtualizing our lives but complementing reality, supporting human connections and interactions with each other and with technology.

The metaverse, in our view, should focus on meaningfully bridging the physical and digital worlds to create spaces that connect people, promoting collaboration and shared experiences, by leveraging the unique features of both realities. It would be for exploring how digital advancements can support real-world interactions, rather than isolating its users in virtual environments and disconnecting them from reality. This means that the metaverse should be digitally supported and enhanced to become a version of our reality rather than a replacement for our daily professional and social interactions. Such an approach requires careful design of interoperable, inclusive, accessible, and safe metaverse applications, developed with the purpose and intent to support and shape social dynamics and experiences.

The development of a metaverse eco-society requires a strategic vision and coordinated efforts among industries, governments, and academia to establish standards, address ethical and social issues, and drive technological innovation. In our opinion, technologically, the emphasis should be placed on the development of open access tools, development kits, platforms, and standards that support the development and fusion of immersive XR with emerging technologies, establishing a seamless and interconnected "network of everything" to address social, environmental, and industrial challenges, and contributing meaningfully to daily life while promoting ethical standards, inclusivity, and accessibility. It is key to create standardized and user-friendly methods of interaction within such an ecosystem. Converging XR with technologies such as AI, robotics, digital twins, blockchain, the Internet of Things, 5G and 6G, edge and cloud computing, dependability, and other emerging disruptive technologies will enable the creation of human-centric metaverse applications, systems, and experiences that support a plethora of domains and applications and meaningfully contribute to our daily lives.

Society and the industry are now getting beyond the phase of digital transformation, entering into the era of digital adaptation and augmented intelligence, with AI and advanced technologies being a part of our everyday lives.

CALL FOR COLLABORATION

Realizing the metaverse as an accessible and interconnected digital ecosystem relies on collaboration across multiple technical disciplines and application domains. The complexity of building an interoperable and sustainable metaverse requires academic institutions, industry leaders, policy makers, standards development organizations, and government entities to align their efforts and share resources. This collaboration must extend beyond technological development to include ethical and societal considerations, ensuring that the metaverse evolves in ways that are beneficial to society.

A key aspect of this collaboration involves defining and adopting core protocols and universal standards. Working groups and related bodies should join forces in establishing the necessary frameworks for seamless cross-technology interaction and integration to enable interoperability and promote security, privacy, and trust within the metaverse ecosystem. Interdisciplinary partnerships and knowledge exchange will enable us to collectively shape a future where an eco-society of metaverse systems and applications becomes reality. We invite researchers, practitioners, policy advisors, and community members to actively engage and contribute to the collective effort to create a resilient, user-centered, and accessible metaverse.

The inaugural ISEMV 2024 marked a milestone in the efforts within the IEEE metaverse initiative in advancing the understanding and development of the metaverse. By bringing together industry leaders, researchers, practitioners, and innovators from diverse fields, the symposium fostered collaboration, contributing to ongoing efforts to shape a resilient, inclusive, and interoperable metaverse ecosystem. The discussions, keynotes, and panels highlighted opportunities for innovation and key challenges. As the metaverse evolves from concept to reality, it is clear that its success requires a multidisciplinary and collaborative approach. ISEMV 2024 highlighted the importance of open standards, protocols, and accessible development tools that enable creators, developers, and researchers to contribute to a unified metaverse. The insights and innovations shared at ISEMV 2024 contribute to the ongoing discourse for a future where physical and virtual realities seamlessly blend into an eco-society of systems that complements and enhances our lives.

Moving forward, the metaverse community must build on this momentum by collaborating across disciplines, promoting equitable access, and ensuring that the metaverse develops and enables meaningful interactions, addressing societal and industrial challenges and shaping the future of our interconnected digital lives.

With the recent significant advancements in technology, concepts that were mere ideas decades ago are now becoming reality, and similarly, the visions we imagine today will materialize in time. In the metaverse space, ongoing efforts are pushing the boundaries of what is achievable, advancing technological breakthroughs in hardware, software, computation, and networking speeds. The cutting-edge research that was presented at ISEMV 2024 is a key example of this forward-thinking momentum. The visionary keynote speakers offered insights into emerging trends and challenges, highlighting the metaverse's disruptive potential across industries and society in general. The comprehensive panel discussions highlighted key areas for fusing technological advancements, creating interoperable platforms and addressing pressing current and future issues through multidisciplinary research to shape the future of immersive ecosystems. The cutting-edge research presented through technical papers at ISEMV 2024 demonstrated the forefront of innovation in the metaverse field, exploring topics ranging from advancements in XR and AI to the development of interoperable frameworks and sustainable digital ecosystems. The diversity of the technical papers showcased the interdisciplinary nature of metaverse research, bridging gaps between engineering, computer science, and social sciences; contributing to existing scientific discourse on practical applications; and driving innovation.

Through the groundbreaking ideas shared at ISEMV 2024, it is clear that the metaverse goes beyond just being a technological frontier. The metaverse is a platform for disrupting the way we understand and interact with real and digital worlds and with each other, building the premises for the vision of today to become the metaverse of tomorrow. C

REFERENCES

1. M. Y. Almoqbel, A. Naderi, D. Y. Wohn, and N. Goyal, "The metaverse: A systematic literature review to map scholarly definitions," in *Proc. Companion Publication Conf. Comput. Supported Cooperative Work Social Comput.*, pp. 80–84, 2022, doi: 10.1145/3500868.3559448.
2. P. Faraboschi, E. Frachtenberg, P. Laplante, D. Milojicic, and R. Saracco, "Virtual worlds (metaverse): From skepticism, to fear, to immersive opportunities," *Computer*, vol. 55, no. 10, pp. 100–106, Oct. 2022, doi: 10.1109/MC.2022.3192702.
3. "Towards the next technological transition: Commission presents EU strategy to lead on Web 4.0 and virtual worlds." European Commission. Accessed: Jan. 16, 2025. [Online]. Available: https://ec.europa.eu/commission/presscorner/detail/en/ip_23_3718
4. "Interoperability in the metaverse," The World Economic Forum and Accenture, Cologny, Switzerland, Jan. 2023. https://www.weforum.org/publications/interoperability-in-the-metaverse/
5. H. Ning et al., "A survey on metaverse: The state-of-the-art, technologies, applications, and challenges," 2022, *arXiv:2201.04888*.
6. K. Yawised, D. Apasrawirote, and C. Boonparn, "From traditional business shifted towards transformation: The emerging business opportunities and challenges in 'metaverse' era," in *Proc. Int. Nat. Conf. Bus. Admin. Accountancy*, 2022, vol. 162, pp. 162–175.
7. L. Nisiotis, "A system architecture for AI-driven interactive systems within a CPSS framework for metaverse applications," in *Proc. Int. Symp. Emerg. Metaverse (ISMEV)*, Seattle, WA, USA, 2024, pp. 9–12, doi: 10.1109/ISEMV63338.2024.00016.
8. P. Faraboschi, E. Frachtenberg, P. Laplante, D. Milojicic, and R. Saracco, "Artificial general intelligence: Humanity's downturn or unlimited prosperity," *Computer*, vol. 56, no. 10, pp. 93–101, Oct. 2023, doi: 10.1109/MC.2023.3297739.

LOUIS NISIOTIS is an assistant professor at the University of Central Lancashire, Cyprus, 7080 Pyla, Cyprus. Contact him at LNisiotis@uclan.ac.uk.

CECILIA METRA is a full professor at the University of Bologna, 40136 Bologna, Italy. Contact her at cecilia.metra@unibo.it.

JYOTIKA ATHAVALE is a director of engineering architecture at Synopsys, Inc., Sunnyvale, CA 94087 USA. Contact her at jyotika@synopsys.com.

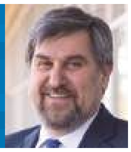
DEJAN MILOJCIC is a Hewlett Packard Enterprise fellow and VP at Hewlett Packard Labs, Milpitas, CA 95035 USA. Contact him at dejan.milojicic@hpe.com.

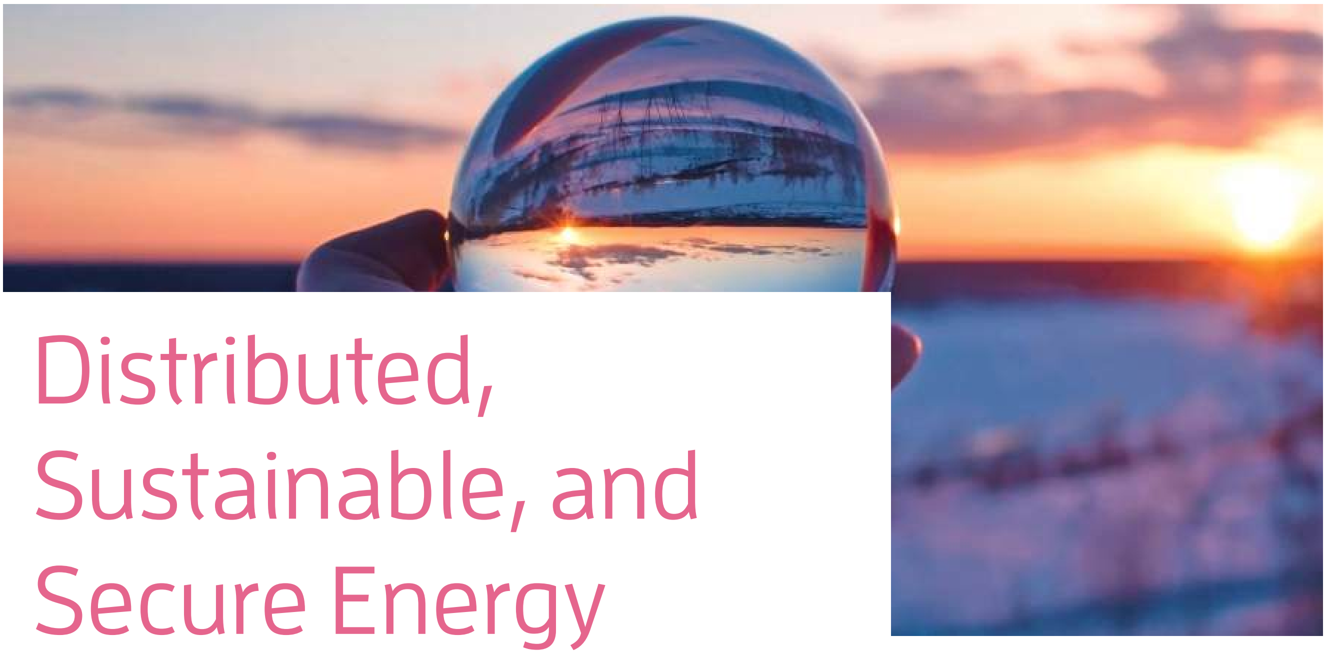
STEPHEN DUKES is an executive and visionary with 47 years of business and technical leadership. He is the president and CEO of Imaginary Universes, LLC, Lake Stevens, WA 98258 USA. You can contact him at stephendukes@ieee.org.

JEEWIKA RANAWEERA is an engineer, author, illustrator, inventor, sustainability champion, frequent conference speaker, and avid volunteer for IEEE and various women-in-tech organizations, San Jose, CA 95128 USA. Contact her at jeewika.ranaweera@ieee.org.

SIN-KUEN HAWKINS is a program director of IEEE Future Directions, Piscataway, NJ 08854 USA. Contact her at sinkuen.hawkins@ieee.org.

EDITOR DEJAN MILOJICIC
Hewlett Packard Labs;
dejan.milojicic@hpe.com





Distributed, Sustainable, and Secure Energy Future Through IT Modernization

Diman Zad Tootaghaj and **Yiqian Wu**, Hewlett Packard Labs and Columbia University

Ting He and **Nilanjan Ray Chaudhuri**, The Pennsylvania State University

Dejan Milojicic, Hewlett Packard Labs and Columbia University

Luka Strezoski, University of Novi Sad and DerMag Consulting

James Anderson and **Puneet Sharma**, Hewlett Packard Labs and Columbia University

The decentralized and digitally interconnected future of power grids will be achieved through IT modernization to meet the unprecedented challenges and opportunities amid the rise of distributed energy resources, climate resilience demands, and energy storage advancement.

Digital Object Identifier 10.1109/MC.2025.3534375
Date of current version: 27 March 2025

The global energy sector is undergoing a groundbreaking transformation,[1] driven by the rapid proliferation of inverter-based resources (IBRs) and distributed energy resources (DERs),[a] the imperative to combat pollution issues and climate change, and advancements in energy storage technology. This evolution marks a fundamental shift from traditional centralized power systems to a decentralized, dynamic, and increasingly complex energy landscape, shown in Figure 1.

This shift requires a collaborative interaction between the centralized and local agents, which complicates the roles of energy management systems/distributed management systems (EMSs/DMSs) at the central level and different participants at the local level. At the core of this decentralized paradigm lies the idea of

[a]All IBRs connected to the distribution grid are DERs. On the other hand, IBRs can be connected to the transmission grid as well, and in that case, they are not DERs.

prosumers—a new concept in power distribution systems that encapsulates the dual role of individuals as both consumers and producers of energy. Prosumers actively participate in energy systems by not only consuming energy but also generating it, typically through DERs like solar panels, wind turbines, electric vehicles (EVs), and energy storage systems. By allowing decisions to be made closer to generation and consumption, localized approaches ensure more efficient use of DERs, reduce transmission losses, and enable faster responses to local energy demands. The central EMS facilitates the real-time coordination and optimization of grid operations, ensuring stability and reliability in the face of decentralized actions. As the integration between transmission and distribution grids becomes increasingly essential, the seamless operation of EMSs alongside advanced distribution management systems (ADMSs)[4] and DER management systems (DERMSs) is paramount.[5]

The ADMS serves as the backbone of modern distribution grid operations, integrating real-time data and advanced analytics to enhance situational awareness. By enabling efficient monitoring and control, the ADMS enables the coordination of distributed resources and operational assets, bridging the gap between transmission-level planning and distribution-level execution.

The DERMS extends this capability further, focusing on optimizing the integration and utilization of DERs. By dynamically managing generation, storage, and demand-side flexibility and incorporating advanced forecasting techniques, the DERMS empowers grid operators to predict energy generation and consumption patterns. This allows for more precise adjustments, enabling operators to meet sustainability goals while maintaining reliability and resilience across increasingly complex networks.

However, effective collaboration between the central EMS and local agents requires new communication protocols, data exchange mechanisms, and control strategies. This includes developing smarter systems to bring forecasting data into the decision-making process.

Effective collaboration between the central EMS and local agents requires new communication protocols, data exchange mechanisms, and control strategies.

The integration of renewable energy sources, such as solar and wind, brings sustainability to the forefront but also introduces challenges of intermittency and operational volatility. Simultaneously, the grid's growing reliance on information and communication technology (ICT) enhances efficiency but exposes this critical infrastructure to cybersecurity risks. Amid these developments, energy storage emerges as a linchpin for grid stability, making it easier to integrate renewable energy and strengthening the system's ability to handle challenges.

This article delves into the challenges and opportunities presented by these trends. It explores the need for advanced technologies and innovative management strategies to ensure reliability, resilience, and sustainability in the face of these transformative changes. As we stand on the brink of a new era in energy, the future of power grids will depend on our ability to navigate these complexities while building systems that can adapt to an unpredictable world.

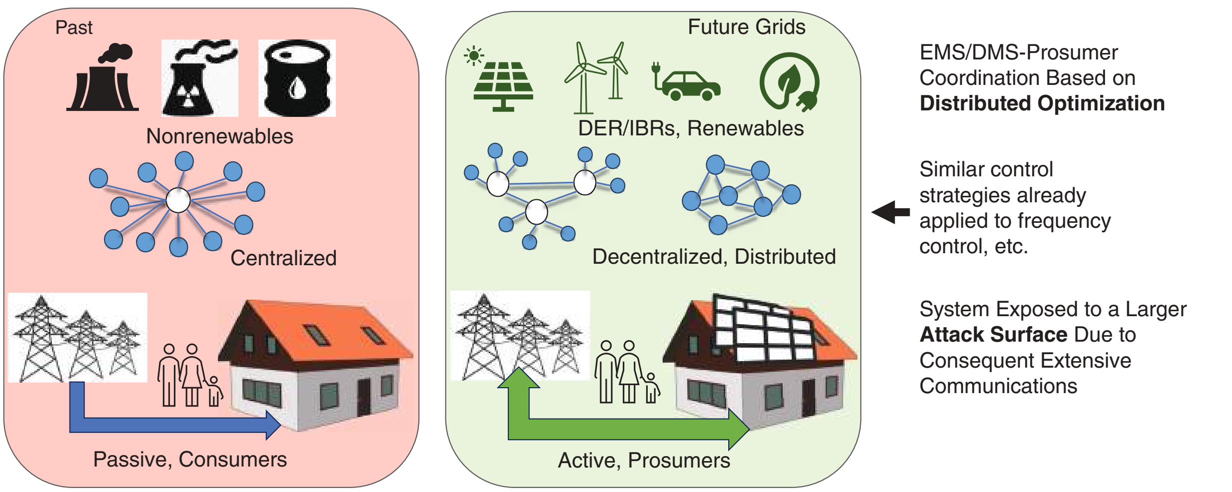


FIGURE 1. The transformation of the power system.

CHALLENGES OF DEEP PENETRATION OF DERS

The integration of DERs, including solar panels, wind turbines, and battery storage, has redefined how we generate and manage electricity. However, this shift introduces several challenges.

Uncertainty and volatility

Renewable energy sources, while sustainable, are inherently variable and uncertain. Solar output depends on daylight and weather, while wind generation fluctuates with wind patterns. This intermittency creates operational challenges in maintaining the balance between load and generation and the so-called (*N*-*k*) security of the grid.[7] Advanced artificial intelligence/machine learning (AI/ML) forecasting techniques and real-time energy markets are crucial for managing these fluctuations and uncertainty.

ICT dependency

The modern grid is increasingly dependent on ICT to enable automation, real-time coordination, and improved efficiency. However, this reliance introduces vulnerabilities that can compromise grid stability. Failures in ICT systems, such as sensor breakdowns and network disruptions, can cascade into power grid failures,[2,3,6,11] creating widespread operational risks, as shown in Figure 2. Additionally, this dependency increases exposure to cyber threats, including false data injection attacks[8,9,10,12] and unauthorized access, which can jeopardize grid integrity. To mitigate these risks, it is essential to invest in robust cybersecurity measures, such as advanced intrusion detection systems, anomaly detection algorithms, and secure communication protocols. Strengthening the resilience of digital infrastructures will be key to safeguarding the grid against both physical and cyber disruptions.[17,18,19]

Need to improve situational awareness

Maintaining grid stability depends on accurate and continuous monitoring to gain situational awareness of the system. With the advent of advanced metering infrastructure, such as phasor measurement units (PMUs) and smart meters, the online monitoring capability has markedly improved. However, due to the vast and complex nature of the power grid, many more sensors need to be placed to gain full observability. PMUs, while highly effective at providing real-time high-precision measurements of voltage and current phasors, are expensive and require significant investment for deployment. As a result, only a limited number are currently deployed in the grid. Given their high cost, it is crucial to determine optimal locations for these devices to maximize their effectiveness and coverage while minimizing deployment costs.

This lack of observability, especially in areas where PMUs are not deployed, can hinder effective decision making and compromise grid reliability. To address this challenge, it's important to invest in improved metering technologies and innovative approaches to optimize their placement for maximum coverage at limited cost.

Strengthening the resilience of digital infrastructures will be key to safeguarding the grid against both physical and cyber disruptions.

Leveraging advancements in sensor technology and data analytics can further enhance observability, ensuring a more resilient and adaptive power grid. State estimation (SE) applications within ADMSs play a crucial role in addressing these challenges by significantly increasing situational awareness, even in cases of low redundancy of measurements and sensors. By processing the limited available data, SE algorithms can provide estimates of unmeasured states, improving grid observability and supporting real-time decision making. However, SE algorithms for distribution grids are still being perfected due to the structural and

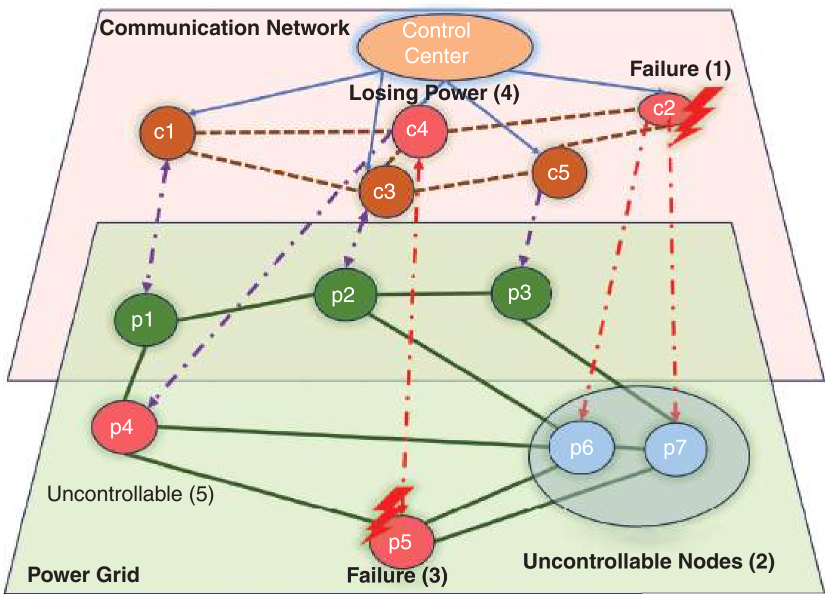


FIGURE 2. ICT dependency leading to cascading failures.

operational differences between distribution and transmission grids. Unlike in transmission systems, where SE algorithms are well established and operate with a high redundancy of sensors, distribution grids face challenges such as greater complexity, lower measurement density, and unbalanced operations. Despite these hurdles, advancements in SE for distribution systems continue to hold significant promise for enhancing grid reliability and resilience.

Power system transformation

The decline of conventional generators, such as coal and nuclear plants, reduces grid inertia, traditionally a key stabilizing factor. Innovative solutions, like grid-forming inverters and synthetic inertia from renewable sources, can help address these challenges. Numerous other challenges are emerging from different fronts, including 1) weak grid issues due to limited short circuit current contributions from IBRs/DERs, which seriously challenge today's protection systems, 2) weak grid stability issues, and 3) transient and voltage stability challenges. Addressing these challenges faces further complications since the IBR/DER vendors refuse to share the models to protect intellectual property. Moreover, these changes are demanding a major effort in introducing new and overhauling existing operating standards of equipment in the power grid.

Small-scale energy resources

Distributed systems composed of small-scale resources, like rooftop solar panels and community batteries, add complexity to grid management. While aggregating these resources through virtual power plants can provide reliable energy and ramping capacity, a significant challenge lies in the fact that these resources are typically "behind the aggregator," making them unobservable to the EMS. This lack of direct visibility can hinder the EMS's ability to manage the grid effectively and requires coordination and data-sharing protocols between aggregators and grid operators.

Decentralized energy management

Traditional unidirectional grids are evolving into bidirectional systems where energy flows between consumers and producers. However, passive distribution grids, which were previously straightforward to monitor, control, and operate, are now becoming active, dynamically changing systems. This shift requires traditional distribution network operators to evolve into much more involved distribution system operators, taking on new responsibilities and complexities. To address these challenges, novel tools, such as ADMSs and DERMSs, have become indispensable.

This transformation places significant demands on utility staff, who are often not well trained to work with these advanced systems. For instance, managing and operating the grid with preset optimization devices, such as load tap changers (LTCs) at predetermined locations, was previously sufficient. Today, such static approaches are no longer viable; the ADMS must continuously monitor grid conditions and react in real time, dynamically adjusting LTCs to ensure optimal performance. Similarly, traditional approaches to grid protection, relying on preset relay protections, are rendered inadequate by the presence of dynamically changing grid conditions and the proliferation of IBRs. Modern grids require real-time monitoring and adaptive adjustments to maintain protection and reliability. Active distribution networks and decentralized control systems, leveraging technologies like distributed optimization, are essential to enable this transformation. These advancements ensure the seamless integration of new energy technologies and the maintenance of grid stability in the face of unprecedented complexity.

Integration of the high penetration of EVs

Global initiatives are actively progressing to integrate large numbers of EVs as a part of efforts to electrify and decarbonize the transportation sector.[23] While this transition is crucial for reducing greenhouse gas emissions and air pollution, the mass integration of EVs presents challenges for electrical power systems, particularly at the distribution level.

One major challenge lies in planning, where traditional methodologies are insufficient to account for the unpredictable nature of EV charging demands and the widespread placement of charging stations. This can lead to inaccurate load forecasts and inadequate grid reinforcement. Operationally, the integration of large numbers of EVs can cause peak load increases, voltage violations, feeder overloads, network congestion, and even transformer aging. The stochastic nature of EV charging—characterized by varying user behaviors and charging schedules—compounds these issues, making it difficult to predict and manage their impact on the grid.

Vehicle-to-grid (V2G) and grid-to-vehicle (G2V) services are increasingly recognized as critical components of the EV ecosystem. V2G technology allows EVs to act as DERs, feeding power back to the grid during peak demand periods, while G2V services focus on optimized and efficient charging of EVs to reduce grid stress during off-peak hours. These bidirectional energy flows can significantly enhance grid flexibility, improve demand-side management, and provide ancillary services such as frequency regulation and voltage support.

The simultaneous charging of EV fleets can lead to localized stresses, particularly at heavily utilized feeders or substations, exacerbating reliability concerns. The lack of comprehensive monitoring and control mechanisms across most distribution grids further hinders the ability to address these technical challenges in real time.

To address these challenges, we need innovative solutions that make it easier to integrate EVs into the grid without sacrificing stability. Advanced systems like ADMSs and

DERMSs are proving to be essential for managing the complexities that come with EV adoption. These tools provide real-time monitoring, dynamic control, and resource optimization, helping to prevent problems like voltage drops, overloaded feeders, and spikes in demand. However, it is important to acknowledge that once a certain threshold of EVs is reached, the existing grid infrastructure will necessitate hardware upgrades to enhance its capacity and adequately support the complete electrification of transportation.

RELIABILITY OF POWER GRIDS AMID CLIMATE CHANGE

The grid must be strengthened to withstand heat waves, storms, and other climate-related events. For instance, heat waves can cause power lines to sag, increasing wildfire risks, as shown in Figure 3. In California, nearly half of the most destructive wildfires have been attributed to power line failures in vulnerable areas.[20] One notable case is the 2017 Nuns Fire, which burned more than 56,000 acres, destroyed 1,355 structures, caused three fatalities, and was traced back to sagging power lines.[21] Notably, sagging power lines owned by Pacific Gas & Electric (PG&E) were identified as the cause of catastrophic wildfires, including the 2018 Camp Fire—the deadliest wildfire in California's history—and the 2021 Dixie Fire. These events resulted in significant loss of life, destruction of property, and widespread environmental devastation.[20]

Such incidents underscore the critical need for proactive measures, such as undergrounding power lines, employing storm-resistant designs, and implementing adaptive protection systems. Additionally, investing in advanced monitoring and predictive systems can help mitigate the risks of sagging lines during extreme heat events and prevent similar tragedies in the future.

In response to this growing risk, utility providers such as PG&E in California have implemented Public Safety Power Shutoffs (PSPSs), a reactive strategy to preemptively cut power in areas at high risk of wildfires.[22] With the current human-in-the-loop decision-making process for such PSPSs, the size of the impacted consumer base can be fairly large. This not only endangers the critical loads but also unfairly impacts certain communities if their power is repeatedly shut off. A well-designed power grid topology analytics engine can quantify the impact of climate change on the risk of wildfire ignition from each power line, enabling automated and fine-grained PSPSs that can minimize the number of impacted consumers.

Climate change has amplified the frequency and severity of extreme weather events, placing extraordinary stress on grid infrastructure. While PSPSs provide short-term mitigation, effectively addressing these challenges requires proactive and forward-looking strategies, including the following:

- *Resilience to extreme weather:* Figure 3 shows how heat waves can cause sagging lines and result in wildfire ignition. Grid infrastructure must be fortified against weather extremes through measures like underground cables, storm-resistant designs, and adaptive protection systems.
- *Microgrid capabilities:* Microgrids can enhance localized energy independence, enabling communities to maintain power even during widespread outages. Scaling up microgrid deployment and improving their interconnectivity will be critical for future resilience. Future grid infrastructure must account for evolving climate conditions. By integrating advanced climate models into grid planning, utilities can design systems capable of withstanding conditions anticipated decades into the future. Proactive approaches such as these will be instrumental in addressing the dual challenges of climate resilience and wildfire risk reduction, ensuring equitable and reliable power delivery for all communities.
- *Long-term climate adaptation:* Planning for future scenarios requires integrating climate models into grid planning. This ensures that new infrastructure can withstand anticipated conditions decades from now.

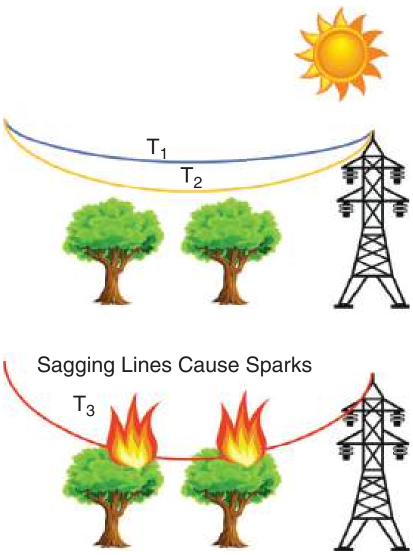


FIGURE 3. The impact of climate change.

DECENTRALIZATION OF ADMSS/DERMS

Decentralized EMSs are poised to play a crucial role in shaping the future grid by enabling efficient and scalable coordination among distributed resources. Leveraging distributed optimization and control techniques, these systems establish a robust framework for communication and collaboration across diverse energy assets. The decentralization of system operations can be broadly categorized into two primary strategies based on communication rules: global broadcast with local response and neighboring consensus-based coordination. These strategies offer varying levels of decentralization and address unique challenges in modern energy systems.

Global broadcast with local response

In this approach, a central coordinating entity broadcasts public information, such as generation capacity, demand forecasts, electricity price signals, and operational constraints, to all agents in the network. Each agent processes this global information to make local decisions independently. This ensures system-wide transparency and coordinated decision making in a hierarchical manner. For instance, real-time pricing mechanisms allow a central controller to broadcast electricity prices, guiding consumption and production decisions. Similarly, demand response programs leverage this approach to coordinate DERs to balance supply and demand during peak or off-peak hours. To strengthen privacy while retaining operational efficiency, distributed optimization algorithms, such as the Alternating Direction Method of Multipliers (ADMM),[16] can be employed. These methods decompose system-wide scheduling problems into smaller localized subproblems, reducing the dependency on central coordination. By distributing computation and decision making, a decentralized energy management scheme with minimal central dependency can be achieved. However, this scheme still faces challenges, including the vulnerability of the central entity as a single point of failure.

Neighboring consensus-based coordination

Here, following a fully decentralized approach, each agent makes local decisions and exchanges information on estimates of shared variables, such as aggregated demand or shared constraints, exclusively with its neighbors. Communication is restricted to interneighbor groups, aligning naturally with peer-to-peer communication structures,[13] promoting privacy and scalability, and reducing communication overhead.[14,15] For example, voltage regulation in distribution networks can be achieved through local agents coordinating reactive power adjustments with neighbors using consensus algorithms or multiagent reinforcement learning (MARL). While this approach eliminates the reliance on central entities and scales efficiently through localized interactions, it faces challenges such as slower convergence in large systems and the computational complexity of advanced techniques like MARL.

Both strategies, global broadcast with local response and neighboring consensus-based coordination, rely heavily on bidirectional and iterative communications, exposing decentralized EMSs to cybersecurity threats. The integration of DERs and supporting ICT infrastructure increases the system's attack surface, creating vulnerabilities at the cyberphysical interface. Potential threats include false data injection, where critical variables such as price signals or demand forecasts are manipulated, and device compromise, where unauthorized control of DERs like inverters or smart meters disrupts grid operations. Denial-of-service attacks that overload communication channels to block or delay critical control signals and eavesdropping, where sensitive operational data are intercepted, are also significant concerns. Mitigating these risks requires robust measures, such as encryption to secure communications, These include AI-based anomaly detection to identify abnormal patterns, resilience planning with redundant systems and fallback mechanisms, and privacy-preserving algorithms like differential privacy to protect sensitive data during communication.

LOCALITY OF ENERGY PRODUCTION AND CONSUMPTION

A microgrid is a localized energy system that can operate either independently or in conjunction with the main grid. It typically consists of interconnected loads and DERs within a defined boundary, capable of providing energy to a specific area. Microgrids represent a significant advancement in the shift toward localized energy production and consumption.

One of the primary benefits of microgrids lies in their ability to enhance reliability and resilience. By operating independently during grid outages (islanding capability), microgrids ensure a stable energy supply for critical infrastructure and communities. Additionally, localized energy generation reduces transmission and distribution losses, contributing to improved overall system efficiency.

Microgrids are particularly well suited for integrating renewable DERs, such as solar panels, wind turbines, and energy storage systems. Their localized nature allows for the effective utilization of intermittent renewable resources, which would otherwise pose challenges in larger centralized grids. As microgrids aim for sustainable energy solutions, small modular reactors (SMRs) are emerging as a clean and reliable option for providing stability and consistent power generation. SMRs complement renewable DERs, ensuring continuous operation during periods of low renewable output.

The concept of networked microgrids takes this approach further by interconnecting multiple microgrids. This interconnected system enhances reliability by enabling resource sharing and mixing diverse energy sources, creating a robust and flexible energy network. Such configurations can provide significant benefits, especially in regions prone to natural disasters or grid instability.

To realize the full potential of microgrids, advanced microgrid management systems (MMSs) are essential. These software solutions optimize energy production, storage, and consumption within the microgrid. MMSs facilitate real-time monitoring, control, and decision making, ensuring efficient and reliable operations while adapting to dynamic energy demands and supply conditions. By leveraging the advantages of microgrids and advancing their management systems, the concept of localized energy production and consumption offers a transformative solution for achieving a sustainable, resilient, and efficient energy future.

THE ROLE OF ENERGY STORAGE IN FUTURE GRIDS

Energy storage plays a critical role in enabling the sustainable and profitable operation of power grids by providing value-adding services as follows:

- *Peak shaving:* Peak shaving refers to the process of reducing the power demand on the grid during periods of peak usage. Energy storage systems play a critical role in this by storing excess energy during off-peak hours (when demand is low and energy is cheaper) and releasing it during peak hours (when demand is high). This reduces the need for utilities to activate expensive and less efficient peaking power plants, lowering operational costs and emissions. Peak shaving benefits both utility providers, by avoiding costly infrastructure upgrades, and consumers, by lowering energy bills through demand charge reductions.
- *Power smoothing:* Power smoothing addresses the variability and intermittency of renewable energy sources like solar and wind. These sources are prone to fluctuations due to changes in weather conditions, which can create instability in the power grid. Energy storage systems help stabilize the grid by absorbing excess energy during periods of high generation and releasing it during periods of low generation. This process ensures a steady and predictable power output, enabling the better integration of renewable energy resources into the grid and enhancing grid reliability.
- *Load balancing:* Load balancing involves maintaining a constant equilibrium between power supply and demand to ensure the stable operation of the grid.

To realize the full potential of microgrids, advanced microgrid management systems are essential.

Energy storage systems contribute to load balancing by dynamically adjusting their charge and discharge cycles based on real-time grid conditions. During times of surplus generation, storage systems absorb excess energy, preventing the overloading of the grid. Conversely, during times of high demand, they release stored energy to meet the shortfall, reducing the reliance on fossil fuel-based backup power sources and enhancing the overall efficiency and sustainability of the grid.

- *Ancillary services:* Storage systems provide essential grid services, such as frequency regulation, voltage support, and black-start capabilities. These services ensure grid stability amid high renewable penetration.
- *Enhancing DER profitability:* Energy storage allows for time-shifting renewable energy production, maximizing revenue by selling power during peak demand periods. This enhances the economic viability of renewable projects.
- *Defense against sequential attacks:* Grid resilience must account for sequential cyberattacks targeting normal operations and recovery phases. Advanced threat detection systems and redundant storage configurations can mitigate these risks.

Figure 4 describes the role of energy storage in future grids. Just like

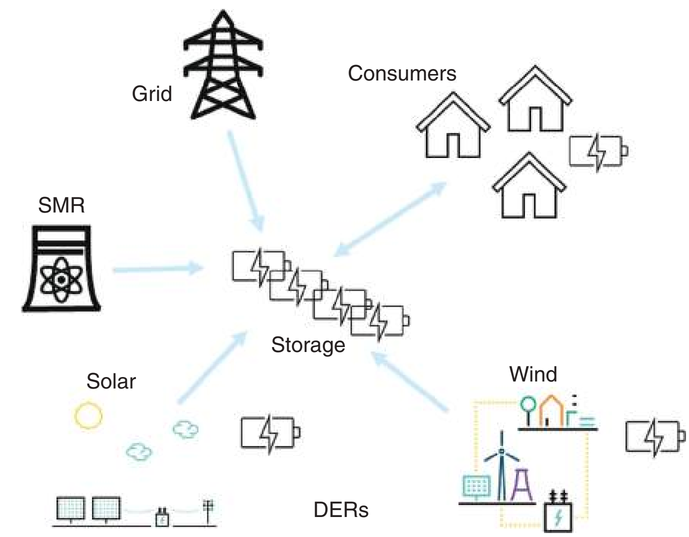


FIGURE 4. Energy storage.

computers and data, storage contains energy that can be utilized according to the aforementioned use cases. Storage can be located in centralized places or close to producers or consumers.

As power grids evolve, their future will be defined by resilience, innovation, and sustainability. Navigating challenges like DER integration, climate change, and cybersecurity will demand close collaboration across engineering, policy, and technology domains. Modernizing EMSs will play a key role in this transition, using advanced technologies to make the grid smarter and more adaptable. By embracing decentralized energy systems, enhancing storage deployment, and investing in advanced technologies, we can create a grid that is not only reliable and secure but also capable of powering a sustainable future.

REFERENCES

1. C. Bash, J. McDonald, D. Milojicic, J. Mueller, T. Wilson, and Y. Zhang, "Convergence of energy and IT sectors," *Computer*, vol. 58, no. 1, pp. 151–158, Jan. 2025, doi: 10.1109/MC.2024.3498572.
2. D. Z. Tootaghaj, N. Bartolini, H. Khamfroush, and T. L. Porta, "Controlling cascading failures in interdependent networks under incomplete knowledge," in *Proc. IEEE 36th Symp. Reliable Distrib. Syst. (SRDS)*, 2017, pp. 54–63, doi: 10.1109/SRDS.2017.14.
3. D. Z. Tootaghaj, N. Bartolini, H. Khamfroush, T. He, N. R. Chaudhuri, and T. L. Porta, "Mitigation and recovery from cascading failures in interdependent networks under uncertainty," *IEEE Trans. Control Netw. Syst.*, vol. 6, no. 2, pp. 501–514, Jun. 2019, doi: 10.1109/TCNS.2018.2843168.
4. L. Strezoski, H. Padullaparti, F. Ding, and M. Baggu, "Integration of utility distributed energy resource management system and aggregators for evolving distribution system operators," *J. Mod. Power Syst. Clean Energy*, vol. 10, no. 2, pp. 277–285, Mar. 2022, doi: 10.35833/MPCE.2021.000667.
5. L. Strezoski, "Distributed energy resource management systems—DERMS: State of the art and how to move forward," *WIREs Energy Environ.*, vol. 12, no. 2, e460, Jan. 2023, Art. no. e460, doi: 10.1002/wene.460.
6. M. Parandehgheibi, E. Modiano, and D. Hay, "Mitigating cascading failures in interdependent power grids and communication networks," in *Proc. IEEE Int. Conf. Smart Grid Commun. (SmartGridComm)*, 2014, pp. 242–247, doi: 10.1109/SmartGridComm.2014.7007653.
7. Y. Yang, X. Guan, and Q. Zhai, "Fast grid security assessment with N- k contingencies," *IEEE Trans. Power Syst.*, vol. 32, no. 3, pp. 2193–2203, May 2017, doi: 10.1109/TPWRS.2016.2608378.
8. L. Che, X. Liu, Z. Shuai, Z. Li, and Y. Wen, "Cyber cascades screening considering the impacts of false data injection attacks," *IEEE Trans. Power Syst.*, vol. 33, no. 6, pp. 6545–6556, Nov. 2018, doi: 10.1109/TPWRS.2018.2827060.
9. G. Morgenstern, J. Kim, J. Anderson, L. Zussman, and T. Routtenberg, "Protection against graph-based false data injection attacks on power systems," *IEEE Trans. Control Netw. Syst.*, vol. 11, no. 4, pp. 1924–1936, Dec. 2024, doi: 10.1109/TCNS.2024.3371548.
10. L. Che, X. Liu, Z. Li, and Y. Wen, "False data injection attacks induced sequential outages in power systems," *IEEE Trans. Power Syst.*, vol. 34, no. 2, pp. 1513–1523, Mar. 2019, doi: 10.1109/TPWRS.2018.2871345.
11. H. Cetinay, S. Soltan, H. Kuipers, G. Zussman, and P. V. Mieghem, "Comparing the effects of failures in power grids under the AC and DC power flow models," *IEEE Trans. Netw. Sci. Eng.*, vol. 5, no. 4, pp. 301–312, Oct./Dec. 2018, doi: 10.1109/TNSE.2017.2763746.
12. J. Liang "False data injection, optimization, state estimation, vulnerability analysis," *IEEE Trans. Power Syst.*, vol. 31, no. 5, Art. no. 3864–3872, Sep. 2016.
13. S. Mhanna, G. Verbič, A. C. Chapman, "Accelerated methods for the SOCP-relaxed component-based distributed optimal power flow," in *Proc. Power Syst. Comput. Conf. (PSCC)*, 2018, pp. 1–7.
14. Z. Cheng and M.-Y. Chow, "Resilient collaborative distributed AC optimal power flow against false data injection attacks: A theoretical framework," *IEEE Trans. Smart Grid*, vol. 13, no. 1, pp. 795–806, Jan. 2022, doi: 10.1109/TSG.2021.3113287.
15. F. Ye, X. Cao, Z. Cheng, and M.-Y. Chow, "CASL: A novel collusion attack against distributed energy management systems," *IEEE Trans. Smart Grid*, vol. 14, no. 6, pp. 4717–4728, Nov. 2023, doi: 10.1109/TSG.2023.3251401.
16. S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein, "Distributed optimization and statistical learning via the alternating direction method of multipliers," *Found. Trends Mach. Learn.*, vol. 3, no. 1, pp. 1–122, 2010, doi: 10.1561/2200000016.
17. F. Schweppe and J. Wildes, "Power system static-state estimation, Part I: Exact model," *IEEE Trans. Power App. Syst.*, vol. PAS-89, no. 1, pp. 120–125, Jan. 1970, doi: 10.1109/TPAS.1970.292678.
18. X. Nian-de, W. Shi-ying and Y. Er-keng, "A new approach for detection and identification of multiple bad data in power system state estimation," *IEEE Trans. Power App. Syst.*, vol. PAS-101, no. 2, pp. 454–462, Feb. 1982, doi: 10.1109/TPAS.1982.317128.
19. Y. Gu, T. Liu, D. Wang, X. Guan, and Z. Xu, "Bad data detection method for smart grids based on distributed state estimation," in *Proc. IEEE Int. Conf. Commun. (ICC)*, 2013, pp. 4483–4487, doi: 10.1109/ICC.2013.6655273.
20. "Public safety power shutoff (PSPS)/de-energization." California Public Utilities Commission. Access: Feb. 6, 2025. [Online] Available: https://www.cpuc.ca.gov/psps/
21. D. Kasler, "PGE gets blamed for another deadly 2017 wildfire, this time from 'Sagging Power Lines'," *The Sacramento Bee*, Oct. 9, 2018.

[Online]. Available: https://www.sacbee.com/news/california/fires/article219731815.html
22. V. Yan, "Early wildfire detection with line sensors," Ph.D. dissertation, Dept. Elec. Eng., California Polytechnic State Univ., San Luis Obispo, CA, USA, 2021.
23. L. Strezoski and I. Stefani "Enabling mass integration of electric vehicles through distributed energy resource management systems," *Int. J. Elect. Power Energy Syst.*, vol. 157, Jun. 2024, Art. no. 109798, doi: 10.1016/j.ijepes.2024.109798.

DIMAN ZAD TOOTAGHAJ is a senior researcher at Hewlett Packard Labs, Milpitas, CA 95035 USA. She is a Senior Member of IEEE. Contact her at diman.zad-tootaghaj@hpe.com.

YIQIAN WU is a Ph.D. student in the Department of Electrical Engineering, Columbia University, New York, NY 10027 USA. She is a Student Member of IEEE. Contact her at yiqian.wu2@columbia.edu.

TING HE is an associate professor at the Pennsylvania State University, University Park, PA 16802 USA. She is a Senior Member of IEEE. Contact her at tinghe@psu.edu.

NILANJAN RAY CHAUDHURI is an associate professor in the Electrical Engineering Department, The Pennsylvania State University, University Park, PA 16802 USA. He is a Senior Member of IEEE. Contact him at nuc88@psu.edu.

DEJAN MILOJICIC is an HPE Fellow and vice president at Hewlett Packard Enterprise, Milpitas, CA 95035 USA. He is a Fellow of IEEE. Contact him at dejan.milojicic@hpe.com.

LUKA STREZOSKI is an associate professor in the Department for Power, Electronics, and Telecommunications Engineering, Faculty of Technical Sciences, University of Novi Sad, 21000 Novi Sad, Serbia. He is a Senior Member of IEEE. Contact him at lukastrezoski@uns.ac.rs.

JAMES ANDERSON is an associate professor in the Department of Electrical Engineering, Columbia University, New York, NY 10027 USA. He is a Senior Member of IEEE. Contact him at james.anderson@columbia.edu.

PUNEET SHARMA is the director of the Networking and Distributed Systems Lab, Hewlett Packard Labs, Milpitas, CA 95035 USA. He is a Fellow of IEEE. Contact him at puneet.sharma@hpe.com.

Digital Object Identifier 10.1109/MC.2025.3546134

Engineering? No Way

Sorel Reisman, California State University, Fullerton

The activity of "prompt engineering" is neither prompt nor does it require any traditional knowledge or basis of the engineering profession. "Experts" speak out!

As a fan of TV detective programs, I have always been fascinated by how detectives interact with suspects in police station interrogation rooms. The rooms are always bare except for a table, two or three chairs, and the always-present two-way mirror. There is usually a team of two interrogators playing the roles of good and bad cop, and their script is almost always the same: start slow and simple and build up to more complex matters to finally get the suspect to reveal information.

There are many circumstances in which professionals (and even nonprofessionals) use similar well-defined (formal or informal) scripts to achieve their ends. Examples include psychotherapists interacting with patients/clients; doctors and nurses communicating with patients to gather medical histories and symptoms; social workers working with individuals and families to address social, emotional, and financial challenges to understand clients' needs and provide support; lawyers and legal advisors consulting with clients to gather information about legal issues and provide legal advice and representation; journalists and interviewers conducting interviews to gather information and stories; customer service representatives engaging in conversations to understand and resolve customer issues; human resources professionals conducting interviews and discussions with employees and job candidates to address workplace concerns; financial advisors discussing financial goals and plans with clients to provide advice based on clients' financial situations and objectives; teachers and educators interacting with students to assess their understanding and learning needs; and even parents communicating with their children.

The processes for all of these are similar, although their duration and objectives may differ. In each case a "conversation leader" employs "rubrics" to accomplish their goals. Except for some very special circumstances, it is notable that the practices employed by those listed above are devoid of and quite different from those inherent in the practices of the Accreditation Board of Engineering and Technology (ABET)'s methodologies inherent in processes deployed by professional engineers:

> "The profession in which a knowledge of the *mathematical or physical sciences* [italics are mine]

Digital Object Identifier 10.1109/MC.2025.3534264
Date of current version: 27 March 2025

EDITORS
GEORGE HURLBURT U.S. Federal Service (Retired), USA;
gfhurlburt@gmail.com
SOREL REISMAN California State University, USA;
sreisman@computer.org

> gained by study, experience, and practice is applied with judgment to develop ways to utilize, economically, the materials and forces of nature for the benefit of mankind."[1]

ABET is not alone in identifying mathematical and physical sciences as foundational to engineering practices. (See also Smith,[2] Wikipedia,[3] and Kirkey and Soltan.[4]) It would be highly problematic, even strange, to incorporate engineering terminology to label the examples above as psychotherapy engineering, medical diagnosis engineering, sociology engineering, education engineering, legal engineering, journalism engineering, customer service engineering, human resources engineering, interview engineering, financial engineering, or parent engineering!

Incorporating the adjective "engineering" into undertakings primarily related to "communicating" seems to be illogical. Consider, for example, a typical "communication" recognizable by virtually any parent (P) with an 11-year-old son, in this case, Jeff (J):

P: Hi, Jeff. How was school today?
J: Fine.
P: What did you do there that made it "fine"?
J: We did some stuff.
P: What kind of stuff?
J: You know, math stuff.
P: Can you explain what you mean? Did you learn how to divide numbers?
J: Yes.
P: What kind of numbers?
J: Big ones.
P: Like?
Etc.

