COMPUTING edge

- Computer Vision Applications
- Automation
- Cyberphysical Systems
- Ethics

NOVEMBER 2024

www.computer.org

IEEE COMPUTER SOCIETY

IEEE

Get Published in the New *IEEE Transactions on Privacy*

This fully open access journal is now soliciting papers for review.

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

Submit a Paper Today!

Visit computer.org/tp to learn more.

IEEE COMPUTER SOCIETY | IEEE

COMPUTING edge

IEEE COMPUTER SOCIETY

IEEE

IEEE COMPUTER SOCIETY computer.org

Printed with inks containing soy and/or vegetable oils

STAFF

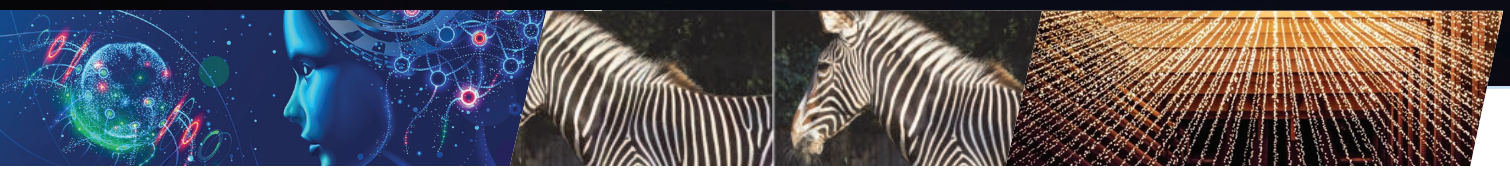
Circulation: *ComputingEdge* (ISSN 2469-7087) is published monthly by the IEEE Computer Society. IEEE Headquarters, Three Park Avenue, 17th Floor, New York, NY 10016-5997; IEEE Computer Society Publications Office, 10662 Los Vaqueros Circle, Los Alamitos, CA 90720; voice +1 714 821 8380; fax +1 714 821 4010; IEEE Computer Society Headquarters, 2001 L Street NW, Suite 700, Washington, DC 20036.

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

IEEE Computer Society Magazine Editors in Chief

NOVEMBER 2024 • VOLUME 10 • NUMBER 11

COMPUTING edge



Computer Vision Applications

Automation

Cyberphysical Systems

Ethics

Departments

Magazine Roundup

The IEEE Computer Society's lineup of 12 peer-reviewed technical magazines covers cutting-edge topics ranging from software design and computer graphics to Internet computing and security, from scientific applications and machine intelligence to visualization and microchip design. Here are highlights from recent issues.

Computer

Language Artificial Intelligence at a Crossroads: Deciphering the Future of Small and Large Language Models

This August 2024 *Computer* article explores the future of language models, focusing on the development and growth of large and small language models. It advocates for interdisciplinary collaboration, responsibility guidelines, educational initiatives, sustainable practices, and effective governance to ensure that these technologies benefit society in the long run.

Computing in Science & Engineering

Deploying Optimized Scientific and Engineering Applications on Exascale Systems

Exascale supercomputers are first-of-their-kind instruments with potentially paradigm-shifting capabilities. However, enabling complex applications at scale and high performance can be difficult, requiring hard work and deep technical understanding. The Application Integration (AppInt) area of the Exascale Computing Project was designed to be the integration point between applications, supporting software, system environments, high-performance computing facilities, and vendors. In this January–March 2024 *Computing in Science & Engineering* article, the authors describe how the AppInt team addressed these challenges while also promoting the use of portable and sustainable programming models and helping harden systems prior to general availability.

IEEE Annals of the History of Computing

Hardware Standardization and State-Socialist Piracy: The Global Reach of the Zilog Z80

The broad contours of the personal computing industry can be traced via contradictory waves of consolidation and fragmentation. For example, the incorporation of diverse systems under the banner of Internet connectivity in the 1990s paradoxically resulted in a narrower range of platforms. This April–June 2024 *IEEE Annals of the History of Computing* article extends this framework backward through the inverse case of the Zilog Z80 microprocessor. Despite platform divergences, the Z80 represents an important illustration of globalizing computational infrastructure prior to the collapse of state socialism and the breakthroughs of the 1990s.