What exactly is going on here? P asks question after question, each new one incrementally probing to elicit details about today's math class. Eventually, with enough patience, P may learn what was taught at school today. Even though P is "prompting" J through the communication, could any reasonable person call this parent/child interaction "parent engineering"? Of course not. This is simply a conversation that might take place while "parenting." As we will see below, J is far less effusive and forthcoming in his responses than a generative artificial intelligence (GAI) tool would be in a typical human/GAI interaction. The article "How to Write Better ChatGPT Prompts to Improve Conversations With AI" advises, "It's as if you are having a normal conversation with a coworker or an acquaintance. You can include some conversation details, like anecdotes, to give your story some texture."[5]

Let us have a look at another conversation, similar to the P/J dialogue above but one between a human (H) and a GAI (Copilot):

H: I went to a concert last night.
GAI: That sounds like a great experience! Concerts can be so energizing and memorable. Who was performing? Was it an outdoor or indoor event? Feel free to share more details—I'd love to hear about it!
H: It was an indoor concert.
GAI: Indoor concerts have a unique ambiance, don't they? The acoustics, lighting, and intimate setting can make the experience truly memorable. Who was the artist or band performing?
H: Taylor Swift.
GAI: Ah, Taylor Swift! She's an incredible artist known for her captivating performances and relatable lyrics. Did she perform any of your favorite songs?
H: No.

Can you see significant differences between this and the earlier P/J interactions? Two entities are conversing with one another, the first between two people and the second between me and Copilot. The first is simply a conversation, and the second is called "prompt engineering"! Wherein lies any notion of *engineering* in the latter conversation? Rather than continue to rant about what appears to be the misappropriation of the adjective describing my hard-earned profession, I decided to ask, in a very biased way, GAIs themselves if newbie AI advocates have misappropriated the term "engineering." Table 1 summarizes the responses of three different GAIs. (A comparison of these GAIs can be found in Rees.[6])

I then input all three response groups to Copilot, requesting (prompting?) an analysis of the similarities and differences among the three sets of responses (Table 2). In summary, while all three groups argue against using the term "engineering" for prompt creation, they approach the topic from slightly different perspectives, with varying degrees of detail, critique, and future-oriented thinking. Clearly, if I had worded the prompt in a different and more positive way, the GAIs would have provided counterarguments. (For example, "The term 'engineering' in 'prompt engineering' is completely *appropriate*. Give me 5 reasons that the term *should be* used in this context.") In fact, in all fairness to the other side of the argument, that is exactly what I did. I fed Claude.ai that prompt, and here is what it came back with:

> "Okay, here are 5 reasons why the term 'engineering' is appropriate to use in the context of 'prompt engineering':
>
> 1) *Structured problem-solving approach*: Like other engineering disciplines, prompt engineering involves a systematic,

step-by-step process of identifying problems, formulating solutions, testing hypotheses, and iteratively refining prompts to achieve desired outcomes from AI systems. This structured approach to problem-solving is a core tenet of engineering.

2) *Optimization and iteration*: Prompt engineers must continuously test, analyze, and refine prompts to optimize the AI's responses. This iterative design process, where ideas are prototyped, evaluated, and improved, is a hallmark of many engineering practices.
3) *Technical understanding and skill*: Crafting effective prompts requires a deep understanding of language, linguistics, cognitive science, and the underlying technical architecture of the AI model. This technical knowledge and skill set is akin to the domain expertise that underpins traditional engineering disciplines.
4) *Interdisciplinary collaboration*: Prompt engineering often involves collaborating with cross-functional teams, including subject matter experts, designers, and developers. This interdisciplinary approach to

TABLE 1. The misuse of "engineering" in "prompt engineering."

Prompt: "The term 'engineering' in 'prompt engineering' is completely inappropriate. Give me 5 reasons that the term should not be used in this context."	
GAI group	Response
Group 1: Copilot (CHAT GPT-4)	1) *Lack of formal methodology*: Traditional engineering disciplines follow rigorous methodologies and standards. Prompt engineering, however, often relies on trial and error without a standardized approach. 2) *Absence of technical design*: Engineering typically involves designing and building complex systems or structures. On the other hand, prompt engineering focuses on crafting effective inputs for AI models, which does not involve the same level of technical design and construction. 3) *Educational background*: Engineers usually require formal education and certification in their field. Prompt engineering has no formal educational requirements, making it more accessible but less structured. 4) *Scope of work*: Engineering projects often have a broad scope, involving multiple stages from planning to execution. Prompt engineering is more narrowly focused on optimizing interactions with AI models. 5) *Regulatory oversight*: Engineering fields are often subject to regulatory oversight to ensure safety and compliance. Prompt engineering lacks such regulatory frameworks as it is a relatively new and evolving practice.
Group 2: Perplexity (CHAT GPT-3, mainly)	1) *Lack of technical complexity*: Traditional engineering involves complex problem-solving, requiring a deep understanding of scientific principles and technical skills. In contrast, prompt engineering often involves crafting sentences or questions to elicit desired responses from AI models, which lacks the technical depth typically associated with engineering disciplines. 2) *Misleading perception of expertise*: Using "engineering" may falsely imply that prompt crafting requires extensive training or expertise akin to fields like software or civil engineering. In reality, effective prompting often relies more on clear communication and understanding of language rather than technical prowess. 3) *Overhyped skillset*: The term "prompt engineering" has been criticized as overhyped, suggesting it is a new and essential skill when it may not be as critical as portrayed. The skills involved are often basic communication skills rather than specialized engineering skills, leading some to view it as a marketing term rather than a genuine field of expertise. 4) *Evolution of AI interaction*: As AI systems evolve, the role of human interaction with these systems is expected to shift toward more strategic and ethical considerations rather than technical prompt crafting. This evolution suggests that the term "engineering" may not accurately represent the future role of those interacting with AI. 5) *Nontechnical nature*: The term "engineering" traditionally implies a technical or mechanical process, whereas prompt engineering involves more of an artful approach to language and communication. This divergence from the technical roots of engineering further supports the argument that the term is not well suited for this context.
Group 3: Claude.AI	1) *Lack of formal methodology*: Traditional engineering disciplines rely on established scientific principles and standardized processes. Prompt creation often lacks such rigorous, formalized methodologies. 2) *Absence of quantifiable metrics*: Engineering typically involves precise measurements and calculations. Prompt effectiveness can be subjective and challenging to quantify consistently. 3) *Limited predictability*: Engineers generally work with systems that behave predictably given specific inputs. AI responses to prompts can be highly variable and context dependent. 4) *No licensing or standardization*: Engineering fields often require professional certification and adhere to industry standards. Prompt creation has no such regulatory framework or universally accepted best practices. 5) *Emphasis on creativity over technical skills*: While engineering involves creativity, it primarily requires technical knowledge and problem-solving skills. Effective prompt writing often relies more heavily on linguistic creativity and an understanding of human communication.

problem-solving is common in various engineering fields.
5) *Measurable outcomes*: While the evaluation of prompt effectiveness may involve some subjectivity, there are measurable metrics that prompt engineers can use to assess the performance and quality of their designs, such as relevance, coherence, and user satisfaction. This focus on quantifiable outcomes is a key aspect of engineering."

It concluded as follows: "Furthermore, the term 'engineering' conveys a sense of rigor, discipline, and intentionality that is highly relevant to the practice of prompt design. It suggests a methodical approach to shaping the interactions between humans and AI systems, which aligns with the goals and responsibilities of prompt engineers. While there may be some differences between prompt engineering and traditional engineering disciplines, the core principles of structured problem-solving, technical expertise, iterative optimization, and measurable outcomes justify using the term 'engineering' in this context."

I really like Claude as a GAI, but for a number of reasons, I think these arguments are weak and maybe even a little "hallucinogenic":

1) There is no reference to math and science as requisite to the practice of engineering, topics that probably took up at least 80% of my own undergraduate electrical engineering education.
2) Almost all the arguments can be made for just about any professional communicator needing to elicit results from their communications; see the list at the beginning of this article.
3) Despite arguments to the contrary, and regarding the only specific GAI item in the list (item 3), there is absolutely no need for most GAI "prompters" to have an understanding of the "underlying technical architecture of the AI model."

If we were to accept Claude's contentions, would we be inclined to call a Google search engineering? So, what would be a more appropriate term to replace the inappropriate "prompt engineering"? I contend that whatever we use should not use the adjective "engineering." What about the word "prompt"? Does it belong in the context of GAI querying? What exactly does this word mean? Without going into its etymology, Merriam-Webster's online dictionary tells us it can be 1) a verb (for example, to incite or cue), 2) an adjective (for example, describe a person or action), or 3) a noun (for example, reminder). Clearly, that word has no specific utility in the context of querying a GAI.

So, with both parts of the phrase "prompt engineering" being completely unapt, I must conclude that the phrase in toto is without merit. But since communicating with GAIs is clearly valuable, what should we call that activity? Who better to ask (again) than the GAIs themselves? (I am sure there are better folks to ask, but since this article is about GAIs, why not ask them?) Table 3 shows the prompt I used and the recommendations of our GAI friends. Claude concluded, "The field

TABLE 2. Differences in GAI analyses.

Prompt: "I will give you 3 groups of 5 'reasons' that the term 'engineering' in 'prompt engineering' is completely inappropriate. Give me a short list of the similarities and differences among the groups."	
Similarities	**Differences**
1) *Lack of formal methodology*: All three groups mention the absence of a standardized, formal methodology in prompt engineering compared with traditional engineering disciplines.	1) *Technical design*: Group 1 specifically mentions the absence of technical design in prompt engineering, which is not explicitly mentioned in groups 2 and 3.
2) *Educational background/requirements*: Groups 1 and 2 highlight the lack of formal educational requirements or extensive training for prompt engineering.	2) *Misleading perception of expertise*: Group 2 discusses how the term "engineering" might falsely imply a level of expertise that prompt engineering does not require, which is not mentioned in groups 1 and 3.
3) *Scope and complexity*: Groups 1 and 2 discuss the narrower scope and lower technical complexity of prompt engineering compared with traditional engineering.	3) *Overhyped skillset*: Group 2 criticizes the term "prompt engineering" as overhyped, a point not raised in groups 1 and 3.
4) *Regulatory oversight/standardization*: Groups 1 and 3 mention the lack of regulatory frameworks or standardization in prompt engineering.	4) *Evolution of AI interaction*: Group 2 mentions the evolving nature of AI interaction and how it might shift away from prompt crafting, which is not discussed in groups 1 and 3.
5) *Creativity vs. technical skills*: Groups 2 and 3 emphasize the nontechnical, creative nature of prompt engineering compared with the technical skills required in traditional engineering.	5) *Quantifiable metrics and predictability*: Group 3 highlights the absence of quantifiable metrics and the limited predictability of AI responses, which are not mentioned in groups 1 and 2.

TABLE 3. Alternative labels for "prompt engineering."

Prompt: "The term 'prompt engineering' is completely inappropriate when interacting with a GAI. Give me 5 alternatives."	
GAI group	Response
Group 1: Copilot (CHAT GPT-4)	1) *Prompt crafting*: This emphasizes the creative aspect of shaping prompts for AI models. 2) *Prompt design*: A more user-friendly term that suggests intentional planning. 3) *Prompt tailoring*: We can tailor prompts to achieve specific outcomes like customizing suits. 4) *Prompt formulation*: Highlighting the thoughtful process behind constructing prompts. 5) *Prompt artistry*: Because crafting prompts is a bit like painting with words!
Group 2: Perplexity (CHAT GPT-3, mainly)	6) *AI dialogue crafting*: Conversational/iterative nature of interacting with AI, focusing on the art of constructing effective dialogues. 7) *Cognitive prompt design*: Acknowledges cognitive aspects of formulating prompts and design thinking in creating effective AI interactions. 8) *Language model interfacing*: Technical term highlights the process of effectively communicating with and leveraging language models. 9) *AI interaction sculpting*: A metaphorical term that suggests the careful/artistic process of shaping interactions with AI. 10) *AI query optimization*: Like search engine optimization, focuses on refining inputs to get relevant/useful outputs from AI systems.
Group 3: Claude. AI	11) *Query crafting*: Emphasizes formulating precise/effective questions or requests. It acknowledges that interacting with AI is more about asking the right questions than "engineering" a response. 12) *AI dialogue optimization*: This phrase recognizes the back-and-forth nature of AI interactions. It suggests a process of refining communication to achieve optimal results through conversation. 13) *Input refinement*: This term focuses on the user's role in honing their input to get better outputs. It implies an iterative improvement process rather than a one-time "engineering" task. 14) *Instruction design*: This alternative highlights the importance of clear, well-structured directions when interacting with AI. It suggests a more thoughtful approach to composing requests. 15) *Conversational guidance*: This phrase aims to steer an AI interaction toward desired outcomes through skilled communication. It emphasizes the dynamic, adaptive nature of the process.

of AI interaction is rapidly evolving, and our terminology will likely continue to adapt as our understanding and use of these systems grow." I also believe that in addition to that contention, with which I agree, as GAIs become more and more "humanlike," at least in the complexity of their interactions with us, it is unlikely that we will have to adapt complex, engineering-like rubrics to carry out our conversations with them, at least not any more complex than we already employ when we converse with other humans.

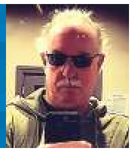
I have taken a lot of words and space above to argue that the term "prompt engineering" is ill conceived and incorporates illegitimate use of the word engineering, but I came upon someone's very pithy Reddit comment that really summarizes my own thoughts about all this. For reasons that will be obvious, I cannot give you the URL because it contains wording that I have to change slightly, to protect the innocent and to work around IEEE publishing policies. Fill in the words and Google it yourself if you do not believe me. I really like this guy! "'Prompt engineering' is easy as s–t, and anybody who tells you otherwise is a f—g clown." ■

REFERENCES

1. W. L. Elden, "What is engineering?" Apr. 11, 2005. Accessed: Aug. 25, 2024. [Online]. Available: https://ewh.ieee.org/cmte/pa/UCF/Engineering.html
2. R. J. Smith. "Engineering." Britannica. Accessed: Aug. 25, 2024. [Online]. Available: https://www.britannica.com/technology/engineering
3. "Engineering." Wikipedia. Accessed: Aug. 25, 2024. [Online]. Available: https://en.wikipedia.org/wiki/Engineering
4. J. Kirkey and S. B. Soltan. "Engineering and technology in society – Canada." BCcampus Pressbooks. Accessed: Aug. 25, 2024. [Online]. Available: https://pressbooks.bccampus.ca/engineeringinsociety
5. "How to write better ChatGPT prompts to improve conversations with AI." AI-Pro.org. Accessed: Aug. 25, 2024. [Online]. Available: https://ai-pro.org/learn-ai/tutorials/how-to-write-better-chatgpt-prompts-to-improve-conversations-with-ai
6. K. Rees. "ChatGPT vs. perplexity vs. claude: AI chatbot tools compared." How-To Geek. Accessed: Aug. 25, 2024. [Online]. Available: https://www.howtogeek.com/chatgpt-vs-perplexity-vs-claude

SOREL REISMAN is a professor emeritus at California State University, Fullerton, Fullerton, CA 92831 USA. He is a Fellow of IEEE. Contact him at sreisman@computer.org.

EDITOR MICHAEL ZYDA
University of Southern California;
zyda@mikezyda.com



Intention Is All You Need to Create Your Own Hollywood Blockbuster Movies

Revolutionizing Filmmaking With Generative AI and Saga

Russell S. A. Palmer and **Andrew M. A. Palmer**, CyberFilm

This article explores the transformative potential of generative artificial intelligence (AI) in democratizing filmmaking, focusing on Saga, an AI-powered platform designed to revolutionize every stage of the creative process, from scriptwriting to visual storyboarding and animation.

When asked why Hollywood feels stagnant, insiders often point to a lack of diverse, original stories and an overreliance on superhero blockbusters and franchises with built-in audiences.

"Hollywood's biggest problem is a lack of original content."

Attention is all that was needed to revolutionize machine learning and effectively invent generative AI, starting with transformers and large language models (LLMs). Today hundreds of millions of people use ChatGPT and Claude in their everyday work and life.

Meanwhile, technology continues to transform the motion picture industry. Generative artificial intelligence (AI), though still facing challenges like *consistency*, has advanced to produce both photo-realistic images and videos. While limitations remain—such as issues with the limited length of the videos, object action and physics, text, fingers, faces speaking, and realism—many are close to being solved.

Millions of people worldwide watch hours of video daily on platforms like TikTok, YouTube, Netflix, and Disney+. Affordable and easily accessible, video content dominates

Digital Object Identifier 10.1109/MC.2025.3534263
Date of current version: 27 March 2025

our attention, with more creators producing content than ever before. Over 3 million YouTubers earn revenue from their channels, with top creators like Mr. Beast alone generating millions yearly. Gen-Z and Gen-Alpha are particularly passionate, with *one in three preteens* naming Video Content Creator "Influencer" as their dream job,[1] and teenagers as well.[2] The youth making a career in product-unboxing, stunts, and dance routines today could be the Academy Award winners of the future, with the right tools and direction.

Using tools like OpenAI Sora, Hailuo AI MiniMax, Metaphysic, and Flawless AI, creating Hollywood-quality films on an indie microbudget is becoming a reality. Small teams can now use multimodal AI to handle nearly every aspect of production, from writing and character creation to visual effects, sound, music, and even motion-capture performances—filming scenes with nothing more than their iPhones.

While not everyone with access to cutting-edge camera technology dreams of making feature films (or would ever want to try), the *opportunity* will soon be within everyone's reach. Generative AI filmmaking tools will empower aspiring creators to produce original content in various formats, from short films and television to music videos, anime, and more. By removing traditional barriers, these technologies promise to address the industry's need for fresh, high-quality stories, allowing anyone with a vision to bring their dream projects to life.

BACKGROUND

The inception of Saga (Figure 1) emerged from a blend of academic exploration, industry experience, and familial collaboration. During the COVID-19 lockdowns of 2020–2021, I had time to take several Stanford courses online under Dr. Ronjon Nag, a renowned AI expert and entrepreneur. These courses, including *Artificial Intelligence: Deep Learning, Human-Centered AI, and Beyond*, explored cutting-edge AI techniques and their practical applications.

As part of the curriculum, I wrote a paper analyzing how AI and machine learning could disrupt the Hollywood film industry—a topic inspired by my brother Andrew's extensive experience on Hollywood and indie movie and TV sets, in roles from *Production Assistant* to *Assistant Director* and *Producer*.[3] The premise was how AI could serve as a powerful *tool* for filmmakers instead of a replacement, a unique perspective in 2021 but now a commonly heard phrase in the AI art community. Discovering firsthand insights from Andrew and the nearly hundred colleagues we interviewed as part of our research, we focused on the industry's challenges that uncovered opportunities for AI-driven transformation. This formed the foundation for Saga's vision, and with Dr. Nag's encouragement we founded our startup in November 2021 among the first AI Media Tooling startups like Runway, Metaphysic, and OpenAI.

> "A+ grade. I really liked this vision on how one would use AI in the cinema industry. A really original commentary and great vision."—Dr. Ronjon Nag (Stanford)

Andrew's journey is equally remarkable. After earning a degree in

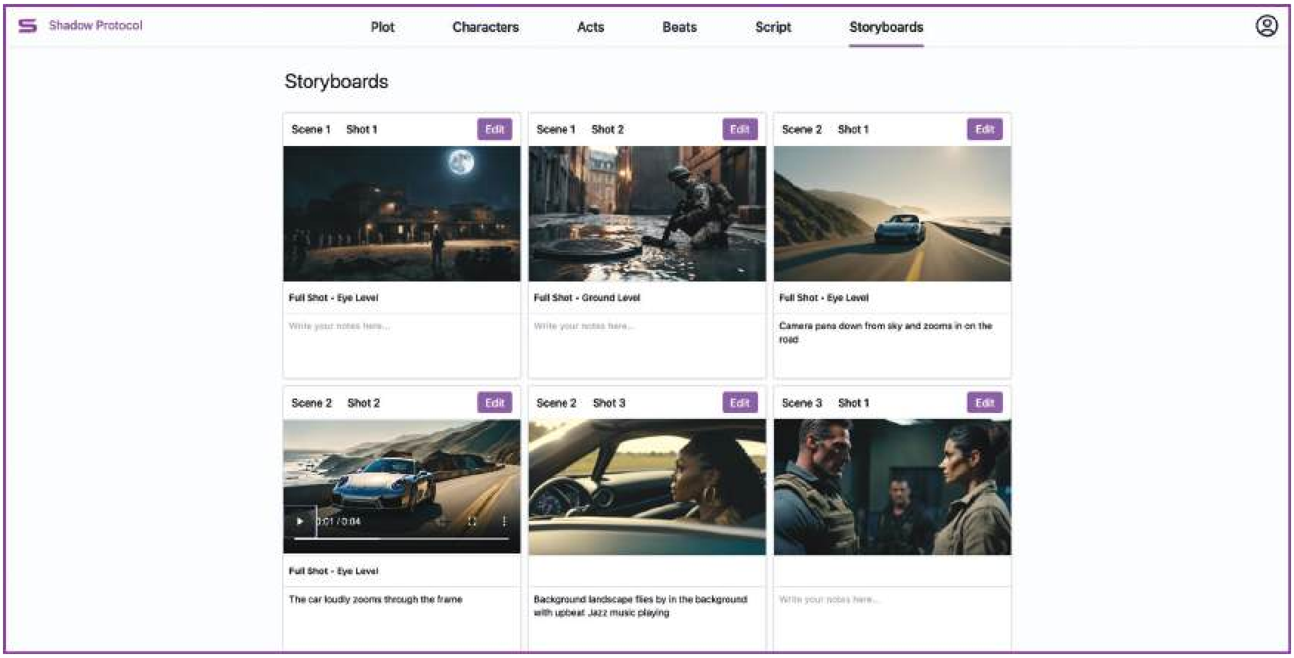


FIGURE 1. Storyboard page on the Saga app.

Computer Engineering, he pursued courses in film school, producing, and creative writing to deepen his understanding of cinematic storytelling and digital visual effects processes. This unique combination of technical expertise and creative vision paved the way for his role as cofounder and Chief Story Officer of Saga, and our founding vision.[4] Together, our experiences—my 15 years as an electrical engineer and computer scientist and AI product manager in Silicon Valley (with roles at Microsoft, Viv Labs AI, Samsung, and the JPMorgan Chase *AI Lab*), and Andrew's extensive on-set experience (shows like *The Boys* and *Suits*) and storytelling expertise—formed the perfect foundation for creating an AI-driven filmmaking platform which we've been evangelizing for years at conferences around the world.[5]

The convergence of these experiences, combined with rapid advancements in generative AI, highlighted the immense potential for AI to democratize and transform filmmaking. This realization sparked the creation of Saga—an application designed to empower creators by streamlining and enhancing every aspect of film production through AI.

The following sections explore Saga's architecture in detail, including its current technical stack, future development plans, and the ethical considerations of integrating AI into Hollywood's creative processes.

METHODOLOGY

Developing Saga required a multidisciplinary approach, combining expertise in artificial intelligence, software engineering, and filmmaking. Our methodology focused on creating a scalable, user-centric platform that leverages advanced AI models while ensuring that the creative process remains intuitive and filmmaker-friendly.

Approach

Human-centered design:

- The platform was designed with filmmakers in mind, ensuring that the tools align with real-world production workflows. Feedback from professional screenwriters, directors, and production teams played a key role in shaping the user interface and feature set.
- Cofounder and Chief Story Officer of Saga, Andrew Palmer, leveraged his film industry experience to bridge the gap between AI capabilities and creative storytelling. Through language and image model prompt engineering and fine-tuning, Saga functions like a filmmaker, using a structured, opinionated film school framework to address key storytelling challenges—such as crafting a B-story that reinforces the theme with secondary characters. Unlike traditional chatbots, which provide less coherent responses by drawing from several of the sometimes-incompatible frameworks it all knows, Saga's approach ensures more focused, consistent, and coherent results.

Iterative development:

- The product was built using agile methodologies, allowing us to incrementally add features, gather feedback from nearly 100 filmmakers, and refine functionality in response to real-world usage.
- Beta testing with dozens of early adopters from the filmmaking community helped validate our core features, including AI-assisted script generation and visual storyboarding.

Focus on scalability and accessibility:

- The architecture was designed to support filmmakers at every level, from film school students to large production studios, ensuring that the tools scale with project complexity.

Technical architecture of Saga

Saga is an AI-powered platform designed to revolutionize the filmmaking process by integrating advanced generative AI models into various stages of film development, including screenwriting, storyboarding, and previsualization animation (previz). Our platform leverages a combination of proprietary algorithms and prompts, with state-of-the-art LLMs and diffusion image and video models to assist filmmakers in crafting compelling narratives and visual content.

High-level system architecture:

1. User Interface (UI):
   a. A web-based application that provides an intuitive interface for users to input their ideas, develop scripts, create storyboards, and generate previz and animatic videos.
2. Application Layer:
   a. *Frontend*: Developed using modern web technologies to ensure responsiveness and seamless user experience.
   b. *Backend*: Implements business logic, manages user sessions, and handles requests between the frontend and the AI services.
   c. *Companion App*: Upcoming iOS iPad app for use on set to show the storyboard previz animations and animatic videos.
3. AI Services Layer:
   a. *Text Generation*: Utilizes LLMs such as OpenAI's GPT-4, Anthropic's Claude 3.5, and open-source models like Meta's Llama to generate and refine screenwriting content. We use different models for different tasks, depending on which is best suited for each task, their cost to run, and to provide users more variety of options for each idea generated inside Saga.
   b. *Image Generation*: Employs generative AI models to create visual content for storyboards, including OpenAI's DALL-E 3 and Stability AI's Stable Diffusion XL. We use multiple

models to provide our users with a varied selection of options, with each model using its own tailored prompt for best results with that model, adding inputs from the user around character description, style, camera level, size, shot type, and more.

c. *Video Generation*: Generates text-and-or-image-to-video using models including Luma Labs AI Dream Machine, for 5-s photo-realistic storyboard pre-viz and soon 40–60-s animatic video clips launching in 2025.

4. Data Storage:
   a. Secure databases store user inputs, generated content, and project metadata, ensuring data integrity and privacy. Our users and their movie projects are completely private, and our users don't even need to mention they used our app.

Workflow overview:

1. Idea Input: Users enter their initial concepts, character details, and plot outlines through the UI.
2. Script Development:
   a. The application processes user inputs and, through the AI services layer, generates script suggestions, dialogue options, scene descriptions, or whatever the user asks for.
   b. Users can iteratively refine the script with AI-assisted recommendations, or by inputting director notes in natural language for fast rewrite options.
3. Storyboarding:
   a. Based on the developed script, the AI models generate visual representations of scenes, assisting in the creation of detailed storyboards, with the user in control selecting from lists of common shot types and camera levels.
4. Pre-visualization and animatics:
   a. The platform offers previz features, allowing users to animate scenes and visualize camera movements, enhancing the planning of complex shots.

TABLE 1. Language models we experimented with.

	Version	Notes
OpenAI GPT	4	
Anthropic Claude	Sonnet 3.5	15+ Languages
Mistral	Large 2	Open source, 12+ Languages
Google Gemini	1.5	100+ Languages
Meta Llama	3.3	Open source, 30 Languages
Hugging Face BLOOM	176B	Open source, 46 Languages
Stability AI Stable LM	2	Open source, 7 Languages

TABLE 2. Image models we experimented with.

	Version	Notes
OpenAI DALL-E	3	
Stability AI Stable Diffusion	XL	Open source
Black Forest Labs FLUX	FLUX.1 [pro]	Open source [dev]
Luma	Photon	
Leonardo	Flux	
Google Gemini	Imagen 2	

TABLE 3. Sound models we experimented with.

	Version	Notes
ElevenLabs	Multilingual v2	Voice in 30+ Languages
Suno	v4	Music

AI large language models. Language models we experimented with, and in some cases are integrating now (or soon) with Saga (Table 1).

AI image generator diffusion models. Image models we experimented with, and in some cases are integrating now (or soon) with Saga (Table 2).

AI sound generator models. Sound models we experimented with, and in some cases are integrating now (or soon) with Saga (Table 3).

AI video generator diffusion models. Video models we experimented with, and in some cases are integrating now (or soon) with Saga. (Table 4).

AI model customization. We continue to use cutting-edge methods to derive the highest-quality results from our multimodal AI generation for filmmakers. This includes trying new scientific approaches, including the following options:

Fine-tuning

We tested fine-tuning of GPT-3 in 2021 on OpenAI's then-new Beta feature. We hand-crafted over 50 data files consisting of top movie synopses and character sheets, effectively filling out Saga's UI structure with existing quality movie plots to put the outputs on rails to fit our app structure. However, it seemed to tilt the model in a worse direction. We would likely have needed tens-of-thousands of such files, with our only option being to generate "synthetic data" movie synopses, and having to personally review each of the, for example, 50,000 movie synopses individually for quality. We also weren't sure if the generative pretrained transformer (GPT) training set already *included* the script and synopsis for every movie publicly available. When GPT-3.5 launched, this seemed to confirm that answer and remove the need for fine-tuning. The high cost of retraining models to fine-tune them does not seem to be worth the price, especially for early-stage startups, given that running fine-tuned models is orders of magnitude more expensive than running the base models which work fine for our use case.

In the future, we want to experiment more with fine-tuning, ideally using datasets of movie scripts (acquired legally, through purchase where necessary) and to create our own rating scale based on combinations of online scores, award nominations and box office success, and other factors. We would try to over-weight the neural network for the best movies, as opposed to, for example, OpenAI's GPT training set which probably includes every script ever published (both the good and bad ones), knowing in the world there are fewer great scripts than bad ones. We believe this is why ChatGPT movie ideas are often cliché and middle-of-the-road in terms of quality because they are the average of all movies written—the lowest common denominator for ideas you could simply get from an amatuer writer at a coffee shop in Silver Lake. While we build on GPT, our goal is to customize it for our vertical use case to improve on its generated results.

Retrieval-augmented generation

We have experimented with retrieval-augmented generation (RAG) especially for our Saga GPT,[6] a first in the ChatGPT Store, which contains RAG documents of Saga's lists of definitions like our opinionated *Character Arcs, Archetypes, Story Types, Beats*, and more.

We don't prefer RAG currently because it doesn't *retrain* the base AI GPT model, it simply gives it a way to pick out existing answers when people ask directly for something contained in one of the documents. As noted previously, building a custom RAG system to include a few documents of ours does not seem to be cost efficient for us at the moment. It *could* be for a well-resourced studio like A24 or Lionsgate [who is working with Runway to train a custom AI model on their proprietary intellectual property (IP)] to build their own custom fine-tuned and RAG AI applications since they have countless quality scripts and video files with more resources.

Prompt engineering

A lot of our success in putting generative AI models on rails has come through prompt engineering. This is considered our own IP [as no one can see our private custom application programming interface (API) calls]. Our business strategy is in constantly refining the model prompts we use across app use cases, as we're constantly testing and verifying which prompts get the best results. We save our users from the hassle of mastering prompt engineering (a role we don't even see existing in the future as chatbots and LLMs get better at understanding what users are looking for). Using Saga is simple, and there is no need to figure out, for example, Discord slash commands to use Midjourney, or learn ComfyUI and LoRAs or other cutting-edge workflows that might not even be relevant in the coming years as the user apps progress. On Saga you can use industry terminology and our simple UI to get what you need, as if you were leaning over the shoulder of an artist at their easel describing what you are looking for.

Multimodel approach

We've learned over the years that, at any given time, certain AI models will be better than others at specific tasks, and that this leaderboard can change from week to week. For example, when we were researching Anthropic Claude, we discovered that they had fostered a large creative writing community, with hundreds of fiction authors and novelists on Discord and other online forums. They would profess that Claude could easily beat GPT at creative writing, until with one version release it couldn't and there was outrage in their writing community. Every day, we read blogs comparing use cases, and reviewing the top models, to make sure we have the best in Saga so our users don't need to keep up on the latest model versions and can trust us to provide them with the best always.

Every quarter, ChatGPT achieves new benchmarks for testing on mathematics, medicine, and law. Our business imperative is to be at the forefront

TABLE 4. Video models we experimented with.

	Version	Notes
Luma	Dream Machine	
Runway	Gen-3	
Google	Veo 2	
Stability AI	SVD	Open source
Hailuo AI	MiniMax	
Kling AI	1.6	

of testing and integrating the top models, and one of our unique design decisions is that *we don't make the user pick which models to use*. Saga will generate multiple options across models for anything they want, be it a storyboard image or a character's name, and we let the user decide which is best across all integrated models. By collecting the data of model preference across user selections, we can learn which models are preferred and continue to optimize and offer the best experience to our users, dropping models that aren't among the top performers and swapping in new ones.

Character consistency

In Saga, when generating storyboards, our users can simply select shot types and camera levels, the size and style they are going for, and even reference characters by name, to which references their likeness from our character page when drawing each new shot. This saves them from having to redescribe each character in every shot prompt, which can number in the thousands for dozens of scenes across a 90-min feature film. We do this from prompt-injecting the character's *physical description* based on fuzzy name matching. A new feature in our Beta launching soon will allow users to upload (or generate) images of the character, including portrait headshots and images wearing all of their main costumes and outfits. We're adding the same for voices, so you can select a character voice to use in our *virtual table reads* and *rehearsal partner* features launching this year. The same will go for set locations, allowing location scouts and artistic directors to upload (or generate) reference images of scene locations, props, and more that reappear throughout the production and name them when generating storyboards.

This is currently limited through model APIs that only allow upload of one image to edit, but we're working with leading AI companies to increase the number of images allowed to upload in an API call (such as the previous shots in a scene) to enhance consistency. If this feature does not become available from them, we plan to build it using open source image generators which allow us to write custom code, in this case allowing for multiple reference images to provide as context when drawing new storyboard images.

RESULTS

In developing the first generative AI application for screenwriters and filmmakers almost four years ago and following the industry closely since the advent of machine learning, we find that despite the rhetoric on social media and in the news, filmmakers tell us that they'll use any tool in their arsenal to succeed, and that they find these AI models for language, image, sound, and video to be extraordinary tools that help them make more of their best work faster. Due to the negotiations between guilds and studios, most choose to remain anonymous in their use of the tools for fear of retaliation and retribution, but this public quote from Paul Schrader (screenwriter of Scorsese's classics *Taxi Driver* and *Raging Bull*) illustrates the point:

> "I'm stunned. Every idea AI came up with (in a few seconds) was good. And original. And fleshed out. Why should writers sit around for months searching for a good idea when AI can provide one in seconds?" *(Deadline.com—19 January 2025)*

Next, we provide qualitative feedback from real Saga users, quotes from our mentors including the cofounder of Netflix and producer of *Breaking Bad*, and an (unscientific) case study showing how fast a screenwriter can use Saga to complete their best work, with evidence of a portion of the written script and a behind-the-scenes video livestreaming the process.

We find that our app can help with what creatives need most: the story and character ideation, planning the structure and beat sheet to build on, and finishing their first rough draft. This makes them more creative, more efficient, and more successful as human filmmakers.

We've heard through nearly 100 public and private interviews that nearly *everyone* in Hollywood and the film industry is testing AI, and that if people *can* use it they will. Hollywood is a lucrative and competitive industry, and especially those "under-the-line" have told us they will use whatever tools are available to help them achieve their dream because the current system is broken and no longer a meritocracy. They prefer using technology to their advantage like self-publishing on YouTube to grow their fanbase and get discovered, as opposed to endless networking and even sometimes suffering abuse at the hands of Los Angeles power players and kingmakers. This is the democratization of filmmaking—not a tool we should try to ban but to use selectively to help produce Hollywood blockbuster-quality films affordably, to tell the world new diverse and original stories, as to paraphrase Scorsese, *everyone has one to tell*.

Saga users are spending over 30 min per session, which can be more than new writers spend typing on a word processor when they first force themselves to sit down and "just write," all while staring at a blank page. Not everyone who uses AI to create a movie will be successful, or be able to create an award-winning work of art. Most of the "AI films" we see today are pointless, lacking in story or characters, are poor quality, and tend to rip off existing franchises or make fake movie trailers for clout. When we premiere our own AI-assisted films, we may decide not to mention that AI tools were even used, to avoid any knee-jerk reactions and partisan reviews, although we believe it's fine for award ceremonies to have separate categories between human-only and AI-assisted works—at least for the time being and to give everyone a fair shot before the industry transforms completely. We also write about this topic on our blog, in the

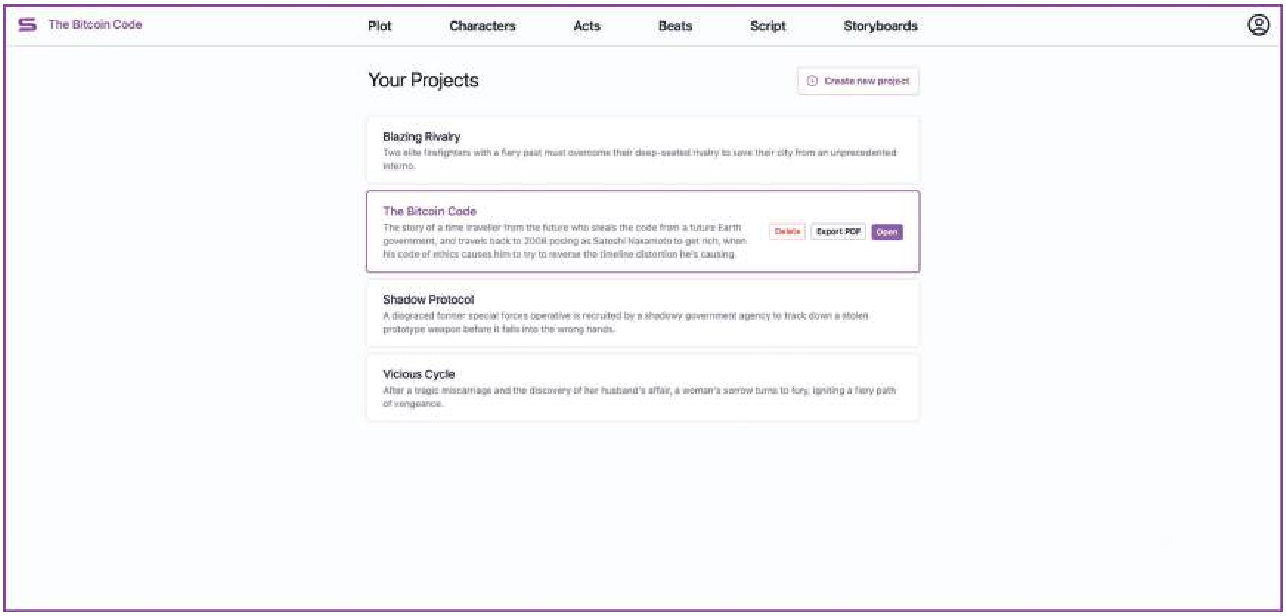


FIGURE 2. Saga Web app home page in the browser.

article The Content Turing Test.[8] We believe movies should succeed on their own merit, not because AI was used as a gimmick to impress people with the state of the art.

Our goal is to match the quality of top Hollywood professionals, including screenwriters and directors. We achieve this by putting the AI models on rails through expert prompt engineering, ensuring that it adheres to best practices used in the top blockbuster US$250 million movies.

Feature demo

When you open Saga, you begin on the Projects page (Figure 2), where you can select which of your existing movies to work on, or create a new project.

The pages of the app are ordered from left to right, traversing the typical ideation steps incubating a film idea. This structure is based on best practices and previously done using analog cue cards on cork or whiteboards. The first page helps ideate Plot (with *Logline*, *Theme*, and *B-Story*) see Figures 3 and 4, then Characters (Figure 5), then Acts and a Beat sheet.

Saga uses GPT-4 and will continue adding new models such as Mistral, helping add variety to the generated text and allowing expanded content ratings (up to R-rated) and additional languages. This enables support for new markets including Japanese and Hindi-speakers (that is, for Anime and Bollywood films), to expand our user base to markets worldwide.

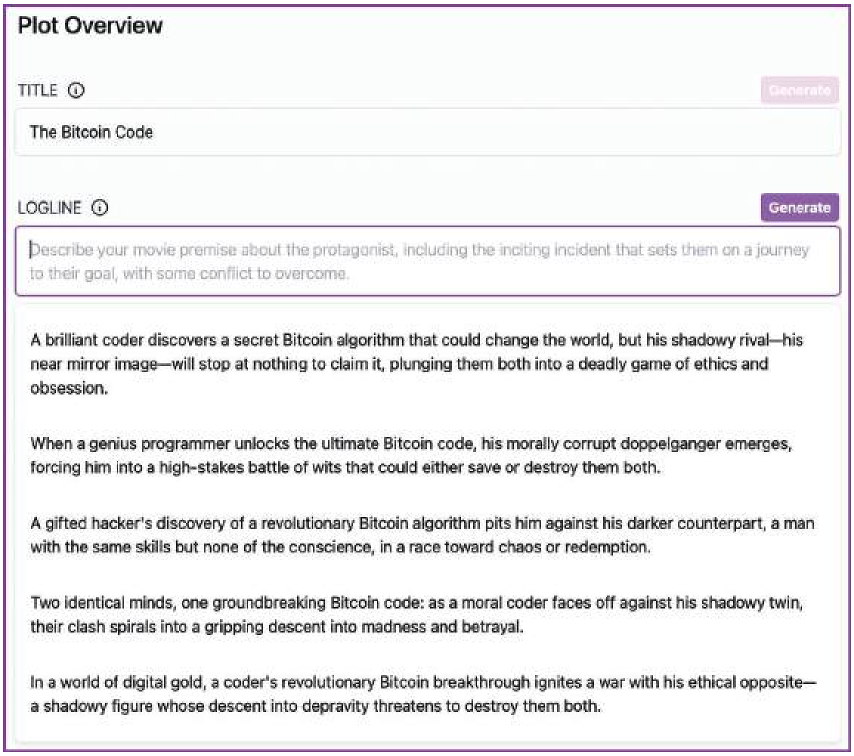


FIGURE 3. Saga generating five ideas for a movie Logline based on the Title.

This includes providing a wide range of archetypes to address one of the current shortcomings with our competitor apps – which seem limited to one dimensional characters, such as the typical villain. We offer extensive lists of antagonist and villain archetype traits, allowing for deeper, more complex characters—even secondary ones—making their character development fast and worth the time investment to improve the depth of your story.

Additionally, we make it easy to combine different story types, drawing from patterns that have been used since the Ancient Greeks and beyond. You can select from examples to create a pitch like "Titanic meets Jaws" and teach the AI what you're going for, a common way professionals pitch ideas in Hollywood and around the world.

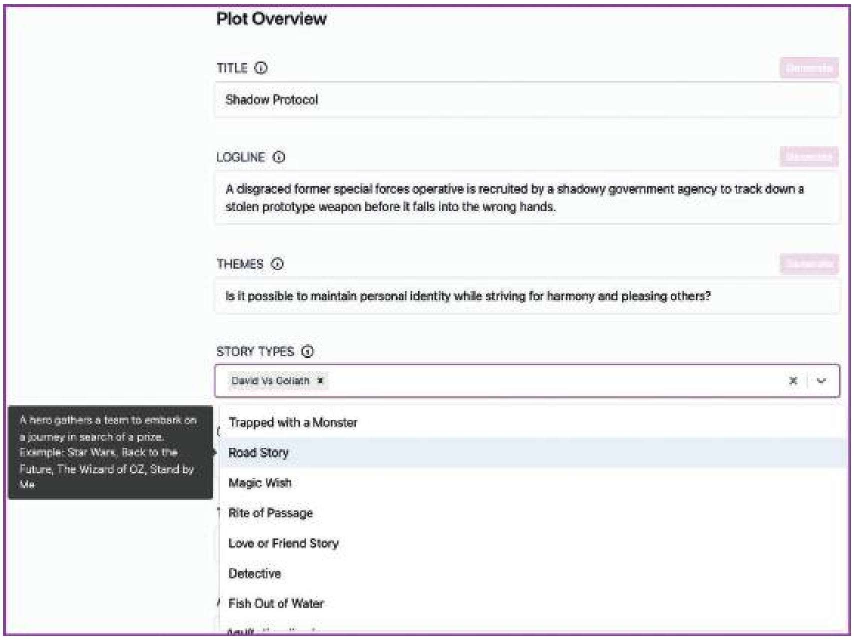


FIGURE 4. Selecting from common story types with examples.

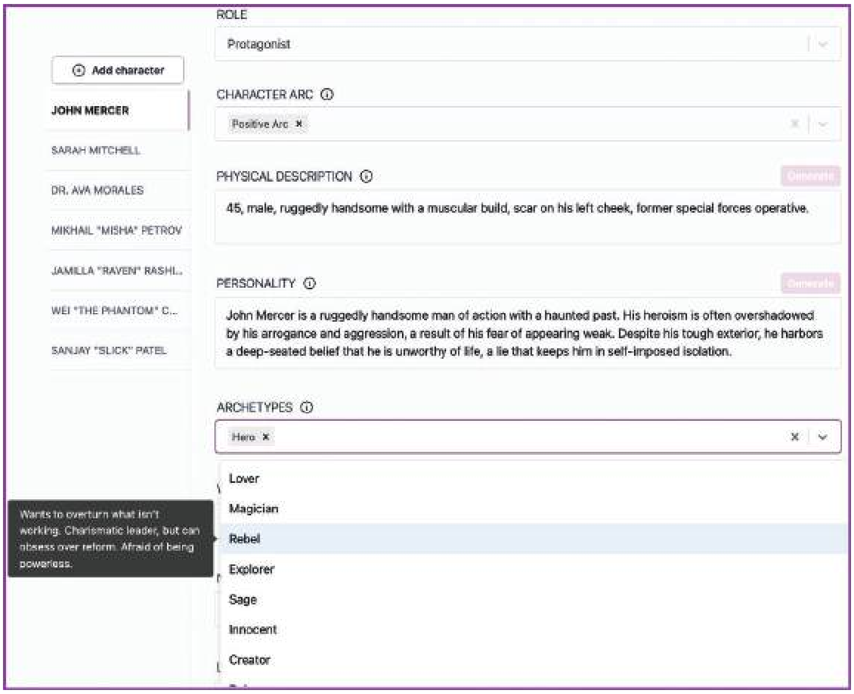


FIGURE 5. Selecting from common character archetypes with examples.

When you start writing or upload a script, we automatically classify all of the scenes for easy and quick access (Figure 6).

Our state-of-the-art Script Editor is like Final Draft but with enhanced features, offering all of the styles and hotkeys screenwriters are used to. Scripts can be uploaded in various text formats, including Final Draft (.FDX) and .Fountain. You can select existing scenes and ask Saga to help rewrite them based on natural language instructions, such as "make this longer and funnier" (Figure 7).

Select any line in your script and click "Rewrite" to enter your feedback or notes from others (Figures 8 and 9).

You cannot generate a 100-page screenplay with a single click in Saga, but based on our research, the vast majority of writers don't *want* their entire script written *for* them. Instead, our app generates partial scenes one at a time, giving you the option to accept ideas, edit them, and make the writing truly your own (Figure 10).

If you're stuck with writer's block at a blank line or page, simply click "Generate" to get a boost from Saga and keep moving forward with fresh text and ideas that you can easily edit. "Just write" is common advice in the industry, and we see AI as a tool to keep you pushing forward, filling more pages with your creativity. Our goal is for you to love and own your movie ideas, using Saga to accelerate your creative process with ideas and first drafts that you can refine and curate.

Here's a 3 min demo video[7] showing how a typical user gets ideas for a movie on Saga: https://www.youtube.com/watch?v=iQ6JhAqeU-g (Figure 11).