IEEE Computer Graphics and Applications

Human-in-the-Loop: Visual Analytics for Building Models Recognizing Behavioral Patterns in Time Series

Detecting complex behavioral patterns in temporal data, such as moving object trajectories, often relies on precise formal specifications derived from vague domain concepts. However, such methods are sensitive to noise and minor fluctuations, leading to missed pattern occurrences. Conversely, machine learning (ML) approaches require abundant labeled examples, posing practical challenges. In this article, featured in the May/June 2024 issue of *IEEE Computer Graphics and Applications*, the authors

 2469-7087/24 © 2024 IEEE

introduce their visual analytics approach, which enables domain experts to derive, test, and combine interval-based features to discriminate patterns and generate training data for ML algorithms.

IEEE Intelligent Systems

Affective Relevance

Today, myriad relevance estimation methods are extensively used in various systems and services, mostly using behavioral signals such as dwell-time and click-through data and computational models of visual or textual correspondence to these behavioral signals. However, behavioral signals can only be used to produce rough estimations of the actual underlying affective states that users experience. In this July/August 2024 *IEEE Intelligent Systems* article, the authors provide an overview of recent alternative approaches for measuring and modeling more nuanced relevance based on physiological and neurophysiological sensing.

IEEE Internet Computing

Protecting Data Buyer Privacy in Data Markets

Data markets serve as crucial platforms facilitating data discovery, exchange, sharing, and integration among data users and providers. However, the paramount concern of privacy has predominantly centered on protecting privacy of data owners and third parties, neglecting the challenges associated with protecting the privacy of data buyers. In this July/August 2024 *IEEE Internet Computing* article, the authors address this gap by modeling the intricacies of data buyer privacy protection and investigating the delicate balance between privacy and purchase cost. Through comprehensive experimentation, their results yield valuable insights, shedding light on the efficacy and efficiency of their proposed approaches.

IEEE micro

High-Performance Cooling for Power Electronics via Electrochemical Additive Manufacturing

In this article, featured in the May/June 2024 issue of *IEEE Micro*, the authors introduce an advanced liquid-cooled thermal management solution for power electronics. Utilizing a novel 3-D metal printing technology called electrochemical additive manufacturing (ECAM), copper cooling structures are printed directly onto the ceramic substrate of the component, thereby eliminating thermal interface materials and significantly reducing the thermal resistance of the system-level stack. The use of ECAM-printed cooling structures in traction inverter applications is shown to have great potential for realizing significant gains in performance, via thermal resistance improvements in the range of 60%–120%.

IEEE MultiMedia

A Convolutional Neural Network Ensemble for Video Source Camera Forensics

In this April–June 2024 *IEEE MultiMedia* article, the authors address the problem of identifying the video source camera of the video data acquired by investigators. They develop a novel convolutional neural network (CNN) ensemble framework to identify the video source camera. In their method, the authors analyze the video data using patches extracted from intracoded frame (I-frame) quadrants (i.e., nonoverlapping squares) using independent CNNs for each quadrant to achieve location awareness. Experimental results demonstrate that their framework is robust for the same device-type classification and outperforms existing deep learning-based techniques.

IEEE Pervasive Computing
MOBILE SYSTEMS | UBIQUITOUS COMPUTING | INTERNET OF THINGS

Re-Envisioning the Role of a User in Sustainable Computing

As more and more computing technologies become pervasive in our daily lives, more and more e-waste is generated. And yet, when designing the next generation of devices, qualities such as speed, usability, and usefulness (a.k.a. user-centered design) are prioritized, rarely exploring options that might not optimize for users but, instead, improve long-term environmental sustainability. Exploring these alternatives is essential for transitioning toward a more sustainable future in computing. To this end, the authors of this April–June 2024 *IEEE Pervasive Computing* article argue that the envisioned roles attributed to users during user-centered design should encompass much more. They discuss this design shift through examples of two interactive systems they built to explore altering the role of the traditional 'user' to that of caretaker and recycler.