Our storyboarding feature was built on the first DALL-E API back in November 2022 and now integrates

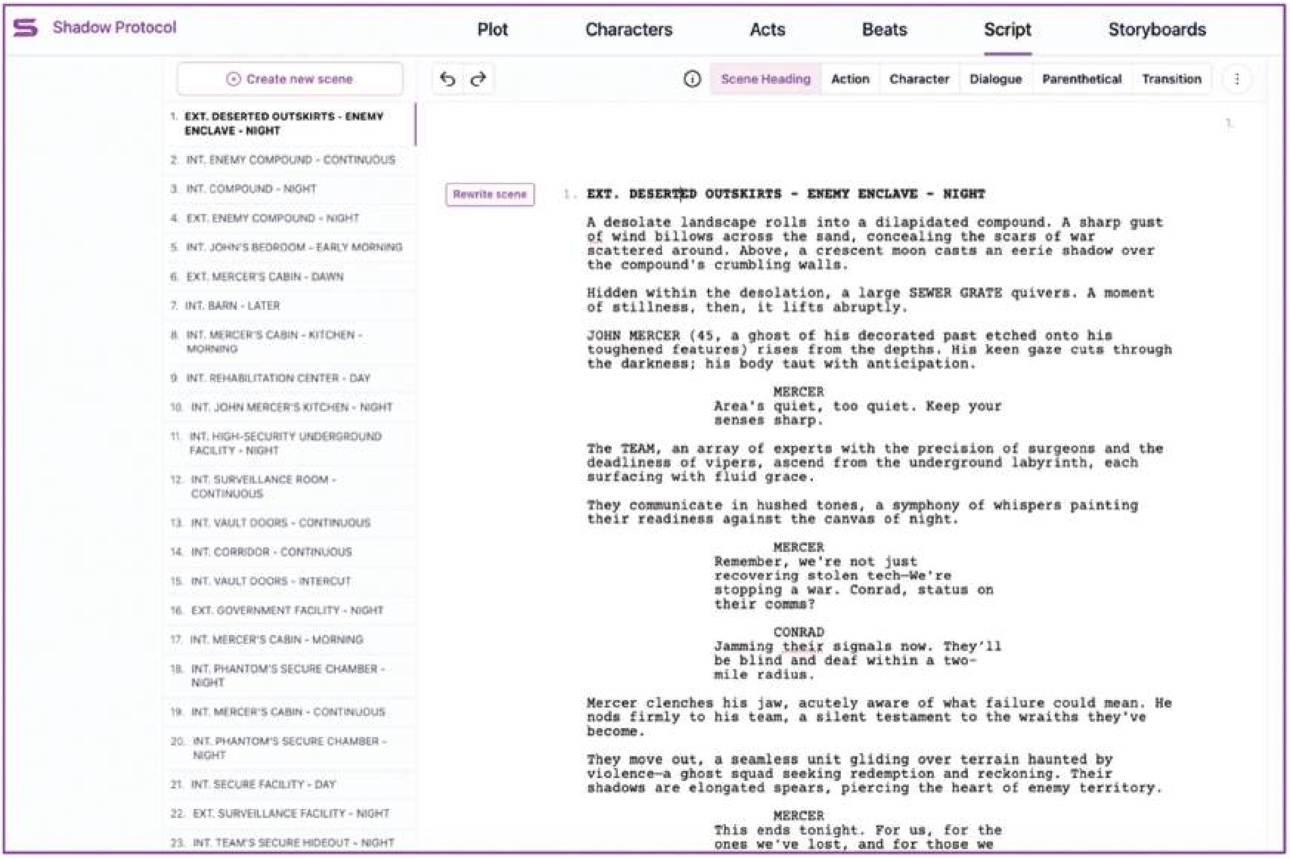


FIGURE 6. The Saga Script editor.

other image models, including Stable Diffusion XL (Figure 12). We continually add new models, such as the cutting-edge FLUX.1, and soon, Sora from OpenAI.

This provides our users multiple options in one place, eliminating the need to personally test each image generator's latest model—something creatives typically want to avoid. We keep Saga updated with the latest and most advanced image and video models (Figures 13 and 14), all fine-tuned for cinema. It's one subscription, with all of the best models, in one app. The following link is a video[9] showing the car animation from Figure 13 (7 s): https://bit.ly/sagavideo.

Our helpful prompt engineering means our users can simply *ask* for what they want using industry terminology they are used to, with helpful teaching guides so anyone can learn and apply new skills like cinematography (Figure 15). Writers enjoy illustrating the scenes as imagined in their head when writing. Saga's animation feature will soon expand to longer previz animation clips, adding sound and voice, with music scoring, animatics, and finally photo-realistic computer-generated imagery scenes.

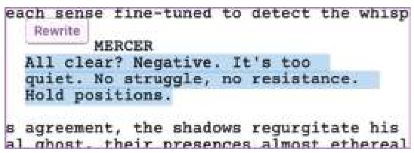


FIGURE 8. Selecting a line to rewrite.

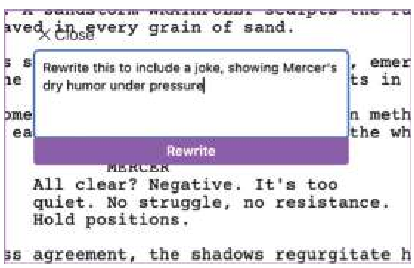


FIGURE 9. Inputting notes from the director for quick rewrites powered by AI with the full movie script and context.

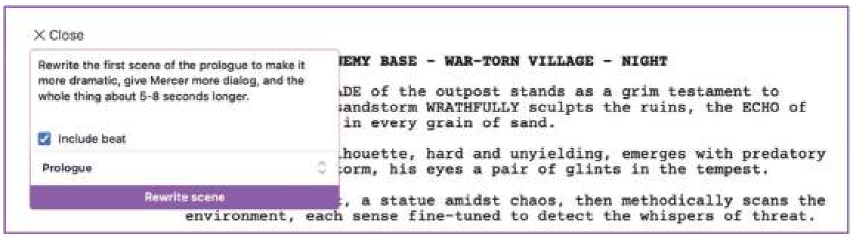


FIGURE 7. Using natural language instruction prompts to ask Saga to rewrite a scene.

Audio

Our Beta users are already enjoying voice features like Virtual Table Reads, where they can hear the characters bring the script to life, helping them feel the dynamic of exchange and pacing of the material. Music will be introduced later, with new models as available from Suno, Udio, and others, allowing filmmakers to score their movies with an original soundtrack that perfectly matches the emotion and tempo of each scene.

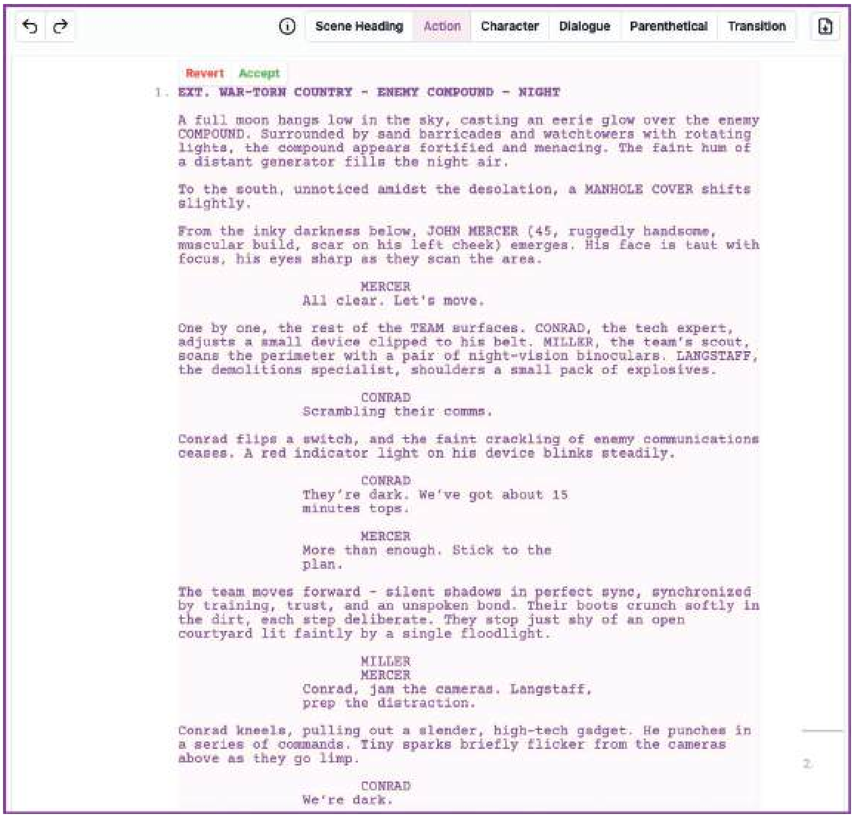


FIGURE 10. A Script page generated by Saga which users can accept and edit.

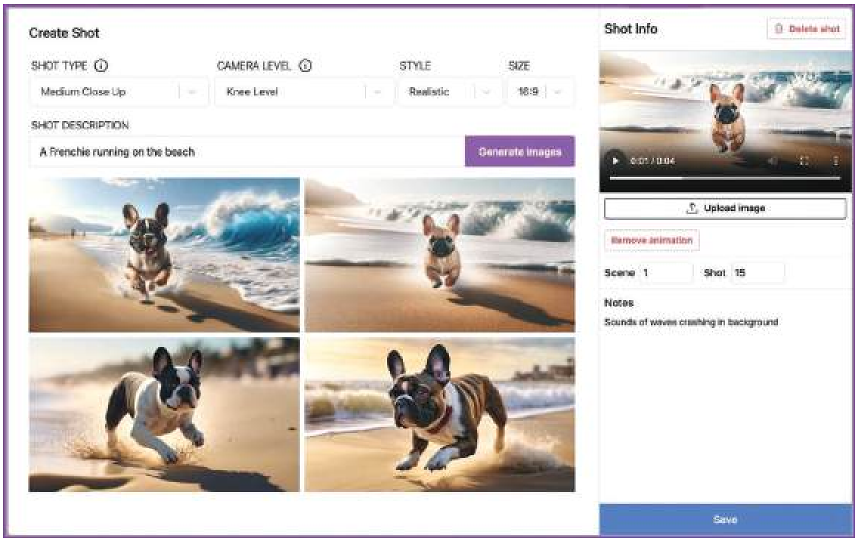


FIGURE 11. Sage 3-min demo video on YouTube cover art.

CUSTOMER TESTIMONIALS

"The app is useful overall and the Logline component and Script tab was quite helpful. The UI was nice to have everything in a centralized place. Interesting suggestions in the Acts that sparked ideas."—Kenny Geiler (indie filmmaker)

"The storyboarding feature is amazing. I loved playing around with it. With a little time spent on prompt engineering/tweaking it captured my vision with ease. It almost distracted me from the writing."—Jared Levine (BS Cinematic Arts at USC)

"First, thanks a lot for Saga, it's an amazing tool. I love it, really. The language, the dialogues, the actions are more reliable than ChatGPT."—Vincent T. (writer)

"The script writer works very well taking Beats and generating scenes, I definitely feel like I could spend hours working on a script now, and if I have writer's block, use to come up with ideas."

"I gotta say, I really like the way it sparks those outside-the-box ideas. In terms of storylining, it's really helpful."

"Saga is next level thank you so much! Honestly what you guys are doing is beyond comprehension. Understanding your background and how it was all set up gives more clarity. I managed to create from idea to concept and process it into a finished product with advanced concepts and good consistency through the generations, creative and amazing outputs. And this was a blitz test. I have been testing as much as I can with various platforms… this is the most impressive thing

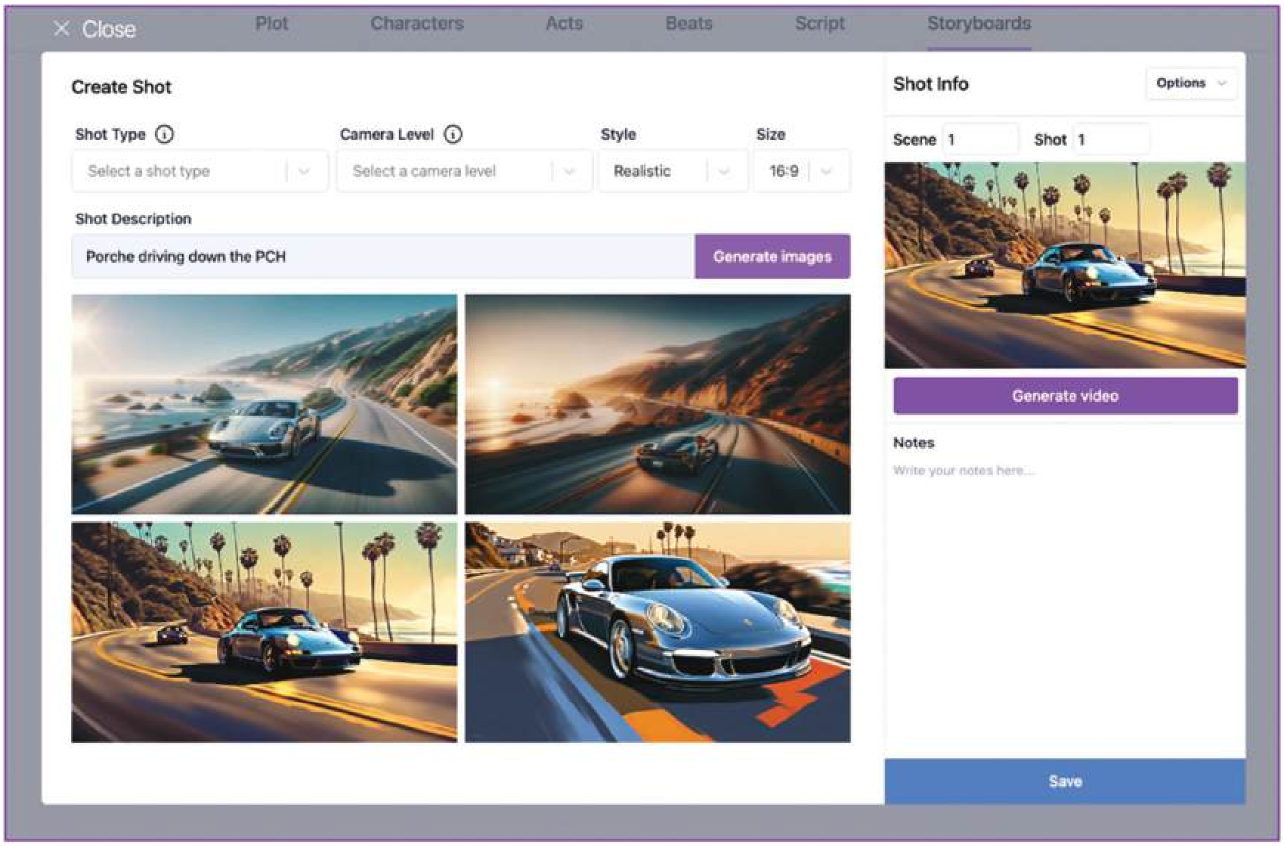


FIGURE 12. Creating a new shot on the Storyboard page.

I have come across FULL STOP. The time and quality of the generations along with the ease of bringing an idea together are seamless… getting access to Saga Cyber Film AI is a game changer!! So much going on this is a beautiful time to be alive."

"I love all the different variables and options that are available to users, and the UI/UX look fantastic, minimal with a leading design, helping the users continue to navigate their story."

"So far Saga is great. I'm fairly new to screenwriting so it's been a huge help."

"I love the options and variety Saga generates that sometimes other AI generated kinda rephrase the words mashed up and I wanted better action verbs or better character development. Also, love hovering over the text to be reminded of certain definitions is nice."

Mentor feedback

"Love it. It asks the questions you need the answers for."—Stew

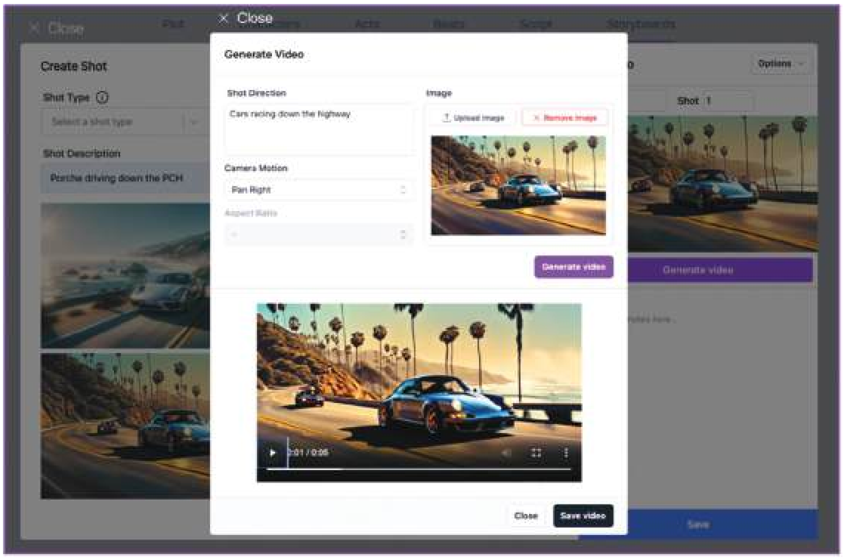


FIGURE 13. Creating previz animation for a shot on the storyboard page.

Lyons (Producer of Breaking Bad, Better Call Saul, Everybody Loves Raymond)

"I tested out your amazing SAGA program and was really impressed with how it works—pretty amazing tool for story development for filmmakers."—Tim Peternel (American Psycho, Buffalo '66)

"The app helped me put together a better story than I could have done alone. It helped me play the what–if game in ways I couldn't have imagined."—Rhys Ryan (Cocreator of Scenechronize)

"This is really intriguing, because you are correct that almost all storytelling (movies, books, tv commercials, plays, speeches) uses a similar structure that has evolved over thousands of years. I would love to play with it. I can't help it. This is just such a cool use case for AI."—Marc Randolph (Cofounder CEO of Netflix)

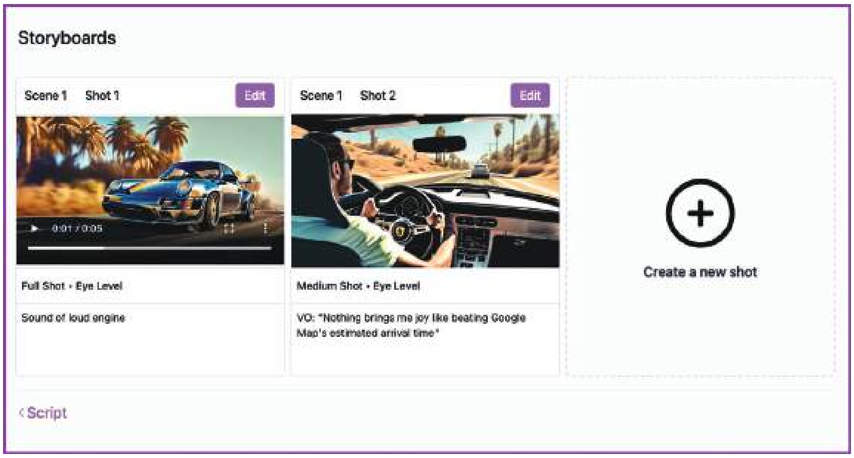


FIGURE 14. Storyboards that come to life with previz panes inline.

Discussion

Interpreting the results of our metrics, our interviews and usability studies with filmmakers, and our inbound feedback, people love what we're building. They request countless new features every week, such as

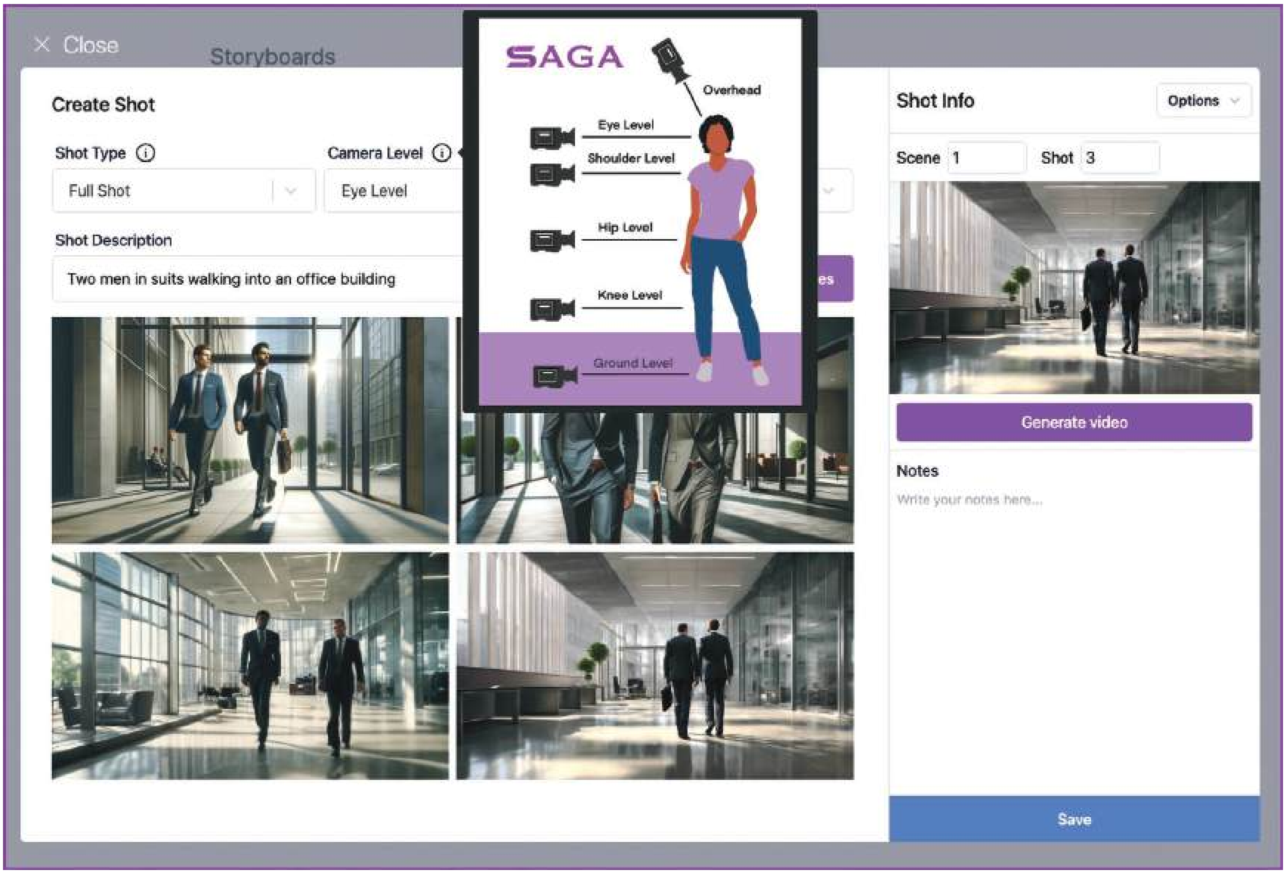


FIGURE 15. Simple teaching guides to explain the various Shot Types and Camera Levels used in cinema.

new templates for writing TV series or commercial advertisements, new languages, and of course longer and more consistent video generation.

Unlike competitor apps that launched soon after Saga resembling our design and roadmap, we have the advantage of years of customer interviews, app Beta tests, and real relationships with Los Angeles-based A-List creators. Our roadmap comes from requests by real filmmakers like them, who want upgrades such as allowing R-rated content (which we're adding through Mistral Mixtral and open source LLMs where we can set our own content filters).

We're launching a custom Chatbot that users can brainstorm ideas with, outside the structure of our app pages. We're adding new styles and formats for Anime, Manga, and other graphic novels, interactive virtual reality storytelling and game experiences, and are even discussing our own streaming service which would take less of a cut from our creators to deliver them more profit. AI translation and lip sync will allow anyone to sell their movie to 8 billion people. Virtual performances will save impressionable young actors from filming scenes of rape and violence, which studies show impact their mental health. Stunts will be safer, as Hollywood has a long history of the untimely demise of stunt performers on set. Actors can play younger or older versions of themselves, even interacting with their clones on screen. Older actors who pass away mid-production can have their final work finished on their behalf and with their permission, and the legacy of voices like James Earl Jones as Darth Vader can live on for generations to come (as he had granted permission before his passing). The 2023 strikes made it clear that artists don't trust—and in some cases have personal animosity—toward studio executives. As artists and filmmakers ourselves, we don't want to see AI abused by executives either, or any jobs lost. As Gandhi said "be the change you want to see in the world," and we're hoping through our leadership that Saga and AI can have a positive impact on the film industry, with new movies and forms of entertainment that set new standards for quality, originality, and diversity—rather than degrade the art form to bland generic "slop" as some fear.

Case Study

Our cofounder Andrew is a skilled writer, having published multiple original novels on Amazon and authored several screenplays. As a case study, he transformed a colleague's idea for a screenplay premise into a completed 100-page script in just 10 days—writing only part-time—using Saga. This demonstrates the remarkable efficiency and creativity that Saga enables for writers and filmmakers.

> "I put some of the beats in the Generate tool, and the stuff it was coming up with was great. It's much faster and to the point, plus sometimes it comes up with unique ideas. With AI it has a cool way of growing into an organic story. The great thing is that usually I would have to go through a script a few times before I did polish pieces, but now I'm kind of polishing it when I'm writing it because of the speed with which I can generate new pieces, so I find SAGA really efficient with writing."—Andrew Palmer

He typically takes two or more years to complete a typical 100-page screenplay for a 90-min feature film. Using Saga is over 70 times faster, while achieving the same- *or better*-quality writing (Table 5).

You can see a livestream recording of Andrew writing a dozen pages in this video,[10] with a selection of pages included for download in the description: https://bit.ly/sagawriting (YouTube—12:16 min) (Figure 16).

You can see more case study videos from others on our YouTube channel @writeonsaga.

ETHICAL CONSIDERATIONS

Machine learning is one of the most transformative inventions of the century. Like previous groundbreaking technologies—such as electricity and atomics—it requires careful attention and regulation. Unlike these past discoveries,

TABLE 5. Using Saga to speed up writing.

	Pages Written	Time
Before Saga	100	2 years
Using Saga	100	10 days

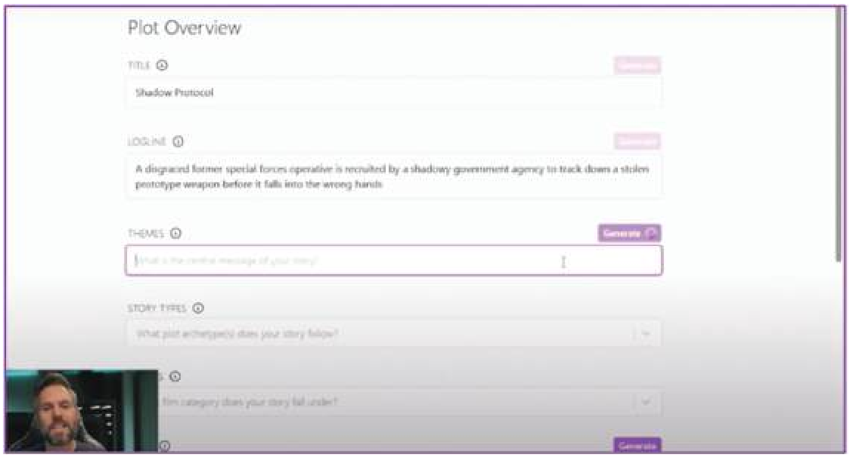


FIGURE 16. YouTube video showing how to write a feature film in Saga.

however, AI introduces a new challenge: the potential for independent goals and intentions, should it ever achieve the ability to think autonomously.

While much has been written about AI ethics, including issues like representation in datasets, we highlight specific challenges within the entertainment industry. For example, when using datasets such as Academy Award nominations, how do we address the underrepresentation and biases evident in earlier decades of the 1920s and 1930s? These historical biases could influence AI-generated outcomes if not critically examined and corrected.

Ethically sourced datasets

The most popular image generators today often rely on datasets like LAION-5b, a vast repository of images sourced from the Internet. While the World Wide Web is intended for image sharing, many creators showcase their work online (such as photography) with the intent of generating income through paid downloads. Watermarks are commonly used to explicitly indicate ownership and a right to copy these images, yet these datasets can include watermarked images, typically under the premise of research—a fair use case.

Fair use is generally accepted for nonprofit research purposes. However, the commercial applications of these training datasets by for-profit companies raise important legal and ethical questions. Should copyright laws and courts reassess their stance on the use of such content in for-profit AI training? Copyright laws, many of which were written long before the advent of modern technology, may need to be updated to better balance protecting intellectual property and encouraging the creation of new content in the 21st century.

Consent, control, and compensation

The Hollywood strikes in 2023 highlighted three essential demands from artists: consent, control, and compensation. We addressed this in our blog post Artists Rights—Getting Paid for your Work in an Age of AI,[11] and proposed a partial solution.

This solution involves creating an opt-in, dividend-paying dataset used to train AI models of for-profit apps—akin to "ethically sourced coffee beans," where contributors are treated and compensated fairly, with the consumer paying extra but knowing they are supporting the artists or bean farmers. It could also include works from the public domain. By integrating this dataset easily by building on open source image-generation reducing further development costs, artists' contributions could be tracked publicly and at mega-scale via blockchain and compensated perpetually through smart contracts on Web3 technologies like Solana.

For this system to succeed, opt-in participation must be respected by all parties – with potential guild and industry boycotts for datasets and apps that break this trust. Another option we were the first to propose back in 2022 could be used for extra assurance, a new or re-purposed opt-out mechanism like robots.txt (e.g. notraining.txt) which has opted websites out of Search Crawlers like Google for decades, and could be used to opt out of AI Training Crawlers to keep content out of their datasets (especially for websites like Getty with watermarked images). This allows content owners who *need* to put their work online to opt out. Both can simply be ignored, especially by bad actors, so we're sure to see open source projects and adversarial countries develop generative AI that does not respect these policies, but if a majority of the world gets behind these new systems with the power and leverage of the Hollywood guilds, it could stand a chance as Starbucks does selling fairtrade coffee beans at a premium.

Our cofounder, Andrew Palmer, is a proud member of the Canadian Writers, Directors, and Producers guilds. This deep connection to the creative industries is part of our founder DNA, driving us to build the kind of AI company we believe the film world needs.

Copyright exists to ensure creators and their estates retain ownership of their works during their lifetimes and for a period afterward, with works eventually entering the public domain. Generative AI, predicted by leading analysts at Goldman and McKinsey to contribute trillions in gross domestic product growth over the coming decade, offers an unprecedented opportunity to direct this value back to the creators and estates whose work is used to create AI, preserving artistic legacies for generations with new revenue streams.

However, if the worst-case scenario unfolds—where human artists abandon creating new works due to disruptions caused by AI, and future training sets consist mostly of outdated human works and piles of AI-generated "synthetic data"—the richness of art and culture risks becoming stagnant and unappealing. Supporting creators and ensuring fair compensation may be the only sustainable path forward for generative AI to foster enduring artistic innovation.

Deepfakes

The term *deepfake* has become infamous due to its association with illicit content, prompting the industry to adopt alternatives like "virtual performances." Tools like Flawless AI have introduced remarkable workflows for virtual reshoots[12] with actor consent and compensation, building innovative features for dubbing, translation, and lip sync, helping studios save millions of dollars and countless hours.

In April 2022, we explored this topic in our series On Actors and Deepfakes,[13] examining both the challenges and opportunities presented by this technology. Our goal is to shed light on the ethical implications, potential benefits, and actionable steps creators and audiences can take to navigate the evolving landscape of virtual performances.

New laws and regulations

It's one thing to meet the bare legal requirements for outdated laws, but it's another to adopt a proactive, artist-friendly approach in our business practices. At Saga, we are committed to supporting filmmakers, which is why we encourage our users to explore and develop their own ideas, not have AI do all of the work for them. In some instances, we may choose to block certain names in user prompts, such as "Write an Aaron Sorkin-style script," out of respect for living artists—a value we share as writers ourselves.

Our company decided when we were founded to give our users 100% ownership of their work, as opposed to artist-unfriendly practices like sneaking ownership clauses in the terms of service so they own your work or a percentage of box-office revenue (as some of our competitors chose to do). Years later during the strikes of 2023 this became a clause in the Writers Guild of America agreement with studios, but it's something we've followed and believed in all along.

Our competitors also make claims on their user's work to retrain the company's AI models and use it to improve ideas for their other users, which is another practice we do not and will not follow at Saga. We believe to be artist-friendly we need to give security and privacy to our users, above and beyond what the law requires, and will continue to push for fair practices through our conference panels, blog writing, and industry evangelism.

In conclusion, while there are risks associated with AI that must be managed, Generative AI tools can help storytellers and filmmakers create more of their best work faster. Self-distribution on platforms like YouTube will allow new creators to showcase their work, build audiences, and possibly get discovered by Hollywood studios to work on bigger-budget films next.

Silicon Valley has seemingly lost the world's trust through a number of recent scandals, from Theranos to FTX. Crypto and Web3 never lived up to the promises evangelized by its proponents in recent years. Though a completely separate technology (that already has a track record of delivering real value), generative AI seems to be caught up in the anger, suffering from a lack of trust and fear it will destroy the film industry as we know it. Technology companies "move fast and break things," and by training their models on copyrighted content, started off on the wrong foot with the art community. Billionaire technology moguls are no longer admired as they once were and are becoming despised (especially as they enter the political sphere). Who wants to see *another* billionaire control Hollywood (simply replacing the old billionaires), while firing all of the artists and below-the-line workers to replace them with robots? Not us.

We started building Saga in 2021 when all anyone would talk about were non-fungible tokens (NFTS) and Bitcoin, refusing to get distracted from our mission. This was long before ChatGPT launched, before "generative AI" was even a term, and before the Hollywood strikes. We've never wavered in our mission to build tools for aspiring filmmakers like us, not so the youth don't need to learn skills or practice writing, but simply to let them make a movie that looks great—as Andrew and I always dreamed of.

There used to be limited space on the shelf at Blockbuster, and a limited number of movie theaters and screens in every neighborhood (in the days before streaming). It made sense for the executives and tastemakers to decide on a limited number of pictures to produce, and make safe bets on proven stars and built-in franchise audiences. Now with everything in the cloud, there is no upper limit to the number of films that can be made every year. With AI preference microtargeting, even with 100 new movies a day releasing on Netflix, each and every one can find its own niche targeted audience—*increasing* the pie of movies and people who create them professionally.

> Machine learning will improve so much in the next few years as we approach AGI, that they will not only master the story structures of today, they will use this training to create new ones, plot twists and characters of the likes we've never seen before, and humans will use these to produce some of the highest-quality cinema ever seen.

We believe a multimodal, multimodel app like Saga will continue to be used by filmmakers to reduce the cost and time of producing movies, increase global sales with dubbing and translation, and grow the number of jobs in the film industry.

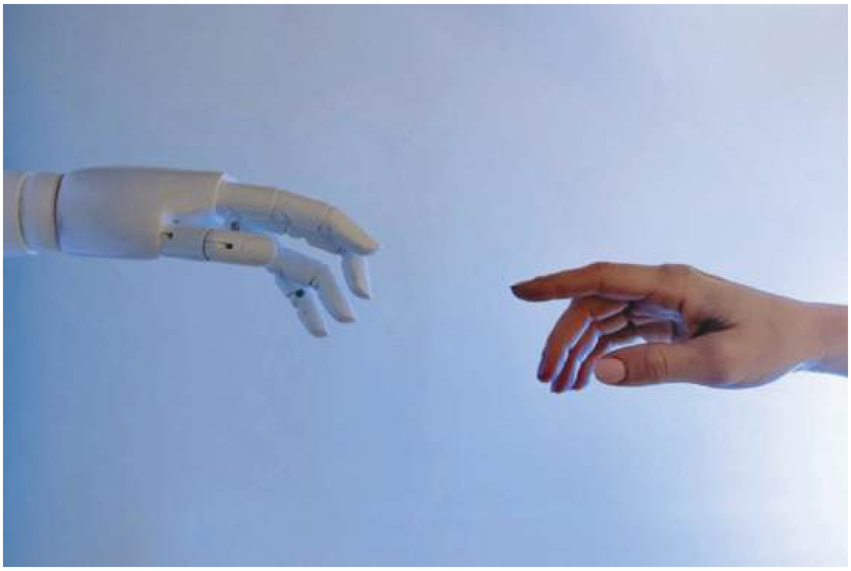
To reach out to the founders for questions, partnerships, or express interest in their upcoming Seed Round for investment, e-mail CEO Russell Palmer at russellp@cyberfilm.ai or use the Contact form on our website. We are building a strong investor and advisory board,[a] and partnerships with studios and other AI companies.

You can also sign up for Saga and try it for free on our site: https://www.writeonsaga.com

[a]Investors: Jason Calacanis (Angel investor; All-In Podcast); Advisors: Alex Jordan (SVP Production at Muse; Directors Guild of Canada) and Dr. Mike Zyda (Emeritus Professor of Engineering at USC; ACM and IEEE Fellow).

COMMENTS?

If you have comments about this article, or topics or references I should have cited or you want to rant back to me on why what I say is nonsense, I want to hear. Every time we finish one of these columns, and it goes to print, what I'm going to do is get it up online and maybe point to it at my Facebook (mikezyda) and my LinkedIn (mikezyda) pages so that I can receive comments from you. Maybe we'll react to some of those comments in future columns or online to enlighten you in real time! This is the "Games" column. You have a wonderful day.



FUTURE 17. Image of a robot touching a human hand in homage to Michelangelo.[14]

Use the following code to test our Premium version free for one month (new accounts only), as our gift to you for reading this far: IEEEFreeMonthSaga

In closing, we chose the image in Figure 17 for our founding vision post nearly four years ago,[4] and posited that "AI is a tool" not a replacement. After millions of views across Silicon Valley, Hollywood, and around the world, this image and statement both seem to have taken off, and we're now hearing it all over the world. The famous image of hands touching in Michelangelo's *Creation of Adam* has come to represent the discussion of human creativity meeting AI, a spark of life given to a creation made in its own image. This meme has since been used by dozens if not hundreds of AI news articles, tech conference logos, and others in various formats, and we're hoping our vision for ethical AI filmmaking can impact the world as well. You can follow Saga on YouTube (https://www.youtube.com/@writeonsaga), Twitter/X (https://x.com/writeonsaga), Instagram (https://www.instagram.com/writeonsaga/), TikTok (https://www.tiktok.com/@writeonsaga), and LinkedIn (https://www.linkedin.com/company/cyberfilm-ai/products/).

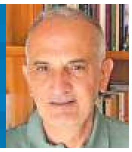
REFERENCES

1. The New York Times. *1 in 3 Preteens Want to Be Influencers. Is It the New American Dream? Op-Docs*. (Dec. 15, 2024). Accessed: Jan. 12, 2025. [Online Video]. Available: https://www.youtube.com/watch?v=SDRh840GrCs
2. F. Zandt. "What do U.S. teens want to be when they grow up?" Statista. [Online]. Available: https://www.statista.com/chart/31014/most-popular-future-jobs-with-united-states-teenagers/
3. "Synapz productions – Home page" Synapz Productions. Accessed: Jan. 12, 2025. [Online]. Available: https://www.synapzproductions.com
4. R. S. A. Palmer and A. M. A. Palmer, "How AI can benefit Hollywood and solve its most pressings needs," *Medium*, Nov. 23, 2021. [Online]. Available: https://bit.ly/AIandHollywood
5. R. S. A. Palmer. *How AI will Benefit Hollywood*. Accessed: Jan. 26, 2024. [Online Video]. Available: https://www.youtube.com/watch?v=B49WjGSnDcM&t=412s)
6. "Saga GPT in the OpenAI ChatGPT store." ChatGPT. Accessed: Jan. 10, 2024. [Online]. Available: https://chatgpt.com/g/g-0X4q4iYYN-saga
7. Saga - Tools for Filmmakers. *From Idea to Previz, Animatics to Animation, Try SAGA and Create for Free Today!* (Jun. 1, 2024). [Online Video]. Available: https://www.youtube.com/watch?v=iQ6JhAqeU-g
8. R. S. A. Palmer, "The content Turing test," *Medium*, Mar. 24, 2022. [Online] Available: https://russellsapalmer.medium.com/the-content-turing-test-d39e9765de19
9. R. S. A. Palmer. *SAGA Advanced Video Generation (New)*. (Oct. 2024). [Online Video]. Available: https://bit.ly/sagavideo
10. A. M. A. Palmer. *Writing a Feature Film Screenplay in Saga*. (Oct. 23, 2023). [Online Video]. Available: https://www.youtube.com/watch?v=n3Dzz-NdKrE
11. R. S. A. Palmer. "Artists rights – On getting paid for your work in an age of AI," *Medium*, Nov. 3, 2022. [Online]. Available: https://russellsapalmer.medium.com/artists-rights-c655b79402bb
12. Bloomberg Live. *Flawless Tech Demonstration on AI & Filmmaking*. (Oct. 12, 2024). Accessed: Oct. 12, 2023. [Online Video]. Available: https://www.youtube.com/watch?v=uPmavUfcMqo
13. R. S. A. Palmer, "On actors and deepfakes," *Medium*, Apr. 2022. [Online]. Available: https://medium.com/@russellsapalmer/on-actors-and-deepfakes-4df58d0f3089
14. T. Winstead. "Person reaching out to a robot (image)." Pexels. Accessed: Nov. 28, 2021. [Online]. Available: https://www.pexels.com/photo/person-reaching-out-to-a-robot-8386434/

RUSSELL S. A. PALMER is the CEO and a cofounder at CyberFilm, San Francisco, CA 94123 USA. Contact him at russellp@cyberfilm.ai.

ANDREW M. A. PALMER is a cofounder and chief story officer at CyberFilm, Hamilton, ON L8K 3A1, Canada. Contact him at andrew@synapzproductions.com.

EDITOR DOMENICO TALIA
University of Calabria; talia@dimes.unical.it



How to Foster Responsible and Resilient Data: The Ethical Data Initiative

Kim M. Hajek, **Paul Trauttmansdorff**, and **Sabina Leonelli**, Technical University of Munich
Stephan Guttinger, University of Exeter
Silvia Milano, University of Exeter and LMU Munich

It is more important than ever to advance data ethics. Informed by science studies, the Ethical Data Initiative uses open and inclusive discussions to foster a responsible and resilient approach to data creation, transmission, and use.

The increasing reliance on and importance of data for almost every aspect of our everyday lives brings with it numerous ethical concerns, from the individual to the global level. Issues range from data protection and privacy to data access, data sovereignty, and multiple inequities and environmental challenges. Several initiatives and movements have emerged that highlight the need to shape technologies according to human values and public interest[1] and to do so in a manner that is not only inclusive but that builds in ethical considerations from the very beginning as digital systems are designed and developed.[2,3]

From our perspective in the social studies of science, we advocate for paying more attention to the various contexts in which data are created, the choices made by the humans involved in the process, and the ways data are configured by material objects, environments, and apparatus. Ethical considerations linked to these choices and their implications should be incorporated into the entire spectrum of data-intensive fields of education and work, including computing, informatics, electrical engineering, data science, and artificial intelligence (AI) development. Data

Digital Object Identifier 10.1109/MC.2024.3522696
Date of current version: 27 March 2025

practices must become more responsive to social settings, more responsible regarding their consequences, and more resilient against possible misuse or misappropriation.

With this in mind, the Ethical Data Initiative (EDI) (https://ethicaldatainitiative.org/) was recently formed as a nonpartisan platform that 1) offers educational tools to increase data literacy and 2) fosters open discussions on data ethics, with a focus on equity and engagement across different domains of data work. Bringing together academic scholarship from the University of Exeter and the Technical University of Munich and building on the remarkable long-term efforts in this space by partner organizations such as the Research Data Alliance (https://www.rd-alliance.org/), CODATA (https://codata.org/), and the Research on Research Initiative (https://researchonresearch.org/), among others, the EDI aims to coordinate and further develop pedagogical, research, and policy efforts to support responsible data practices. Our efforts prioritize education as a means of enabling and promoting a proactive ethical approach to all aspects of data work. We direct attention to the multiple environments in which data come to matter, with particular emphasis on the distinct choices involved in creating, collecting, storing, and using data.

> The EDI aims to coordinate and further develop pedagogical, research, and policy efforts to support responsible data practices.

WHAT ARE DATA?

Astute readers may have noticed that in this column, we refer to data in the plural: data are, not data is. This is a conscious choice that stems from our conception of data not as a completed product, not as isolated or decontextualized pieces of information, but rather as the continuously mutable outcome of processes and decisions undertaken in a specific context for specific reasons. Those who create and work with data will develop and perform a plurality of "standards and criteria of best practice … to suit their specific goals and working conditions."[4]

Data—and particularly digital data—tend to move around; they escape their place of origin and are reused in a variety of different ways, which often affects their format, interpretation, and the ways in which they are visualized. However we think of such data journeys[5] as data "enrichment," "interpretation," "analysis," "clustering," or the like—doing something with data involves intervening on their key characteristics, making it possible for data to be used (to become "actionable") for new purposes as required. Data provenance matters; only if we understand the technical, scientific, practical, and circumstantial motivations behind these diverse interventions can we ensure that key information accompanies the data on their further travels.

Moreover, data are not just digits. They can take many different forms, including material objects, symbols, sounds, observations, text, and images. All forms are subject to different constraints when put to use, particularly when attempts are made to cross-analyze and integrate diverse data types.[6] Recognizing data as plural, diverse, and in motion allows us to understand better how power operates and is contested through data, how hierarchies and binaries are upheld, and how pluralism can be promoted.[7] Viewing data as a neutral, fixed, and immutable set of objects does not allow us to capture the reality of data practices.

WHY WORRY?

Calling for an ethical approach to data work means calling for critical reflection on the realities of data work and the significant challenges they entail. In the following, we outline our most pressing worries.

Injustice and bias

Current practices for collecting, sharing, and interpreting data frequently include forms of bias and discrimination (not always intentionally) that reinforce existing social and global hierarchies and power imbalances.[8] This constitutes a form of injustice as it disproportionally impacts already marginalized groups. A well-known example is the underrepresentation of people of color in datasets used to train facial recognition algorithms. This leads to increased errors and potentially fatal consequences in contexts such as law enforcement and security. People, particularly from marginalized groups, are often misrepresented, made invisible, or subject to large-scale surveillance practices.[9] The contexts in which data, algorithms, and technological systems operate matter and have wide-ranging discriminatory implications for social, political, and economic outcomes.

Inequity and access

The push toward "bigger data" and data-driven innovation does not affect all regions and populations equally but instead tends to deepen global inequalities and the existing digital divide. Large-scale databases and infrastructures are mostly concentrated in wealthy well-resourced areas and countries, with data access primarily limited to well-funded institutions. Data monopolies mean that closed companies or organizations collect and control large amounts of data, often for commercial purposes, without being transparent about whose data are collected or how they are stored and (re)used. Access to reliable digital infrastructures is limited in many parts of the world and segmented along class, ethnic, and gender lines. What is needed is a digital environment that values regional and local autonomy and avoids replicating colonial-era power structures.

Privacy and confidentiality

Privacy and confidentiality in data work are essential for safeguarding individuals' autonomy and control over their personal information. Without effective principles of data privacy, the risks of misuse, unauthorized tracking and sharing, data breaches, and manipulation increase, potentially exposing individuals and creating significant vulnerabilities. Even when information is anonymized, machine learning practices and AI algorithms can compromise confidentiality, revealing personal details through indirect proxy information or patterns in aggregated data.

Transparency and trust

In today's increasingly automated data environments, many data practices lack transparency, with digital subjects left unaware of what information is collected, in what settings it is collected, or how it will be used—a concern that affects personal data as well as environmental and administrative data, which can also be misused in ways that harm humans, nonhumans, and/or the planet. Trust and confidence in data suffer as a result. Regulatory efforts to improve the situation are difficult to agree upon and to implement in meaningful terms. Particularly significant are efforts to build trustworthy institutions and infrastructures to steward and safeguard data; however, those require much higher levels of investment and public engagement than those available at present.

Openness and ownership

Advocates for greater openness have long called for increased sharing of resources, enhanced access to data pools and infrastructures, and the reuse of data. Yet unreflective forms of openness may also have unwanted effects, such as limiting epistemic diversity and fostering epistemic injustice, for instance, if "open information" is harvested without informed consent or proper attribution. Bigger datasets and unlimited access do not necessarily provide better scientific or societal results. Rather, they are often less representative and less reliable than smaller datasets that are carefully curated and produced under a responsible ethos of data work. In this way, treating data ethically overlaps with efforts to improve the quality and trustworthiness of datasets for research and other purposes—and again to invest in intelligent openness strategized and mediated by expert data stewards. The work of creating and maintaining data needs to be recognized and rewarded adequately—all too often, this does not happen.

At the same time, we note that a focus on ownership obscures the question of whether data on individuals and communities should be treated as tradable assets at all rather than as a "common good." Concrete measures are needed to preserve autonomy in data usage and sharing without data becoming commercialized and dominated by existing profit-oriented economic and regulatory regimes. This is particularly urgent given the difficulties in controlling data flows, especially when it comes to personal digitalized data—which are so easily copied, traded, and mobilized that it is hard, if not impossible, to track their travels and identify who may be accountable for the use of those data.

Misuse and error

The rapid expansion of large datasets and the growing automation of data analysis are not accompanied by proper auditing mechanisms to ensure data quality. Mistakes and errors are hard to track, especially in AI systems, and even when they are identified, correcting them can be challenging and costly across interconnected systems. The rise of automation further increases the risk of producing mistakes at a large scale, undermining trust and harming individuals or communities affected by such errors, particularly in highly sensitive areas like health care, migration, or criminal justice. Even small error rates can have significant consequences for thousands of people. The risk of misuse is amplified when data are presented or interpreted out of context and without a proper understanding of the limitations of data technologies and systems. There is significant diversity in the expertise and practices used to produce and make sense of data. Applying one standard across the board has the potential to degrade trust in certain areas of research and even hamper scientific and technological advancement.

> Viewing data as a neutral, fixed, and immutable set of objects does not allow us to capture the reality of data practices.

Environmental damage and sustainability

Current data ecosystems are dramatically unsustainable both in terms of their durability—significant resources are needed for ongoing maintenance and repair—and from an environmental standpoint. Yet these issues have been largely ignored in public discussion, with digital solutions often proposed as the "clean alternative." The increasing volume of data storage and demand for fast processing require immense energy and material resources and depend on technologies (such as batteries and chips) that are unevenly produced and distributed across the globe. At the same time, electronic waste piles up in landfills. The rise of AI only intensifies these demands, with current machine learning models using vast amounts of energy and producing significant carbon emissions. Future datafication must balance technical imperatives against environmental harms—all while ensuring the financial and material outlays needed to keep data infrastructures operating reliably in the long term.

What is needed is a digital environment that values regional and local autonomy and avoids replicating colonial-era power structures.

AI

AI models tend to rely on huge amounts of data originating from the Internet and online platforms, which are harvested without clear regulations, attention to boundaries, or ethical guidelines, prompting significant worries around fair use, proprietary content, and proper attribution.[10] Data mined for AI purposes are furthermore frequently stripped of contextual content or metadata, which makes it hard to distinguish reliable data from unreliable information. In turn, the uncertain data quality and representativeness of AI models and applications have ethical consequences, such as when biased or unrepresentative information patterns are replicated on a large scale and produce one-sided insights and potentially harmful outcomes.

PATHS FORWARD: THE EDI

The aforementioned challenges are not insoluble, nor is the kind of global coordination required to address them an insurmountable task. Cooperation across nations to tackle complex technological challenges has been successfully achieved in other cases, such as trade and food distribution networks. Precisely such cross-cutting cooperation is also crucial in the case of ethical data. The EDI aims to enable these concerted efforts, connecting academia and research, policy, education, and relevant local actors.

When discussions around ethics are led by a few select experts, this risks turning ethics into a standardized set of external principles that have little connection with the realities of data work. We contend that ethical data practices are not merely a technical matter for a few experts to evaluate or oversee. Also, formulating ethical principles is not sufficient because they are often too abstract; they acquire concrete meaning only when they are interpreted in practice. Ethical data practices need to be supported by all sectors of society and engage the communities most affected by data policies and practices. In the EDI, we thus construe data ethics as a living *ethos*: a responsible way of approaching the creation, transmission, storage, and (re-)use of data.

This ethos is necessarily dynamic; it needs to be continuously informed by local settings and framed by context-dependent methods and practices. It entails paying attention to the processes of decision making and the particularities that make up every step along a data journey—the steps data travel from the initial moment of creation in a lab, field, or digital encounter through inscription in, for instance, an open source repository to interpretation and reuse by a third-party actor. When assessed in terms of specific contextual steps in that journey, ethical principles and guiding values become meaningful and actionable.

Without attention to choices, context, and consequences, the data and knowledge we produce "threatens to blindly privilege specific ways of knowing,"[4] often to the benefit of powerful groups in society. The contexts in which data are handled are hugely diverse, and the implications of data practices should be assessed in light of that diversity with the recognition that each domain has its specific forms of expertise. However, recognizing diversity does not mean that anything goes. Rather, it means ensuring that the criteria used to evaluate data work are relevant and appropriate to the research in question in light of its goals, situation, and methods. In other words, to approach data work ethically is also to consider what characteristics of data allow for producing more reliable and trustworthy scientific knowledge.

Attending carefully to the consequences and contingencies of each step on a data journey requires targeted ways of integrating data ethics within a given situation, and particularly, of dealing with the conflicts that often arise between different values. Examples of such conflicts include the need to preserve *privacy* while still fostering *openness* in research on medical data as well as efforts to *share* data fairly among researchers while also recognizing the *ownership* claims of those who may have invested the most in creating and disseminating the data. In trying to address such diverging expectations in concrete cases of data work, it is crucial to be alert to what choices are made and why, who benefits from those choices, and how they may affect the relevant public. Such analysis helps to unearth data-related discrimination and bias and makes data practices more accountable to scrutiny and critique.

Approaching data ethically is about prioritizing human judgment over what may be technologically feasible or best adapted to computational analysis. As the use of data evolves constantly, so must our awareness of how they are created and processed, what constraints that entails, and who is benefited or disadvantaged. We advocate making data work a reflective process, which engages with relevant stakeholders and environmental concerns, rather than rushing toward technological convenience or the appearance of innovation. Ethical paths to data must address how technologies and practices can contribute to more democratic data environments, equitable collaborations, and forms of solidarity that envision and build more inclusive data ecosystems.

For the EDI, responsible data work ultimately aims to improve the living conditions of all creatures on Earth, which involves being attuned to the plurality of human experience and the preservation of our environment. This means valuing the diversity of the humans who take part at every stage of data work and building mechanisms to address persistent injustices and inequities in their visibility and ability to access resources. Harms affecting nonhuman organisms and the environment as a whole will also be effectively reduced and prevented only when they are given attention throughout data practices. The EDI sees its mission as providing a politically neutral yet intellectually potent platform where these perspectives can be shared, can be integrated into actual data practices, and can seed partnerships for a future of sustainable and ethical data work. ◼

REFERENCES

1. H. Werthner, A. Stanger, V. Schiaffonati, P. Knees, L. Hardman, and C. Ghezzi, "Digital humanism: The time is now," *Computer*, vol. 56, no. 1, pp. 138–142, Jan. 2023, doi: 10.1109/MC.2022.3219528.
2. P. Arora and U. Raman, "The case for Femwork: Feminist design principles in tech and tech-enabled work," *Computer*, vol. 57, no. 1, pp. 100–104, Jan. 2024, doi: 10.1109/MC.2023.3325593.
3. L. Floridi and M. Taddeo, "What is data ethics," *Philos. Trans. R. Soc. A*, vol. 374, no. 2083, doi: 10.1098/rsta.2016.0360.
4. S. Leonelli, *Philosophy of Open Science*. Cambridge, U.K.: Cambridge Univ. Press, 2023.
5. S. Leonelli and N. Tempini, Eds., *Data Journeys in the Sciences*. Cham, Switzerland: Springer-Verlag, 2020.
6. R. Kitchin, *The Data Revolution. A Critical Analysis of Big Data, Open Data and Data Infrastructures*, 2nd ed. London, U.K.: Sage, 2022.
7. C. D'Ignazio and L. F. Klein, *Data Feminism*. Cambridge, MA, USA: MIT Press, 2020.
8. S. U. Noble, *Algorithms of Oppression: How Search Engines Reinforce Racism*. New York, NY, USA: NYU Press, 2018.
9. P. Trauttmansdorff, *The Digital Transformation of the European Border Regime: The Powers and Perils of Imagining Future Borders*. Bristol, U.K.: Bristol Univ. Press, 2024.
10. S. Vallor, *The AI Mirror: How to Reclaim Our Humanity in an Age of Machine Thinking*. New York, NY, USA: Oxford Univ. Press, 2024.

KIM M. HAJEK is a research fellow in the Ethical Data Initiative and in the Chair for History and Philosophy of Science and Technology, Technical University of Munich, 80333 Munich, Germany. Contact her at kim.hajek@tum.de.

PAUL TRAUTTMANSDORFF is a research fellow in the Ethical Data Initiative and in the Chair for History and Philosophy of Science and Technology, Technical University of Munich, 80333 Munich, Germany. Contact him at paul.trauttmansdorff@tum.de.

SABINA LEONELLI is the director of the Ethical Data Initiative and Chair of History and Philosophy of Science and Technology, Technical University of Munich, 80333 Munich, Germany. Contact her at sabina.leonelli@tum.de.

STEPHAN GUTTINGER is a lecturer in the philosophy of data and data ethics at the University of Exeter, EX4 4PJ Exeter, U.K., and serves as the head of research for the Ethical Data Initiative at Exeter. Contact him at s.m.guttinger@exeter.ac.uk.

SILVIA MILANO is a senior lecturer in philosophy at the University of Exeter, EX4 4PJ Exeter, U.K., and Humboldt Fellow at the Munich Center for Mathematical Philosophy, LMU Munich, 80333 Munich, Germany. Contact her at s.milano@exeter.ac.uk.

©SHUTTERSTOCK.COM/VARUNYUUU

AI for Cloud and SaaS: Technologies and Business Models

Christof Ebert, Vector Consulting Services
Sriram Panyam, Omlet
Praveen Gujar, LinkedIn

Artificial intelligence (AI) is transforming cloud services, and specifically, software as a service (SaaS). Business models will evolve, driven by AI-native products, adaptive cloud fabrics, and tracking and observability. The article evaluates technologies and business of AI for SaaS.

Software as a service (SaaS) platforms are structured on the cloud infrastructure, typically using a multitier architecture consisting of front-end interfaces, application logic, and database layers. Figure 1 shows a typical layout. This architecture allows for scalability, modularity, and secure data management.

THE EVOLVING SAAS LANDSCAPE

Cloud-based SaaS platforms provide benefits such as easy remote accessibility, streamlined updates, and high availability. By leveraging cloud elasticity, these platforms can scale resources dynamically based on demand, making them ideal for businesses of all sizes. SaaS simplifies software distribution, eliminating on-premises installations, and offers easy update and integration capabilities, thus reducing operational burdens and mitigating cybersecurity risks.

Artificial intelligence (AI) will reshape cloud services with intelligent automation, predictive capabilities, and adaptive resource management. AI will enable more efficient decision making and proactive system optimization, which are critical for handling

Digital Object Identifier 10.1109/MC.2024.3522697
Date of current version: 27 March 2025

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

FROM THE EDITOR

Cloud computing and software as a service (SaaS) are the basis of practically all IT infrastructures, empowering businesses with scalable resources and services. With artificial intelligence (AI) deeply integrating into cloud ecosystems, enterprise technologies and business models are heavily impacted. This article explores major AI trends and recent innovations, the challenges faced in adopting and provisioning, and the calls to action for technology leaders and professionals. —*Christof Ebert*

dynamic workloads and personalized user needs.

AI-NATIVE SaaS: INTELLIGENT FEATURES FOR SELF-OPTIMIZING SYSTEMS

Traditional SaaS products have long used AI for smarter features, such as heuristics for load balancing, recommendation engines, and automated insights. With AI-native SaaS, AI doesn't merely enhance functionality but defines the entire product. The shift is toward a model where AI capabilities, like natural language processing or predictive analytics, are embedded deeply into the core architecture.[1] Products can now continuously learn from usage data, autonomously adapt their workflows, and even predict customer needs before they articulate them. This is changing the nature of what SaaS solutions can provide—going from static configurable tools to dynamic self-optimizing systems. For example, AI can transform patient management systems by predicting patient inflow, optimizing resource allocation, and proactively suggesting staffing schedules to meet demand. This makes healthcare management more dynamic, continuously adding value by improving patient care and reducing wait times while optimizing resource utilization.[2]

Challenges and mitigations

The adoption of AI-native SaaS comes with challenges such as integrating AI into existing architectures, ensuring data privacy, and addressing concerns about overreliance on AI systems. The risk of "model drift," where the AI model's performance degrades over time, is also a significant concern.[3] It's crucial to invest in training toward AI integration and implement robust governance frameworks. Professionals should focus on building skills in AI model evaluation and understanding how to collaborate effectively with AI-driven systems.[4]

AI-DRIVEN CLOUD FABRICS

Cloud infrastructure must evolve with the rise of AI workloads. AI models demand more computational power, lower latency, and greater parallel processing, leading to the development of cloud fabrics that are inherently AI optimized. AI will dissolve the traditional boundaries between compute, storage, and networks, giving rise to distributed fabrics where AI can dynamically optimize workloads across different hardware and regions. Imagine a cloud infrastructure that actively learns which components should be used and scales autonomously in response to evolving workload needs—that's the power of AI-driven cloud fabrics. It is changing the economics of cloud computing as well, allowing cost models that are far more predictable and efficient.[5]

AI-driven cloud fabrics can be used to manage peak traffic during major sales events like Black Friday. By dynamically analyzing customer behavior and historical sales data, AI can predict spikes in demand and automatically scale resources accordingly. Unlike traditional cloud scaling that relies on predefined thresholds, AI can optimize the use of compute and storage in real time, ensuring seamless performance during peak loads while avoiding unnecessary costs during off-peak times. This helps e-commerce platforms maintain fast and reliable user experiences without the need for constant manual intervention.

Challenges and mitigations

Adopting AI-driven cloud fabrics requires significant investment in new infrastructure. The complexity of managing distributed dynamic environments can lead to operational challenges, and there is a risk of vendor lock-in due to proprietary AI optimizations. AI-based cloud business models must weigh the costs of transitioning against the benefits of improved scalability and efficiency. Professionals should become adept at understanding how AI-driven optimization works to both leverage it effectively and troubleshoot when things go wrong.[7]

AI-ENHANCED OBSERVABILITY AND SELF-HEALING SYSTEMS

Observability in the cloud has always been crucial, but AI is making it revolutionary. Instead of merely collecting metrics and events, AI can help correlate thousands of data points across services, identify anomalies, and suggest or even perform remediation autonomously. This means not only the faster detection of issues but also the possibility of preventing incidents before they occur. This trend toward AI-powered observability is pushing us toward a future of true self-healing cloud services. AI-enhanced observability could involve monitoring machinery performance and detecting anomalies like unexpected temperature spikes or unusual vibration patterns. The system could autonomously resolve minor issues, such as adjusting cooling parameters, or alert maintenance teams when significant threats are detected, ensuring uninterrupted production and improved operational reliability.[8]

Challenges and mitigations

AI-enhanced observability brings a learning curve involved in trusting AI recommendations. There is also the risk of false positives or negatives, which could lead to unnecessary actions or missed issues. Training implies, for instance, understanding how AI detects anomalies in user activities and how to use these insights to enhance security and reliability. IT professionals should focus on learning the intricacies of these tools, such as how the AI algorithms identify patterns or correlations and understanding how AI recommendations are generated to effectively validate or override them when needed.

DEMOCRATIZATION OF AI WITH NO-CODE AND AI AS A SERVICE

AI has recently become more accessible, uplifting cloud and SaaS capabilities by providing businesses of all sizes with advanced tools. No-code AI platforms and AI as a service (AIaaS) will significantly reduce the barriers to using machine learning, allowing even nonexperts to harness AI for their operations, like what we observed with GenAI copilots.[3,9] This empowers small business owners to leverage AI-driven analytics without needing a dedicated data science team, transforming SaaS from a simple software provider into a critical enabler of smarter business decisions. AI-driven no-code platforms can, for instance, allow utility companies to forecast energy demand and automate grid management using intuitive drag-and-drop interfaces. Instead of just providing basic monitoring, these platforms leverage AI to continuously learn from consumption data, predict peak usage times, and optimize energy distribution.

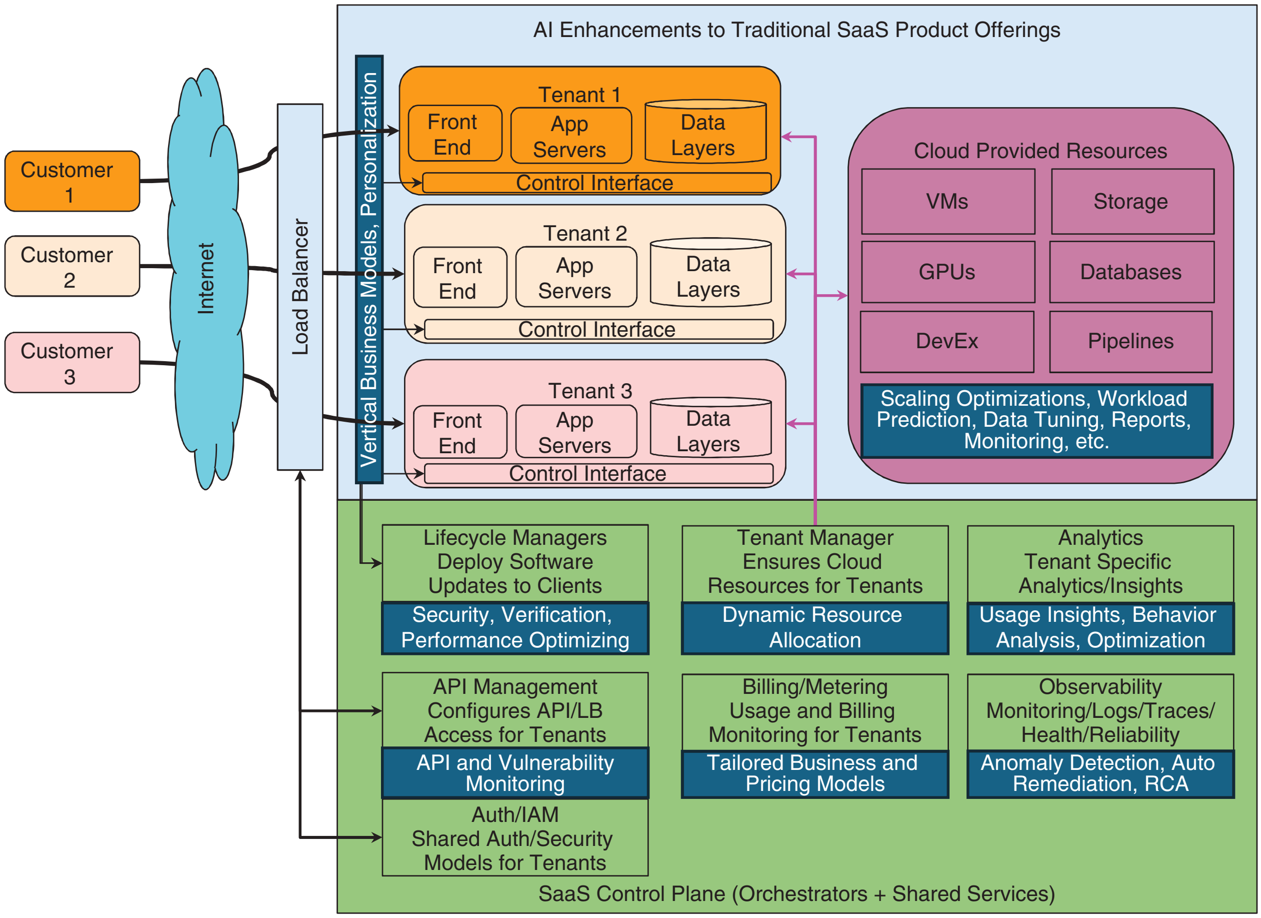


FIGURE 1. The reference SaaS product architecture with multiple tenants and AI enhancements (blue boxes). VM: virtual machine; DevEx: developer experience; API/LB: application programming interface/load balancing; IAM: identity and access management; RCA: root cause analysis.

Challenges and mitigations

Without adequate understanding, businesses may misuse AI, leading to flawed decision making.[9] Checks should establish approval workflows for automated decisions and implementing audits of AI outputs. Professionals should seek to understand AI-driven analytics, such as how it classifies transactions or predicts trends, to use these tools effectively and responsibly.

HUMAN–AI PARTNERSHIP—COPILOT FOR EVERYTHING

The rise of AI copilots within SaaS platforms might be the most visible shift. With AI taking on a collaborative role, cloud and SaaS applications are moving from being passive tools to being active partners in productivity. SaaS products are now evolving into platforms where human creativity and machine intelligence work together seamlessly. AI copilots can assist technicians by providing step-by-step repair guides, suggesting the right tools for specific tasks, and reminding them of maintenance schedules. Such copiloting saves time, reduces manual errors, and focuses technicians on solving complex problems rather than routine procedures.

Challenges and mitigations

There are concerns about the overdependence on AI copilots as well as the risk of users losing creativity and decision-making skills by relying too heavily on automated suggestions.[4] Additionally, data privacy and bias in AI-generated content pose significant risks. Professionals should develop a mindset of collaboration with AI, enhancing their own skills by understanding when to trust and when to question AI suggestions,[3,4,9]

BUSINESS AND TECHNOLOGY TRENDS

AI is reshaping technology offerings, service adjustments, and the business models of the cloud and SaaS. Leading SaaS providers adjust their offers to AI. Table 1 provides an overview of some of the leading SaaS providers on a global scale. AI can analyze usage patterns and predict future needs, providing a more personalized pricing model. This shift not only changes how businesses consume cloud resources but also how SaaS products generate value. Imagine a SaaS product that adjusts its pricing dynamically based on the specific return on investment it delivers to a business.[6] For instance, in the travel and mobility industries, AI-driven pricing models will dynamically adjust ticket prices based on real-time factors, such as demand, weather conditions, and customer booking behavior.

Challenges and mitigations

AI-driven pricing models can lead to concerns about fairness, transparency, and potential customer pushback. Ways to mitigate this include ensuring transparency in pricing

TABLE 1. Leading SaaS providers adjust their offers to AI.

Company	SaaS Services	Leadership	AI Strengths	AI Trendsetter
Google GCP	Offers AI-powered tools like Google Workspace and industry-specific solutions.	Strong in AIaaS with platforms like Vertex AI.	Integrates AI across services, enhancing data analytics and machine learning capabilities.	AI-powered analytics, ML platforms, and enterprise-ready AI tools.
Amazon AWS	Provides AI-driven services such as Amazon Connect and AWS SaaS Factory.	Leading in AI-driven cloud fabrics with services like SageMaker	Offers a broad range of AI services, enabling scalable machine learning deployments.	Scalable, elastic, developer-centric AI services with SageMaker and cloud efficiency.
Microsoft Azure	Includes AI-enhanced offerings like Dynamics 365 and Office 365.	Excels in AI-native SaaS and copilot functionalities, integrating AI deeply into productivity tools.	Pioneers in embedding AI across enterprise solutions, enhancing user productivity.	AI copilots, productivity integration, and full-stack AI across enterprise SaaS.
Alibaba	Offers AI-infused services tailored for e-commerce and retail sectors.	Growing presence in AI-driven cloud fabrics within the Asia-Pacific region.	Leverages AI to optimize supply chain and retail operations, enhancing efficiency.	Leading AI adoption in e-commerce optimization, supply chain forecasting, and retail intelligence.
Tencent Cloud	Provides AI solutions focused on gaming, social media, and entertainment.	Emerging in AIaaS, particularly within China's digital ecosystem.	Utilizes AI to enhance user engagement in gaming and social platforms.	Trendsetter in entertainment AI, gaming optimizations, and social engagement algorithms.
SAP	Delivers AI-integrated enterprise solutions like SAP S/4HANA.	Innovates in new business models, incorporating AI to transform business processes.	Focuses on AI to drive intelligent enterprise resource planning and management.	Pioneering AI-driven ERP solutions, intelligent process automation, and enterprise-level business innovation.

models and communicating clearly how pricing adjustments actually benefit the customer. We highly recommend, based on our own experiences with GenAI, that companies build trust with customers by explaining AI-driven decisions and demonstrating value.

WHERE DO WE GO FROM HERE?

AI is fundamentally transforming cloud and SaaS by facilitating innovative business models. It enhances segment focus, automation, predictive capabilities, and dynamic scalability. AI-driven automation accelerates workflows in enterprise systems, like enterprise resource planning (ERP) and customer relationship management (CRM), while generative AI enhances creative tasks such as content production and software coding. Vertical SaaS will allow industry-specific AI, such as diagnostics in health care and demand forecasting in retail, travel, and transport. Predictive and data-driven innovations, like data as a service (DaaS), make curated datasets widely accessible, while developer productivity tools powered by AI are streamlining software development. Cloud AI services provide scalable machine learning tools with reduced in-house expertise. Multicloud AI deployments will offer flexible solutions for businesses needing scalability and redundancy in hybrid environments.

With AI being integrated with cloud infrastructures and SaaS, IT departments and companies across industries must challenge their existing business models. Master detective Sherlock Holmes observed that "it is a capital mistake to theorize before one has data." Today, we have the data, and AI helps us to better capitalize data. Go ahead, challenge your business models, and deliver innovative services. See "Guide to Grow" for additional take-aways and questions, to transfer some insights to your own business and technologies. ■

GUIDE TO GROW

TAKE-AWAYS FOR AI IMPACT ON SAAS

- AI will fundamentally reshape SaaS technology and business models. Examples include generative AI for content creation and AI-driven personalization; tailored customer experiences in e-commerce, CRM, and marketing; predictive analytics, such as forecasting sales, demand, and customer behavior in retail, finance, and health care; and AI-powered data automation in ERP, CRM, and robotic process automation tools.
- Vertical AI solutions will change traditional industries. Examples include AI for health care (diagnostics), logistics (route optimization), and law (contract review).
- Novel business models allow freemium services with advanced AI tools at a premium; subscription models based on AI resource usage, such as OpenAI application programming interface (API); outcome-based pricing with fees tied to measurable results like cost savings or growth, such as fraud detection and client base growth; and DaaS, such as selling access to tailored datasets and pretrained models.

TRANSFER QUESTIONS

- How can you enhance your technology offerings and products with generative AI to create a competitive advantage along content-driven SaaS platforms?
- How might AI-driven personalization conflict with data privacy regulations, and what solutions could a SaaS company implement?
- What vertical AI solutions could you deploy to create deeper customer loyalty compared to general-purpose AI tools?
- What freemium models can you imagine with the need to monetize advanced AI features?
- What outcome-based pricing models could you introduce, such as leveraging DaaS to build an ecosystem of partners and increase recurring revenue?
- Which subscription-based usage model for AI tools would give value to your company, such as customer retention?

REFERENCES

1. N. L. Rane, S. K. Mallick, O. Kaya, and J. Rane, "Machine learning and deep learning architectures and trends: A review," in *Applied Machine Learning and Deep Learning: Architectures and Techniques*, J. Rane, N. L. Rane, and S. K. Mallick, Eds., Nottingham, U.K.: Deep Science Publishing, 2024, pp. 1–38.
2. A. S. Pillai, "AI-enabled hospital management systems for modern healthcare: An analysis of system components and interdependencies," *J. Adv. Analytics Healthcare Manage.*, vol. 7, no. 1, pp. 212–228, 2023.
3. C. Ebert and P. Louridas, "Generative AI for software practitioners," *IEEE Softw.*, vol. 40, no. 4, pp. 30–38, Jul./Aug. 2023, doi: 10.1109/MS.2023.3265877.
4. C. Ebert and U. Hemel, "Grow your artificial intelligence competence," *Computer*, vol. 57, no. 10, pp. 144–150, Oct. 2024, doi: 10.1109/MC.2024.3436168.
5. S. Poonguzhali and A. Revathi, "AI-driven cloud computing to

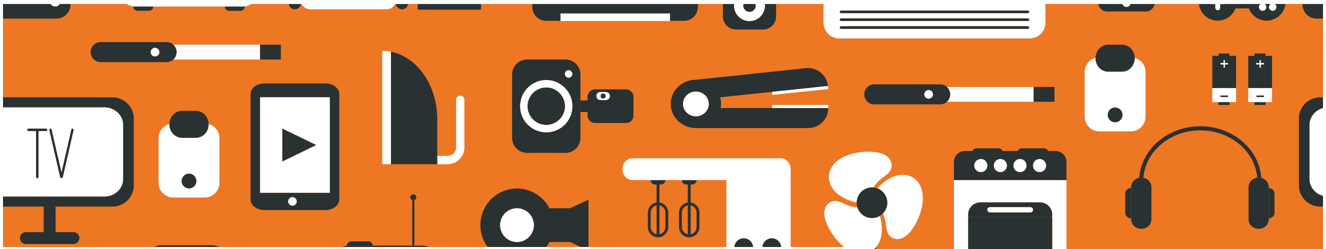
revolutionize industries and overcome challenges," in *Emerging Trends in Cloud Computing Analytics, Scalability, and Service Models*, D. Darwish, Ed., Hershey, PA, USA: IGI Global, 2024, pp. 395–410.
6. T. Chavan, "Optimizing customer value: The role of AI in the usage-based pricing model," *Forbes*, Dec. 21, 2023. [Online]. Available: https://www.forbes.com/councils/forbesbusinessdevelopmentcouncil/2023/12/21/optimizing-customer-value-the-role-of-ai-in-the-usage-based-pricing-model/
7. O. D. Segun-Falade, O. S. Osundare, W. E. Kedi, P. A. Okeleke, T. I. Ijomah, and O. Y. Abdul-Azeez, "Assessing the transformative impact of cloud computing on software deployment and management," *Comput. Sci. IT Res. J.*, vol. 5, no. 8, pp. 2062–2082, 2024.
8. E. Brusa, L. Cibrario, C. Delprete, and L. G. Di Maggio, "Explainable AI for machine fault diagnosis: Understanding features' contribution in machine learning models for industrial condition monitoring," *Appl. Sci.*, vol. 13, no. 4, 2023, Art. no. 2038, doi: 10.3390/app13042038.
9. "How no-code AI development platforms could introduce model bias." VentureBeat. Accessed: Jan. 7, 2025. [Online]. Available: https://venturebeat.com/uncategorized/how-no-code-ai-development-platforms-could-introduce-model-bias/

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

SRIRAM PANYAM is the chief architect at Omlet Inc., an observability startup, San Ramon, CA 94583 USA. Contact him at sri.panyam@gmail.com.

PRAVEEN GUJAR is a director of product specializing in enterprise data products for digital advertising at LinkedIn, Mountain View, CA 94043 USA. Contact him at praveen.gujar.s@gmail.com.

Digital Object Identifier 10.1109/MC.2025.3546135



Smart Farming for Poultry: Enhancing Growth and Efficiency With Low-Cost Internet of Things Solutions

Roberto Oliveira and **Carla Nébele Ferreira Hanau**, State University of Goiás

Valdemar Vicente Graciano-Neto and **Eliomar Araújo de Lima**, Federal University of Goiás

Vinícius Carvalho Lopes, Notre Dame University

José Maria David, **Regina Braga**, and **Wagner Arbex**, Federal University of Juiz de Fora

Mohamad Kassab, Boston University and New York University Abu Dhabi

This article investigates the impact of a low-cost Internet of Things system for autonomous environmental regulation in poultry farming, demonstrating its potential to optimize growth, welfare, and operational efficiency in small-scale production within the Brazilian agricultural context.

Digital Object Identifier 10.1109/MC.2024.3522703
Date of current version: 27 March 2025

Poultry production has steadily increased in response to the rising global demand for chicken meat. The Food and Agriculture Organization estimates that global chicken consumption will reach approximately 140 million metric tons by 2024.[1] This growing demand has spurred an increase in poultry producers at both the family and industrial scales. However, managing poultry in these environments presents significant challenges. Poultry are highly sensitive to environmental factors within poultry houses, such

as ambient temperature and relative humidity, which directly impact their health and productivity.[2]

For instance, excessively high temperatures can reduce feed intake by increasing thirst and decreasing appetite, ultimately lowering weight gain.[3] High relative humidity, on the other hand, weakens poultry and increases the risk of disease.[4] These challenges are particularly pronounced in tropical countries with consistently high temperatures as well as in northern countries that experience extreme cold for part of the year, complicating the environmental management within poultry houses.

While advanced technologies for monitoring and controlling these environmental factors have been developed, their high cost makes them inaccessible for many small- and medium-sized producers.[5] This financial barrier creates a significant gap between large-scale producers and smaller operators in terms of efficiency and productivity.

In this article, we present the results of a study evaluating a low-cost Internet of Things (IoT)-based information system for poultry management. Our goal is to determine how this system could automate the regulation of key environmental variables, reduce the effort required from producers, and ultimately improve weight gain in poultry. The system's architecture and technological configuration were carefully designed to address these challenges while maintaining affordability for small- and medium-scale producers.

THE IoT-BASED LOW-COST SYSTEM FOR POULTRY RAISING

The proposed low-cost, IoT-based system is designed to effectively improve poultry farming by integrating monitoring, automation, and data analytics into a unified, scalable architecture. This system comprises four core components (Figure 1): the Monitoring Device, the LoRa/Wi-Fi Gateway, the ThingSpeak Cloud, and a custom-built Web Platform. Each component has been meticulously selected and optimized to address the unique challenges of poultry farming, particularly in rural areas with limited connectivity and a critical need for real-time environmental monitoring.

System components

The system components are as follows:

- *Monitoring Device*: At the core of this architecture lies the Monitoring Device, which is responsible for capturing and transmitting key environmental parameters, such as light intensity, temperature, humidity,

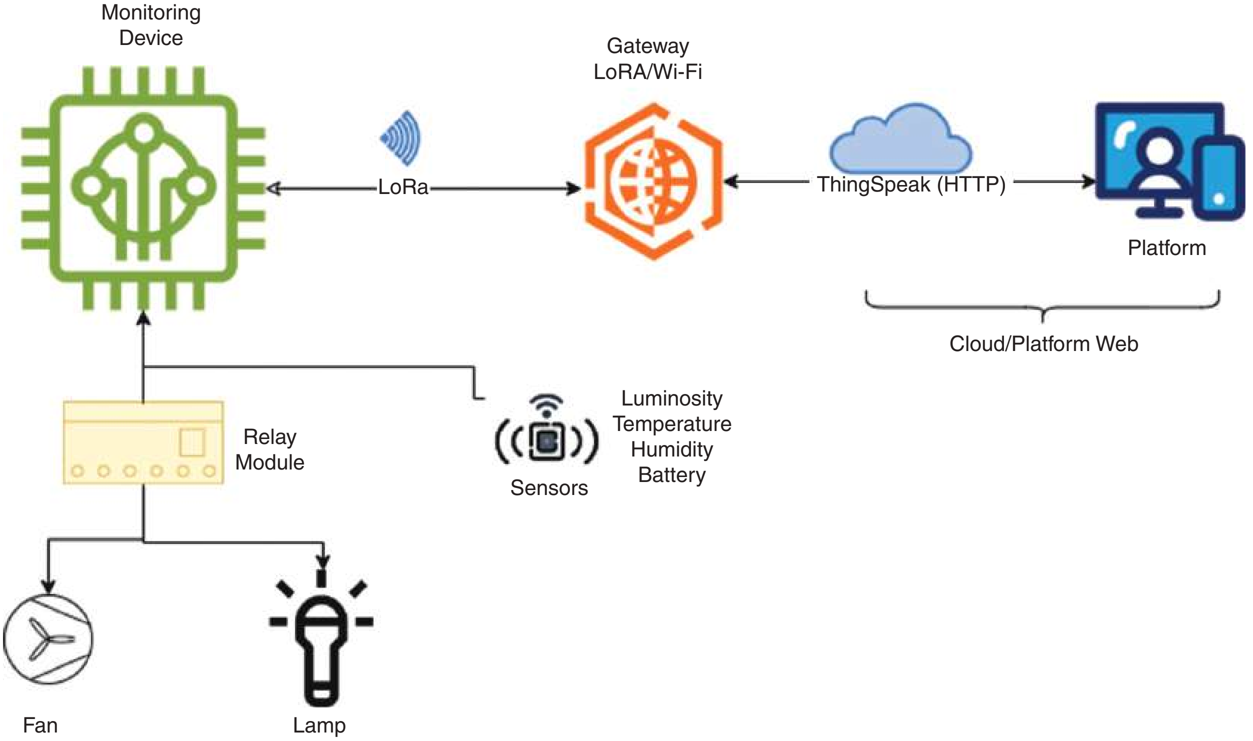


FIGURE 1. Design of the IoT-based system architecture; adapted from Lopes et al. 2021.[6]

and device battery levels. Built around the JARM ESP32 microcontroller—a cost-efficient platform equipped with integrated Wi-Fi and Bluetooth capabilities—the device includes a suite of sensors and communication modules. Specifically, the SHT20 sensor ensures accurate temperature and humidity readings, while a LoRa module (explained in the next paragraph) facilitates long-range data transmission. An SD card provides local data storage, ensuring redundancy in case of connectivity issues. To guarantee continuous operation, the device is designed with dual power sources: a dc plug and a battery holder, safeguarding functionality during power outages. The firmware, developed using Arduino IDE and C++, is compatible with a wide range of sensors and communication protocols and benefits from extensive community support, which simplifies both development and future expansion.

- *LoRa/Wi-Fi Gateway*: Given the frequent limitations in Wi-Fi coverage in farm environments, a LoRa/Wi-Fi Gateway bridges the Monitoring Device to the cloud. Also built on the JARM ESP32 platform, the gateway integrates both LoRa and Wi-Fi communication technologies. It receives data over long distances via LoRa from the Monitoring Device and transmits it via Wi-Fi to the ThingSpeak Cloud for further processing. This setup ensures reliable data transmission, even in areas distant from Wi-Fi access points, while maintaining the real-time availability of environmental data crucial for decision making.
- *ThingSpeak Cloud Platform*: The ThingSpeak Cloud serves as the system's data repository and analytics engine. As an open source IoT platform, it facilitates the collection, processing, and visualization of data transmitted from the farm. The platform supports standard communication protocols, such as HTTP and MQTT, making integration with various IoT devices straightforward. Real-time visualization of sensor data enables immediate insight into environmental conditions within the poultry house, allowing for rapid interventions when necessary to maintain optimal conditions for poultry health and productivity.
- *Custom Web Platform*: User interaction and data visualization are further enhanced through the development of a custom Web Platform. Constructed using JavaScript, HTML, Cascading Style Sheets, and WordPress plug-ins, the platform delivers a user-friendly interface displaying real-time sensor readings alongside control mechanisms for critical systems, such as fans and lighting. Users can set environmental thresholds (for example, temperature limits) and receive automated alerts when these thresholds are breached, enabling proactive environmental management. Additionally, the platform supports historical data visualization through graphs, which aid in trend analysis and informed decision making for long-term operational improvements.

The platform supports historical data visualization through graphs, which aid in trend analysis and informed decision making for long-term operational improvements.

Automation and control

The system implements automation via a 5-V, four-channel relay module, currently utilizing two channels to control a fan and a lamp. The relay module operates in response to signals from the Monitoring Device, based on real-time environmental data. This automated control ensures optimal environmental conditions within the poultry house, reducing manual intervention and enhancing both animal welfare and operational efficiency.

Design philosophy: modular and independent operation

A key strength of the system is its modularity and independent operation, ensuring resilience against individual component failures. For example, the Monitoring Device continues to collect and store data locally even if the LoRa/Wi-Fi Gateway or Internet connection fails. This autonomy is critical for maintaining uninterrupted monitoring and control functions, particularly in remote or underresourced farming environments. By preventing cascading system failures, the design enhances reliability, a key requirement for agricultural applications.

Technological choices and scalability

The system's architecture is built around principles of cost-efficiency, reliability, and scalability. The use of the ESP32 board within the Arduino ecosystem ensures affordability without sacrificing flexibility or performance, supporting a wide array of sensors and communication protocols. LoRa technology is chosen for its low-power consumption and long-range capabilities, which are particularly beneficial in the context of geographically dispersed farm environments.

The integration of open source platforms, like ThingSpeak, combined with widely used web technologies for platform development, reduces both initial costs and the barriers to future scalability.

Study conduction and key results

The present study sought to evaluate the efficacy of a cost-effective IoT application specifically designed to optimize poultry farming through the automation of environmental monitoring and control. The experimental research was conducted at a poultry farm located in Iaciara, Goiás, Brazil, utilizing 20 one-day-old, properly vaccinated chicks. These chicks were evenly divided into two distinct groups: one group of 10 chicks housed in an IoT-assisted environment (the monitored group) and a control group of 10 chicks without the benefit of such automation. The region's tropical savanna climate, characterized by temperatures fluctuating between 18 °C and 34 °C, served as a suitable setting for assessing the IoT system's capability to sustain optimal environmental conditions for poultry rearing.

The experimental infrastructure comprised stalls situated within the farm's poultry house, each measuring 1.9 m by 1.65 m. These stalls were enclosed with a 1-in-thick mesh, shielded by tarpaulin to offer protection against excessive wind and cold. All stalls were uniformly outfitted with heating lamps, consistently operating to maintain an average ambient temperature of 31 °C, which is crucial for ensuring thermal comfort in the early stages of chick development (Figure 2).

In the monitored stall, supplementary equipment was integrated with the IoT system, including an additional heating lamp and a fan. The secondary heating lamp was triggered automatically when the temperature dropped below the desired range, while the fan was activated upon exceeding the thermal threshold. This IoT-driven apparatus provided real-time environmental data and made necessary adjustments autonomously, thus minimizing the need for manual interventions and the constant physical presence of farm personnel.

The experiment ran from 23 August to 20 September 2022, during which time both groups of chicks were fed and watered twice daily for 28 consecutive days, in the early morning and late afternoon, ensuring nutritional parity. Oak wood bedding was utilized uniformly across all stalls to absorb moisture, manure, and feathers, and it was replaced on a weekly basis. The used bedding material was collected in labeled plastic bags, and we carefully documented the collection dates and corresponding stall information for future reference.

Notably, the monitoring procedures differed significantly between the groups. The control group was subjected to routine physical checks approximately four to six times daily to identify signs of thermal or environmental stress. Conversely, the monitored group benefited from continuous oversight through the IoT system, which provided automated environmental adjustments without the need for direct human intervention.

Upon conclusion of the experiment, the chickens were weighed to measure growth performance. The results were statistically significant: the IoT-monitored group achieved an average weight increase 25% greater than that of the control group. A p value of 0.045 affirmed the statistical significance of the difference, suggesting that the IoT-based system had a demonstrably positive effect on weight gain.

The results of this study indicate that the IoT-based environmental monitoring system was successful in maintaining optimal conditions, leading to enhanced growth rates in the monitored group. Automated environmental adjustments contributed to a healthier, less stressful environment, which in turn facilitated better feed conversion efficiency. Furthermore, the system's automation significantly reduced the need for manual monitoring, offering substantial labor savings and operational efficiency, which is

> The use of the ESP32 board within the Arduino ecosystem ensures affordability without sacrificing flexibility or performance.

FIGURE 2. A photograph of the real environment with the system deployed.

especially advantageous for small-scale poultry producers.

The integration of automated environmental control systems, when combined with the use of poultry breeds possessing high productivity potential, appears to generate synergistic benefits for overall production. In this study, the monitored chickens not only exhibited accelerated growth but also likely enjoyed superior health outcomes because of the stable environmental conditions maintained by the IoT system. Although the magnitude of improvement may vary across different growth phases and poultry species, the observed positive trend underscores the potential for enhancing poultry production through IoT-driven automation.

However, while the study demonstrates promising results, several avenues for future research emerge. First, further experimentation with larger sample sizes and extended monitoring periods would provide more robust data to generalize these findings across different poultry breeds and growth stages. Additionally, future studies could explore the adaptability of the IoT system to diverse climatic conditions as well as its effectiveness in managing other critical environmental factors, such as humidity, air quality, and lighting, which also impact poultry health and productivity.

Moreover, the integration of machine learning algorithms into the IoT system could enable predictive adjustments based on historical data, further improving environmental management by anticipating climate fluctuations or animal behavior. Research could also investigate the cost–benefit analysis of IoT implementation on various scales of poultry farming, offering insights into the economic viability for both small-scale and large-scale producers. ■

ACKNOWLEDGMENT

We thank FUNAPE-UFG for the financial support to conduct this research.

REFERENCES

1. “FAO: Global meat production to grow in 2024, led by poultry,” *AviNews.com*, Jun. 24, 2024. Accessed: Nov. 18, 2024. [Online]. Available: https://avinews.com/en/fao-global-meat-production-to-grow-in-2024-led-by-poultry/
2. V. V. Graciano Neto and M. Kassab, *What Every Engineer Should Know About Smart Cities*, 1st ed. Boca Raton, FL, USA: CRC Press, 2023.
3. V. V. Graciano Neto, M. Kassab, V. C. Lopes, R. Oliveira, and R. B. Neto, “The state of IoT for agribusiness in Brazil,” *Computer*, vol. 55, no. 12, pp. 140–144, Dec. 2022, doi: 10.1109/MC.2022.3205934.
4. Y. Xiong, Q.-s. Meng, J. Gao, X.-f. Tang, and H.-f. Zhang, “Effects of relative humidity on animal health and welfare,” *J. Integrative Agriculture*, vol. 16, no. 8, pp. 1653–1658, 2017, doi: 10.1016/S2095-3119(16)61532-0.
5. H. Wang, N. Bolan, M. Hedley, and D. Horne, “Potential uses of fluidised bed boiler ash (FBA) as a liming material, soil conditioner and sulfur fertilizer,” in *Coal Combustion Byproducts and Environmental Issues*, K. Sajwan, I. Twardowska, T. Punshon, and A. Alva, Eds., New York, NY, USA: Springer-Verlag, 2006, pp. 202–215.
6. V. Lopes, R. de Oliveira, and V. V. Graciano Neto, “Towards an IoT-based architecture for monitoring and automated decision-making in an aviary environment,” in *Proc. 13th Brazilian Congr. Agri-Inform. (SBIAgro)*, 2021, pp. 320–328.

ROBERTO OLIVEIRA is an associate professor at the Academic Institute of Technological Sciences at the State University of Goiás, Posse-GO 73902-000, Brazil. Contact him at roberto.oliveira@ueg.br.

CARLA NÉBELE FERREIRA HANAU is an alumna in agriculture of the State University of Goiás, Posse-GO 73902-000, Brazil. Contact her at carlanebelle@gmail.com.

VALDEMAR VICENTE GRACIANO-NETO is an associate professor at the Instituto de Informática Universidade Federal de Goiás, Goiânia 74690-900, Brazil. Contact him at valdemarneto@ufg.br.

ELIOMAR ARAÚJO DE LIMA is an assistant professor at the Instituto de Informática Universidade Federal de Goiás, Goiânia 74690-900, Brazil. Contact him at eliomar.lima@ufg.br.

VINÍCIUS CARVALHO LOPES is a research assistant in the Computer Science and Engineering Department at the University of Notre Dame, Notre Dame, IN 46556 USA. Contact him at vlopes@nd.edu.

JOSÉ MARIA DAVID is an associate professor in the Computer Science Department at the Federal University of Juiz de Fora, Juiz de Fora-MG 36036-900, Brazil. Contact him at jose.david@ufjf.br.

REGINA BRAGA is a full professor in the Computer Science Department at the Federal University of Juiz de Fora, Juiz de Fora-MG 36038-330, Brazil. Contact her at regina.braga@ufjf.br.

WAGNER ARBEX is a professor at the Federal University of Juiz de Fora, Juiz de Fora-MG 36038-330, Brazil, and a scientific analyst at Brazilian Agricultural Research Corporation, Brasilia, Brazil. Contact him at wagner.arbex@embrapa.br.

MOHAMAD KASSAB is an adjunct professor of computer science at Boston University and a visiting associate professor of computer science at New York University Abu Dhabi, Abu Dhabi 02215, United Arab Emirates. Contact him at mkassab@bu.edu.

EDITOR JAMES BRET MICHAEL
IEEE Fellow; bmichael@nps.edu

The Origins of Software Errors: Where Do They Come From? Where Do They Go?

John McHugh, Assurance Labs

This article explores the need to further investigate the origins and nature of software errors so that we can be adequately prepared to eliminate them.

Software errors enable many security exploits and are instrumental in many program failures. In spite of this, relatively little attention has been paid to their origins and nature. This article explores the surface of the problem, abstracting errors into two nondisjoint classes: blunders and communications failures. It looks at two developmental methods, N-version programming and Cleanroom, that might shed light on the problem, and it concludes with a discussion of an additional confounding issue, the increasing use of artificial intelligence (AI) to help generate software, and a plea for the community to raise the importance of understanding errors.

HOW LITTLE WE KNOW

Henry Petroski's classic book, *To Engineer Is Human*,[1] focuses on the role of errors in advancing engineering, noting that, through the analysis of failures, we advance our capabilities. Software engineering has yet to adopt a culture of error analysis and learning from failure, which may (at least partially) explain why, despite more than 50 years of large-scale programming, we are still unable to prevent errors from appearing in released code. One explanation may be that we do not understand how or why they occur. I have been interested in this problem for some decades and have made little progress. Recently, I revisited

Digital Object Identifier 10.1109/MC.2024.3524936
Date of current version: 27 March 2025

the problem and discussed it with a number of colleagues. In this article, I review several approaches that attempted to reduce software errors and suggest that the community revisit one of them, adding detailed analyses of the residual errors as they are discovered.

I have been writing programs and doing research in computer security and software engineering for government, industry, and academia since the

Despite its name, the field of software engineering lacks the corpus of *proven* practices that characterize most engineering disciplines.

1970s. This includes building data acquisition, data analysis, and program analysis and verification tools. I have written production code in Rust, C, Fortran, PL-I, Lisp, and various assembly languages. I have made my share of errors, but I believe that I have produced a few significant error-free programs.

Based on personal experience and the work discussed next, I divide software errors into two overlapping classes: *programming blunders* and *communication failures*. (A possible third class is due to broken language definitions—material for another article.) Programming blunders include typos, initialization errors, iterator mismatches (off by one), and so on. Many of the blunders are prevented or detected by better language designs, static analysis tools, and modern compilers. The latter class includes failures to understand or honor requirements, specifications, etc. It also includes cases where the program's client does not realize the limits of the programmer's expertise. Despite its name, the field of software engineering lacks the corpus of *proven* practices that characterize most engineering disciplines, although one might argue that the corpus exists, encoded in languages and tools in ways that render it generally inaccessible, per se.

In 1994, I gave a talk titled "Faith and Hope: Methodologies for Building Trusted Systems," which is one basis for this article. I gave updated versions of the talk for 15 years (including the one in 2003[2]), and I am currently resurrecting it. The talk notes that there is little or no evidence that most of the frameworks, methodologies, and even standards put forward as solutions to make programming easier, faster, more secure, and/or less error prone actually help. There was no evidence that applying the standards of the day actually produced software with fewer errors or better security than software built without them. This holds true to the present. Current standards, such as the IEEE 1012-2016 standard,[3] are unsupported by evidence that following them results in more secure or reliable software. Knowledge compendiums, such as the Software Engineering Body of Knowledge (SWEBOK),[4] are, effectively, catalogs of ways to do things, but they provide no advice that would allow one to choose the best way to meet a specific reliability or security goal. (It is interesting to note that the uniform resource locator referenced in the citation seems to explicitly disclaim any prescriptive role for the SWEBOK, deferring to the decisions of the management team for any given project.) We have some evidence to support the view that reducing errors improves security in the multitude of Common Vulnerabilities and Exposures entries that attribute specific vulnerabilities to software errors.