IEEE Security & Privacy

Comprehensive Memory Safety Validation: An Alternative Approach to Memory Safety

Comprehensive memory safety validation identifies the memory objects whose accesses provably comply with all classes of memory safety, protecting them from memory errors elsewhere at low overhead. In this article, featured in the July/August 2024 issue of *IEEE Security & Privacy*, the authors assess the breadth and depth of comprehensive memory safety validation.

IEEE Software

Explainability for Property Violations in Cyberphysical Systems: An Immune-Inspired Approach

A systematic approach is essential to help understand the system behaviors that lead to critical cyberphysical system failures. The authors of this article in the September/October 2024 issue of *IEEE Software* present a methodology that identifies and isolates crucial anomalous behaviors that can hamper the system and are often challenging to capture.

IT Professional

Trajectory Analysis in UKF: Predicting Table Tennis Ball Flight Parameters

In this May/June 2024 *IT Professional* article, the authors aim to develop a sophisticated system capable of accurately predicting the 3-D trajectory of a ball in sports and conducting an in-depth analysis of the obtained trajectory. The proposed system comprises three key components: a binocular vision system, an unscented Kalman filter trajectory and velocity prediction system, and an algorithmic data analysis system. Compared to traditional binocular image triangulation, the proposed system improves accuracy by 25%, with an error margin reduced to only 86 mm.

Editor's Note

Keeping an Eye on AI

The rapid progression of the accuracy and precision of artificial intelligence (AI) is tangibly impacting engineering and systems as well as people's livelihoods. For instance, computer vision allows AI to make decisions based on their collection and analysis of images and videos, enhancing surveillance and facial recognition technology. This issue of *ComputingEdge* discusses the ethics behind software engineering and AI, including the rising influence of AI and automation and the need to explain and understand how they work. The articles also explore how to make image coding more efficient, how to create effective automated systems, and the differences between cyberphysical systems (CPSs) and the Internet of Things (IoT).

AI tools are used to make important assessments based on imagery. To employ this technology responsibly and effectively, engineers must enhance image storage and processing and understand how AI makes decisions. *Computer*'s article, "What's in an AI's Mind's Eye? We Must Know," emphasizes the importance of explaining how AI "thinks" as it analyzes images. The authors of "The JPEG AI Standard: Providing Efficient Human and Machine Visual Data Consumption," from *IEEE MultiMedia*, present an image coding tool that can facilitate more efficient transmission and storage.

While automation can replace many mundane tasks, human oversight is still essential. "Automation Doesn't Work the Way We Think It Does," from *IEEE Software*, argues that human employees must be trained in coordination with automated systems to ensure that those systems function effectively. *IEEE Software* article, "Automating a Massive Open Online Course's Production," demonstrates how to automate an online course with tools and techniques that can be applied to other types of projects.

There has been a lot of debate over the relationship between CPSs and the IoT. In "The 12 Flavors of Cyberphysical Systems," from *Computer*, the authors define CPSs and contrast CPS and IoT technologies. *Computer* article "Should Cyberphysical Systems and the Internet of Things Get Married?" includes a roundtable discussion on distinguishing CPSs and the IoT and their key research areas.

With the increasing influence of AI tools comes the increasing need to consider ethics in both the process and products of software engineering. "Ethics: Why Software Engineers Can't Afford to Look Away," from *IEEE Software*, explains why ethics-based discussions should be crucial to the software engineering profession. The author of the *IT Professional* article "What If Ethics Got in the Way of Generative AI?" questions whether generative AI is accessible, equitable, and fair.

EDITOR: Hsiao-Ying Lin, IEEE Member, hsiaoying.lin@gmail.com

ARTIFICIAL INTELLIGENCE/MACHINE LEARNING

This article originally appeared in
Computer
vol. 57, no. 7, 2024

What's in an AI's Mind's Eye? We Must Know

Moshe Sipper, *Ben-Gurion University of the Negev*
Raz Lapid, *Ben-Gurion University of the Negev and DeepKeep*

We discuss explainability and understandability in artificial intelligence (AI) and offer an experiment and a discussion of responses, challenges, and obstacles. The pursuit of AI explainability and understandability is crucial and ignored at our peril.