EXPERIMENTS WITH TWO PARADIGMS

Little proper experimentation is done in software engineering. Small-scale experiments seldom produce relevant results. Significant software systems are expensive and time consuming to build, and we cannot justify the cost of building a system several ways to see which is better. Even if we could, controlling for all of the relevant factors is likely to be impossible. Some experimental work, essentially case studies, has been done with two paradigms, *N-version programming* and the *Cleanroom*. Both provide some insights into the origins of software errors, but, as we will see, they are not very satisfying.

N-version programming

N-version programming was inspired by the success of redundant hardware in surviving component failures, but it requires that replica failures be randomly distributed in the input space rather than in time. This means that failures of multiple replicas on the same input, coincident failures, must be very rare.[5] Two experiments, the Knight–Leveson experiment (KL)[6] and the subsequent Four University experiment (4U),[7] showed that coincident failures negated the potential benefits of N-version software. They also provided some limited insights into software error origins.

KL developed 27 versions of a ballistic missile interceptor launch program. The article gives a reasonable accounting of the faults that resulted in coincident failures but does not discuss the remaining faults. 4U developed 20 versions of a sensor management program for an inertial navigation platform similar to those used in commercial aircraft of the time. There were seven faults that manifested a total of 16 times. Only one version exhibited a software blunder, the use of an uninitialized variable.

The development methodologies in the two experiments differed, but both isolated the replica developers in the hope of avoiding common errors. In both cases, the coincident failures were examined and reported, but the remaining failures were not. Most of the coincident failures in both experiments seem to be due to either misunderstanding of the provided specification or failure to follow explicit instructions in the specification. In both experiments, most of the programs produced were of very high quality (six KL and 10 4U replicas had no failures during testing), but the potential reliability gain was limited by coincident failures.

Cleanroom software engineering

Most programming paradigms test at the unit, subsystem, and system levels, consuming much of the development effort. The objective of Cleanroom development is to approach zero defects prior to code execution by using lightweight formal methods and structured team reviews. Initial execution and testing occur after (sub)system integration. The Cleanroom methodology,[8] developed by Harlan Mills and others at IBM in the 1980s, depends on developing formal specifications for the system to be constructed, creating a top-level design from them and refining the design, top down, until the final system code is produced. Each step of the refinement is accompanied by a proof that the refinement is correct.

This development hierarchy is the heart of Cleanroom. However, there is another key aspect: system testing driven by a realistic simulated workload and guided by statistical measures, for instance, reaching an acceptable failure rate. Faults discovered during testing are repaired using rerefinement and reproof and repeated testing until the reliability goal is reached. Faults found this way seem to be easier to remediate than those missed in low-level tests. Case studies showed that software developed with Cleanroom techniques had comparable costs and schedules and up to an order-of-magnitude fewer faults at release than conventionally developed software. In spite of this, Cleanroom did not achieve widespread adoption. Programmer pushback may be one reason[9] and it seems to be unused today. The reduction in faults is impressive, but there are no data on the causes of those that remain.

WAYS FORWARD

Both N-version results indicate that the programmers did not understand crucial aspects of the application domain (trigonometry, nonorthogonal coordinate systems, and floating-point comparisons). This is a communications problem. The specifier(s) overestimated the knowledge of the implementers. It also indicates a weakness in programmer education. The details of computations that are critical to many applications areas are not covered in the required courses of many computer science curricula. I have long thought that programmer education concentrates too much on "computer sciency" things and not enough on solving problems that appear in the real world. A strong computer science minor that concentrated on giving people from applications disciplines the programming abilities needed to solve their own problems (or at least to communicate them better to professional programmers) would be useful. The core lesson from the N-version experiments is that programmers who do not understand the application domain will make errors.

> The core lesson from the N-version experiments is that programmers who do not understand the application domain will make errors.

Why does Cleanroom reduce errors? Does the proof process find and remove errors in the code? Does the discipline and thought required to develop provable code make the programmer less error prone? My experiences with machine and manual verification indicate the latter is more likely, though the answer could lie elsewhere. There are no relevant data. I contacted Victor Basili[10] in 2022, and he replied: "... I would tend to agree with your assumption that thinking about correctness while programming is more important than proving the code correct afterward." In 2024 May, David Parnas[11] agreed that the thought that goes into creating a provable program likely results in error reduction.

Cleanroom and other specification- and design-intensive methodologies force this thought up front. Reviving it and collecting detailed error data might help. In general, collecting good data would require a shift in programming culture. Programmers seem to enjoy the cycle of code, test run, debug, code. Both the open/crowd source programming paradigms and the Git release paradigm appear to value code production over design and analysis. Most tests are exploratory, that is, directed toward discovering behavior, rather than validatory, confirming expected behavior. Cleanroom limited programmer activities to design, refine, prove, review cycles with proofs replacing unit testing. No fun there. Most programmers are not good at introspection, but we might be able to train them to be more aware of what they are doing so that they can report it—a substantial departure from the usual developer culture and contrary to many current methodologies and practices. Given the opportunity, the rewards could be substantial.

BUT WHAT ABOUT AI?

Programming and also engineering are largely human activities that rely on unambiguous communications. Discovering the causes of programming errors requires observing and working with programmers. The use of AI to perform or support programming may change this. The use of AI in this way is evolving so rapidly that any observations made today (mid-December 2024) may be irrelevant when this article is published in April 2025. In addition, results obtained using an AI specifically trained to aid programmers, such as GitHub Copilot, are likely to be different from those from an AI trained on a general large language model (LLM), for example, one of the many ChatGPT versions. In early 2024, a student I worked with discovered that, if an example matching the prompt was at the training, it was returned verbatim. Some samples would not even compile. If there was no obvious answer, the answer was synthesized and often had fairly obvious errors. Recently, I have seen fairly complex programs synthesized from carefully designed prompts. They appear to compile and run successfully. There is a growing body of information on the successful "engineering" of prompts. GitHub has a large amount of tutorial and other material; no specific citations are given because the material appears to be changing rapidly. Recent work involving chain-of-thought prompting appears promising in this area.[12] Kevin E. Valakuzhy[13] of Georgia Tech recently built a simple server using ChatGPT o1-mini (see https://platform.openai.com/docs/guides/reasoning/advice-on-prompting?reasoning-prompt-examples=coding-planning). Since o1-mini appears not to work well with previous prompting techniques, he did not give much thought to prompt engineering but started with basic functionality and then requested added features. Several errors resulted when the AI made what may have been a valid interpretation of the prompt but not one that he had considered. A more serious flaw occurred when the AI implemented a request for an "asynchronous" solution using threads in an environment in which threads cannot run in parallel. More disturbing is the fact that as the complexity of the application increased with the addition of new features, the effort of developing prompts to generate the required code exceeded the effort required to code the solution directly. This, in turn, caused additional problems as the code, as modified by a human programmer, was out of sync with the LLM, and subsequent modifications by the AI required starting a new conversation supplying only the portions of the code thought relevant to the feature request. While these are "communications errors," careful studies and effort will be required to understand and avoid them.

In a comic strip published on Earth Day in 1970, the fictional character Pogo observed: "We have met the enemy and he is us."[14] We will not eliminate software errors until we truly understand where they come from. An effort to do this will require careful examination of the causes of software errors. Only when we understand them can we begin to eliminate them. Until that time, we are stuck in a cycle of debug, patch, and debug some more. ■

ACKNOWLEDGMENT

Over the years, many people have helped me develop my thoughts on this topic. Among them in the early days were Dave Parnas, Harlan Mills, Vic Basili, John Knight, Nancy Leveson, Don Good, and James Jerome Jenkins. More recently, Mike Collins, Rick Linger, Charles Martin, and Terry Benzel have encouraged this effort. My thanks to you and all the others who have expressed interest in the problem.

REFERENCES

1. H. Petroski, *To Engineer is Human: The Role of Failure in Successful Design*. New York, NY, USA: St. Martin's Press, 1985.
2. J. McHugh, "Faith and hope revisited: Rethinking methodologies for building trusted systems," Apr. 2003. [Online]. Available: https://www.umiacs.umd.edu/docs/FaithandHopeRevisited.pdf
3. *IEEE Standard for System, Software, and Hardware Verification and Validation*, IEEE Standard 1012-2016 (Revision of IEEE Standard 1012-2012/Incorporates IEEE Standard 1012-2016/Cor1-2017), pp. 1–260, Sep. 29, 2017, doi: 10.1109/IEEESTD.2017.8055462.
4. H. Washizaki, "Guide to the software engineering body of knowledge (SWEBOK Guide)," *IEEE Comput. Soc.*, version 4.0 ed., 2024. [Online]. Available: https://www.computer.org/education/bodies-of-knowledge/software-engineering
5. D. E. Eckhardt et al., "An experimental evaluation of software redundancy as a strategy for improving reliability," NASA Tech. Memorandum 102613, NASA Langley Research Center, Hampton, VA, USA, May 1990. [Online]. Available: https://ntrs.nasa.gov/api/citations/19900014642/downloads/19900014642.pdf
6. S. S. Brilliant, J. C. Knight, and N. G. Leveson, "Analysis of faults in an N-version software experiment," *IEEE Trans. Softw. Eng.*, vol. 16, no. 2, pp. 238–247, Feb. 1990, doi: 10.1109/32.44387.
7. L. Lauterbach, "Development of N-version software samples for an experiment in software fault tolerance," NASA Contractor Rep. 178363, NASA Langley Research Center, Hampton, VA, USA, Sep. 1987. [Online]. Available: https://ntrs.nasa.gov/api/citations/19870020663/downloads/19870020663.pdf
8. H. D. Mills, M. Dyer, and R. C. Linger, "Cleanroom software engineering," *IEEE Softw.*, vol. 4, no. 5, pp. 19–25, Sep. 1987, doi: 10.1109/MS.1987.231413.
9. R. Platek, private communication, 1990.
10. V. Basili, private communication, 2022.
11. D. Parnas, private communication, May 2024.
12. J. Li, G. Li, Y. Li, and Z. Jin, "Structured chain-of-thought prompting for code generation," *ACM Trans. Software Eng. Method.*, Aug. 2024. [Online]. Available: https://doi.org/10.1145/3690635
13. K. E. Valakuzhy, private communication, 2024.
14. W. Kelly, *Pogo: We Have Met the Enemy and He is Us*. New York, NY, USA: Simon & Schuster, 1972.

JOHN McHUGH is a computer scientist consulting on a part-time basis with Assurance Labs, Gaithersburg, MD 20886 USA. Contact him at mchugh@cs.unc.edu.

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

The Software Bill of Materials

Dirk Riehle, Friedrich-Alexander-Universität Erlangen-Nürnberg

Cybersecurity threats and software supply chain attacks are at an all-time high. Customers and agencies keep tightening the requirements for their software. An important recent development is the practical use of software bills of materials.

A bill of materials (BOM) is a list of components ("materials") that make up some artifact. A software BOM (SBOM) is a BOM where all the components are software components. It is important to have an SBOM for a software project or product that is as complete and correct as possible, for two main reasons.

- *Critical data structure:* Complete and correct SBOM data are critical for a host of engineering functions. The three most important functions that require an SBOM are
  - *Open source governance:* To decide which components are acceptable to a project or product, you first need to know whether they are included and then what licenses and other conditions they come with.
  - *License compliance:* To deliver your project to clients or your product to customers, you need to comply with the licenses of any open source code included in the software. The SBOM tells you what those are.
  - *Security and vulnerability management:* To manage operational risk, you need to understand what components are doing their job in the given software, whether there are known vulnerabilities, and whether new vulnerabilities have been discovered.
- *Non-functional requirement:* A complete and correct SBOM has become a purchasing requirement of many customers.

Originally, large customers in a software supply chain would require from their suppliers that they provide SBOMs together with any software they were supplying. In the case of custom software projects, large customers would even request to receive signoff authority on the use of open source components before they were incorporated into the software being built.

Digital Object Identifier 10.1109/MC.2025.3530276
Date of current version: 27 March 2025

FROM THE EDITOR

Welcome back! This month's "Open Source" column continues discussing open source use in and by organizations. We turn to a fundamental data structure that anyone developing or using project or products built from open source needs: the software bill of materials, that is, an inventory of components in software. This data structure has become so important that governments have made it a requirement of professional software.—*Dirk Riehle*

In 2020, the European Union (EU) announced the Cyber Resilience Act (CRA).[a] This regulation complements the previous NIS-2 legislation to improve product security across the EU. The CRA entered into force in 2024. Vendors of products that include software are required to provide an SBOM to customers as well as to proactively track and respond to any vulnerabilities that become known about their products.

In 2021, the U.S. American government issued an executive order requiring, among other things, that any U.S. federal purchaser of software be provided an SBOM for the software being purchased.[b] While previous motivations for an SBOM were mostly about license compliance, the U.S. government cares more about cybersecurity and the risks from vulnerabilities in software. It is safe to assume that other governments will follow suit.

For any given software, the SBOM needs to list the original code of the supplier, presumably with their proprietary license, as well as any third-party components. A third-party component is any code, including open source code, not owned by you. Such third-party components come in two main forms.

1. Standalone components are the traditional libraries and components you are including in your project or product.
2. Code snippets are chunks of source code that have been copied and pasted into your code or into the third-party code you are using.

The two prominent (and competing) specifications for representing SBOMs are the SPDX and the CycloneDX specifications. These specifications allow the presentation of an SBOM in a linear format (list) of records with each record representing a component and some of its metadata.

There are different types of SBOM, created for different purposes. The most common SBOM is the one given to customers as a part of selling a product. Other types of SBOMs add tooling information to document how the software is being built or include verification information to comply with regulatory requirements.

To create an SBOM, you need to identify and gather all third-party components your code is using, whether a standalone component or a code snippet. For any such component, you need to gather the necessary metadata for each component.

It is impossible to create a complete and correct SBOM for a nontrivial piece of software. Too much copy and paste without tracking lineage in both open and closed source software has ruined this opportunity.

THE DEPENDENCY GRAPH

The process of creating an SBOM is called *software composition analysis* (*SCA*). An SCA first creates the so-called dependency graph of your software and then derives the SBOM from it.

A dependency is a software component that some other component depends on. A component depends on another component if the component can't perform its function without the depended-on component. In the common case, this is a code dependency, like being able to call the functions of the depended-on component.

A dependency graph is a graph of software components as the nodes connected by depends-on (dependency) relationships as the edges (links). In any modern software, most of these components will be third-party components, including open source components, which are components owned and licensed to you by someone else.

There are many different types of components that can become nodes in a dependency graph, depending on how broadly or narrowly the dependency graph is to be used.

- In the original narrow sense, the components in a dependency graph are all code components. There are two types of components.
  - *Traditional standalone components or libraries:* These are components that have a clear boundary with their context (they come as their own package, ideally with a well-defined interface).
  - *Code snippets:* Code snippets are pieces of code that have been copied and pasted into your code by your developers or into open source dependencies by the open source developers. Legally speaking, such code snippets are components separate from the embedding component because they usually have a different copyright holder and a different license.
- In a more recent broader sense, with the goal of completely documenting everything that goes into the building of

[a] https://digital-strategy.ec.europa.eu/en/policies/cyber-resilience-act.
[b] https://www.nist.gov/itl/executive-order-14028-improving-nations-cybersecurity

software, components can also be tools that build the software, resources that provide the necessary information, etc.

A dependency graph is a directed graph; incoming links to a component originate from other components that depend on this component, and outgoing links from a component go to the other components that this component depends on. As a matter of good software architecture, the graph is ideally also an acyclic graph.

Dependencies have levels. The level number is the number of steps removed from the root of the graph. This leads to the following definitions:

- The root component of a dependency graph has the level zero and is usually your own original code. There may be one or more root components.
- The first-level dependencies are the immediate dependencies of the root component. They are noteworthy because they are present in the minds of your developers and they are explicitly specified in your build system instructions. They are also often called the *direct dependencies*.
- Second- and higher-level dependencies are the dependencies of your first-level dependencies. They are also called *indirect dependencies*. They are noteworthy because they are not present in the minds of your developers and they are not very visible in their day-to-day work. Yet they constitute the largest part of the code that your project or product is built from.

> The two prominent (and competing) specifications for representing SBOMs are the SPDX and the CycloneDX specifications.

As a rule of thumb, the size relationship between your original code, your direct dependencies, and your indirect dependencies is one to nine to 90 in parts. In other words, 90% of your vulnerabilities stem from code you are not thinking much about. The indirect dependencies are the proverbial iceberg under the waterline.

SBOMs are created from a dependency graph. The nodes of a dependency graph correspond to the component entries in the SBOM. While the dependency graph remains a graph structure, the SBOM drops the relationships and is (mostly) a flat list of components. For this reason, the dependency graph and SBOM are not the same. Figure 1 shows a dependency graph, including our term definitions.

SCA

SCA is the analysis of your project or product's source code to identify the component structure of the software, also known as its *dependency graph*. As discussed, components may be standalone components, or they may be code snippets. The code of a component may be owned by you or by a third party, then called *third-party code*. Open source code is the most prominent example of third-party code.

The main motivation for SCA, originally, was to ensure license compliance. Any third-party code is legally separate code that comes with its own licenses. You need to comply with these licenses when you are delivering your projects to clients and delivering your products to customers.

Legally separate does not necessarily mean technically separate. Most notably, source code snippets that have been copied into your source code or into your dependencies are

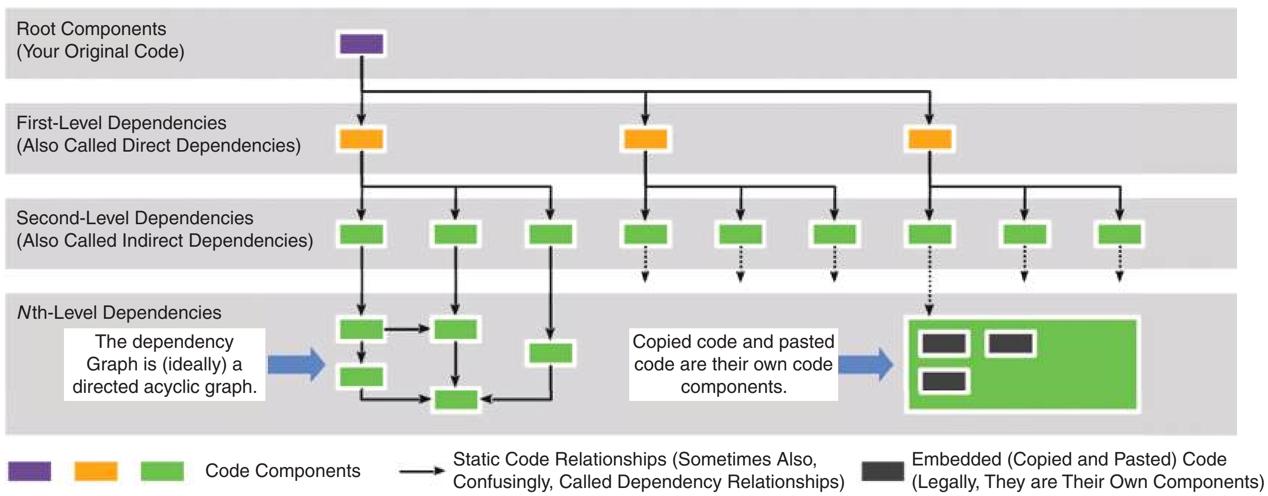


FIGURE 1. Illustration of a dependency graph.

legally separate code components even though they are embedded in your code or third-party code. You still need to identify these snippets, even in your dependencies, if you want to deliver license-compliant software.

SCA is typically performed using specialized tools. These tools read through the whole source code base of the software and try to identify any third-party code. An SCA tool needs access to the full source code, so in addition to providing your original code, you also have to either download the dependencies yourself or direct the tool on how to do so.

An SCA first creates the so-called dependency graph of your software and then derives the SBOM from it.

Examples of open source SCA projects are FOSSology,[c] a complete solution, and ScanCode,[d] a scanning tool to be embedded into a larger custom toolchain. An example of a free but comprehensive service is SCA Tool.[e] In addition, there are many commercial tools on the market.

For an SCA tool, the software consists of a hierarchical folder structure with files and code snippets in files. Source code outside this folder structure is not considered. An SCA tool does not and should not make assumptions about the folder structure mapping to the dependency graph in a particular way. As the result of an SCA, you will be presented with the folder, file, and snippet structure rather than the dependency graph. An export of this information in SBOM form will provide a flat list (rather than a graph).

SCA is not a fully automatic process. Existing SCA tools will analyze the source code and present their findings to their users for signoff. The key findings presented to users are

- *Component identification:* For a given software component, an SCA tool will suggest a specific origin component, ideally using a unique component identifier like a package URL (PURL).

For a given code snippet, an SCA tool will also suggest the origin component and to this add the location of the source code within the component that the snippet may have been copied from.

- *Legal information:* Originally designed for license compliance, SCA tools will try to determine the component's legal information: which licenses, which copyright holders, and any other notices that a user needs to know about.
- *Vulnerabilities:* More recently, SCA tools started adding known information on vulnerabilities, though this is often considered a follow-on step in a toolchain and not part of SCA.

In addition to source code analysis, binary analysis tools let you analyze the software composition of binary files. Binary files can be found anywhere; they might be hiding in a source code folder or be part of a container image. Like source code, they need to be found, identified, and analyzed.

WORKING WITH SCA TOOLS

SCA is a tool-based process that cannot be fully automated. An SCA tool expects or downloads a hierarchical structure of all relevant artifacts. Typically, this is a folder hierarchy of source code files pulled from version control, but it can also be container images with nondescript binary files included.

Different SCA tools naturally provide different functionalities. At its core, however, there are three different types of source code analysis results that an SCA tool might provide to its users (not all tools do).

- *The dependency graph:* A core output is the actual dependency graph (not just the SBOM). This includes correctly identifying both components and their linkage (forming the components into a graph).

Modern package managers have made it easy to determine a software's dependency graph, but many older software systems written in languages without established package managers resist any automation of creating the dependency graph. Package managers help SCA tools identify a component. The metadata provided by package managers, for example, component licenses and owners, is more often incorrect than not.

- *Meta-data from source code analysis:* Another core output of an SCA tool is the analysis of the source code. Most commonly, SCA tools look for legal information to help users ensure license compliance.

Identifying legal information is commonly performed in a simple and straightforward way by using regular expression matching against defined terms and databases like license text databases. Code quality analysis and identifying unknown vulnerabilities are also useful analysis functions available in some tools.

- *Snippet matching of your and third-party source code.* The final core output of some SCA tools is the identification of code snippets that may have been copied from the web into your code or any third-party code, including open source components.

Free-to-use open source SCA tools usually don't offer a snippet matching

[c]https://github.com/fossology/fossology.
[d]https://github.com/aboutcode-org/scancode-toolkit.
[e]https://scatool.com.

feature because to perform this function, the tool needs to compare any code snippet against the whole wide world of third-party code. This requires the creation and continuous updating of a large database of such third-party code, which can become rather expensive.

A tool like SCA Tool works through the artifact hierarchy and collects its findings for review and signoff by its users. It is not advisable for users to just accept what an SCA tool is suggesting. More often than not, the findings will be wrong. To this end, SCA tools provide users with a workflow in which they can review each finding for correctness.

There are many challenges to a human review.

- *Erroneous data:* An SCA tool may pull in erroneous data, for example, from package managers. Users need to review and correct these data.
- *Laborious process:* The developers of an SCA tool typically don't want to be on the hook for overlooked third-party code. Hence an SCA tool is set to be highly sensitive, often suggesting third-party code, in particular copied and pasted snippets, where there is none. This leads the tool to declare a large number of findings, many of which, if not most of them, will be false positives. Working through all these findings is a significant time sink for SCA tool users.
- *Error-prone process:* The review process is highly error prone because it is mind-numbingly boring. Reviewers have to work through a large set of findings, many of which are similar and repetitive yet may vary in minor but important details. As humans work, attention may wane, and a desire to move forward will get its way, leading to sloppy work and, ultimately, errors in the analysis and review process.
- *Expensive review:* The review is often delegated to the original developers, who would rather be writing new code and shipping features than reviewing old code and cleaning up legal debt. Using your developers to review SCA tool findings is rather expensive labor and often better to be delegated to third parties.

Creating a dependency graph and deriving the SBOM for the first time, therefore, is often a laborious, expensive, and error-fraught process. Ideally, changes to your project and product lead only to an incremental adjustment of the dependency graph and SBOM data.

BASIC SBOM REQUIREMENTS

An SBOM captures which code components are included in the software. There are two original uses.

- The first use of the SBOM information is to ensure that only code components that both the developer and any recipient would find acceptable were included; most notably, developers generally prefer to keep copyleft-licensed components out of their products.
- The second use of this information is to create proper legal notices for the third-party code in the software. A developer, when distributing the software, has to provide these legal notices about the included open source components to comply with their licenses.

Customers in a supply chain often make the provision of an SBOM a purchasing requirement, as discussed before. Governments have followed suit, mostly driven by the need to make software more secure.

A report by the U.S. Department of Commerce details the basic requirements for an SBOM.[f] Any SBOM should name its author and the time it was created. Each component (material) in an SBOM should provide the component's name, its version number, and the supplier of the component. Interestingly, the report also states that the component should list its relationship to other components, which I would have considered helpful but not critical.

The report sees SBOMs as hierarchical structures. At the root is the SBOM for the software being described. The components in the SBOM can then have their own SBOM, potentially creating a hierarchical structure. You cannot, however, map the dependency graph into a hierarchy, at least not without creating significant redundancy. I argue that the components in an SBOM should simply be captured as a flat list; if preserving the dependency graph is important, each component can reference the components it depends on.

Also, an SBOM should be machine readable for automated processing. The report lists SPDX, CycloneDX, and SWID tags as established format specifications for capturing SBOM information. The report notes that the industry so far has failed at providing unique identifiers for components and that the supplier and component names should therefore be human readable, for human interpretation, but not necessarily machine interpretable.

The grassroots PURL effort is offering help to uniquely identify components.[g] The supplier of the component and its name (and version number, etc.) are encoded into one heterogeneous name value, the PURL. It consists of seven components structured using the following syntax:

```
scheme:type/namespace/name@
version?qualifiers#subpath
```

While not directly a traditional URL, a PURL uniquely nevertheless identifies a location. The location then becomes the supplier of the component.

[f] https://www.ntia.doc.gov/report/2021/minimum-elements-software-bill-materials-sbom.

[g] https://github.com/package-url/purl-spec.

Therefore, identical copies of the same code base in different locations are treated as different components.

An SBOM that fulfills these basic requirements can already be delivered with the software to its users to fulfill a purchasing requirement. That said, there are many more types and uses of SBOMs.

TYPES AND USES OF SBOMs

The original and still primary use of an SBOM is to list what components are included in the software when provided to a customer. As explained, there are two main uses.

1. *Governance and compliance:* Large companies in a supply chain wanted and still want to know what components they have to deal with, often in advance of a delivery. The primary reasons are open source governance (ensuring that only desired components are included) and license compliance (making sure that the licenses can be complied with).
2. *Security:* More recently, fueled by worries about cybersecurity, governments, including the United States and the EU, have put forth requirements that any product is to come with an SBOM. This way, users can identify any security issues with the product as vulnerabilities of the included components become known. SBOMs have become a purchasing requirement.

This is not the only type of SBOM; there are several more. The most prominent classification of SBOM types is provided in a 2023 white paper by the U.S. Cybersecurity & Infrastructure Security Agency (CISA).[h] CISA identifies six different types of SBOM, which can be broken down into two sets of three depending on how the SBOM is created.

- The first set is SBOMs created from the supplier's development process.
  1. *Design:* A Design SBOM is created from planning documents like prospective product architectures. As a consequence, a Design SBOM may not be an accurate reflection of what will be shipped eventually. It may be helpful to buyers in a supply chain to prepare for what's to come their way.
  2. *Source:* A Source SBOM provides a static picture of the supplier's source code and its dependencies, as found in the repositories. It can be helpful to identify vulnerabilities but does not provide a complete picture as it omits any build or runtime dependencies.
  3. *Build:* A Build SBOM is created from the build process of the supplier as it compiles source code and assembles the final package for delivery to customers. Aimed at operations, it does not include components needed for building and testing. It may still miss dynamic dependencies, though.
- The second set is SBOMs created by the buyer (or others) through analysis.
  1. *Analyzed:* An Analyzed SBOM is created from SCA of the static delivered software. This is almost always a binary analysis of the artifact. As such, an Analyzed SBOM will miss much, but it may discover components that the suppliers may have overlooked.
  2. *Deployed:* A Deployed SBOM is created by analyzing the deployed software. After deployment, additional components may have been loaded or may have become visible that were not identifiable before. Like Analyzed, Deployed SBOMs complement the supplier's SBOMs.
  3. *Runtime:* A Runtime SBOM is created from observing the running software (often requiring instrumentation). Of the SBOMs created through analysis, a Runtime SBOM provides the most comprehensive picture, but it will miss components that have not been activated and are not visible yet.

The original type of SBOM mentioned in the beginning corresponds to the Build SBOM created by the supplier. Other SBOM types have other uses; for example, the developer may want to track and document details of testing and staging their products for various reasons—for example, debugging, auditability, or certification.

There is a logic of progression in the two classes of SBOMs. A Build SBOM is by and large more comprehensive and more accurate than a Source SBOM than a Design SBOM, and a Runtime SBOM is by and large more comprehensive and more accurate than a Deployed SBOM than an Analyzed SBOM.

Both classes complement each other; SBOMs created by the supplier may miss some dynamically loaded components, knowingly or unknowingly, and SBOMs created by user analysis may miss some or many of the components that the system has not yet run into or that were obscured otherwise.

Taken together, a Build SBOM and a Runtime SBOM can provide a comprehensive picture, one that is needed for safe and secure operations. of software by organizations of any size. C

DIRK RIEHLE is the professor for open-source software at Friedrich-Alexander-Universität Erlangen-Nürnberg, 91058 Erlangen, Germany. Contact him at dirk@riehle.org.

[h] https://www.cisa.gov/sites/default/files/2023-04/sbom-types-document-508c.pdf.

EDITOR NIR KSHETRI
The University of North Carolina at Greensboro;
nbkshetr@uncg.edu

©ISTOCKPHOTO.COM/SITTIPONG PHOKAWATTANA

From Predictive and Generative to Agentic AI: Shaping the Future of Marketing Operations and Strategies

Nir Kshetri, The University of North Carolina at Greensboro

This article explores how marketing is being redefined by agentic artificial intelligence (AI), discussing the integration of agentic AI across marketing research, strategy, and actions, emphasizing its potential to enhance efficiency, personalization, and decision making within organizations.

While many organizations are still experimenting with predictive artificial intelligence (PAI) and generative AI (GAI), early adopters are advancing with smarter, cognitively superior AI agents, surpassing the capabilities of current AI bots.[1] As of the second half of 2024, a Capgemini survey of 1,100 executives revealed that 10% of large enterprises had already implemented AI agents. More than 50% planned to adopt them within the following year, and 82% anticipated integrating them within three years.[2]

Unlike AI assistants, which follow user commands, autonomous agents initiate actions based on events, handling processes like customer e-mail responses and inventory checks independently.[3] Vendors like Microsoft, SAP, Salesforce, Slack, and Meta are increasingly focusing on the value of AI agents rather than AI assistants, highlighting their potential to autonomously manage complex tasks and enhance business efficiency.[4] In October 2024, Microsoft introduced 10 new autonomous agents in Dynamics 365 to support sales, service,

Digital Object Identifier 10.1109/MC.2025.3530304
Date of current version: 27 March 2025

finance, and supply chain teams. The company plans to release more agents in the future.[5] In December 2024, Salesforce reported closing 200 deals for its agentic AI (AAI) platform Agentforce within a single week, with thousands more deals in progress, highlighting strong market interest in its AAI platform.[6] Agentforce has been implemented by Accenture, IBM, and Indeed, showcasing its utility in streamlining workflows through AAI.[7] These trends point to the fact that AAI is evolving into digital labor, allowing companies to automate and streamline complex workflows across multiple departments.[8]

Among organizational functions, marketing stands out as a striking example of transformation through AAI. This article thus explores the transformative impact of AAI on marketing strategies, highlighting its ability to automate tasks, enhance personalization, and improve decision making. It compares the roles of predictive, generative, and agentic AI in optimizing marketing functions, offering insights into the evolving role of AAI in driving efficiency and customer engagement.

In 2022, marketing spending in the United States reached US$481 billion.[9] Specifically, this article is expected to improve our understanding of the powerful economic impact AAI is likely to have on the marketing industry and market.

EMERGENCE OF TURNKEY, NO-CODE AI AGENTS

A notable trend is the emergence of turnkey, no-code AI agents that of simplifying the process of creating and training AI for sales interactions. These agents, designed for complex conversations, automate the lead-to-opportunity journey, allowing marketers to focus on customer engagement rather than setup. With ready-to-use features, these AI solutions are increasingly efficient, reducing the time and effort required to implement effective sales strategies.[10]

Table 1 highlights examples of AI agents for marketing launched by Microsoft, Salesforce, and Accenture. These include tools like Dynamics 365 for lead qualification, Agentforce for sales coaching, and Accenture's AI Refinery for automating workflows and improving campaign outcomes.

Microsoft

Key agents in Microsoft's Dynamics 365 include sales qualification, supplier communications, and customer intent/knowledge management agents. These agents help prioritize sales opportunities, optimize supply chains, and resolve customer issues by autonomously managing tasks and learning from interactions. The sales qualification agent automates lead research and prioritizes opportunities, while the supplier communications agent tracks supplier performance and mitigates disruptions. The customer intent and knowledge management agents help customer service teams by resolving issues and scaling best practices.[5] As of November 2024, Toyota Motor Corporation was leveraging AAI technology with Microsoft, implementing a system of around nine AI agents to enhance knowledge storage and sharing. These agents streamline organizational processes.[4]

Salesforce

Salesforce's Agentforce was launched at the Dreamforce conference in September 2024 (https://www.salesforce.com/agentforce/), which integrates large language models (LLMs) with business data, enabling teams to build agents that drive predictions, scale operations, reduce errors, and improve consistency while cutting costs. This integration enhances decision making and boosts operational efficiency.[11]

In December 2024, Salesforce launched Agentforce 2.0, an enhanced version of its digital labor platform designed to augment teams with autonomous AI agents, seamlessly integrated into workflows. Agentforce 2.0 has introduced new agent skills like Sales Development and Sales Coaching for sales, broadening its utility in sales operations, alongside existing skills, such as lead development and personal shopper.[12] The Sales Coach Agent, for instance, uses AI and customer relationship management data to analyze sales pitches and role-play sessions, providing tailored feedback to help sales representatives close deals more effectively.[13] Other AI agents help with marketing campaign, commerce merchant, and service-related scheduling skills. These additions enable enterprises to nurture leads, join prospect calls, provide feedback, and tailor skills to various use cases, including field service work. Skills can be customized to meet specific business needs.[12]

TABLE 1. Some turnkey, no-code AI agents.

Company	Platform	Example of AI agent for marketing
Microsoft	Dynamics 365	Sales Qualification Agent: Automates lead research and prioritizes opportunities.
Salesforce	Agentforce/Agentforce 2.0	Sales Coach Agent: Analyzes sales pitches and role-play sessions, providing tailored feedback to help sales representatives close deals more effectively.
Accenture	AI Refinery for Industry	B2B Marketing Agent: Automates workflows, integrates data insights, and uses preconfigured agents to boost efficiency, accelerate deployment, and enhance campaign outcomes.

Agentforce 2.0 enables enterprises to build and deploy AI agents more efficiently across systems and workflows. Salesforce's new Agentforce update enables enterprises to create AI agents using natural language. The Agent Builder tool autogenerates relevant topics and instructions, while also pulling from a library of existing skills and actions to streamline the development process.[12]

Accenture

In January 2025, Accenture launched the AI Refinery for Industry, introducing 12 agent solutions to help organizations rapidly deploy AI agents that enhance workforce capabilities, tackle industry challenges, and accelerate business value creation. These industry agents integrate business workflows and industry expertise, enabling rapid deployment of customizable multiagent networks tailored to an organization's data.[14] For instance, the B2B Marketing Agent solution simplifies campaign management by automating workflows and integrating data insights. With preconfigured agents, it reduces manual effort, accelerates deployment, and improves campaign efficiency, supporting better outcomes and growth.[15]

AAI IN RELATION TO PREVIOUS AI GENERATIONS

AAI can be viewed as the third wave of AI, which evolves from the foundations of PAI (first wave) and GAI (second wave). PAI analyzes data to make informed forecasts about future events based on patterns in past data and trends and using analyses to predict the most likely outcome.[16]

PAI automates routine tasks, such as remote sensing and machine translation, processing unstructured data to uncover patterns through methods like text mining and speech recognition. Additionally, it analyzes human emotions and facilitates interactions using technologies, such as sentiment analysis, natural language processing (NLP), chatbots, and emotional robots.[17] These capabilities enhance user engagement and decision making by interpreting and responding to human behavior and language.

GAI leverages data to produce novel outputs, such as text, images, or other creative content. Its outputs are limited to the data it has trained on, meaning it can't create entirely new or original content.[18] GAI boosts efficiency, allowing marketing teams to complete tasks more quickly.

AAI uses independent AI agents that collaborate, reason, and solve complex problems, with LLMs guiding decision making.[19] Agentic workflows enhance organizations by autonomously processing and analyzing data in real time, learning from past customer interactions, and recommending goal-aligned marketing decisions.[20] The key differences between AAI and earlier AI systems are discussed in the following sections.

Autonomy

Whereas GAI's autonomy is limited, requiring external prompts to produce responses and cannot function without human guidance, AAI operates independently. AAI tools autonomously process and analyze data in real time[20] and handle complex tasks, like data analysis and decision making, with minimal or no human oversight.[21] Agents can be scheduled to run at fixed intervals or programmed to trigger actions in response to events from other applications or tools.[22]

In the agentic economy, a marketing agent can autonomously manage a campaign, create specialized agents for specific tasks (like social media or e-mail marketing), and even generate new agents for roles, such as data analysis or copywriting. This structure allows for continuous scaling and optimization, where each agent enhances the capabilities of others, leading to highly efficient and dynamic operations.[23] Optimizing tasks becomes more effective with multiagent systems, as each agent is fine-tuned for its purpose. This approach avoids the compromises seen in single LLMs, where focusing on one task can reduce overall versatility.[2]

Goal-directed and proactive behavior

Whereas GAI is task-oriented and reactive (that is, it responds to prompts to generate content like text or images, but it lacks long-term goals and doesn't pursue overarching objectives, completing each task based solely on immediate input). AAI operates with a specific goal in mind, actively working toward it by taking purposeful actions, such as a self-driving car that makes decisions like steering and braking to ensure safe arrival at its destination. AAI evaluates variables, balances tradeoffs, and aligns decisions with business goals, surpassing rule-based systems by analyzing trends and behaviors.[14] Agentic workflows enhance organizations, learning from past customer interactions and recommending goal-aligned marketing decisions.[20] In an e-commerce scenario, an AAI chatbot streamlines returns by instantly checking the customer's purchase history and return eligibility. It then offers personalized responses and may even suggest the optimal time for returns or extend discounts on future purchases.[24]

Complex decision making

Whereas GAI makes decisions by selecting outputs based on learned patterns, it does not evaluate multiple alternatives or consider long-term consequences. AAI evaluates multiple options and outcomes before making decisions, such as in stock-trading algorithms, where it analyzes data, predicts trends, and determines whether to buy or sell based on that information. For instance, AI agents can assist sales teams by combining input like target audience and outreach channels with historical data to generate and test multiple lead-generation strategies. This real-time adaptation increases efficiency, allowing teams to focus on high-conversion channels and optimize their efforts, ultimately improving deal closure rates.[23]

Adaptation and learning

AAI learns from its experiences, adapting and improving over time. Spotter learns from users and organizations to improve context, delivering personalized responses based on its growing understanding. This capability allows it to provide more relevant insights tailored to specific needs.[25] For instance, it can learn from past customer interactions.[21] Leveraging agents' adaptive learning and collective intelligence, businesses can refine operations continuously. Thus, while GAI has primarily focused on saving time and money, AAI adds value by adapting its thinking to continuously improve and refine its processes.[1]

> These trends point to the fact that AAI is evolving into digital labor, allowing companies to automate and streamline complex workflows across multiple departments.

STRATEGIC PLANNING FRAMEWORK IN MARKETING AND THE ROLES OF AAI

Figure 1 present a three-stage marketing strategic planning framework that follows a circular process. It starts with understanding the market, competitors, and customers (marketing research), followed by strategy development for segmentation, targeting, and positioning (marketing strategy development). The cycle culminates in marketing actions (marketing action execution), which feed back into research, creating a continuous loop of data that refines strategies and actions. AI plays pivotal roles in each stage.[17]

Table 2 compares predictive, generative, and agentic AI across marketing stages. PAI focuses on improving data analysis and decision making, GAI enhances creative tasks like content generation and insights, while AAI autonomously identifies trends, adapts strategies in real time, and delivers personalized experiences without human intervention.

Marketing research

Marketing research is a systematic, objective process for generating information to support marketing decision making.[26] At this stage, PAI analyzes large datasets to uncover consumer insights.[17] PAI can identify competitors and uncover competitive advantages using supervised learning in stable markets or unsupervised learning in new, uncertain markets. This approach helps marketers understand market structures and trends to better meet customer demands.[17]

GAI tools, such as LLM, assist in data generation, analysis, and conducting interviews, and perform repetitive tasks like summarizing reviews, producing richer, more insightful data than humans alone. For instance, by analyzing company profiles, news, past purchases, and consumer research, Salesforce's Einstein Assistant aids in crafting tailored recommendations for customers. GAI also plays a key role in marketing research by generating synthetic data to support analysis.

The ability to autonomously perform marketing research is a significant advantage of AAI compared to previous AI generations. For example, Mountain View, CA-based ThoughtSpot's Spotter has autonomous capability to find trends and generate insights, streamlining the process without needing prompts. Spotter and similar agents enhance marketing research by understanding context and incorporating reasoning capabilities, enabling more accurate analysis and insights that inform strategy.[24]

Marketing strategy

A marketing strategy outlines an organization's key decisions on target markets, marketing activities, resource allocation, and the delivery of value to customers, enabling the achievement of specific objectives.[27] At this stage, marketers can use PAI for segmentation, targeting, and positioning. This type of AI processes data for decision making, such as segmenting markets by gender or price, and excels in identifying patterns through mining and grouping techniques.[17]

Once trained on large unstructured data, GAI can efficiently summarize inputs and generate novel insights that contribute to marketing research. By processing vast amounts of information, it uncovers patterns, trends, and opportunities, offering valuable support for decision making, strategic planning, and competitive analysis.

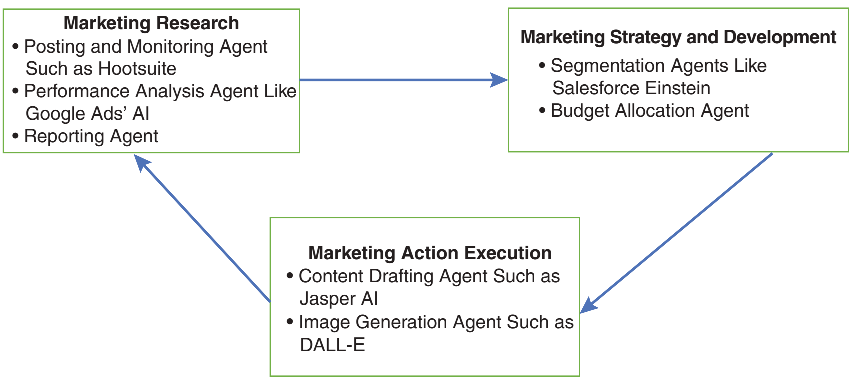


FIGURE 1. Marketing strategic planning framework and the roles of AAI.

For instance, as noted above, understanding competitors and their actions is a cornerstone of effective marketing research. GAI systems play a pivotal role in this by analyzing competitor reports, detecting shifts in strategies, and identifying trends. These insights not only signal market changes but also uncover emerging opportunities. Compared to previous generations of digital technologies, GAI tools provide higher-quality customer and competitor insights, enhancing behavioral targeting.[28] For example, Microsoft's Dynamics 365 Customer Insights, integrated with GPT-4 in Office products, helps companies identify new customer groups to target.[29]

AAI outperforms PAI and GAI in forming marketing strategies by providing strategic insights[14] and integrating automation, continuous learning, and innovation to dynamically enhance engagement and optimize processes.[30] Unlike PAI, which forecasts trends, or GAI, which creates content, AAI leverages orchestrated AI agents to autonomously analyze market data, adapt to changes, and make real-time, informed decisions. For example, Accenture uses AAI to streamline promotional decision making, optimizing revenue growth across product lines and markets, driving more strategic and adaptive top-line outcomes.[14] AI agents centralize data on culture, clients, competitors, and companies to design targeted marketing strategies while offering actionable insights on campaign performance, sales growth opportunities, and optimization for future planning.[14]

> These capabilities enhance user engagement and decision making by interpreting and responding to human behavior and language.

Marketing actions

Marketing actions refer to the decisions regarding the marketing mix a company uses to influence and meet the needs of its target customers. The 4Ps and 4Cs frameworks are used to analyze strategies and understand consumer behavior. The 4Ps focus on the marketer's perspective: product (what is being sold), price (its cost), place (where it's sold), and promotion (how it's advertised). The 4Cs take the consumer's viewpoint: consumer (their needs and wants), cost (what they pay beyond price), convenience (ease of purchase), and communication (two-way interaction instead of one-way promotion). Together, these frameworks guide effective marketing strategies. The focus is on ensuring that the actions taken within

TABLE 2. Comparison of AI Types in marketing research, strategy, and actions: PAI, GAI, and AAI.

Stage	PAI	GAI	AAI
Marketing research	Enhancing data collection, market analysis, and customer understanding. Identifying competitors and uncovering competitive advantages.	Facilitating consumer research to generate customer insights. Providing company overviews, summarizing news articles, and analyzing past product purchases (for example, Einstein Assistant).	Identifying trends and generating insights autonomously using (for example, Spotter).
Marketing strategy	Making improved decisions related to segmentation, targeting, and positioning. Processing data for decision making, such as segmenting markets by gender or price. Identifying patterns through mining and grouping techniques.	Identifying strategic shifts from competitors' reports. Performing repetitive tasks like summarizing reviews.	Developing dynamic marketing strategy by integrating automation, learning, innovation and adapting to real-time data.
Marketing actions	Enhancing product and branding efforts by creating conversational AI with brand personalities. Recommending content based on user moods. Monitoring brand reputation through sentiment analysis. Using chatbots to mimic customer communication styles for emotional support.	Generating text-based marketing content. Creating realistic visuals and artwork. Generating innovative product ideas and solutions. Helping create personalized recommendations.	Delivering adaptive personalization at scale autonomously: Enabling systems to dynamically adjust user experiences based on real-time data and interactions without human intervention.

the 4Ps deliver meaningful consumer benefits, aligning with the broader goal of satisfying customer needs and creating value.[31]

PAI optimizes product innovation, pricing, distribution, and promotional strategies, ensuring precision and customer-centricity throughout the process.[17] For instance, text-based chatbots handle routine customer service, while NLP chatbots address more complex issues, including those from diverse customers. AI systems, like Cogito's emotional AI, analyze conversation dynamics, providing real-time guidance to agents for more natural and engaging interactions.[17] PAI can enhance product and branding efforts by creating conversational AI with brand personalities, recommending content based on user moods, monitoring brand reputation through sentiment analysis, and using chatbots to mimic customer communication styles for emotional support.[17]

Regarding GAI systems' roles in marketing actions, tools like Copy.ai, Jasper.ai, and Peppertype.ai generate text-based content, while DALL-E 2 creates realistic visuals from text prompts. These systems contribute to ideation by helping develop innovative product ideas and solutions, enhancing marketing strategies.[27]

Traditional personalization methods, constrained by limited data and static rules, often lack depth. In contrast, AAI leverages vast datasets to craft highly personalized experiences, analyzing real-time customer behavior like browsing history, purchase patterns, and prior engagement, such as past interactions and behavior.[20] AAI enables adaptive personalization to operate autonomously, allowing systems to dynamically tailor user experiences based on real-time data and interactions without requiring human intervention. LuxStyle, a high-end fashion retailer, implemented AAI to enhance personalized marketing. The AI created detailed customer profiles, identifying trends, like seasonal demand for "statement jackets," and tailored recommendations accordingly. It dynamically adjusted campaigns based on engagement metrics, such as replacing stylized images with relatable visuals. Real-time personalization was reported to boost sales, with tailored suggestions increasing average order value by 25%. Additionally, efficient resource allocation reduced ad spend by 20% (https://tinyurl.com/yafc73ht).

> Agentic workflows enhance organizations by autonomously processing and analyzing data in real time, learning from past customer interactions, and recommending goal-aligned marketing decisions.

One major advantage of AAI is its capacity to improve customer service and overall satisfaction by delivering faster and more accurate responses to inquiries. In the aforementioned Capgemini survey, 64% of respondents anticipated improved customer service and satisfaction.[2] McKinsey highlights successful use cases of AI agents in customer service, such as providing real-time assistance to customer service agents during calls. These agents suggest the best knowledge articles and recommend next steps, enhancing the support experience.[32] AAI also ensures a smooth transition between human agents by providing all relevant context, including customer profiles, past interactions, and steps already taken. This minimizes repetition for the customer and enhances efficiency.[33] In addition, these agents can serve as analysts, allowing customer support teams to query data in natural language for instant answers, streamlining the process and improving customer satisfaction.[23]

Multiagent systems to handle all stages of marketing strategic planning

Companies can integrate multiple agents to handle complex tasks that usually require coordination across various functions.[1] Multiagent systems can thus revolutionize marketing strategies by automating processes, personalizing interactions in real time, and enabling data-driven decisions. These systems optimize customer engagement and deliver superior outcomes by adapting dynamically to market trends and preferences.[34] Each agent autonomously handles specific tasks and integrates into larger workflows, drawing data from LLMs, internal business systems, and external sources. This seamless integration ensures real-time insights and maintains operational efficiency across systems.[35]

In a multiagent system, different AI agents could assist marketers in exploring new markets and mode of engagement (such as social media campaigns), tailoring campaign content and channels to customer microsegments while adapting based on gained insights to drive growth opportunities.[1] According to the aforementioned Capgemini survey, 71% of respondents believed AI agents will enhance workflow automation.[2]

By utilizing specialized agents and automating workflows, marketing tasks, such as e-mail campaigns and advertisements, become more efficient, leading to improved engagement and higher conversion rates. These agents work in unison to ensure consistent messaging across channels, optimizing return on investment (ROI) and refining future strategies. This collaboration results in more targeted, effective campaigns, ultimately driving higher performance and continuous improvement. The roles of different AI agents at each stage of marketing strategic planning are explained in the following[33]:

- *Marketing strategy development:* Segmentation agents, like Salesforce Einstein, leverage customer data to create precise audience segments based on behavior, preferences, and demographics. For instance, they can identify eco-conscious customers for tailored marketing campaigns.[33] The Budget Allocation Agent uses AI-driven tools to adjust bids and reallocate marketing budgets in real time based on performance data.
- *Marketing action execution:* Content Drafting Agent, using Jasper AI, generates social media posts that resonate with the audience based on trending topics. DALL-E's Image Generation Agent creates visuals that match the content.
- *Marketing research:* Hootsuite's Posting and Monitoring Agent schedules and tracks posts, analyzing engagement, such as likes, shares, and comments. Hootsuite schedules posts during peak engagement times and tracks audience responses to optimize marketing strategies. Performance Analysis Agent, like Google Ads' AI, continuously tracks key ad performance metrics, such as click-through rates and conversion rates. In addition, the Reporting Agent uses AI tools to generate detailed performance reports, offering valuable insights into campaign effectiveness. These reports help businesses identify trends, strengths, and areas for improvement. By compiling data on key metrics, such as conversions, click-through rates, and ROI, the agent enables more informed decision making.

CHALLENGES AND WAYS FORWARD

Despite the potential of AAI in transforming marketing, however, a number of major challenges need to be overcome. Deploying AAI involves overcoming challenges, like ensuring data compliance, maintaining data quality, and managing errors.[36] The lack of mature security and governance in AAI systems raises significant concerns, especially as AI adoption grows.[7] Gartner forecasts that by 2028, one-quarter of all enterprise security breaches will be linked to AI agent misuse, originating from both external and internal malicious sources.[37]

Because AI agents operate autonomously and make rapid decisions, errors can cascade, especially when multiple agents make sequential decisions. To mitigate such risks, organizations should anticipate potential issues by incorporating time delays, adding human checkpoints, and ensuring sensitive data remains secure. While AAI may automate certain tasks, human oversight is essential to build trust and prevent errors. Leaders should educate teams about AAI's benefits and potential disruptions, emphasizing its value to employees and customers.[1]

While many organizations acknowledge the transformative potential of autonomous agents, their adoption is still in its infancy. Most implementations are limited to demonstrations and have not yet achieved full production readiness.[38]

Cultural resistance poses significant challenges to adopting AAI, as organizations may hesitate to embrace new systems. Additionally, difficulties in measuring ROI further complicate its implementation.[35]

Furthermore, while some systems demonstrate agent-like capabilities, defining genuinely "agentic" AI remains a challenge. While AI agents excel at enhancing automation, they cannot yet fully handle complex tasks independently.[39] Many current technologies still depend on a blend of human input and machine-driven actions.[21]

Despite the challenges mentioned, the advantages of AI agents are clear. In the aforementioned Capgemini survey, 57% of respondents felt the productivity benefits of AI agents outweigh the associated risks.[2] Overcoming adoption challenges for AAI requires companies to establish robust data frameworks and implement real-time monitoring systems. Transparent communication about capabilities and the creation of frameworks for performance evaluation are essential to address these limitations. By building trust and providing clear benchmarks, organizations can mitigate concerns and pave the way for broader acceptance and effective deployment of AAI solutions.[35]

> AAI evaluates variables, balances tradeoffs, and aligns decisions with business goals, surpassing rule-based systems by analyzing trends and behaviors.

AAI is revolutionizing industries by turning concepts into practical implementations. The rapid growth of AAI is reshaping marketing, driving innovations in automation, personalization, and decision making. This transformation offers businesses new opportunities to enhance strategies, streamline operations, and foster deeper customer engagement. For instance, AAI enhances customer service by enabling real-time, personalized interactions tailored to customer behavior and preferences, boosting satisfaction and fostering brand loyalty through 24/7 tailored support and recommendations. As organizations increasingly adopt these AI agents, they are transforming strategic planning and execution across functions.

The integration of turnkey, no-code solutions has democratized access to these technologies, enabling faster implementation and streamlining processes. By leveraging AAI, businesses can achieve greater operational efficiency, improved customer service, and dynamic, data-driven marketing strategies. Its adaptive learning and collective intelligence can help businesses refine operations continuously. This shift signals a future where AI agents are central to driving innovation and success in the marketing domain. By experimenting with various use cases, organizations can uncover how AI agents provide maximum value. ◼

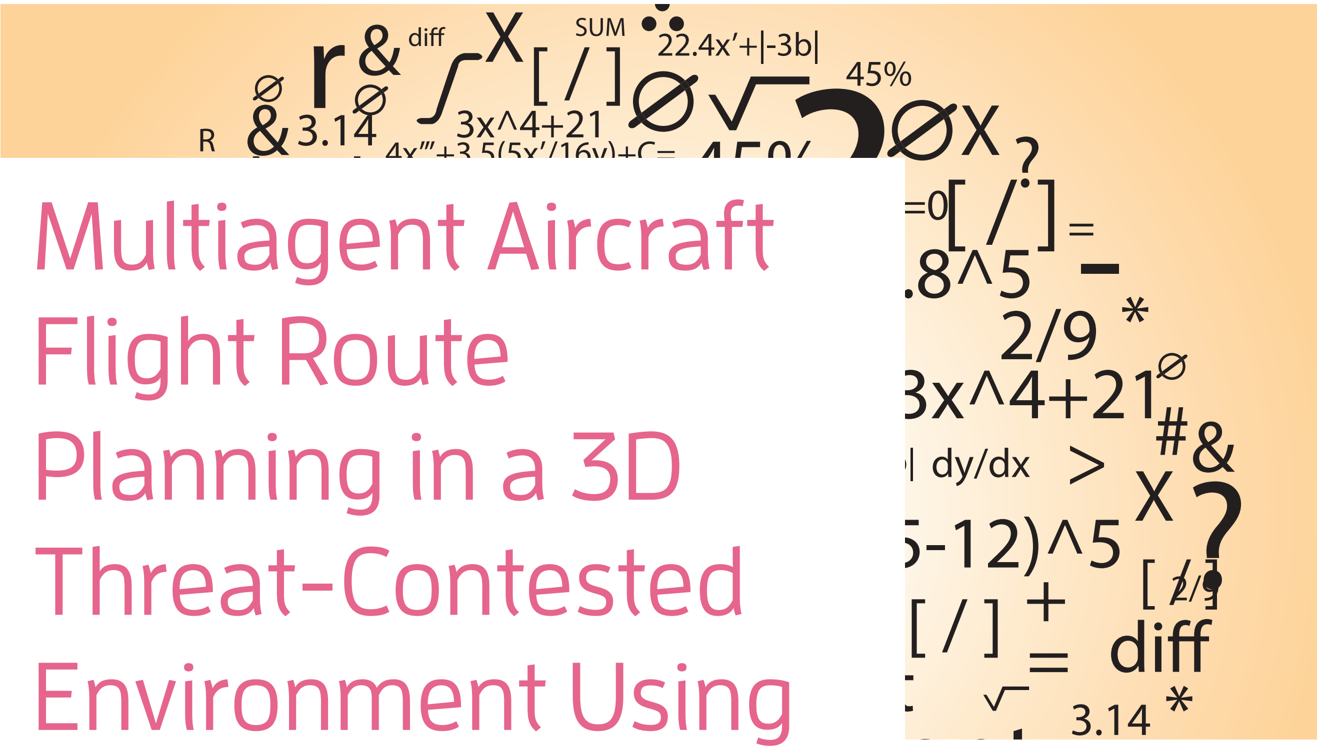
REFERENCES

1. C. Cutten, J. Kelly, and L. Teegarden, "Agentic AI is the next frontier in autonomous marketing," *The Wall Street Journal*, Dec. 10, 2024, Accessed: Jan. 5, 2025. [Online]. Available: https://deloitte.wsj.com/cmo/agentic-ai-is-the-next-frontier-in-autonomous-marketing-1d39e441
2. M. Korolov, "AI agents will transform business processes — and magnify risks," *CIO*, Aug. 21, 2024. Accessed: Jan. 5, 2025. [Online]. Available: https://www.cio.com/article/3489045/ai-agents-will-transform-business-processes-and-magnify-risks.html
3. M. Finnegan, "Microsoft to launch autonomous AI agents in November," *Computerworld*, Oct. 21, 2024. Accessed: Jan. 5, 2025. [Online]. Available: https://www.computerworld.com/article/3574341/microsoft-to-launch-autonomous-ai-agents-in-november.html
4. L. Wilkinson, "Microsoft readies Copilot Studio for agentic AI," *CIO Dive*, Nov. 19, 2024. Accessed: Jan. 5, 2025. [Online]. Available: https://www.ciodive.com/news/Microsoft-Copilot-Studio-AI-agents-Ignite-conference/733352/
5. A. Zaid Shirbeeni, "Microsoft rolls out AI agents to help businesses with everyday tasks," *Marketing Interactive*, Oct. 23, 2024. Accessed: Jan. 5, 2025. [Online]. Available: https://www.marketing-interactive.com/microsoft-ai-agents-businesses-everyday-tasks
6. L. Wilkinson, "Salesforce teases next-gen agentforce, parades early wins," *CIO Dive*, Dec. 4, 2024. Accessed: Jan. 5, 2025. [Online]. Available: https://www.ciodive.com/news/Salesforce-Agentforce-agent-AI-earnings-report/734674/
7. L. Wilkinson, "Salesforce introduces Agentforce 2.0, unveils pre-built use cases," *CIO Dive*, Dec. 17, 2024. Accessed: Jan. 5, 2025. [Online]. Available: https://www.ciodive.com/news/Salesforce-Agentforce2-agentic-AI-platform/735850/
8. Fortune Editors, "Agentic AI is a 'new labor model, new productivity model, and a new economic model,' says Salesforce's Marc Benioff," *Fortune*, Dec. 18, 2024. Accessed: Jan. 5, 2025. [Online]. Available: https://fortune.com/2024/12/18/agentic-ai-salesforce-marc-benioff/
9. "Marketing." Statista. Accessed: Jan. 5, 2025. [Online]. Available: https://www.statista.com/markets/479/topic/680/marketing/#overview
10. D. Greenberg, "Delivering A 1:1 Brand experience at scale with agentic AI," *Demand Gen Report*, Jan. 7, 2025. Accessed: Jan. 5, 2025. [Online]. Available: https://www.demandgenreport.com/demanding-views/delivering-a-11-brand-experience-at-scale-with-agentic-ai/48736/
11. A. Bridgwater, "The new tablestakes, AI tools that go to work," *Forbes*, Nov. 24, 2024. Accessed: Jan. 5, 2025. [Online]. Available: https://www.forbes.com/sites/adrianbridgwater/2024/11/24/the-new-tablestakes-ai-tools-that-go-to-work/
12. A. Ghoshal, "Salesforce's Agentforce 2.0 update aims to make AI agents smarter," *CIO*, Dec. 18, 2024. Accessed: Jan. 5, 2025. [Online]. Available: https://www.cio.com/article/3628094/salesforce-adds-skills-to-its-ai-agents-and-agentic-platform-to-serve-more-enterprise-use-cases.html
13. "Coach sales reps at scale with agentforce sales coach." Salesforce. Accessed: Jan. 11, 2025. [Online]. Available: https://help.salesforce.com/s/articleView?id=release-notes.rn_sales_agents_coach.htm&release=252&type=5
14. "Accenture launches AI refinery for industry to reinvent processes and accelerate agentic AI journeys." Accenture. Accessed: Jan. 5, 2025. [Online]. Available: https://newsroom.accenture.com/news/2025/accenture-launches-ai-refinery-for-industry-to-reinvent-processes-and-accelerate-agentic-ai-journeys
15. "AI refinery: Tame complexity, unlock scale." Accenture. Accessed: Jan. 5, 2025. [Online]. Available: https://www.accenture.com/us-en/services/data-ai/ai-refinery
16. Onome, "Generative vs predictive AI: Breaking it down," *AutoGPT*, Sep. 10, 2024. Accessed: Jan. 5, 2025. [Online]. Available: https://autogpt.net/generative-vs-predictive-ai-breaking-it-down/
17. M. H. Huang, and R. T. Rust, "A strategic framework for artificial intelligence in marketing," *J. Acad. Marketing Sci.*, vol. 49, no. 1, pp. 30–50, 2021, doi: 10.1007/s11747-020-00749-9.
18. S. J. Bigelow. "Generative AI vs. machine learning: Key differences and use cases." TechTarget. Accessed: Jan. 5, 2025. [Online]. Available: https://www.techtarget.com/searchenterpriseai/tip/Generative-AI-vs-machine-learning-How-are-they-different#:~:text=Limitations%20of%20generative%20AI,unique%2C%20but%20not%20necessarily%20original
19. K. Garvey, P. Gupta, D. Propson, B. Z. Zhang, and H. Sims, "How agentic AI will transform financial services with autonomy, efficiency, and inclusion," World Economic Forum, Dec. 2, 2024. Accessed: Jan. 11, 2025. [Online].

Available: https://www.weforum.org/stories/2024/12/agentic-ai-financial-services-autonomy-efficiency-and-inclusion/

20. S. Jensen. "How agentic AI is revolutionizing digital marketing." Single Grain. Accessed: Jan. 11, 2025. [Online]. Available: https://www.singlegrain.com/artificial-intelligence/how-agentic-ai-is-revolutionizing-digital-marketing/
21. A. Mathews, "Agentic AI is here and these 8 startups are leading the way in 2024," AIM Research, Sep. 10, 2024. Accessed: Dec. 28, 2024. [Online]. Available: https://aimresearch.co/ai-startups/agentic-ai-is-here-and-these-8-startups-are-leading-the-way-in-2024
22. G. Rama, "Microsoft to introduce AI agent building tools in Copilot Studio," *The Journal*, Oct. 31, 2024. Accessed: Jan. 5, 2025. [Online]. Available: https://thejournal.com/Articles/2024/10/31/Microsoft-to-Introduce-AI-Agent-Building-Tools-in-Copilot-Studio.aspx
23. K. Gomez, "The agentic economy is coming," *Medium*, Aug. 30, 2024. Accessed: Dec. 28, 2024. [Online]. Available: https://medium.com/@kyeg/the-agentic-economy-is-coming-ecf789a370f2
24. V. Shah. "5 agentic AI examples: Inspiring use cases shaping the future." ThoughtSpot. Accessed: Jan. 5, 2025. [Online]. Available: https://www.thoughtspot.com/data-trends/ai/agentic-ai-examples
25. E. Avidon. "ThoughtSpot AI agent Spotter enables conversational BI." TechTarget. Accessed: Jan. 5, 2025. [Online]. Available: https://www.techtarget.com/searchbusinessanalytics/news/366615693/ThoughtSpot-AI-agent-Spotter-enables-conversational-BI
26. W. Zikmund, S. D'Alessandro, B. Lowe, H. Winzar, and B. J. Babin, *Marketing Research*, 4th Asia-Pacific ed. Melbourne, Victoria, Australia: Cengage Learning Australia, 2017.
27. R. Varadarajan, "Strategic marketing and marketing strategy: Domain, definition, fundamental issues and foundational premises," *J. Acad. Marketing Sci.*, vol. 38, no. 2, pp. 119–140, 2010, doi: 10.1007/s11747-009-0176-7.
28. N. Kshetri, Y. K. Dwivedi, T. H. Davenport, and N. Panteli, "Generative artificial intelligence in marketing: Applications, opportunities, challenges, and research agenda," *Int. J. Inf. Manage.*, vol. 75, Apr.2024, Art. no. 102716.
29. M. Reines. "Microsoft's new dynamics 365 Copilot uses generative AI." TechTarget. Accessed: Jan. 5, 2025. [Online]. Available: https://www.techtarget.com/searchcustomerexperience/news/365532058/Microsofts-new-Dynamics-365-Copilot-uses-generative-AI
30. C. Dube, "What is agentic AI and will it replace your job?" *Fast Company*, Dec. 31, 2024, Accessed: Jan. 5, 2025. [Online]. Available: https://www.fastcompany.com/91248551/what-is-agentic-ai-and-will-it-replace-your-job
31. C. Beard and C. Easingwood, "New product launch: Marketing action and launch tactics for high-technology products," *Ind. Marketing Manage.*, vol. 25, no. 2, pp. 87–103, Mar. 1996, doi: 10.1016/0019-8501(95)00037-2.
32. "McKinsey and Microsoft drive business value creation with gen AI agents enabled by Copilot Studio," *McKinsey & Company*, May 29, 2024. Accessed: Jan. 5, 2025. [Online]. Available: https://www.mckinsey.com/about-us/new-at-mckinsey-blog/mckinsey-expands-alliance-with-microsoft-to-scale-copilot-solutions-across-enterprises
33. "What is agentic AI? Everything you need to know about AI's next buzzword and how it's set to transform CX," *CX Today*, Jan. 3, 2025. Accessed: Jan. 5, 2025. [Online]. Available: https://www.cxtoday.com/contact-centre/what-is-agentic-ai/
34. "AI-assisted marketing with multi-agent systems," *Integrail.ai*, June 25, 2024. Accessed: Jan. 10, 2025. [Online]. Available: https://integrail.ai/blog/ai-assisted-marketing-with-multi-agent-systems
35. S. Kaufman, "Beyond ChatGPT: The rise of agentic AI and its implications for security," *CSO Online*, Oct. 22, 2024. Accessed: Jan. 5, 2025. [Online]. Available: https://www.csoonline.com/article/3574697/beyond-chatgpt-the-rise-of-agentic-ai-and-its-implications-for-security.html
36. P. Akkiraju, S. Beshar, H. Korn, A. Tutika, E. DeSilva and C. Gyra. "The state of the AI agents ecosystem: The tech, use cases, and economics." Insight Partners. Accessed: Jan. 5, 2025. [Online]. Available: https://www.insightpartners.com/ideas/state-of-the-ai-agent-ecosystem-use-cases-and-learnings-for-technology-builders-and-buyers/
37. "Gartner unveils top predictions for IT organizations and users in 2025 and beyond," *Gartner*, Oct. 22, 2024. Accessed: Jan. 5, 2025. [Online]. Available: https://www.gartner.com/en/newsroom/press-releases/2024-10-22-gartner-unveils-top-predictions-for-it-organizations-and-users-in-2025-and-beyond
38. J. Rodriguez, "AI agents can help crypto become the currency of AI," *CoinDesk*, Nov. 19, 2024. Accessed: Jan. 11, 2025. [Online]. Available: https://www.coindesk.com/opinion/2024/11/19/ai-agents-can-help-crypto-become-the-currency-of-ai
39. E. Pounds, "What is agentic AI?" *NVIDIA Blog*, Oct. 22, 2024. Accessed: Jan. 5, 2025. [Online]. Available: https://blogs.nvidia.com/blog/what-is-agentic-ai/

NIR KSHETRI is a professor at the Bryan School of Business and Economics, University of North Carolina at Greensboro, Greensboro, NC 27412 USA. Contact him at nbkshetr@uncg.edu



Multiagent Aircraft Flight Route Planning in a 3D Threat-Contested Environment Using Cross-Entropy Search

Christopher (Bud) Daniel, **Doron Drusinsky**, and **Loren E. Peitso**, Naval Postgraduate School

We present cross-entropy search for multiple-aircraft facing inter- and intra-aircraft constraints, a problem that cannot be solved by traditional Dynamic Programming algorithms such as A.*

Path-finding algorithms can assist aviators in generating flight routes that balance efficiency and surface-to-air threat avoidance. Traditional dynamic programming (DP) algorithms, such as A*, are appropriate for a single aircraft but fall short in multiagent cooperative environments.

Path-finding is an example of a quintessential combinatorial optimization problem. Many algorithms have been developed and applied to various basic path-finding problems. While vanilla versions of these puzzles can be solved quickly using algorithms, such as Dijkstra's or A*, they struggle when the complexity is increased. Additionally, they are forms of DP, which function by breaking a problem into smaller subproblems from the bottom up. Since DP algorithms are guaranteed to achieve optimality, every subproblem is also optimal. This property of DP algorithms is known as the *optimal substructure property* (*OSP*), meaning the optimal solution to a problem is the aggregate of optimal solutions to its subproblems. There are benefits and restrictions to the OSP, which will be discussed in detail in Single-Agent Search Algorithms.

Digital Object Identifier 10.1109/MC.2025.3533153
Date of current version: 27 March 2025

DISCLAIMER

The views expressed in this article do not necessarily represent the views of the DoN, DoD, or the United States.

EDITOR DORON DRUSINSKY
Naval Postgraduate School; ddrusins@nps.edu

Multiagent path-finding problems require that multiple agents must coordinate paths (i.e., satisfy inter-agent constraints), while optimizing their own objectives (i.e., satisfy intra-agent constraints and goals). However, these problems are not OSP-compliant. As a result, traditional DP methods cannot guarantee optimality in these settings, necessitating alternative approaches. We offer one solution through the primary contribution of this article: cross-entropy (CE) search to efficiently explore the expanded solution space.

In addition to the challenges posed by multiagent coordination, most path-finding algorithms have traditionally been applied to 2D space. These include GPS-based navigation systems, video game nonplayer character navigation, robotic and autonomous vehicles, and network routing. While research and applications into 3D space are present and ongoing, it is a newer domain and less well developed. The addition of the third axis increases the search space and, perhaps more importantly, the branching factor for DP-based algorithms. While still incredibly fast and optimal in reasonably sized search spaces, Dijkstra and A*, along with similar algorithms, cannot guarantee optimality when applied to scenarios that do not adhere to the OSP.

Flight planning is fundamentally a path-finding problem. Both military and civilian air forces develop flight plans for every sortie, usually executed manually. This is a time-intensive and frustrating task, and depending on the complexity of the airspace, there are no guarantees of optimality. Additionally, the tools available to planners paint an incomplete picture or make it difficult to obtain more comprehensive knowledge. Individual flight planners often leverage their personal knowledge and experience, which may lead less experienced individuals to develop inferior routes. An automated route-generating aid could shorten planning time, level the playing field for different experience levels, and possibly lead to the development of better, unique routes.

In this article, we model a 3D search space representing contested airspace with line-of-sight (LOS) weapons engagement zone (WEZs) and explore various path-finding approaches for flight route planning. We then apply these algorithms to single-agent scenarios in 3D environments with threats. While this is an interesting problem in its own right, we go on to use the optimality of A* to validate the quality of solutions generated by the potentially suboptimal CE search method, which we then apply to the more complex multiagent problem where routes are interdependent.

We distinguish between the operational problem domain and the broader research problem addressed in this study. The operational problem requires the development of automated routing strategies in a complex, 3D, threat-contested environment. This includes the task of modeling real-world 3D hostile airspace as a search space suitable for the application of various pathfinding algorithms. The research problem focuses on solving multiagent non-OSP pathfinding problems in general.

This article is structured as follows. Operational Context and Problem Domain introduces the operational challenges and developing the search domain. Single-Agent Search Algorithms provides an overview of single-agent search algorithms. Single-Agent 3D Path-Finding for Threat Avoidance extends these algorithms to 3D and includes comparative results. Multiagent Search Algorithms focuses on the multiagent application of CE search to address the research problem of optimizing paths for multiple agents in non-OSP environments. Finally, we conclude with insights and suggestions for future research.

OPERATIONAL CONTEXT AND PROBLEM DOMAIN

In military aviation, pilots and mission planners currently use a software called the *Joint Mission Planning System* (*JMPS*) to model the operational landscape and design aerial solutions to achieve specific objectives. Every end product is a flight route that optimizes mutual support, threat avoidance, terrain avoidance, airspace utilization, and time-to-completion. JMPS, while a powerful tool, requires manual construction of routes based on human interpretation of visual plots of integrated air defense systems (IADS) and other relevant data. While JMPS employs some automation to validate routes against platform flight limitations, it does not generate routes automatically.

For simplicity, we focus solely on JMPS's threat modeling capability to create threat coverage maps, disregarding other functionalities. Our task is to automate flight route generation to avoid IADS and still arrive at a given target point. To achieve this, our software replicates the threat output produced by JMPS. In JMPS, planners can lay down specific surface-to-air threat systems with unique WEZ distances. Each threat includes up to three user-specified altitudes in either height above ground level (AGL) or mean sea level (MSL). JMPS then plots LOS detection zones for these altitudes, referred to as *bugsplats* or *threat masks* (Figure 1). Altering these bugsplats is time-intensive and error-prone, as the user must edit each threat individually. Additionally, it is impossible to view complete coverage across altitude layers simultaneously.

Building the threat coverage map

There is no easy way to export the base function that JMPS uses to create its bugsplats. We needed to replicate the logic that could plot LOS coverage from a specific point, given the terrain, curvature of Earth, and a defined threat WEZ.

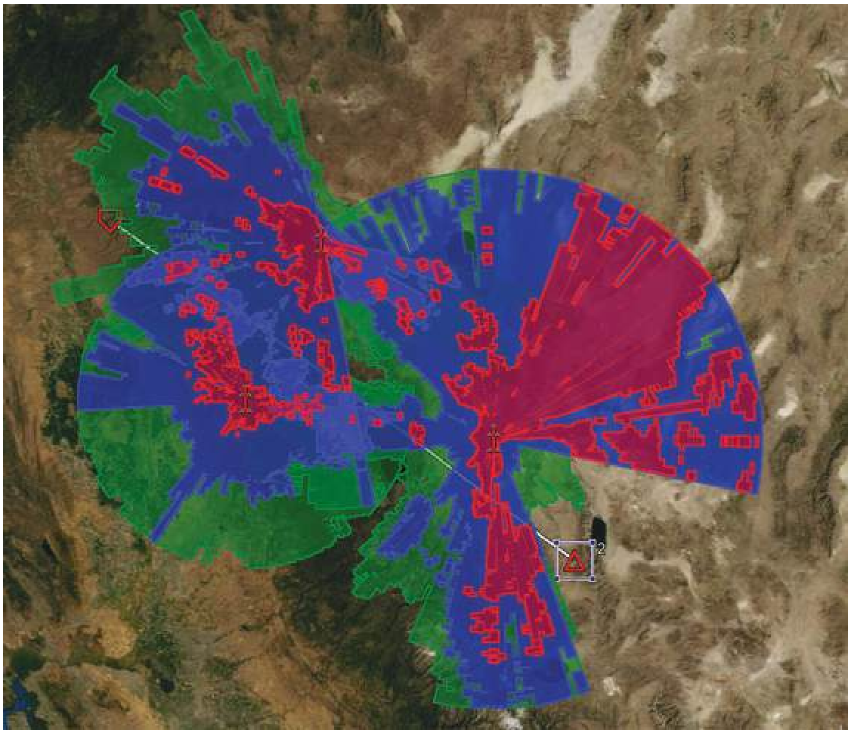


FIGURE 1. JMPS-produced IADS bugsplats using notional surface-to-air threat systems. The colors represent user-selected altitudes of 1,000 feet AGL (red), 5,000 feet AGL (blue), and 10,000 feet AGL (green).

To model the terrain, we need accurate ground elevation data. Global surface elevation above MSL is available for government purposes via digital terrain elevation data (DTED) on various levels through the National Geospatial-Intelligence Agency. For our purposes, we used DTED level 0 (DTED-0), which represents surface elevation in meters above MSL at specific latitude/longitude points spaced approximately 900 m apart.

Pilots navigate using nautical miles (NMs) as a distance measurement. Since DTED-0 is a database containing elevations at specific geospatial coordinates, we need to be able to quickly calculate distances in terms of NMs between these coordinates to model threat maps with known coverage distances. We do so using the Haversine formula [shown in (1) at the bottom of the page] to determine the great circle between two points on a sphere (Earth), with the average radius of Earth being $R = 20{,}902{,}231$ ft.[1]

To model a single threat mask over the terrain, we use an iterative process that broadcasts a ray in a single horizontal direction from the threat origin outward in discrete steps at a vertical "view angle" from the threat until it hits the threat's maximum viewing distance. At each step, the ray's height above the ground (pulled from the coordinate's DTED elevation) is stored in an array. If the height is less than or equal to zero (the ray intersects with the ground), a value of zero is stored in the array, and the view angle is updated to account for ground obstruction. The ray is then shifted horizontally by a set amount and the process begins again. The threat mask is complete when the ray has completed 360° of horizontal coverage. See Figures 2 and 3 for a visual depiction of this process.

$$\text{distance} = 2R \cdot \arcsin\left(\sqrt{\sin^2\left(\frac{\Delta\text{lat}}{2}\right) + \cos(\text{lat1})\cos(\text{lat2})\sin^2\left(\frac{\Delta\text{long}}{2}\right)}\right). \quad (1)$$

Each threat in the coverage map has defined geospatial coordinates given in latitude and longitude, and a corresponding surface elevation above MSL at that location, which we define as *threat height* (h_t). Additionally, we assume each threat's broadcasting and receiving antenna has a height above the surface, which we call *the antenna height* (h_a). Therefore, the total distance from Earth's center to the top of the threat can be written as $T = R + h_t + h_a$, where R is the radius of Earth.

The threat's *view angle* (ψ), or vector ray along one horizontal radial, is defined as the angle between a line from the threat to Earth's center and its current lowest visible area. To identify undetectable areas between the ground and the view angle, we progressively step the ray outward a discrete distance δ along a vector away from the threat to a coordinate point p. If the ray collides with the surface, the view angle is updated so that the ray is tangent to the surface at the point. At each new p, we pull the elevation above MSL of the ground at that point from DTED, which we call the *height of the ground* (h_g). The height can also be defined as the distance from the Earth's center, or $G = R + h_g$.

Each element of the threat mask array will be populated with a "safe" altitude (s), which is the highest altitude above the ground at a specific coordinate location where the threat does not have visual coverage. This is done by calculating the safe altitude above MSL ($h_{\min}$) and subtracting the ground elevation at that point, or $s = h_{\min} - h_g$.

The process of populating a threat's bugsplat given the terrain can be executed using Algorithm 1.

Threat coverage maps are created for all threats in the environment, having the arrays superimposed to build a complete IADS picture. When coverage of two or more threats overlaps, the lowest safe altitude is chosen. Finally, start and target points, specified in latitude/longitude format, are converted to grid locations within the array.

To obtain accurate LOS coverage, the threat maps have to be high fidelity, leading to a total coverage map of approximately 500 × 500 grid squares (Figure 4), depending on the start point, target point, and threat locations. This presents a large but manageable search space for some of the faster algorithms, but adding the third dimension increases it at least 10-fold. To reduce computation time and enable methods like CE to operate efficiently, we discretize the z-axis by sampling specific altitudes to generate binary detection masks. These masks are then compressed into smaller 2D arrays based on a user-specified NMs per grid-square resolution. At each selected altitude, the search space is divided into a coarse grid, where each grid square represents a level of threat coverage between 0% and 100%. This coverage value is transferred to the corresponding location in the compressed array. The resulting smaller arrays contain detection probabilities ranging from 0 to 1 (Figures 5 and 6). Together, these arrays form a 3D data structure of size $i \times j \times z$ (where z is the number of discrete altitudes), referred to as `comp_grids`.

Start and target points are represented by 2D locations, as we want to have the freedom to specify the start altitude, and we are treating the aircraft's altitude at the target as arbitrary.

With our search domain `comp_grids`, we can use any single- or multi-agent search algorithm to develop and compare generated flight routes. For verification and validation purposes, we will compare the single-agent (aircraft) CE search method to a known optimal search method: A* search with a single agent (aircraft) and single target. Then we extend our methods to show that the CE method can perform integrated searches with multiple agents and multiple target points.

SINGLE-AGENT SEARCH ALGORITHMS

Dijkstra/A*

Dijkstra's algorithm[3] is a classic shortest path discovery method. It is a DP-based single-source shortest path algorithm guaranteed to find the shortest distance between a given source vertex on a graph and all other vertices on that graph.[3] As any map or navigation space can be reproduced as a graph, this process can prove to be advantageous. However, Dijkstra's algorithm is

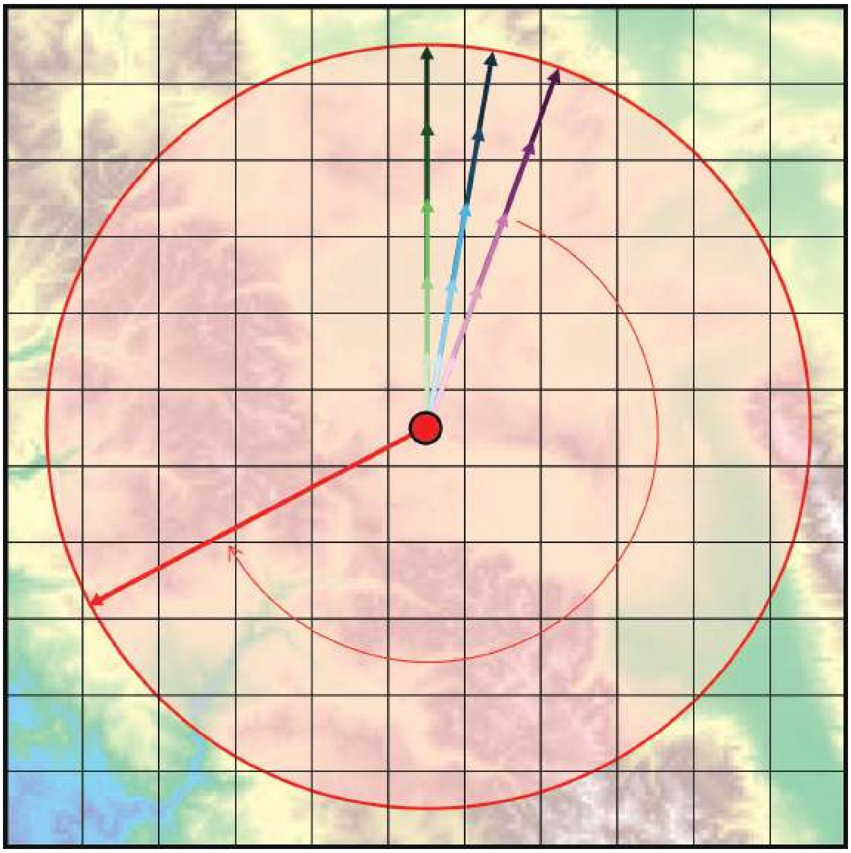


FIGURE 2. Visual representation of how a threat mask is determined, based on terrain elevation.

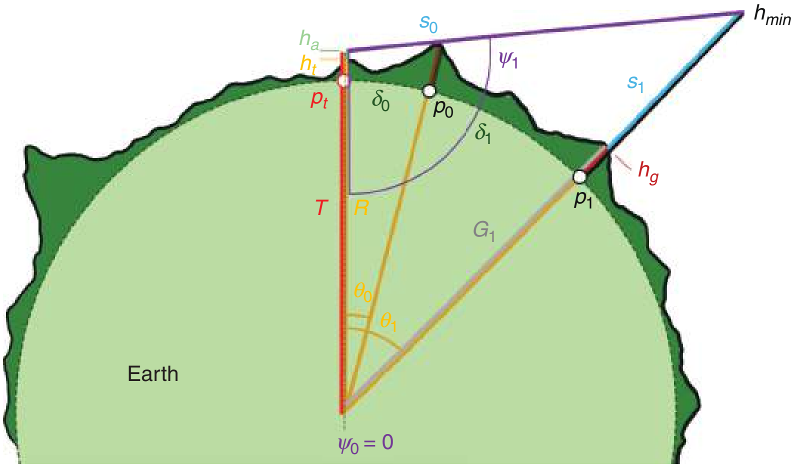


FIGURE 3. Slice of the Earth showing single ray from a threat, with geometry to calculate threat coverage given the terrain.

time- and compute-intensive compared to contemporary path-finding algorithms because it visits all, or a large portion of the entire graph even when the intended use is to find the shortest path between two points. Dijkstra's algorithm operates as a single source, multiple (the entire graph) destination algorithm. However, most physical route-generating problems involve a single start point and a single end point, which suggests that the shortest path can often be found without necessarily having to explore the entire graph.

In 1968, an extension of Dijkstra's algorithm was published that guarantees optimality and is at least as fast as Dijkstra's, and regularly significantly more so.[4] A* combines Dijkstra's DP approach with a "look-ahead" heuristic function during each iteration of the DP's bottom-up path construction process, thereby quickly discovering the shortest path between two points on a graph.[4] Both A* and Dijkstra algorithms use a priority queue (implemented as a min-heap[3,4]) to select the best vertex to extend a subpath from one vertex n' to the next vertex n during every iteration of the algorithm. However, the plain Dijkstra algorithm uses $g(n)$, the "backward" cost of the path from the source to vertex n, as the priority queue's key, whereas A* uses $f(n) = g(n) + h(n)$ as that key, where $h(n)$ is a "forward" cost for extending the path from n to the goal vertex, a cost that is estimated using a heuristic. See Algorithm 2 for the pseudocode of the A* search algorithm.

To ensure optimality, A* must use a forward-looking heuristic $h(n)$ that is admissible and consistent at every point. Admissibility means that the heuristic must never overestimate the true cost to reach the goal. If the heuristic is not admissible, then A* might not consider a path that should have been considered, hence resulting in a suboptimal path. For the heuristic to be consistent, every node n and its successor n', the estimated cost $h(n)$ must satisfy the inequality:

$$h(n) \leq c(n, n') + h(n')$$

where $c(n, n')$ is the step cost between n and n'. The simplest method to ensure $h(n)$ is admissible is to use the Euclidean distance between n and the target point. However, other heuristics may save time by expanding fewer nodes. The A* search algorithm is used extensively in many path-finding applications, from transportation to robotics to the video game industry.[5] Traditionally, the cost of a path can be defined as purely the 2D distance or total time to reach the destination. By introducing other factors to the cost function, we can tailor the algorithm to more accurately describe our problem space, albeit at the cost of added complexity. Both Dijkstra and A* are examples of DP and therefore must follow the OSP.

Variations of A* have been developed and implemented for both offline and online flight path planning problems. Li et al.[6] improved upon the classic A* algorithm by replacing the open and closed tables with a value table to speed up computation time when calculating optimal obstacle-avoidance routes for fixed-wing aircraft. Wang

ALGORITHM 1: Algorithm for populating an individual threat mask.

```
Determine T ← R + h_t + h_a at the threat's coordinates
Define threat's maximum effective range D
Initialize horizontal radial rad ← 0, signifying North
Define maximum horizontal radial rad_max ← 2π
Define horizontal radial step size step_rad

while rad < rad_max:
    Set threat view angle to maximum visibility, ψ ← 0
    Initialize horizontal distance δ ← 0
    Define horizontal step distance step_d

    while δ ≤ D:
        δ ← δ + step_d
        Determine point p using (1)
        Determine ground height h_g
        G ← R + h_g
        Calculate radial distance θ ← δ / R
        % Calculate h_min using equation A.6 from Appendix A in Daniel^2:

            h_min = T sin ψ / sin(θ + ψ) − R

        Calculate s ← h_min − h_g

        if s ≤ 0:
            s ← 0
            % Update ψ using equation A.9 from Appendix A in Daniel^2:

                ψ = arctan( G sin θ / (T − G cos θ) )

        Store s in mask array at point p
    end while
    rad ← rad + step_rad
end while
return Final mask array
```

et al.[7] developed the enhanced bidirectional search A*, which uses bidirectional search to speed up search time, obstacle expansion distance to allow for a buffer or safety zone around obstacles, and smoothing to allow for realistic paths. While Wang et al. only addresses the 2D path planning problem, the methods introduced can be logically expanded to a 3D realm. Xue et al.[8] blend discrete and continuous search methods to detect valleys in a 3D terrain space with their Hybrid-A* algorithm. This method is directly applicable to military flight missions, as valleys offer concealment from radar and visual acquisition threats.

CE search algorithm

The incentive for using the CE search method is that our underlying 3D path

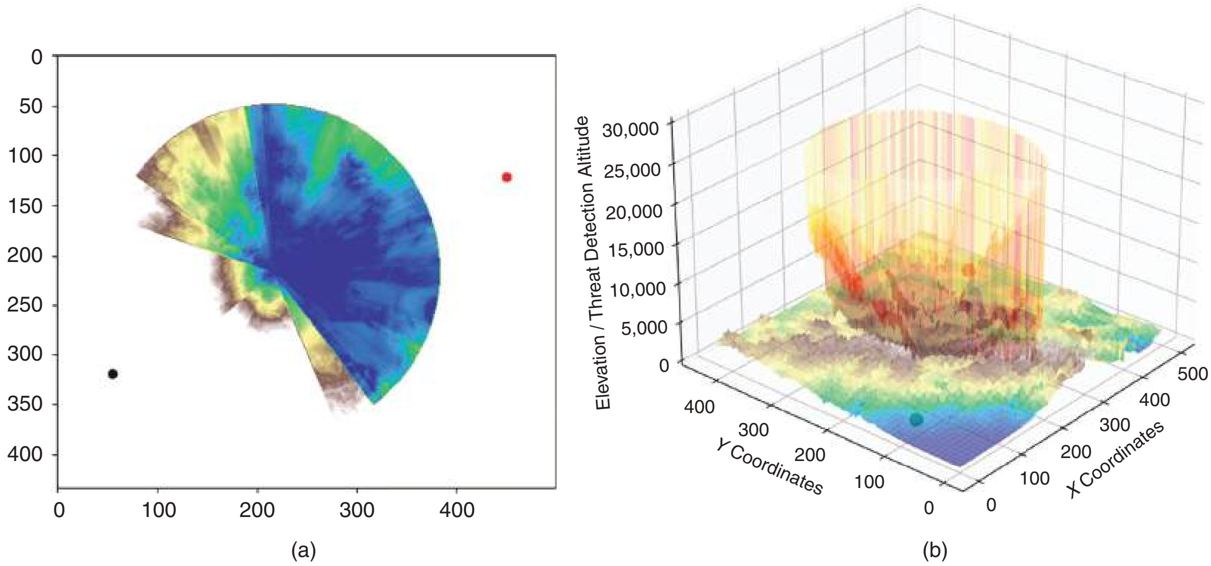


FIGURE 4. High-fidelity threat coverage generation for one threat and start and target points given the terrain. (a) Two-dimensional coverage shows lower detection altitudes in blue and higher detection altitudes in brown-white. (b) Three-dimensional coverage of the same coverage overlayed on the same terrain.

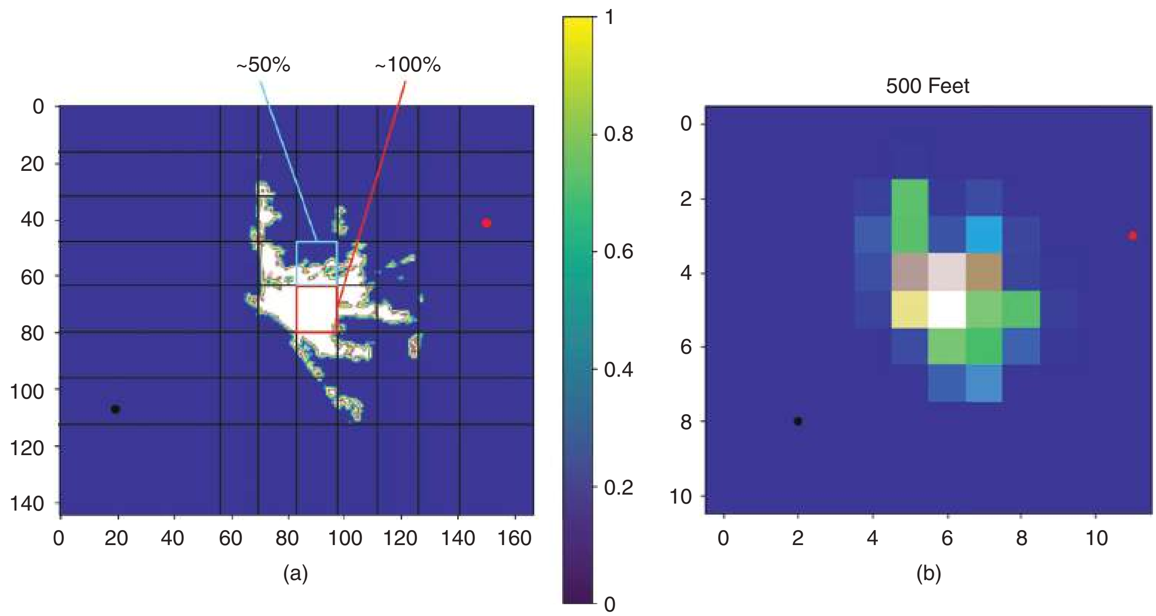


FIGURE 5. (a) AGL sample (500 feet) from high-fidelity threat coverage and (b) the same coverage compressed to a 10 NM/grid square (GS) scale. The blue-white scale indicates probability of detection according to the scale in the middle.

finding problem involves multiple agents (aircraft), a non-OSP problem that cannot be solved with A* when there are interaircraft constraints.

An optimal path in a large search space with many possible paths can be viewed as a rare event. In 1997, Reuven Rubinstein[9] developed a technique to simulate rare events by leveraging a change in probability measure to reduce the variance of estimators, making rare events more frequent in

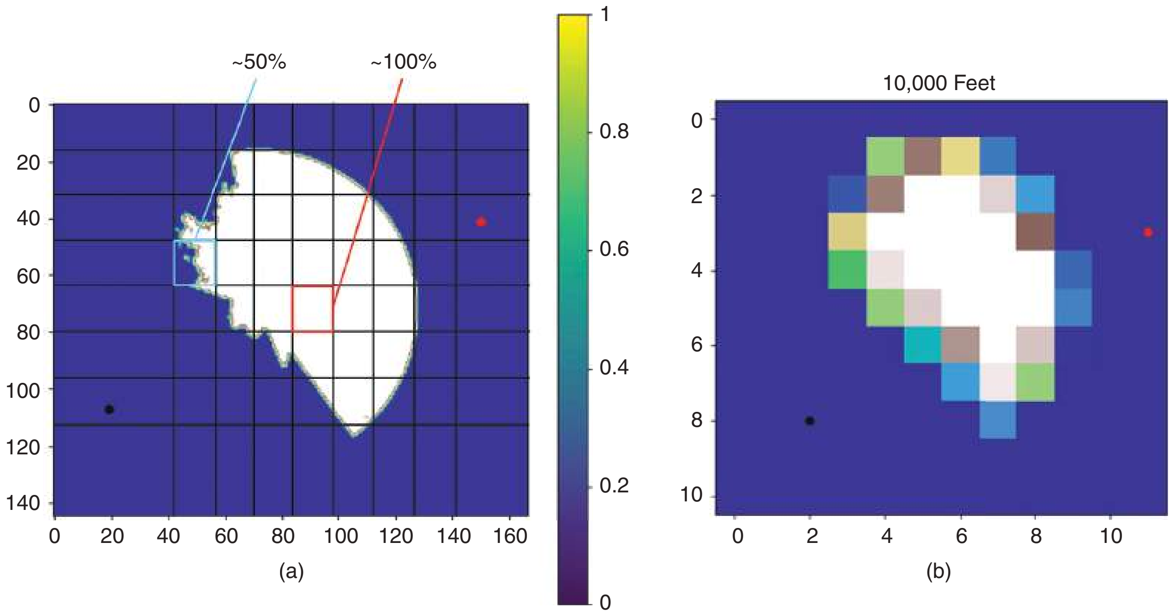


FIGURE 6. (a) AGL sample (10,000 feet) from high fidelity threat coverage and (b) the same coverage compressed to a 10 NM/GS scale. The blue-white scale indicates probability of detection according to the scale in the middle.