One of the major challenges of modern-day artificial intelligence (AI) is the striving to understand the "mind" of an AI (in particular, deep networks), that is, to comprehend how such networks "think" (some would argue the quotation marks are superfluous).

AI EXPLAINABILITY IS HARD, YET CRUCIAL

We humans often ask "why"; we *need* explanations. This is not just some abstract desire. It is part of our makeup, part of our evolutionary ancestry whose survival depended on modeling and understanding the world. Thus, it is natural that we seek explanations from AIs. Why did the AI reject a college application? Why did the AI classify the image as an elephant? What are the reasons? Are they sound? Or do they evidence bias or some other form of erroneous thinking? Why did the AI recommend a certain medical procedure? Indeed, in the medical domain, explainability is legally required.

An entire subfield of explainable AI (XAI) has arisen in recent years,[1] focusing on finding explanations for the reasoning of deep networks. However, we find that usually, such explanations are "shallow" in some sense, when compared with deeper explanations that humans offer upon being questioned about their thoughts and reasoning.

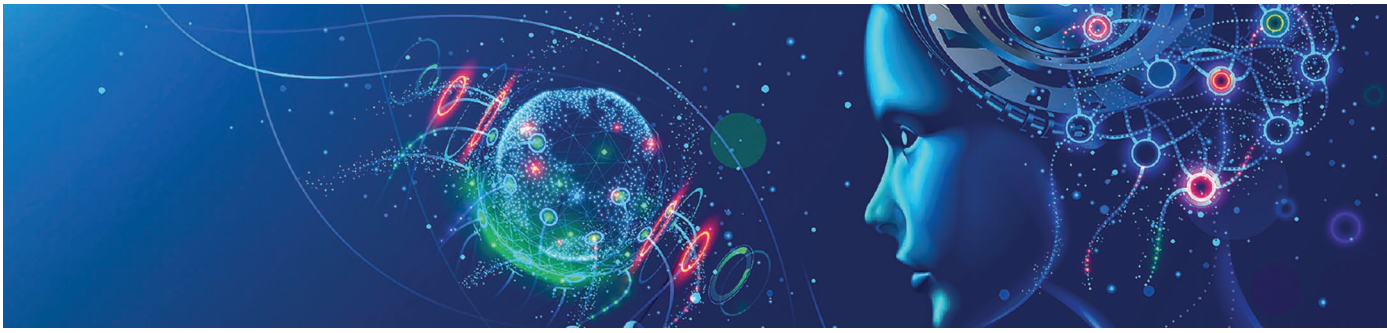
A typical example of XAI in images involves a so-called "explanation map," which shows the pixels most responsible for producing a network's output, for example, classification of a bird in a nature image [Figure 1(a)]. Essentially, these are the pixels that had the most influence on the network's predictions.[2] While such commonly used explanation maps offer insight into a network's workings, we think the explanations are shallow and local, being, as they were, about low-level pixels and not about high-level concepts, such as "This is a bird because of its beak and feathers." XAI is used not only to explain images but also for other forms of data: tabular, time series, linguistic, and more. Yet from what we have seen, there is always some sense of not being quite up to par with human explanations.

There have been interesting attempts to go beyond shallow explanations. For example, Feather et al.[3] focused on metamers, "stimuli that produce the same responses at some stage of a network's representation," showing that metamers from early network layers were recognizable to human observers, but those from deeper layers were not.