ALGORITHM 2: A* search algorithm.

```
Initialize the open set OpenSet ← {s} containing only the start node s.
Initialize the closed set ClosedSet ← {} as an empty set.
Calculate f(s) ← g(s) + h(s), where g(s) ← 0 and h(s) is a heuristic estimate of the cost to the goal.
while OpenSet is not empty:
    Select the node n ∈ OpenSet with the smallest f(n). Resolve ties arbitrarily, but prefer goal nodes when possible.
    if n is a goal node:
        Terminate
        return the path to n.
    Remove n from OpenSet and add it to ClosedSet.
    for each successor n′ of n:
        if n′ ∈ ClosedSet:
            continue % Skip already processed nodes.
        Calculate g(n′) ← g(n) + c(n,n′), where c(n,n′) is the cost of the edge from n to n′.
        if n′ ∉ OpenSet or g(n′) is better than the previously recorded cost:
            Update f(n′) ← g(n′) + h(n′), where h(n′) is the heuristic estimate to the goal.
            Record n as the predecessor of n′.
            if n′ ∉ OpenSet:
                Add n′ to OpenSet.
    end for
end while
if no path is found:
    return failure
```

simulations and thus easier to analyze and optimize. Two years later, he published an optimization approach that modified his simulator by introducing a CE measure, leading to the CE method. It is a sampling-based optimization method used for solving combinatorial and continuous optimization problems that estimates probabilities of rare events and guides the search for constraint-satisfying solutions. It does so by defining a score function, iteratively sampling and updating a probability distribution toward areas of higher score, while minimizing the CE between the current distribution and the optimal one.[10] While similar to other random search algorithms, such as simulated annealing, tabu search, and genetic algorithms—in that it adjusts the initial probability distribution so that rare events are more likely to occur in regions where the global optimum of the problem is expected to be—the CE method uses a fundamentally different idea. Other techniques employ local search heuristics, whereas CE focuses on a global search of the space.[10]

The CE method finds an optimal solution to a given problem by iteratively solving secondary optimization problems using Kullback-Leiber (KL) CE and importance sampling.[10] The procedure is summarized in Algorithm 3. In our case, there are two random variables whose distributions (denoted P in Algorithm 3) are learned: one for the choice of a move from a grid cell to one of its eight neighbors, and one for the choice of an elevation. Samples in our case are 3D paths.

CE has been used in a variety of applications, including buffer allocation, DNA sequence alignment, vehicle routing, network reliability, and navigation problems.[11] These all share the same characteristic in that they are combinatorial optimization problems. Path-finding can be formulated as an optimization problem that uses a weighted graph, where each possible location is a node connected to its neighbor nodes by weighted edges. In standard path-finding problems, the weights and connections between nodes are known and fixed, making the problem deterministic. By using the CE method, this deterministic optimization problem can be transformed into a stochastic one by introducing randomness into the edges, allowing for more robust exploration and estimation of rare events (optimal paths) in the network.[11]

Helvik and Wittner[12] and Chepuri and Homem-de Mello[13] have successfully applied CE to approximate optimal solutions to the traveling salesman problem (TSP), which is a hard (NP-Complete) path-finding combinatorial optimization problem. Helvik and Wittner[12] extended the TSP for multiagent path-finding and CE shows comparable results to the ant colony system method. Chepuri and Homem-de Mello[13] additionally used CE to solve the vehicle routing problem with stochastic demands, a much harder problem than a traditional TSP due to the presence of random demands and penalties. They achieved promising results and note that the method remains general in terms of implementation to specific problem formulations.

Drusinsky and Michael[14] also solved a vehicle routing problem using CE to discover a set of optimal routes for multiple agents in a complex environment. They applied the technique to a multiagent setting with uncertainty constraints modeled as Bayesian networks. To employ CE in their environment, Drusinsky and Michael created a transition matrix `Mat`, which contains initially equal probabilities of moving from every grid point (or node) to each of its neighbors. The CE algorithm then runs an inner loop, creating N random paths, scoring each path, and updating `Mat` according to a specified percentage of the highest-scoring paths (the elite set). The process iterates until appropriate stopping criteria are met. Individual paths are created by sampling from `Mat` for each generated point in the path until a goal is encountered or the path is stuck.

The core of their algorithm, as with any CE algorithm, lies in the score function. Each path is scored per a set of rigid, optimization, and uncertainty constraints. If a path does not pass all rigid constraints, it is given a proportionally bad score such that it is deemed an unacceptable path. It is important to note that some

ALGORITHM 3: Pseudocode for the CE method.

1: Initialize random distributions P **for** the problem variables, Set iteration count $t \leftarrow 0$.
2: **while** convergence criteria are not met:
3: Generate a sample S_t of random data points using the current distributions P.
4: Evaluate each sample in S_t using the scoring function $f(x)$.
5: Select the top-performing samples S_t^{top} based on a predefined threshold $p\%$.
6: Update the parameters of P using S_t^{top} by minimizing the KL divergence.
7: Increment $t \leftarrow t+1$.
8: **end while**
9: **return** the parameters and solution corresponding to the best sample in S_t^{top}.

ALGORITHM 4: A* for single-agent (aircraft) 3D threat-avoiding search.

```
% Initialization of constants and parameters
Initialize constants α,γ,δ,ϵ.
Define start, target, alts, comp_grids, start_alt, NM_per_grid_square.
Calculate alt_diff and max_climb_descent.
Initialize min-heap future ← [(f = 0, (pos = (start, start_alt), g = 0), parent = −1)].
Initialize alt_nodes ← {start : 0}.
Initialize visited ← [ ], parent_pos ← [ ].

% Main loop: Expand nodes from the min-heap
while future is not empty:
    Pop n from future.
    Append (n.pos,n.g.) to visited.
    Append n.parent to parent_pos.

    if n.pos = target:
        break while loop % If target is found, stop search.

    % Process each neighbor of the current node
    for neighbor ∈ n.neighbors:
        mc ← Euclidean_distance(n, neighbor).
        ac ← α / (n.alt + 1).
        ec ← ϵ · comp_grids[neighbor].

        if neighbor.alt > n.alt:
            cdc ← γ · (neighbor.alt − n.alt).
        else:
            cdc ← δ · (n.alt − neighbor.alt).

        neighbor.g ← mc + ac + ec + cdc.
        neighbor.h ← Euclidean_distance(neighbor, target).
        neighbor.f ← neighbor.g + neighbor.h.

        if neighbor.g < all_nodes [neighbor]:
            Push (neighbor.f, (neighbor, neighbor.g), visited. index (n)) to
              future.
            all_nodes [neighbor] ← neighbor.g.
    end for
end while

% Backtracking to reconstruct the optimal path
last_pt ← visited [−1].
parent_pos ← last_pt.parent.
Initialize route ← [last_pt.pos].

while parent_pos ≠ −1:
    point ← visited [parent_pos].
    Append point.pos to route.
    parent_pos ← point.parent.
end while

return route.
```

unacceptable paths are worse than others, and the score function must be flexible enough to distinguish between "bad" and "worse" so that it can progress toward better outcomes. Once acceptable paths are generated, the optimization constraints should refine the results to achieve improved routes until the process converges on a single "best" path. Since their environment also contains a measure of inherent stochasticity, the uncertainty constraints use a simple Bayesian network to model that uncertainty.[14]

SINGLE-AGENT 3D PATH-FINDING FOR THREAT AVOIDANCE

A* algorithm implementation

Our A* algorithm (see Algorithm 4) fundamentally works the same as traditional A*, using a priority queue (future) to track the open set of possible future points in the path and storing the closed set in a list (visited). Additionally, we use a list to track parent nodes (parent_pos) and a dictionary to track the lowest accumulated weight (g) for each node in the open set (all_nodes). Although our backwards distance (g) uses a 3D Euclidean distance, as discussed subsequently, the heuristic (h) is a simple 2D Euclidean distance from a node (n) to the target point (target), where the target's altitude is set to equal the node's altitude. As arguments, the function uses comp_grids, the start point altitude (start_alt) (default is start_alt = 0), the start (start) and target point (target) locations, the searchable altitudes (alts) in feet, and the specified NMs per grid square (NM_per_grid_square).

When a node n with the lowest $g+h$ is popped from future, all possible neighbor nodes are populated and stored in future if their g value is less than any precomputed g, as referenced in all_nodes.

ALGORITHM 5: CE search for single-agent 3D pathfinding.

```
% Initialization of parameters and data structures
Define grid size W × L × H; let the sample set size be N = W × L, and the elite set size be E = N * 0.01.
% Mat is an N dimensional array that encapsulates two arrays, prob_ij and prob_z, for lateral and vertical
  movement, respectively
For each position d within grid: initialize the probabilities of the two sub-arrays of Mat[d].
Define start and target, the two ends of the desired path. start is always at elevation step 0 (ground level);
  target is only defined laterally.
Define alts, max_climb_descent, and NM_per_grid_square,
Calculate lower_bound and upper_bound to determine allowable altitude transitions.
Set initial iteration count t ← 0.

while convergence criterion is not met:
% Outer loop: Generate and refine paths
    for k = 1 to N:
        Set current lateral node d ← start.
        Initialize path p_k ← d.
    % Inner loop: Construct random paths
        while d ≠ target and path length < L_max:
            Let (i,j) be the lateral cell coordinates of d, and z the elevation step component of d.
            Sample next lateral move from i to i′, using Mat[d].prob_ij[i]
            Compute available_z, a list of probabilities for the elevation steps within the elevation step
              margins lower_bound and upper_bound of the current elevation step z.
            Normalize available_z to obtain valid probability measures.
            Sample next elevation step z′ from Mat[d].prob_z[i]
            Update d: new lateral position (i′, j′) and elevation z′
            Extend p_k by appending d to it.
        end while
        Add p_k to paths.
    end for

    % Score the N sampled paths, where smaller path score means better path, and negative scores mean
      admissible(good) paths
    for p_k ∈ paths:
        if target ∉ p_k:
            scores[p_k] ← +(very large number) % Penalize paths that do not reach the target.
        else:
            Calculate g(p_k) using A* cost function.
            % Smaller score is better, hence we use the negated inverse of the A* cost function.
            scores[p_k] ← −1/g(p_k)
    end for

    % Select elite paths
    Sort paths by scores (ascending order).
    Select top E paths into elite_set.

    % Update transition matrix
    for p_k ∈ elite_set:
        for each transition (d,d′) ∈ p_k :
            Let i and i′ be the lateral component of d and d′, respectively.
            Let z be the elevation step component of d and d′, respectively.
            Update Mat[d].prob_ij[j] = Mat[d].prob_ij[j] + 1 if i′ is i′s neighbor on side j.
```

```
48:             Update Mat[d].prob_z[z] = Mat[d].prob_z[z] + 1 if z is the elevation step taken when in lateral position i.
49:          end for
50:       end for
51:       Normalize Mat to maintain valid probability distributions.
52:       Increment iteration count t ← t + 1.
53:    end while
54:
55:    % Return final path
56:    return Elite path with the best score.
```

A neighbor node is any of the eight lateral positions around n and any altitude determined by ± max_climb_descent, which is a function of max climb/descent rate (climb_rate).

$$\texttt{max\_climb\_descent} = \left\lfloor \frac{\texttt{climb\_rate} \times \texttt{NM\_per\_grid\_square}}{\texttt{alt\_diff}} \right\rfloor.$$

The variable alt_diff is calculated as the difference between two adjacent discrete altitudes (e.g., alts = [0, 1,000, 2,000], where alt_diff = 2,000 – 1,000 = 1,000).

A node's g value is equal to its parent's g value (denoted g') plus the weighted distance from its parent. The weighted distance is comprised of a movement cost (mc), altitude cost (ac), exposed cost (ec), and climb/descent cost (cdc). Constants α (altitude), γ (climb), δ (descent), and ϵ (exposed) are set and can be adjusted to promote varying aspects of a target path. Therefore, the g value is computed as:

$$g = g' + mc + ac + ec + cdc$$

With each variable defined as follows, where n is the node and n' is its parent node:

$$
\begin{aligned}
&mc = \text{Euclidean\_distance}(n', n)\\
&ac = \frac{\alpha}{n'.alt + 1}\\
&ec = \epsilon \cdot \texttt{comp\_grids}[n]\\
&\text{if } n.alt > n'.alt:\\
&\qquad cdc = \gamma \cdot (n.alt - n'alt)\\
&\text{else:}\\
&\qquad cdc = \delta \cdot (n'.alt - n.alt)
\end{aligned}
$$

Since h is a 2D Euclidean distance and all weight factors are nonnegative, any accumulated weighted distance from a node n to the target point is guaranteed to be greater than h, making h admissible. Therefore, this implementation of A* is guaranteed to return the least-cost path from the start to the target point, as specified by the given weights.

Our version of A* is unique in that it operates in 3D space, using altitude-specific weights to promote some elevations over others, while discouraging deviations by penalizing climbs and descents.

TABLE 1. Comparison of search algorithms.

	Dijkstra	A*	CE Search
Optimality	Optimal	Optimal	Approaches optimal
Speed	Fast	Fastest	Slow
Grid size	Medium	Large	Small
OSP-reliant	Yes	Yes	No
Multiagent capable	No	No	Yes

CE method implementation

For our implementation of the CE search method, we adapted the algorithm used in Drusinsky and Michael[14] to work in 3D space (Algorithm 5). The search domain is 3D with no hard obstructions, and contains at least one start point and one target point. Similar to the algorithm used in Drusinsky and Michael,[14] a transition matrix Mat is used to steer the creation of random paths in an inner loop, and the transition probabilities are updated in the outer loop.

In our algorithm, Mat integrates a third-dimension z with dedicated transition probabilities. For each entry d, Mat[d] contains a transition array of size 8, named prob_ij, which holds probabilities of moving laterally to any of the eight neighbor cells. In addition, Mat[d] contains an array of size num_altitudes, called prob_z, which holds the probabilities of transitioning to a discrete altitude block in the next cell. With this, lines 16 to 24 of Algorithm 5 sample a path by extending it probabilistically from node d at location (i, j, z) to node d'. The new lateral position is one of locations $(i-1, j-1)$, $(i, j-1)$, $(i+1, j-1)$, $(i-1, j)$, $(i+1, j)$, $(i-1, j+1)$, $(i, j+1)$, and $(i+1, i+1)$, while its new vertical position z' changes z to one of the altitudes specified in the alts array. To be true to the real-life the capabilities of an aircraft, we restrict the available altitude moves based on the current z of node d.

As discussed in the "Operation Context and Problem Domain" section, the environment is built with each grid square equaling some defined NMs in width. This allows us to establish a max_climb_descent using the same logic as in the A* algorithm (Algorithm 4). Therefore, when the last node d in the current path has a vertical value z, it can only select altitude blocks within the range [lower_bound, upper_bound], where

$$\texttt{lower\_bound} = \max(1, z - \texttt{max\_climb\_descent})$$
$$\texttt{upper\_bound} = \min(\texttt{length}(\texttt{alts}), z + \texttt{max\_climb\_descent}).$$

This ensures that no altitude below the second-lowest altitude is selected (the lowest is always 0 feet, and aircraft cannot fly on the ground), and no altitude above the highest available altitude is selected. Since prob_z contains probabilities of transitioning to every altitude, and these values sum to 1, we normalize the values in the available range (available_z). For every z_tran in available_z, we compute

$$\texttt{z\_tran}' = \frac{\texttt{z\_tran}}{\texttt{sum}(\texttt{available\_z})}.$$

This normalization allows the pathbuilder to construct realistic paths in the inner loop without spoiling the integrity of the overall prob_z.

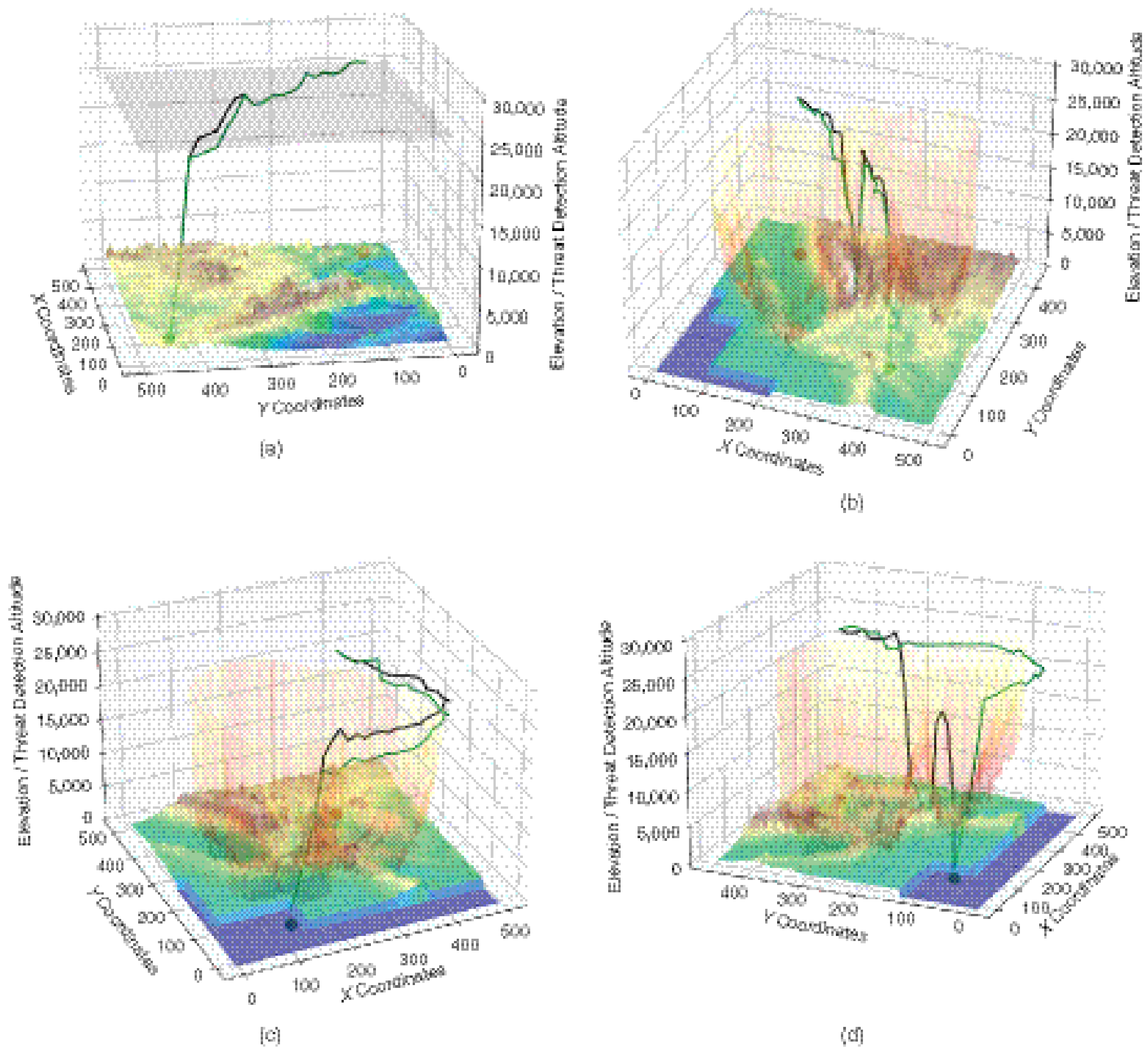


FIGURE 7. Single-agent, single-target paths generated by A* (black) and the CE method (green). (a) Almost identical paths in a zero-threat environment, (b) similar paths in a three-threat environment, and (c) similar paths in a five-threat environment. (d) An interesting case where vastly different paths with similar quality scores are produced by the two search methods.

In the outer loop of Algorithm 5 (lines 12 to 26), the algorithm creates N paths, with the default set to $n \times m \times z$, where n is the 2D size of the grid space, m is the number of possible movements at a given point in a path (for our purposes, $m = 8$), and z is the number of altitude blocks. After one iteration, each path p is given a score via our scoring function.

We follow the approach taken by Drusinsky and Michael,[14] which use rigid and optimization constraints to rate the quality of a given path. Paths that do not satisfy the rigid constraints receive large positive values, with larger scores equating to worse paths. Once "acceptable" paths are produced, the scores are negative, with values further from zero representing better paths. Our score function has only one rigid constraint: The path must reach the target. Similarly, we impose only one optimization constraint. We score each path according to its g value (path cost) from our A* algorithm. To fit the CE method, a path's score is the negated inverse of its g cost, i.e. $-(1/g)$.

This ensures that lower-cost paths are rewarded with better CE scores. Additionally, we can directly test CE-generated routes against A*-generated ones.

After N paths are scored, they are sorted from best to worst scores, and an elite set is selected. Using the elite set, we smoothly update the transition probabilities in `Mat` based on the transitions made in each path of the elite set. Doing so moves the probability distribution closer to the target distribution in terms of its KL divergence. Eventually, the algorithm converges, and the elite sets of every outer loop return the same path and associated score.

> Our version of A* is unique in that it operates in 3D space, using altitude-specific weights to promote some elevations over others, while discouraging deviations by penalizing climbs and descents.

Comparative results

We ran all search algorithms through various domains and aggregated the results. Table 1 highlights the relative benefits of each method. For our purposes, Dijkstra's algorithm was applied by using A* without incorporating a heuristic.

For the A* and CE methods, we designed 46 unique domains comprised of different start and target points and contained between zero and five threat systems. The start and target points were separated by distances ranging from 220 to 390 NMs, with the grid size set at 10 NM/grid square. Vertically, we used 14 elevation levels ranging from 0 to 24,000 feet AGL.

Since our (single agent, single target) problem adheres to the OSP, we determined that our A* search method returned the optimal flight path, given our specified requirements and constraints. Therefore, the A*-generated paths served as baselines against which to compare the CE-generated paths.

As predicted, the CE method approached optimality in terms of path-quality scores for the paths generated, validating its functionality for use in non-OSP problems. Figure 8 illustrates the relationship between CE-discovered quality scores compared to the baseline optimal A* search for single-agent, single-target. The near-linear trend between CE and A* quality scores suggest a systematic bias, with a near-fixed difference between all CE quality scores against those produced by A*. Additionally, the high coefficient of

ALGORITHM 6: Pseudocode for running multiple CE pathfinders.

```
Initialize A CE pathfinders CE_i, one corresponding to each agent i∈[1, A].
Set bestPath_i ← NIL to each i.

while convergence criteria are not met:
    for each CE_i:
        Perform one iteration of the generic CE algorithm (Algorithm 3) lines 3–7, where the scoring function
          takes into account the scores of all other bestPath_j ∈ bestPath.
        Update bestPath_i with the best admissible path found by CE_i.
    end for
    Construct bestPath ← [bestPath_1, bestPath_2,..., bestPath_A].
end while

return bestPath containing the best paths across all agents.
```

determination, R^2, indicates that the relationship between the two relative quality scores is highly predictable, enforcing the notion that fixed margin exists between the two search methods. A paired t-test further confirms this, with $t(45) = -11.29$ and $p \approx 10^{-14}$, indicating a highly statistically significant difference between the scores.

Visually, we observe that paths generated by the CE method typically align closely with their A*-produced counterparts, as shown in Figure 7(a)–(c). However, in some cases, significantly different paths yielded similar quality scores, as illustrated in Figure 7(d). This is likely a result of the score function, which may need further refinement in future applications of this research.

The tight correlation between the routes and relative scores produced by the two methods allows us to conclude that while the CE method may not return strictly optimal paths, they are predictably suboptimal with only a slight margin of difference. The CE method therefore generates adequate flight routes for any single aircraft and target. This enables us to feel confident that each route for multiple aircraft in a multiagent domain will be satisfactory.

MULTIAGENT SEARCH ALGORITHM

In our multiagent path finding problem the goal is to discover the best paths for all agents subject to the following inter-agent constraint: All agents must stay as close as possible to their leader (agent 1) during the first half of their respective flight paths.

Background

As mentioned in the "CE Search Algorithm" section, Drusinsky and Michael[14] used the CE method to conduct a multiagent search to optimize all paths and avoid collisions within a 2D search space. The general CE algorithm (Algorithm 3) and our implementation of CE for a single agent in a 3D search space (Algorithm 5) both depend on scoring based purely on the environment and a set of rigid and optimization constraints. However, we use the multiagent CE approach taken by Drusinsky and Michael,[14] where an individual agent's score depends not only on that agent's path, but also on the paths produced by the other agents in the domain. See Algorithm 6 for a general construct of a multiagent implementation of the CE search method.

The dependence of one agent's path on all other agents' paths allows us to produce solutions that approach Nash's bargaining solution or a Nash equilibrium, depending on whether the agents are cooperative or competitive, respectfully. For further explanation of Nash bargaining solution and equilibrium, see Chapter 2 in Daniel.[2] It is important to note that algorithms such as Dijkstra's and A*, discussed in the "Dijkstra/A*"

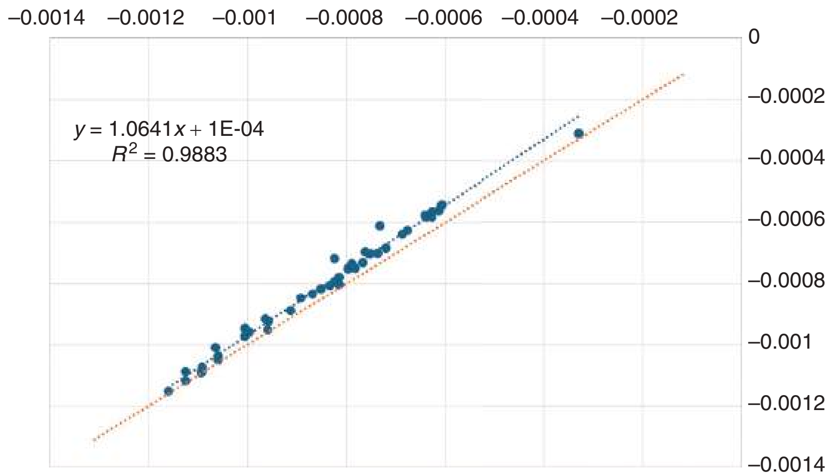


FIGURE 8. Path-quality scores of CE-generated single agent paths (blue) compared to the baseline optimal A* paths (red). Lower (more negative) values indicate better performance, demonstrating that CE consistently produces slightly suboptimal paths with a stable deviation from A*, even as search space complexity increases.

ALGORITHM 7: Scoring paths for multiagent (aircraft) CE method

for each path $p^i \in paths$:
    Compute $g(p^i)$ as per Algorithms 4 and 5.

    if length$(p^i) < \frac{\text{length}(p^o)}{2}$:
        Assign a large (bad) score to p^i: score $\leftarrow +$ large number.
        continue to the next path.

    Initialize **distance_score** $\leftarrow$ 0.
    for each node d^i in the first half of p^i:
      Let d^o be the corresponding node in p^o.
      distance_score $\leftarrow$ **distance_score** + Euclidean_distance(d^i, d^o).
    end for

    Update the score of p^i: score $\leftarrow g(p^i)$ + distance_score.
    score $\leftarrow -\frac{1}{\text{score}}$.
end for

section in Single-Agent Search Algorithms, cannot produce these types of solutions for a multiagent search environment.

Drusinsky and Michael[14] used a vehicle routing/ground shipping example to showcase the multiagent CE method, coordinating routes of four shipping trucks to maximize deliveries to multiple drop-off locations, given certain rigid, optimization, and uncertainty constraints. Ma and Lebacque[15] also modeled a version of a vehicle routing problem to deconflict travelers in a congested traffic environment, minimizing travel time and maximizing activity accomplishment. The work by Drusinsky and Michael[14] can be seen as a cooperation problem, whereas Ma and Lebacque's[15] research falls under a more competitive category.

Our work seeks to build upon these ideas by implementing multiple target locations with associated aircraft into the 3D threat-avoidance search area discussed in "Operational Context and Problem Domain" section. We reward cooperation between all aircraft as well as successful individual flight routes via the score function in the CE search method.

CE multiagent 3D path finding for threat avoidance

In the multiagent (multi-aircraft) problem, all aircraft have the same start

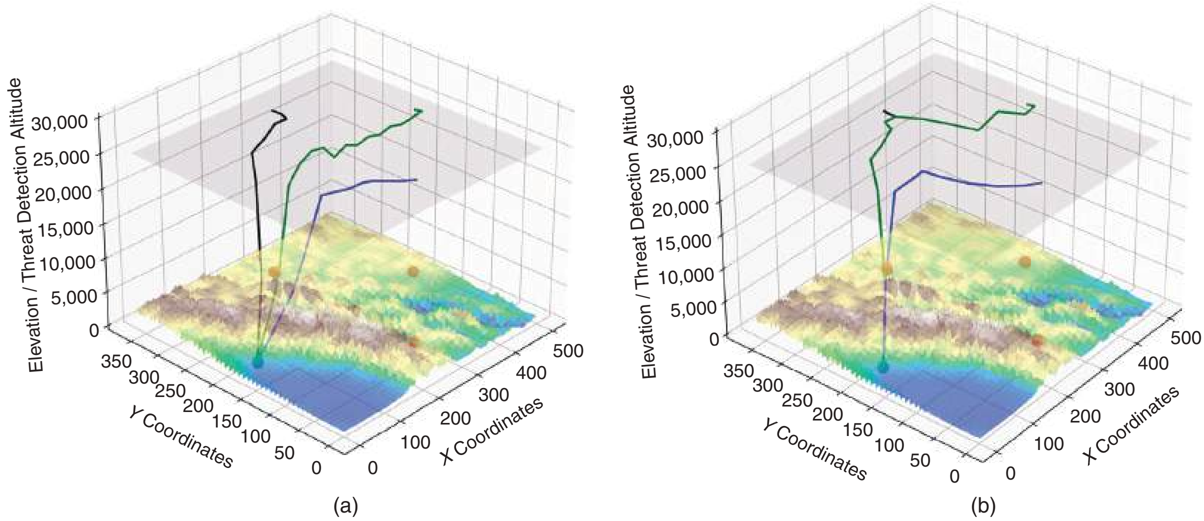


FIGURE 9. Three agents (aircraft) and their routes to three targets with no threats. (a) Paths generated by running A* for each aircraft individually and (b) paths generated using the multiagent CE method.

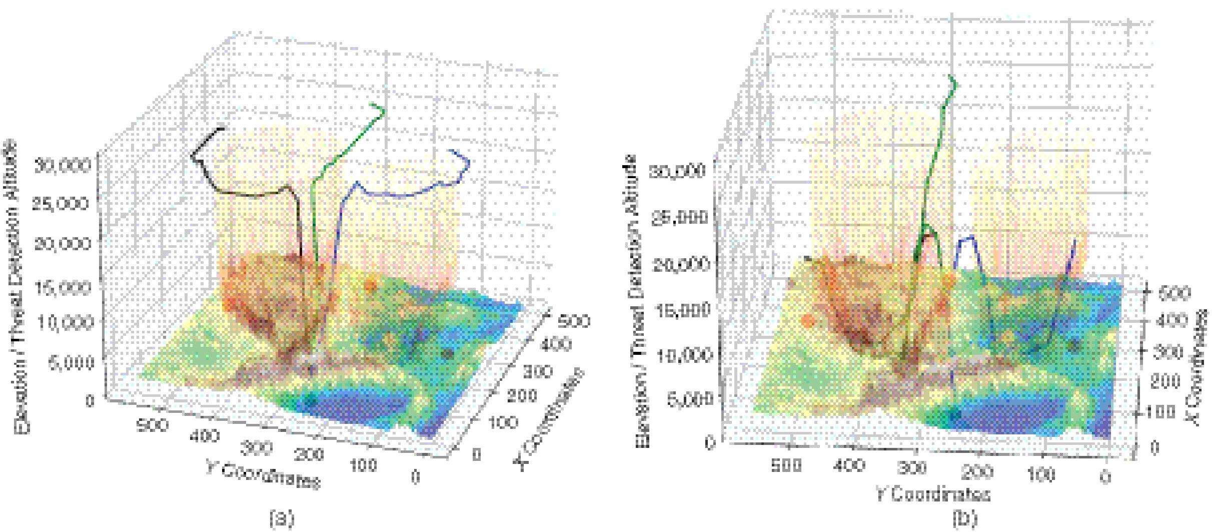


FIGURE 10. Three agents (aircraft) and their routes to three targets with two threats. (a) Paths generated by running A* for each aircraft individually and (b) paths generated using the multiagent CE method.

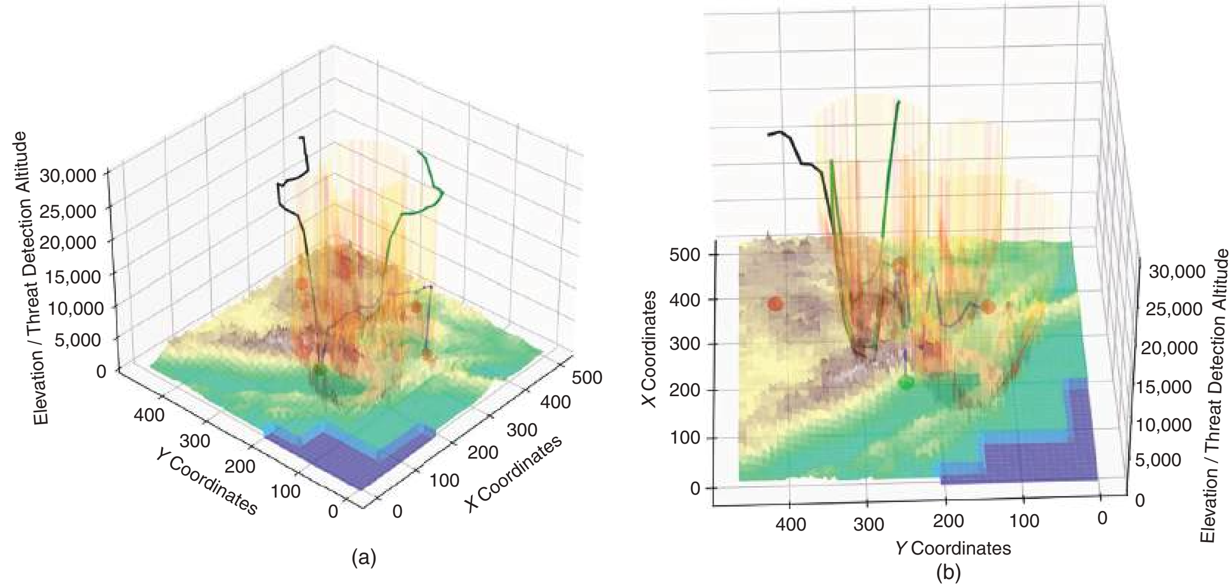


FIGURE 11. Three agents (aircraft) and their routes to three targets with five threats. (a) Paths generated by running A* for each aircraft individually and (b) paths generated using the multiagent CE method (b).

location, and each aircraft is assigned its own target. For our implementation of the multiagent CE search method, we chose to test cooperation between aircraft routes by imposing a constraint that rewards close proximity of routes for the first half of each path. This approach loosely mirrors the desire for mutual support between aircraft for a reasonable amount of time in a flight. To do so, we altered Algorithm 5, specifically in the scoring of paths: lines 33–35. An overview of the updated score function is detailed below and in Algorithm 7.

Each path p^i in a generated set `paths` gets its initial score value using the same the A* g function detailed in Algorithms 4 and 5. However, in addition to that score value, which reflects the quality of an aircraft's path independent of the others, we need to account for the above-mentioned interaircraft constraint. To that end, one agent is assigned the role of a "leader," with its path denoted as p^0. The additional component of p^i's cost function is its distance from the p^0 during the first half of p^i. More specifically, it is the Euclidean distance between each node d^i in the first half of p^i and the corresponding d^0 in p^0.

This cooperation between agents' paths to receive the best score for the set of paths results in a Nash bargaining solution. While each individual path may not be optimal for that specific agent, the combination of all paths is best for all agents as a whole. Figures 9–11 show the comparisons in different threat scenarios between three optimal paths for each aircraft, generated by A* (i.e., paths that ignore inter-agent constraints), compared to the sets of paths that are best for the group. It is obvious that the paths generated by the CE method do much better at prioritizing proximity between them at the beginning of the routes than those generated by A*. This shows that the CE method produces a resultant set of paths that is best for the group.

Our research has shown that the CE search method can effectively discover both single- and multiagent flight routes in a threat-populated 3D environment. Additionally, it provides a baseline for generating realistic LOS threat detection areas given accurate terrain data. This work integrates advanced path-finding algorithms with detailed threat modeling to automate route planning in complex scenarios. The CE method has shown a promising ability to facilitate cooperative planning across multiple unique platforms for the seamless integration of diverse operational goals.

Both A* and CE search algorithms demonstrate potential for further refinement and development, offering the capability to address increasingly complex and dynamic flight planning challenges. With enhancements, these methods could provide even greater accuracy, efficiency, and adaptability, further solidifying their role in mission-critical applications in contested airspaces. ■

REFERENCES

1. R. A. Azdy and F. Darnis, "Use of haversine formula in finding distance between temporary shelter and waste end processing sites," *J. Phys., Conf. Ser.*, vol. 1500, no. 1, Apr. 2020, Art. no. 012104, doi: 10.1088/1742-6596/1500/1/012104.
2. C. Daniel, "Multiagent aircraft flight route planning in a 3-dimensional threat-contested environment using cross-entropy search," M.S. thesis, C.S. Dept., Naval Postgraduate School, Monterey, CA, USA,

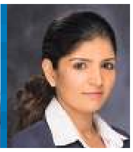
2025. [Online]. Available: https://libguides.nps.edu/citation/ieee-examples#unpublished
3. E. W. Dijkstra, "A note on two problems in connexion with graphs," *Numerische Mathematik*, vol. 1, no. 1, pp. 269–271, 1959, doi: 10.1007/BF01386390.
4. P. E. Hart, N. J. Nilsson, and B. Raphael, "A formal basis for the heuristic determination of minimum cost paths," *IEEE Trans. Syst. Sci. Cybern.*, vol. 4, no. 2, pp. 100–107, Jul. 1968, doi: 10.1109/TSSC.1968.300136.
5. S. P. Gudari and G. Vadivu, "A study on the performance of the a-star algorithm with various heuristics in grids and graphs," unpublished, 2023, doi: 10.13140/RG.2.2.31372.49287.
6. J. Li et al., "Improved a-star path planning algorithm in obstacle avoidance for the fixed-wing aircraft," *Electronics*, vol. 12, no. 24, 2023, Art. no. 5047, doi: 10.3390/electronics12245047.
7. H. Wang, S. Lou, J. Jing, Y. Wang, W. Liu, and T. Liu, "The EBS-A* algorithm: An improved A* algorithm for path planning," *PLoS One*, vol. 17, no. 2, Feb. 2022, Art. no. e0263841, doi: 10.1371/journal.pone.0263841.
8. T. Xue, Y. Cao, Y. Zhao, J. Ai, and Y. Dong, "Hybrid A*-based valley path planning algorithm for aircraft," *Aerospace*, vol. 11, no. 7, 2024, Art. no. 516, doi: 10.3390/aerospace11070516.
9. R. Y. Rubinstein, "Optimization of computer simulation models with rare events," *Eur. J. Oper. Res.*, vol. 99, no. 1, pp. 89–112, 1997, doi: 10.1016/S0377-2217(96)00385-2.
10. R. Rubinstein, "The cross-entropy method for combinatorial and continuous optimization," *Methodol. Comput. Appl. Probability*, vol. 1, no. 2, pp. 127–190, 1999, doi: 10.1023/A:1010091220143.
11. P.-T. De Boer, D. P. Kroese, S. Mannor, and R. Y. Rubinstein, "A tutorial on the cross-entropy method," *Ann. Oper. Res.*, vol. 134, no. 1, pp. 19–67, Feb. 2005, doi: 10.1007/s10479-005-5724-z.
12. B. E. Helvik and O. Wittner, "Using the cross-entropy method to guide/govern mobile agent's path finding in networks," in *Mobile Agents for Telecommunication Applications* (Lecture Notes in Computer Science), vol. 2164, G. Goos, J. Hartmanis, J. Van Leeuwen, S. Pierre, and R. Glitho, Eds. Heidelberg, Germany: Springer-Verlag, 2001, pp. 255–268.
13. K. Chepuri and T. Homem-de Mello, "Solving the vehicle routing problem with stochastic demands using the cross-entropy method," *Ann. Oper. Res.*, vol. 134, no. 1, Jan. 2005, Art. no. 153, doi: 10.1007/s10479-005-5729-7.
14. D. Drusinsky and J. B. Michael, "Multiagent pathfinding under rigid, optimization, and uncertainty constraints," *Computer*, vol. 54, no. 7, pp. 111–118, Jul. 2021, doi: 10.1109/MC.2021.3074264.
15. T.-Y. Ma and J.-P. Lebacque, "A cross-entropy based multiagent approach for multiclass activity chain modeling and simulation," HAL, Working Papers Halshs-00310903, May 2011 [Online]. Available: https://ideas.repec.org/p/hal/wpaper/halshs-00310903.html
16. D. Drusinsky, M. Litton, and J. B. Michael, "Machine-learned verification and advance notice oracles for autonomous systems," *Computer*, vol. 56, no. 7, pp. 121–130, Jul. 2023, doi: 10.1109/MC.2023.3265732.
17. M. D'Emidio, "Faster algorithms for mining shortest-path distances from massive time-evolving graphs," *Algorithms*, vol. 13, no. 8, Art. no. 191, Aug. 2020, doi: 10.3390/a13080191.
18. T. Akiba, Y. Iwata, and Y. Yoshida, "Dynamic and historical shortest-path distance queries on large evolving networks by pruned landmark labeling," in *Proc. 23rd Int. Conf. World Wide Web*, New York, NY, USA: Association for Computing Machinery, 2014, pp. 237–248, doi: 10.1145/2566486.2568007.
19. B. Xuan, A. Ferreira, and A. Jarry, "Evolving graphs and least cost journeys in dynamic networks," Mar. 2003.
20. Z. Gong, Y. Zeng, and L. Chen, "Querying shortest path on large time-dependent road networks with shortcuts," 2023, *arXiv:2303.03720*.
21. "Proof of the law of sines." Math Open Reference. [Online]. Available: https://www.mathopenref.com/lawofsinesproof.html
22. S. Karaman and E. Frazzoli, "Sampling-based algorithms for optimal motion planning," *Int. J. Rob. Res.*, vol. 30, no. 7, pp. 846–894, Jun. 2011, doi: 10.1177/0278364911406761.
23. L. Babel, "Online flight path planning with flight time constraints for fixed-wing UAVs in dynamic environments," *Int. J. Intell. Unmanned Syst.*, vol. 10, no. 4, pp. 416–443, Nov. 2022, doi: 10.1108/IJIUS-11-2020-0063.
24. J. F. Nash, "The bargaining problem," *Econometrica*, vol. 18, no. 2, Apr. 1950, Art. no. 155, doi: 10.2307/1907266.
25. J. F. Nash, "Equilibrium points in n-person games," *Proc. Nat. Acad. Sci.*, vol. 36, no. 1, pp. 48–49, Jan. 1950, doi: 10.1073/pnas.36.1.48.
26. J. Nash, "Non-cooperative games," *Ann. Math.*, vol. 54, no. 2, Sep. 1951, Art. no. 286, doi: 10.2307/1969529.

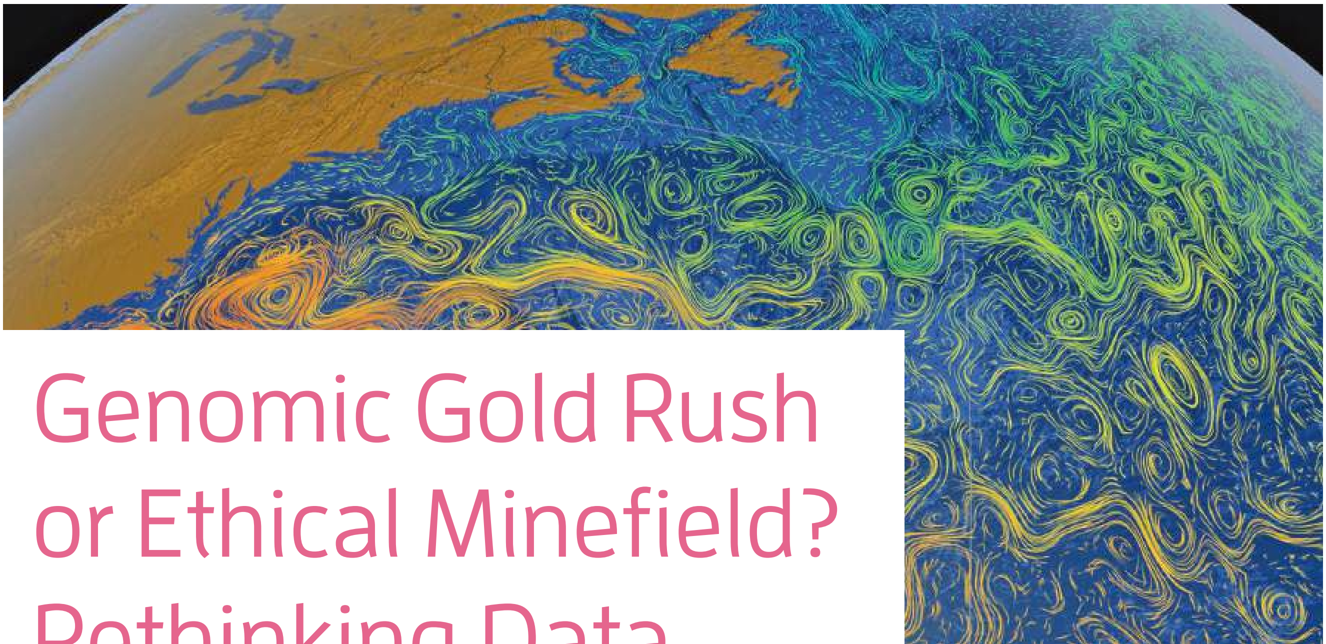
CHRISTOPHER (BUD) DANIEL is a U.S. Marine Harrier II+ pilot and M.S. student at the Naval Postgraduate School, Monterey, CA 93943 USA. Contact him at christopher.daniel@nps.edu.

DORON DRUSINSKY is a professor in the Department of Computer Science, Naval Postgraduate School, Monterey, CA 93943 USA. Contact him at ddrusins@nps.edu.

LOREN E. PEITSO is a senior lecturer in the Department of Computer Science, Naval Postgraduate School, Monterey, CA 93943 USA. Contact him at lepeitso@nps.edu.

EDITORS
NORITA AHMAD American University of Sharjah; nahmad@aus.edu
PREETI CHAUHAN IEEE Reliability Society; preeti.chauhan@ieee.org





Genomic Gold Rush or Ethical Minefield? Rethinking Data Practices in Health Tech Giants

Aqilah Julaihi, Warwick Medical School
Norita Ahmad, American University of Sharjah

Direct-to-consumer genomic testing offers unprecedented access to genetic insights but raises significant ethical challenges. Addressing these issues requires transparent data practices, stronger informed consent mechanisms, and ethical governance to ensure equitable and responsible use of genomic innovations.

Digital Object Identifier 10.1109/MC.2025.3533264
Date of current version: 27 March 2025

"Genetics is not just about the genes we inherit but how we use them." This profound statement by Richard Dawkins[1] highlights the transformative potential of genetic information in reshaping human health and disease management. At the heart of this transformation lies genomic data, a comprehensive blueprint housed within every cell of the human body. This data, composed of approximately six billion DNA letters,[2] contains unique variations that can reveal critical insights into an individual's health, ancestry, and predisposition to disease.[3] Advances in technology such as artificial intelligence (AI) have turned this wealth of information into a valuable asset, powering breakthroughs in personalized medicine and precision health care.[4] Yet, the increasing commodification of genomic data by health-care systems and tech giants has brought ethical concerns to the forefront, particularly around privacy, ownership, and the equitable use of this sensitive information.[5]

The rapid growth of direct-to-consumer (DTC) genomic testing has made this once-exclusive knowledge broadly accessible, fostering what some describe as the "democratization" of genetic information. Companies like 23andMe, Ancestry.com, and Nebula Genomics have enabled millions of individuals to explore their genetic profiles, uncovering health risks, ancestry details, and personal traits from the comfort of their homes.[5] However, this ease of access comes with significant ethical challenges. As genomic data becomes increasingly commodified, it is often treated as a lucrative asset by corporations, raising concerns over privacy, data security, and informed consent.[6] Many consumers remain unaware of the risks associated with sharing their DNA, including data breaches, unauthorized usage, and the potential for genetic discrimination. Addressing these challenges requires robust ethical frameworks that prioritize transparency, protect user autonomy, and ensure equitable use of genomic resources.

PRIVACY AND OWNERSHIP OF GENOMIC DATA

Genomic data are deeply personal, offering insights into health, ancestry, and familial connections. However, when individuals submit their genetic information to DTC testing companies, they often unknowingly surrender control over this data. Companies like 23andMe have faced criticism for sharing anonymized genetic data with pharmaceutical firms without explicit consumer consent. This raises significant questions about who truly owns and controls genomic data.[7]

The issue of ownership lies at the heart of ongoing ethical debates. Should genomic data be considered personal property, or do companies have the right to commercialize it once submitted? The legal landscape surrounding this question is fragmented and inconsistent. Table 1 shows a comparative analysis of key data protection laws that highlights this disparity.

As illustrated in Table 1, gaps remain even in regions with robust data protection frameworks, such as the European Union's GDPR. For instance, genomic data are not explicitly categorized as unique personal information, leaving its interpretation to varying legal jurisdictions.[8] In the United States, GINA provides protections against genetic discrimination in employment and health insurance, yet it excludes critical areas like life and disability insurance, exposing individuals to potential misuse.[9]

These regulatory gaps have allowed companies to prioritize their commercial interests. For example, 23andMe's partnerships with GlaxoSmithKline raised concerns about transparency, as many users were unaware that their anonymized data could be sold to pharmaceutical companies for drug development.[7] This lack of informed consent underscores the need for stronger regulations and clearer definitions of ownership and control over genetic data.

Ethical concerns are particularly pressing for marginalized populations, such as lower-income or minority groups. These communities may have their genomic data disproportionately used for profit without fair representation or benefits. To address these concerns, experts have suggested treating genomic data as personal property, akin to intellectual property, allowing individuals to retain control over how their data are used, shared, or monetized.[5] This approach aligns with privacy laws like GDPR, which give individuals greater control over their personal data.

INFORMED CONSENT AND TRANSPARENCY

Informed consent is a cornerstone of ethical medical and research practices, yet it remains a significant challenge in the DTC genomic testing industry. Many companies employ complex, jargon-filled consent forms that obscure how genetic data will be collected, used, and shared. As a result, consumers often sign agreements without fully understanding the potential implications, including data sharing with third parties or the monetization of their information.[11]

To better understand this process and its challenges, it is helpful to visualize the lifecycle of genomic data in DTC testing. Figure 1 outlines the key stages, showing critical points where transparency and consent mechanisms are essential:

- *Data collection*: Genetic samples are collected and processed, often with minimal consumer

TABLE 1. Comparative overview of genomic data protections across major regions.

Region	Applicable Laws	Protections Offered	Gaps/Challenges
European Union	General Data Protection Regulation (GDPR)	Comprehensive data privacy for personal information, but genomic-specific gaps remain.	Does not explicitly address genomic data as unique; interpretation varies across jurisdictions.
United States	Genetic Information Non-Discrimination Act (GINA)	Protects against genetic discrimination in employment and health insurance.	Excludes life insurance, disability insurance, and other nonhealth-related uses.
Global	Various national and regional frameworks	Patchy protections; often lacks specificity for genomic data.	Inconsistent enforcement; no universal standard for genomic data handling.

awareness of how their information may be used.

- *Data storage*: Information is stored in databases, which may lack adequate security measures.
- *Data usage*: Companies analyze data for health insights, ancestry reports, or other consumer-facing services.
- *Data sharing*: Genetic data may be shared with third parties, such as pharmaceutical companies, often without explicit user consent.
- *Data monetization*: Companies may profit from selling anonymized data, raising ethical concerns about ownership and autonomy.
- *Potential risks*: Risks include data breaches, re-identification of anonymized data, and misuse leading to genetic discrimination.

This lifecycle highlights the importance of robust informed-consent mechanisms at every stage. True informed consent should empower consumers to make decisions based on clear, accessible information about how their genetic data will be handled.[11] Companies should adopt practices that prioritize simplicity and transparency,[12] such as simplifying legal jargon to ensure consumers can easily understand what they are agreeing to, using diagrams or infographics to illustrate data flows and potential uses, allowing users to modify their consent preferences over time, such as opting in or out of specific uses or research projects, and keeping consumers informed of new developments, such as changes to how their data are stored or shared.

Blockchain technology offers promising solutions for enhancing transparency and data security. For example, companies like Nebula Genomics use blockchain protocols to give users more control over data access and sharing. This technology enables an auditable record of who accesses data and for what purpose. However, critics argue that blockchain alone cannot address challenges like ensuring users fully comprehend their rights or the irreversible nature of some data-sharing agreements.[10]

Transparency is also crucial for fostering trust in the DTC genomic testing industry. Consumers must have confidence that their data are being handled ethically and securely. Without clear and accessible consent mechanisms, public trust in these services could erode, limiting participation and undermining the potential benefits of genomic innovation.

> This lack of informed consent underscores the need for stronger regulations and clearer definitions of ownership and control over genetic data.

MISUSE AND POTENTIAL HARM

The DTC genomic testing industry presents a range of ethical challenges, particularly concerning the misuse of data and its potential harm to individuals. Genetic data are profoundly personal, revealing sensitive information about not only the individual but also their family members.[13] These insights, while invaluable for advancing personalized medicine, can also lead to significant risks if misused.

One of the most pressing concerns is the potential for privacy breaches. Data breaches, such as the 2018 MyHeritage incident that exposed information on over 92 million users, show the vulnerabilities in current genomic data storage practices.[14] Even anonymized data are not immune to reidentification, as advanced algorithms can cross-reference datasets to deduce personal identities. Breaches like these erode consumer trust and expose individuals to unexpected risks.

The commercialization of genomic data raises complex ethical questions. Many DTC companies share or sell anonymized genetic data to pharmaceutical companies and other third parties without obtaining explicit user consent. While this practice supports drug development and other research, it often occurs without adequate transparency, leaving consumers unaware of how their data are monetized.[11] This commodification of genetic information shifts control away from the individual, creating imbalances in data ownership and benefit distribution.

Another significant risk is genetic discrimination. Employers, insurers, or government entities could misuse genetic data to make decisions that disadvantage individuals. For example, an insurer might adjust premiums or deny coverage based on a customer's

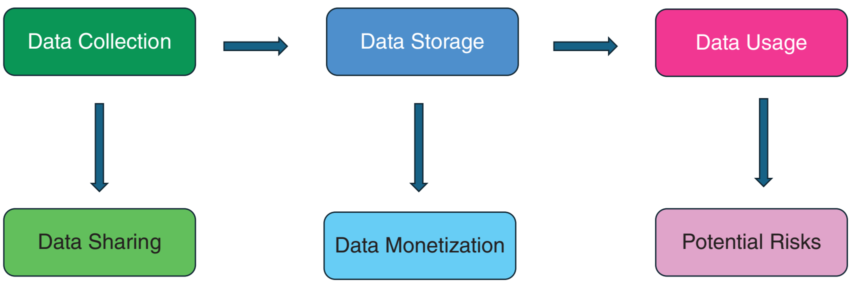


FIGURE 1. Lifecycle of genomic data in DTC testing.

genetic predisposition to certain diseases.[15] While laws like the GINA offer some protections, they are often limited in scope, excluding areas like life and disability insurance. This regulatory gap leaves individuals vulnerable to exploitation.[9]

By aligning technological advancements with these standards, the DTC genomic testing industry can safeguard consumer rights while achieving its potential.

The psychological effects of genomic testing can also be profound. Learning about a predisposition to severe or incurable conditions may cause anxiety, depression, or feelings of helplessness. Without proper counseling and clear communication of results, individuals may misinterpret their risk levels, leading to unnecessary health interventions or distress.[16] This issue is especially evident in cases where DTC companies fail to provide adequate support resources for interpreting results. Marginalized populations, including low-income or minority groups, are particularly at risk of exploitation in the genomic testing industry. These communities may lack the resources or education to fully understand the implications of submitting their genetic data.[16] Consequently, they may be disproportionately targeted for data collection without receiving equitable benefits from the resulting advancements.

There are several measures that can be adopted to address these challenges. For example, governments should consider expanding existing legal frameworks to include genomic data protections, ensuring equitable treatment and reducing the risk of misuse. Companies must also clearly communicate how genetic data will be used, stored, and shared, empowering consumers to make informed decisions. Additionally, they should also simplify consent forms and providing ongoing updates about data usage. Finally, both governments and companies should engage marginalized communities in discussions about genomic testing and its implications in order to ensure trust and equitable participation.

IMPLICATIONS FOR PUBLIC TRUST

The rapid expansion of the DTC genomic testing industry has brought ethical concerns into sharp focus, particularly regarding its impact on public trust. Trust is essential for the continued success and growth of the industry, as it ensures consumer participation and the responsible use of genomic data. However, the lack of transparency in data usage, insufficient informed consent mechanisms, and frequent privacy breaches have eroded consumer confidence.[17] When companies fail to clearly disclose how they collect, store, and share genetic data, they increase the risk of misuse and discrimination. For instance, fears about data being sold to third parties or used for purposes beyond what was initially agreed upon are common among consumers.[11] Addressing these concerns requires companies to adopt robust transparency measures, such as detailed consent processes, regular updates on data usage, and clear communication about security protocols.

The relationship between informed consent and trust is equally critical. Consumers need to fully understand the implications of sharing their genetic data, including potential emotional and psychological risks. Without accessible and user-friendly consent mechanisms, consumers may feel coerced into agreements they do not fully comprehend. This lack of clarity can lead to harmful consequences and further diminish trust in the industry. Furthermore, public trust hinges on the industry's ability to protect the privacy and security of genetic data. High-profile data breaches, such as the MyHeritage incident, have highlighted vulnerabilities in existing security frameworks.[14] Companies must prioritize strong data protection measures and communicate their efforts transparently to reassure consumers that their information is secure.

Corporate accountability and adherence to ethical standards also play a key role in fostering trust. Exaggerated claims or unsupported predictions in genetic testing can mislead consumers, damaging the industry's credibility. Regulatory oversight can help ensure that companies meet established scientific standards and avoid practices that exploit consumers.[17] Ultimately, building and maintaining public trust requires a collaborative approach. Policymakers, technologists, health-care providers, and ethicists must work together to establish ethical guidelines and regulatory frameworks. By aligning technological advancements with these standards, the DTC genomic testing industry can safeguard consumer rights while achieving its potential.

TOWARDS ETHICAL GENOMIC INNOVATION

Genomic data holds the potential to revolutionize human health and disease management, offering insights that were once inconceivable. DTC genomic testing has democratized access to this information, enabling millions to uncover details about their ancestry, health risks, and personal traits. However, with these advancements come significant ethical challenges, such as privacy risks, informed consent issues, and the erosion of public trust.

As Richard Dawkins famously noted, "*Genetics is not just about the genes we inherit but how we use them.*" This sentiment underscores the dual responsibility of using genomic data for progress while safeguarding its ethical use. Genomic data are more than just a scientific resource; it represents an intimate map of human identity.[3] Addressing the ethical concerns associated with its use requires balancing innovation with accountability. The commodification of genetic information underscores the need for transparency, robust governance, and equitable practices to ensure that advancements in genomics benefit society as a whole.

Emerging technologies like blockchain and AI present opportunities to enhance data security and improve personalization. However, these tools must be complemented by strong regulatory frameworks and ethical oversight. Collaboration among stakeholders, including policymakers, researchers, and industry leaders is vital to establishing standards that protect individual rights while fostering innovation.

In conclusion, the future of the DTC genomic testing industry depends on its ability to address ethical concerns proactively. By establishing a foundation of trust, transparency, and accountability, the industry can continue to innovate responsibly, unlocking the transformative potential of genomic data while safeguarding individual and societal well-being. ■

ACKNOWLEDGMENT

Norita Ahmad is the corresponding author.

REFERENCES

1. R. Dawkins, *The Selfish Gene*. Oxford, U.K.: Oxford Univ. Press, 1976.
2. S. Kim-Hellmuth et al., "Cell type–specific genetic regulation of gene expression across human tissues," *Science*, vol. 369, no. 6509, Sep. 2020, doi: 10.1126/science.aaz8528.
3. M. J. Khoury, "The shift from personalized medicine to precision medicine and precision public health: Words matter!," *CDC Genomics and Precision Health Blog*, Apr. 21, 2016. [Online]. Available: https://blogs.cdc.gov/genomics/2016/04/21/shift/
4. Y. Duan, J. S. Edwards, and Y. K. Dwivedi, "Artificial intelligence for decision making in the era of big data – Evolution, challenges and research agenda," *Int. J. Inf. Manage.*, vol. 48, no. 1, pp. 63–71, Oct. 2019, doi: 10.1016/j.ijinfomgt.2019.01.021.
5. J. Moran, "Privacy perspectives on direct-to-consumer genetic testing in the era of big data: Role of blockchain technology in genomics," *Tulane J. Technol. Intellectual Property*, vol. 22, pp. 185–204, Spring 2020.
6. B. Berger and H. Cho, "Emerging technologies towards enhancing privacy in genomic data sharing," *Genome Biol.*, vol. 20, no. 1, Jul. 2019, Art. no. 128, doi: 10.1186/s13059-019-1741-0.
7. J. Ducharme, "A major drug company now has access to 23andMe's genetic data. Should you be concerned?" *Time*, Jul. 26, 2018. [Online]. Available: https://time.com/5349896/23andme-glaxo-smith-kline/
8. K. Pormeister, "Genetic data and the research exemption: Is the GDPR going too far?" *Int. Data Privacy Law*, vol. 7, no. 2, pp. 137–146, May 2017, doi: 10.1093/idpl/ipx006.
9. Y. Joly, C. Dupras, M. Pinkesz, S. A. Tovino, and M. A. Rothstein, "Looking beyond GINA: Policy approaches to address genetic discrimination," *Annu. Rev. Genomics Human Genetics*, vol. 21, no. 1, pp. 491–507, Aug. 2020, doi: 10.1146/annurev-genom-111119-011436.
10. D. Hofman and A. Novin, "Blocked and chained: Blockchain and the problems of transparency," in *Proc. Assoc. Inf. Sci. Technol.*, 2018, vol. 55, no. 1, pp. 171–178, doi: 10.1002/pra2.2018.14505501019.
11. A. E. Raz, E. Niemiec, H. C. Howard, S. Sterckx, J. Cockbain, and B. Prainsack, "Transparency, consent and trust in the use of customers' data by an online genetic testing company: An exploratory survey among 23andMe users," *New Genetics Soc.*, vol. 39, no. 4, pp. 1–24, May 2020, doi: 10.1080/14636778.2020.1755636.
12. L. M. Beskow, and K. P. Weinfurt, "Exploring understanding of 'understanding': The paradigm case of biobank consent comprehension," *Amer. J. Bioethics*, vol. 19, no. 5, pp. 6–18, May 2019, doi: 10.1080/15265161.2019.1587031.
13. X. Shi and X. Wu, "An overview of human genetic privacy," *Ann. N.Y. Acad. Sci.*, vol. 1387, no. 1, pp. 61–72, Jan. 2017, doi: 10.1111/nyas.13211.
14. Reuters Staff, "Security breach at MyHeritage website leaks details of over 92 million users," *Reuters*, Jun. 5, 2018. [Online]. Available: https://www.reuters.com/article/business/security-breach-at-myheritage-website-leaks-details-of-over-92-million-users-idUSKCN1J1301/
15. C. D. Zick, C. J. Mathews, J. S. Roberts, R. Cook-Deegan, R. J. Pokorski, and R. C. Green, "Genetic testing for Alzheimer's disease and its impact on insurance purchasing behavior," *Health Affairs*, vol. 24, no. 2, pp. 483–490, Mar. 2005, doi: 10.1377/hlthaff.24.2.483.
16. J. Collmann and S. A. Matei, *Ethical Reasoning in Big Data an Exploratory Analysis*. Cham, Switzerland: Springer-Verlag, 2016.
17. M. Majumder, C. Guerrini, and A. Mcguire, "Annual review of medicine direct-to-consumer genetic testing: Value and risk," *Annu. Rev. Med.*, vol. 72, no. 1, pp. 151–166, 2020, doi: 10.1146/annurev-med-070119-114727.

AQILAH JULAIHI is a master of public health candidate at the University of Warwick, CV4 7AL Coventry, U.K. Contact her at aqilaanis@gmail.com.

NORITA AHMAD is a professor of information systems and business analytics at the American University of Sharjah, Sharjah, United Arab Emirates. Contact her at nahmad@aus.edu.

Uncertainty in Machine Learning and Future Computers

Shinya Takamaeda-Yamazaki, The University of Tokyo

Unfortunately, computers are subject to uncertainty. If we can quantitatively evaluate the uncertainty in the output of ordinary computers, we can realize efficient and trustworthy computers that actively utilize fluctuating computation principles.

What do people see computers as? Most people probably think of a computer as a device that is fast, performs calculations without error, and answers accurately. This is the expectation and trust in computers. In reality, however, there are bugs in computer hardware and software, and they can behave in ways that designers and users do not anticipate. Also, hardware will eventually fail. There are not only permanent failures but also temporary errors caused by external factors. However, various countermeasures, such as hardware redundancy, error correction techniques, and circuit design with margins that account for variations in semiconductor devices, aim to minimize computer uncertainty as much as possible to ensure that calculations are performed accurately. Computer accuracy is not free, and we pay a cost to remove uncertainty in exchange.

DO COMPUTERS HAVE UNCERTAINTY?

One of the important uses of such computers today, with as much uncertainty removed as possible, is in artificial intelligence, especially machine learning. Ideally, the behavior of the machine learning model would be fair and accurate at all times, but this is not the case in practice. Machine learning models can be considered as software that is programmed in a data-driven manner, training models with a set of correct answers corresponding to the input, but the training data are not always perfectly correct. In addition, the hypotheses

Digital Object Identifier 10.1109/MC.2025.3536056
Date of current version: 27 March 2025

EDITORS
TIMOTHY JONES University of Cambridge, U.K., timothy.jones@cl.cam.ac.uk
ROBERT MULLINS University of Cambridge, U.K., robert.mullins@cl.cam.ac.uk

assumed during training are not always equally rigorous during inference. Therefore, the output of machine learning is not always accurate and contains uncertainty. This uncertainty in machine learning can be divided into two main categories: "aleatoric uncertainty," which is the uncertainty that comes from the noise and variability inherent in the data themselves, and "epistemic uncertainty," which originates from the model or training process. The former cannot be completely eliminated, no matter how many training data are collected. The latter can be reduced to some extent by enhancing training data and devising model structures, but it cannot be completely eliminated, because the scale of training data and models is finite. In other words, even if we pay various costs to reduce the uncertainty of the computer itself as much as possible, we cannot completely remove uncertainty from the machine learning system realized on such computers.

UNCERTAINTY ESTIMATION IN MACHINE LEARNING

To realize a trustworthy machine learning system, it is desirable to quantify the level of confidence (and uncertainty) in the inference of the machine learning model. Let us take Bayesian deep learning as an example. Bayesian deep learning quantitatively estimates uncertainty by expressing network weights and estimation results in terms of probability distributions. This has the great advantage that the model can clearly show, for example, how fuzzy the prediction results are for an unknown sample. However, the act of estimating uncertainty in Bayesian deep learning has the disadvantage that it requires sampling a sufficient number of times for the probability distribution, which significantly increases the computational cost. The increase in computational cost naturally leads to an increase in execution time and energy consumption.

Via Gaussian distribution

Domain-specific architectures that efficiently estimate "uncertainty" in Bayesian deep learning have been studied. An early accelerator study on Bayesian deep learning is VIBNN[1]; VIBNN is an accelerator that efficiently handles variational inference. For variational inference in Bayesian deep learning with a Gaussian distribution as a prior, a pseudorandom number generator with Gaussian distribution and an accelerator architecture for efficient weight sampling are proposed. By embedding the random number generator based on Gaussian distribution inside the accelerator, high memory access and computational efficiency are achieved. We have also proposed an even more efficient weight generation scheme for Bayesian deep learning using an approximation based on the inverse function method.[2]

Bayesian deep learning based on variational inference generally repeats forward propagation multiple times for the same input data, using different weight instances from sampling. The predictive distribution and uncertainty are then estimated from the output histogram. In fact, the computational process can find similarities in the values. We have developed accelerator for sparse Bayesian neural networks (ASBNN),[3] a cooperative algorithm and hardware architecture that accelerates computation while preserving output quality by exploiting the nature of intermediate results that appear during Bayesian deep learning computation. ASBNN is an approximate computation accelerator that eliminates computation by exploiting the similarity latent in sampling computations. The weight instances are sampled for each forward propagation calculation, and each layer's calculation proceeds. Figure 1 shows how much the output of an intermediate layer changes with each forward propagation. The horizontal axis is the absolute difference value from the initial forward propagation, and the vertical axis is the number of occurrences. The observation shows that most of the outputs of each layer have similar values to those of the first forward propagation, and only a

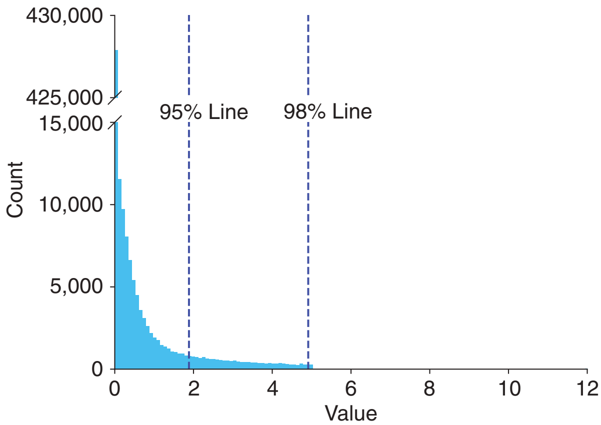


FIGURE 1. The distribution of value differences of an intermediate layer for different forward propagations (cited from Fujiwara and Takamaeda-Yamazaki[3]).

few values differ significantly, even though the sampling takes different weight instances.

ASBNN is a method to reduce the number of calculations by increasing the sparsity (fraction of zeros) that appeared during the variational inference by taking advantage of the similarity latent in multiple forward propagations. The key idea is that calculations of only some limited outputs depending on significantly changing nodes in the previous layer are sufficient to estimate a final output distribution adequately.

In general computing, other than machine learning, making effective use of uncertainty, rather than just hiding it, will be an important theme in future computer architecture.

By Monte Carlo dropout

Instead of sampling based on Gaussian distribution, there is an easy-to-realize perturbation method using Monte Carlo dropout (MC-Dropout).[4] In the original dropout, nodes are randomly disabled to suppress overlearning of the neural network model during training. MC-Dropout is used as a perturbation in the inference phase to estimate the uncertainty of the model output. Different nodes will be dropped for each forward propagation, and the output fluctuates correspondingly. The predictive distribution of output is estimated based on this fluctuation. Compared with variational inference based on Gaussian distribution, sampling for weights is not required, which simplifies the random number generator and the circuit configuration of the accelerator.

The interesting point here is that the value of a node to be dropped out is treated as zero, so there is no need to compute the predropout value of the node if it is known in advance that it will be dropped out. Fast-BCNN[5] has been proposed as an accelerator that finds the dropout pattern first and computes only the nodes that are not dropped out. As a related approach, Masksembles[6] has been proposed as a method to obtain predictive distributions using multiple fixed dropout patterns instead of random dropouts, and an efficient field-programmable gate array (FPGA)-based accelerator[7] that supports Masksembles has also been proposed.

By moment propagation

A major drawback of Bayesian deep learning variational inference based on sampling is its high computational cost. Increasing the number of sampling times will yield a more stable predictive distribution, but this increases the computational cost. Therefore, deterministic variational inference (DVI), a sampling-free method,[8] has been proposed to obtain the predictive distribution with one forward propagation. In DVI, each layer is extended to input moments of the probability distribution, and each layer propagates the moments directly to the subsequent layers.

In this case, Taylor expansion is required to calculate moments for nonlinear functions such as rectified linear unit (ReLU) and Sigmoid, which are widely used in neural networks, and the computational cost is high. In BYNQNet,[9] moments can be calculated using only polynomial operations by using the quadratic activation function. We also proposed Gaussian error moment propagation (GeMP),[10] a hardware-friendly nonlinear function for moment propagation based on the Gaussian error linear unit, which is used in transformer models. As shown in Figure 2, compared with ReLU-based moment propagation, the GeMP-based moment propagation circuit is lightweight and low latency.

DO NOT CONCEAL UNCERTAINTY, EXPLOIT IT

So far, we have introduced computer architecture techniques based on general digital circuit techniques that deal with uncertainty in machine learning and do not tolerate uncertainty. However, a question arises here. Should we pay a significant cost to remove uncertainty from a computer that processes a program (machine learning) that, in principle, cannot remove uncertainty in the first place? If uncertainty in computers is tolerated, what kinds of uncertainty are tolerated and to what degree? Can computational principles that have uncertainty in their behavior be used to compute machine learning to achieve computer systems that are superior in energy efficiency and machine learning capability?

In the field of machine learning acceleration, simplified numerical representations such as quantization and low-bit-precision floating point, as well as approximate computations such as weight pruning, are widely used. Approximate multipliers and adders that tolerate deterministic computational errors have also been proposed. However, these are nonaccurate rather than uncertain computations.

Stochastic computing, known for a long time, represents numerical values using the probability of occurrence of zeroes and ones. Thus, it is a computation principle with uncertainty. The fluctuation in stochastic computing is realized by pseudorandom numbers, not by natural mechanisms. Similarly, digital annealing machines inspired by quantum annealing are computation accelerators that use such pseudo fluctuations. For example, STATICA[11] is a combinatorial optimization accelerator based on a mathematical approach called stochastic cellular automata, which solves combinatorial optimization problems by stochastically updating binary spin states.

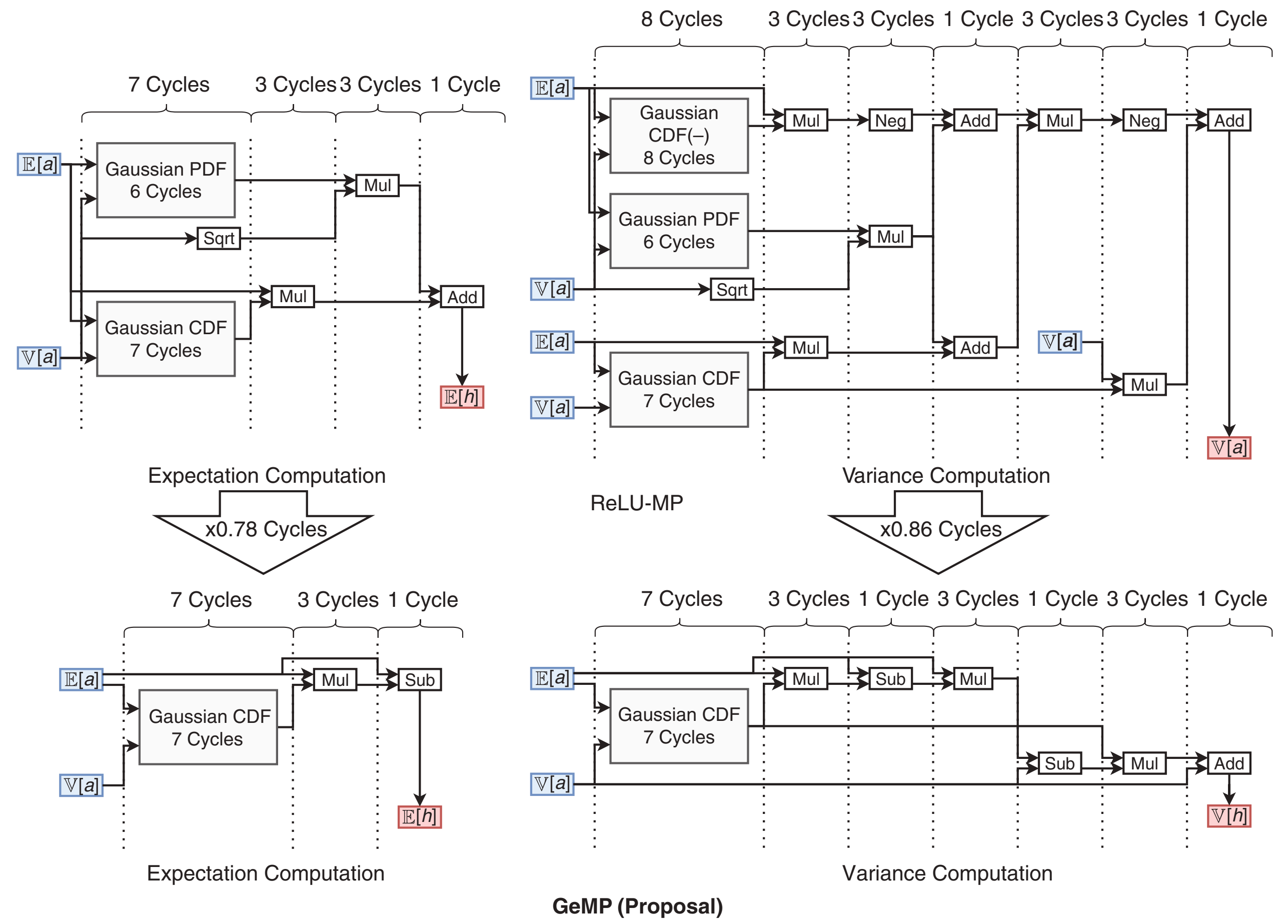


FIGURE 2. An architectural comparison between ReLU-based MP and GeMP-based MP (cited from Hirayama et al.[10]).

The above computational devices are designed and manufactured using digital circuit technology to remove uncertainty. In contrast, thermodynamic computing (TC)[12] has attracted attention as a method that directly exploits the phenomenon of fluctuations in physical devices and takes advantage of their propensity to transition to minimum energy states according to thermodynamic principles. An external classical computer provides constraints of a corresponding target problem and input data to a physically fluctuating device, and the device will eventually transition to an energetically stable state. Because TC actively exploits the uncertainty of the device, it eliminates the energy cost paid to hide the uncertainty in conventional computers.

Quantum computers have been actively studied in recent years with the aim of realizing a fault-tolerant quantum computer. In a noisy intermediate-scale quantum (NISQ) computer, it is assumed that there is an error in the quantum computer side of the computation; at NISQ, a quantum-classical hybrid algorithm is being studied in which the quantum computer performs the appropriate processing on the quantum side, and the classical computer, which is capable of exact computation, performs the rest of the processing. This kind of right-performer approach has the potential to achieve something better than a classical computer-only system despite the existence of uncertainties in the NISQ machine.

In the area of machine learning, methods for handling uncertainty, such as Bayesian deep learning, have been widely studied. Understanding the inherent uncertainty helps us to make trustworthy and safe decisions based on fluctuated machine learning outputs.

In general computing, other than machine learning, making effective use of uncertainty, rather than just hiding it, will be an important theme in future computer architecture. It may become possible to achieve both energy efficiency and reliability of calculation results by introducing the viewpoint of "optimizing the system based on the assumption of uncertainty," rather than just pursuing "perfect accuracy."

As a related topic, approximate computing is the widely known direction that allows the existence of deterministic and/or probabilistic computing errors, and the research provides various specific calculation methods with feasible circuit technologies. The important point for going beyond approximate computing is that error-prone approximate computers should still provide trustworthy results for users, even if the raw computation results contain uncertainty.

Our key assumption is that there is a substantial tradeoff between the amount of supplied energy and calculation accuracy, but not all calculations need to be rigorous to provide trustworthy results. Under this assumption, we believe the realization of such trustworthy computers will require technology in which different levels of applications, software, and hardware cooperate to "observe," "interpret," and "control" uncertainty.

The "observe" process is the fundamental process of applying a computation operation to a fluctuated computer device and measuring raw computation outputs. The computer might return a fluctuated result for every operation, and we can reissue the operation and measure the result again, if required, according to decisions by the following "interpret" and "control" processes to obtain a final computation result with the required quality. Note that the entity and granularity of an operation can be arbitrary. We can think of both hardware and software for the entity. A larger operation will be beneficial in terms of control overheads, but it decreases the applicability of the system to applications.

The "interpret" process is the advanced process of estimating the degree of uncertainty in the obtained raw computation outputs to provide a trustworthy result. The uncertainty can be estimated based on probability distributions via multiple variational operations, as well as Bayesian deep learning, but more suitable approaches for general fluctuated computers might exist.

The "control" process is the dynamic feedback process used to calibrate the operation setting, such as supplied energy (that is, electric voltage and current), to obtain the next variational output. Based on the current uncertainty estimation and the required trustworthiness, a higher energy supply might be preferred to obtain a more confident raw computation, or a lower energy supply might be selected for energy saving.

This direction will not work well if you prefer complete computers with perfect accuracy. However, we hope we can find a narrow but hopeful way to realize energy-efficient and trustworthy computers if we become a friend of uncertainty. ■

REFERENCES

1. R. Cai et al., "VIBNN: Hardware acceleration of Bayesian neural networks," in *Proc. 33rd Int. Conf. Architectural Support Program. Lang. Operating Syst. (ASPLOS)*, New York, NY, USA: ACM, 2018, pp. 476–488.
2. Y. Hirayama, T. Asai, M. Motomura, and S. Takamaeda, "A hardware-efficient weight sampling circuit for Bayesian neural networks," *Int. J. Netw. Comput.*, vol. 10, no. 2, pp. 84–93, 2020, doi: 10.15803/ijnc.10.2_84.
3. Y. Fujiwara and S. Takamaeda-Yamazaki, "ASBNN: Acceleration of Bayesian convolutional neural networks by algorithm-hardware co-design," in *Proc. IEEE 32nd Int. Conf. Appl.-Specific Syst., Archit. Processors (ASAP)*, 2021, pp. 226–233, doi: 10.1109/ASAP52443.2021.00041.
4. Y. Gal and Z. Ghahramani, "Dropout as a Bayesian approximation: Representing model uncertainty in deep learning," in *Proc. 33rd Int. Conf. Mach. Learn.*, M. F. Balcan and K. Q. Weinberger, Eds., PMLR, 2016, vol. 48, pp. 1050–1059.
5. Q. Wan and X. Fu, "Fast-BCNN: Massive neuron skipping in Bayesian convolutional neural networks," in *Proc. 53rd Annu. IEEE/ACM Int. Symp. Microarchit. (MICRO)*, 2020, pp. 229–240, doi: 10.1109/MICRO50266.2020.00030.
6. N. Durasov, T. Bagautdinov, P. Baque, and P. Fua, "Masksembles for uncertainty estimation," 2021, *arXiv:2012.08334*.
7. H. M. Chen et al., "Enhancing dropout-based Bayesian neural networks with multi-exit on FPGA," 2024, *arXiv:2406.14593*.
8. A. Wu, S. Nowozin, E. Meeds, R. E. Turner, J. M. Hernandez-Lobato, and A. L. Gaunt, "Deterministic variational inference for robust Bayesian neural networks," in *Proc. Int. Conf. Learn. Representations*, 2019, pp. 1–24.
9. H. Awano and M. Hashimoto, "BYNQNet: Bayesian neural network with quadratic activations for sampling-free uncertainty estimation on FPGA," in *Proc. Des., Autom. Test Europe Conf. Exhib. (DATE)*, 2020, pp. 1402–1407, doi: 10.23919/DATE48585.2020.9116302.
10. Y. Hirayama, K. Suga, and S. Takamaeda-Yamazaki, "GeMP-BNN: High-performance sampling-free Bayesian neural network accelerator with Gaussian error moment propagation," in *Proc. IEEE 16th Int. Symp. Embedded Multicore/Many-Core Syst.-on-Chip (MCSoC)*, 2023, pp. 522–529, doi: 10.1109/MCSoC60832.2023.00083.
11. K. Yamamoto et al., "7.3 STATICA: A 512-spin 0.25m-weight full-digital annealing processor with a near-memory all-spin-updates-at-once architecture for combinatorial optimization with complete spin-spin interactions," in *Proc. IEEE Int. Solid-State Circuits Conf. (ISSCC)*, 2020, pp. 138–140, doi: 10.1109/ISSCC19947.2020.9062965.
12. T. Hylton, T. M. Conte, and M. D. Hill, "A vision to compute like nature: Thermodynamically," *Commun. ACM*, vol. 64, no. 6, pp. 35–38, May 2021, doi: 10.1145/3431282.

SHINYA TAKAMAEDA-YAMAZAKI is with The University of Tokyo, Tokyo 113-8656, Japan. Contact them at shinya@is.s.u-tokyo.ac.jp.

CALL FOR SPECIAL ISSUE PROPOSALS

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

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

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

Proposal guidelines are available at:

www.computer.org/csdl/magazine/co/write-for-us/15911

IEEE COMPUTER SOCIETY

IEEE

Digital Object Identifier 10.1109/MC.2025.3546136

Functional Safety Standards: IEEE P2851 Road Map

Jyotika Athavale, Synopsys
Darren Galpin, Infineon Technologies

This article describes the functional safety-related standards development initiatives sponsored by the IEEE Computer Society's Functional Safety Standards Committee (FSSC), in particular IEEE P2851, focusing on interoperable activities related to functional safety and its interactions with reliability, security, operational safety, and time determinism within a dependability lifecycle. The P2851 standards describe methods, description languages, data models, and database schema that have been identified as necessary or critical to enable the exchange/interoperability of data across all steps of the lifecycle, encompassing activities executed at intellectual property (IP), system-on-chip (SoC), system, and item levels in a way that allows integration in different application domains, such as automotive, industrial, medical, and avionics safety-critical systems.

Digital Object Identifier 10.1109/MC.2025.3536796
Date of current version: 27 March 2025

THE FSSC

The IEEE Computer Society's FSSC was established in 2021, focusing on architectures, methodologies, tools addressing functional safety, and other safety-related aspects of the intended functionality at the different levels of abstraction [system of systems, systems, hardware, or software (SW) component] and across application fields, such as automotive, industrial, avionics, and high-performance computing. It also covers the relationships of functional safety with contiguous domains such as system safety, cybersecurity, reliability, real-time interactions, and artificial intelligence (AI).

The broad nature of the domains means that there are many intersections with other Societies within IEEE, and to this end, the FSSC has been establishing links with other IEEE Societies, such as the Reliability Society, whose work of system safety also touches on functional safety. We want to ensure that the IEEE ecosystem of standards is consistent with each other and does not conflict with other well-established standards, such as those from the ISO.

FSSC KEY ACTIVITIES

Since the last overview of FSSC activities published in *Computer* in 2023,[1] the FSSC has worked on two standards activities and two white papers and has looked at emerging technologies, such as AI, and how they might affect functional safety.

EDITOR JYOTIKA ATHAVALE
SYNOPSYS; jyotika@synopsys.com

An example of an emerging technological issue is the issue of plasticity loss and forgetting in deep continual learning.[2] Continual learning models have been used in the automotive space to learn from real traffic situations and to improve the responses over time. However, neural networks can begin to forget previously learned information when they learn new information and can lose plasticity over time, reducing their ability to learn. These two failure modes act in different ways, with forgetting affecting performance on old data and plasticity affecting performance on new data, but the implication is that if you have a functional safety case reliant on these operating correctly, then the issues that the failure modes could create also need to be taken into account. There is evidence that in real-life neural networks such as the brain, large numbers of synapses are "reset" to help avoid biological plasticity problems,[3] so similar approaches for AI neural networks may aid the situation. But this means intentionally forgetting some previous learning, which would mean understanding the boundaries of any changes to maintain a functionally safe operating space. This is an emerging area of study that the FSSC will need to keep monitoring closely.

The numerous domains that are affected by functional safety mean that there are a number of domain-specific standards, and unfortunately, the differing standards use the same terminology but with subtly different meanings. This is a problem that was explored in Mariani et al.,[1] and since that article was published, the FSSC has successfully published a white paper on the subject.[4] The white paper takes an overview across domains and across standards from different organizations to present a number of definitions of key terms that will be used moving forward in new standards sponsored by the FSSC.

The safety-critical community is urgently asking for a solution to accelerate the safety engineering process while reducing risks and costs.

The second white paper published by the FSSC is in the field of prognostics and was published in June 2024,[5] the work having been previewed in Mariani et al.[1] The white paper investigates the methodology of prognosis and preventative maintenance, identifies potential uses and gaps in the state of the art of current standards, and presents recommendations to stakeholders. The contributions that this white paper makes to the subject are being used in the latest P2851.1 standard, which will be discussed in a later section of this article.

Apart from P2851 and its child standards, work is also well underway on P3332 on control-oriented system safety analysis, which was featured in the IEEE Computer Society Standards webinar series in November 2024.[6] The approach being developed in this standard is to consider the system and operating environment together, with the system comprising hardware and SW. The combined behavior is represented as a network of cause-and-effect chains with forks and joins, called a *control structure*. The analysis identifies unsafe control paths that lead to potential sources of harm, which are then iterated around so that they can be designed out of the control structure. Work is well underway in creating a first draft of the proposed standard, and it is hoped to publish this within the next two years.

IEEE P2851

The IEEE 2851 standardization initiative was started to define a dependability lifecycle as well as methods and formats for the exchange/interoperability of data across all steps of the lifecycle in a consistent way. Additionally, the work products of this standardization initiative will help enable interoperability between tools. The goal of this work is to provide structures and directions to allow a seamless exchange of information and interoperability between activities at the same or different level of abstraction as well as activities across lifecycles related to different dependability attributes across application domains such as automotive, industrial, medical, and avionics safety-critical systems. This standards working group (WG) was kicked off in early 2020, under the leadership of the Design Automation Standards Committee, and then, in 2021, its reporting was transferred to the newly formed FSSC.

The development of safety- and cybersecurity-critical systems is rapidly growing due to the expansion of new applications, such as automated driving or autonomous mobile robotics. Standards such as ISO 26262, IEC 61508, and many others define the complete set of activities that need to be performed, requiring companies at different levels of the supply chain to tailor the lifecycle activities that apply to them and deliver results to other levels of the chain for which other requirements are applicable. However, there are no common methods, languages, or formats to exchange the relevant data. As a result of this gap, companies are struggling with many different types of methods and description languages and are investing valuable time and effort to reconsolidate, compare, integrate, and combine the data. For this reason, the safety-critical community is urgently asking for a solution to accelerate the safety engineering process while reducing risks and costs.

The WG's first standard (2851-2023 – IEEE Standard for Functional Safety Data Format for Interoperability within the Dependability Lifecycle) was published in December 2023. This standard defines a dependability lifecycle of products with a focus on interoperable activities related to functional safety and its interactions with reliability, security, operational safety, and time determinism. The standard also describes methods, description languages, data models, and databases that have been identified as necessary or critical to enable the exchange/interoperability of data across all steps of the lifecycle encompassing activities executed at IP, SoC, system, and item levels in a technology-independent way across application domains such as automotive, industrial, medical, and avionics safety-critical systems.

Last year, the P2851 WG began work on derivative standards, beginning with P2851.1. At different levels of supply chains, companies struggle with many different types of methods and description languages for the exchange of functional safety and reliability information. Therefore, the purpose of this standard is to enable a seamless exchange of data and interoperability between functional safety and reliability engineering activities at the same or different levels of object abstraction. This standard defines a method of seamlessly exchanging functional safety- and reliability-related information in a technology-independent manner. Therefore, the standard helps enable interoperability between tools used by functional safety and reliability engineers at the same or different levels of object abstraction. The standard guides implementers on key methods, description languages, and database topics identified in IEEE Standard 2851-2023. It covers topics that relate to radiation testing for soft errors, vulnerability factors measurements, base failure rate estimation, reliability, availability, serviceability system architecture, and prognostics activities. In addition, this standard defines an alignment flow of functional safety and reliability in the dependability lifecycle context, as described in IEEE Standard 2851-2023.

In addition, a Project Authorization Request (PAR) request was recently submitted for P2851.2. The proposed standard would cover implementation guidance for key methods, description languages, and database topics identified in the IEEE 2851-2023 standard, related to functional safety and its interactions with cybersecurity, to enable the exchange/interoperability of data. These identified topics include SW architectural design, safety assessment for SW updates, and confirmation reviews. The activities are executed at IP, SoC, system, and item levels in a technology-independent way across application domains.

More AI-based components in future highly automated vehicles impose new risks in the context of automotive safety, and there has been a growing effort recently in the development of newer standards for dependable computing. New topics being initiated by FSSC and IEEE P2851 will help address challenges related to functional safety interoperability in the context of dependable technologies.

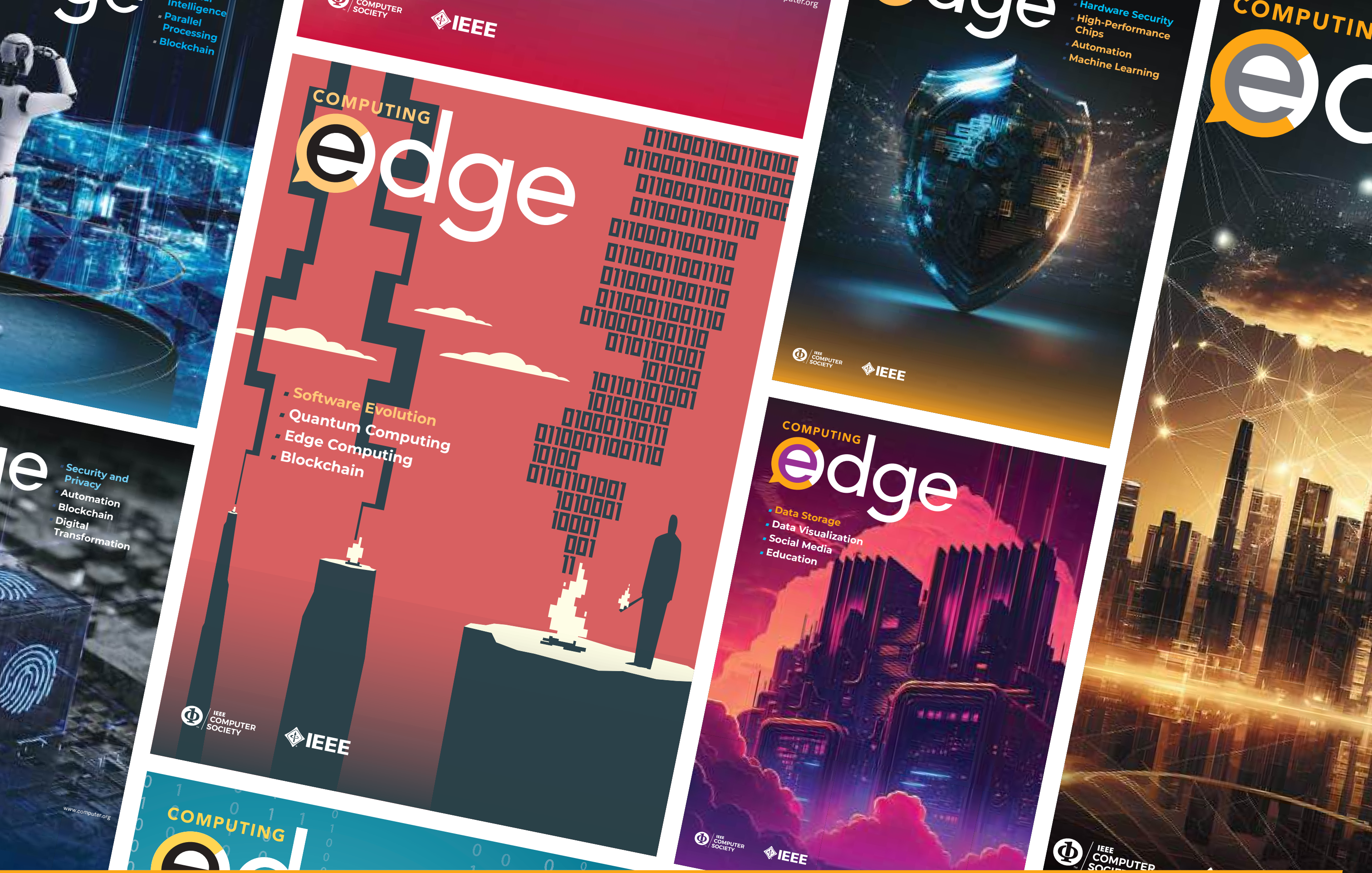
REFERENCES

1. R. Mariani, J. Athavale, A. Matteucci, D. Galpin, and R. Schaaf, "Functional safety standards committee: Results and perspectives," *Computer*, vol. 56, no. 2, pp. 143–148, Feb. 2023, doi: 10.1109/MC.2022.3229482.
2. S. Dohare, J. Hernandez-Garcia, Q. Lan, P. Rahman, A. R. Mahmood, and R. S. Sutton, "Loss of plasticity in deep continual learning," *Nature*, vol. 632, no. 8026, pp. 768–774, Aug. 2024, doi: 10.1038/s41586-024-07711-7.
3. H. Kasai, N. Ziv, H. Okazaki, S. Yagishita, and T. Toyoizumi, "Spine dynamics in the brain, mental disorders and artificial neural networks," *Nature Rev. Neurosci.*, vol. 22, no. 7, pp. 407–422, 2021, doi: 10.1038/s41583-021-00467-3.
4. FSSC, "The functional safety terminology landscape," IEEE, Piscataway, NJ, USA, Jan. 2023. [Online]. Available: https://ieeexplore.ieee.org/servlet/opac?punumber=10042203
5. FSSC, "Predictive maintenance in the context of functional safety," IEEE, Piscataway, NJ, USA, Jun. 2024. [Online]. Available: https://ieeexplore.ieee.org/servlet/opac?punumber=10555002
6. IEEE Computer Society. Control-Oriented System Safety Analysis. (Jan. 4, 2025). Accessed: Jan. 4, 2025. [Online Video]. Available: https://www.youtube.com/watch?v=1wU5DrqHN4w

JYOTIKA ATHAVALE is the director of RAS architecture at Synopsys, Sunnyvale, CA 94085 USA. Contact her at jyotika@synopsys.com.

DARREN GALPIN is lead principal microarchitect/concept engineer at Infineon Technologies, BS34 8HP Bristol, U.K. Contact him at darrengalpin43@gmail.com.

Digital Object Identifier 10.1109/MC.2025.3546073


COMPUTING edge
Software Evolution
Quantum Computing
Edge Computing
Blockchain
IEEE COMPUTER SOCIETY
IEEE
COMPUTING edge
Data Storage
Data Visualization
Social Media
Education
Hardware Security
High-Performance Chips
Automation
Machine Learning
Security and Privacy
Automation
Blockchain
Digital Transformation
Parallel Processing
Blockchain