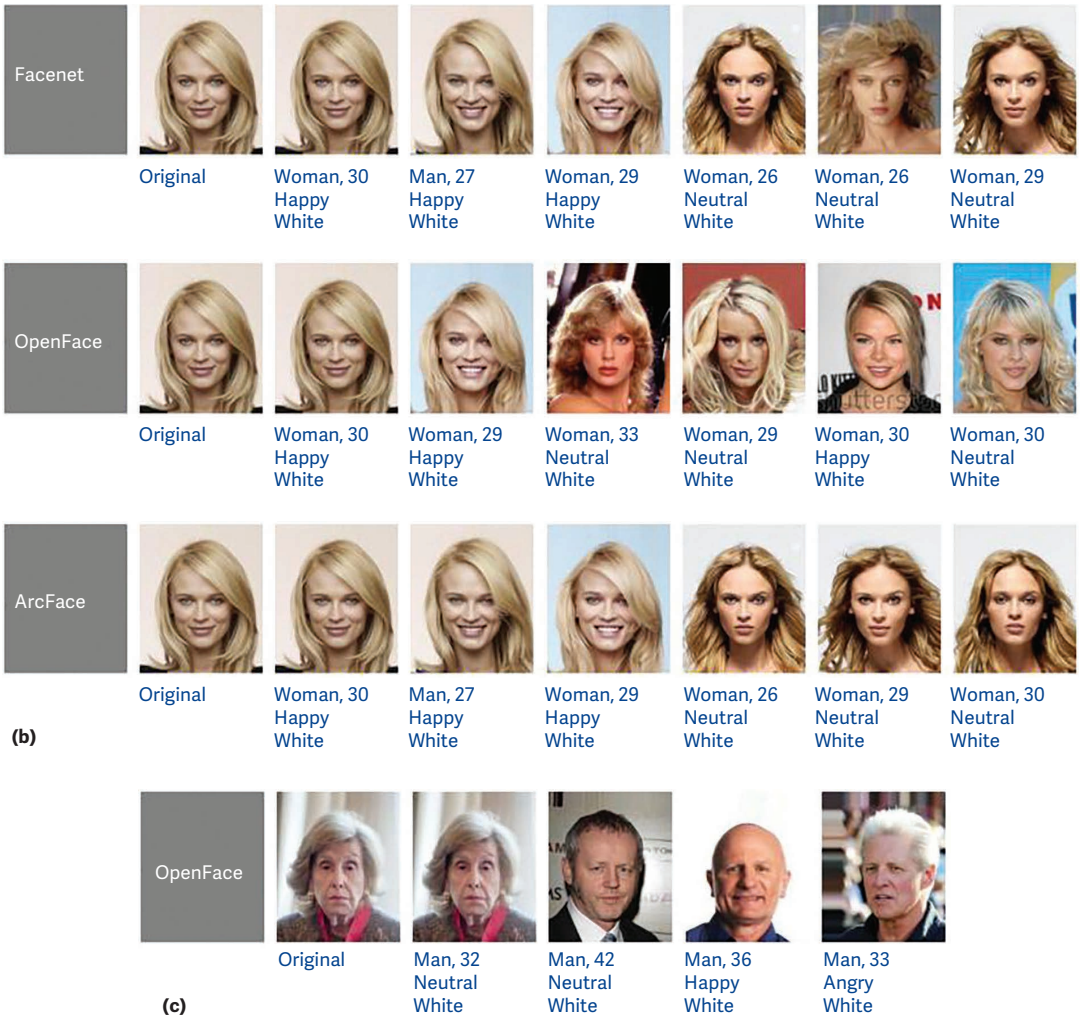
A recent work introduced the idea of a "probe," a neural network that is simpler than the one under study, trained to decode the original network's internal activations.[4] While undoubtedly a step forward, this still does not quite provide a full-blown, humanlike explanation. Very recently, thanks to advances in multimodal AI,[5] new XAI approaches have begun delivering conceptual explanation capabilities. Another recent approach used a large language model (GPT-4) to explain neurons in another large language model (GPT-2XL), focusing on what they termed the "explanation score": a measure of a language model's ability to

Digital Object Identifier 10.1109/MC.2024.3356820
Date of current version: 26 June 2024

 Published by the IEEE Computer Society



(a)



(b)

(c)

FIGURE 1. Understanding AI thinking. **(a)** An image with a sample explanation map. **(b)** A sample output panel. Each row shows 1 + 1 + 5 images: the original, the original again, and the five most similar to the original. **(c)** An example of an interesting error.

compress and reconstruct neuron activations using natural language.[6] Schwettmann et al.[7] introduced Function Interpretation and Description, a benchmark suite for evaluating the building blocks of automated interpretability methods. These approaches are resource intensive since they rely on large, deep networks.

EXPLANATION AT EYE LEVEL: AN APPLIED GEDANKENEXPERIMENT

We wish herein to advocate a higher level of explanation modeling and understanding. Toward this end we designed and performed a simple yet thought-provoking setup focusing on faces in the CelebA dataset. We deployed the DeepFace software package, which includes both facial recognition and facial attribute analysis.[8] The package offers several models, trained *independently* and on *different* datasets. For facial recognition we chose three of the available models: Google FaceNet, OpenFace, and ArcFace. For facial attribute analysis, DeepFace offers four models: gender, age, facial expression, and ethnicity.

Crucially, we now have access to completely different models, trained on different datasets, performing different tasks. We then perform the following steps:

- Select a random image from CelebA; designate it the "original."
- Call DeepFace.find with the original image and the entire CelebA dataset. This function finds the most-similar images to the original by generating latent representations (embeddings) of all dataset images and then comparing those with the embedding of the original image through the cosine-similarity measure.
- The most-similar image should be the same as the original. For the next five most similar images, call DeepFace.analyze, a function that deploys the four models that asses gender, age, expression, and ethnicity.

This simple procedure is repeated to produce multiple outputs. Note that find (facial recognition) and analyze (facial attributes) use different models, as explained above. (The code is available online.[9])

Figure 1(b) shows a sample output panel. There are three rows, per three facial recognition models. For each model we considered the six most similar images. The most similar is the original, followed by five additional images to the right. *Independently* of face recognition, now come the attribute models and analyze the images. The analysis results are given below each image, showing the outputs produced by the gender model, the age model, the expression model, and the ethnicity model.

As we've emphasized above, there are several independent models at work here. Face recognition is done separately from face analysis, and within each of these two categories the models are different.

Observing the sample panel of Figure 1(b), we note that the analyses of similar images tend to agree to some extent or another. Indeed, to gather statistics we ran 1,000 random images, which amounted to 15,000 images (1,000 × 3 models × 5 images). For each image we then asked whether the analyzed attributes agreed with those of the original (for three attributes this is a simple true/false assessment; for age, we defined "agreement" as being within three years either way). The results were as follows: age, 59% agreement with original; emotion, 49%; gender, 88%; and ethnicity, 68%.

We find it interesting that when one model outputs images it considers similar, a completely different model tends to view *high-level concepts* (and human at that), gender, age, emotion, and ethnicity, similarly.

Another intriguing phenomenon we observed is that now and again, similar images found by the recognition models caused the analysis models (again, independently) to make similar mistakes. This is demonstrated in Figure 1(c): the analysis models seem to have misjudged the original image with respect to gender and age. We then obtain similar images through the recognition model. They do not look quite similar (like humans, AI is not perfect), yet curiously, when you hand them over to the analysis models, gender and age coincide with the original mistakes.

RESPONSES, CHALLENGES, AND OBSTACLES

In response to the fundamental challenge of insufficient to no explainability in most contemporary

AI solutions, the literature presents several strategic avenues. Each approach comes with pros and cons.

Interpretable models, such as decision trees and linear models, inherently offer transparency in the decision-making process. This transparency, however, comes at the expense of a tradeoff between interpretability and predictive accuracy: the more interpretable the model, the simpler it needs to be, and thus, its predictive accuracy declines. That said, for some tasks, these oft-overlooked models are the perfect choice.

Rule-based systems define decision rules explicitly, thus offering inherent transparency. However, manual crafting of rules may be impractical for complex tasks, and automated rule generation encounters challenges in capturing subtle decision boundaries.

Explainability techniques for black-box models, such as local interpretable model-agnostic explanations (LIMEs), provide locally good explanations for complex, black-box models. However, global interpretability is not guaranteed at all, and fidelity with respect to the overall model behavior may be compromised.

Visualizations, such as saliency maps, offer intuitive insights into model decisions. Challenges lie in designing effective visualizations, and interpretation by humans may greatly vary, potentially leading to misconceptions. Further, it has been shown that it is possible to manipulate these maps, so-called adversarial attacks.[2]

The pursuit of XAI is often driven by the desire to enhance trust and understanding in AI systems. However, while XAI holds the potential to address these concerns, it is important to recognize possible unforeseen challenges and unintended consequences. We think there are (at least) four key obstacles that warrant careful consideration:

The *tradeoff between accuracy and interpretability* is always an intricate balancing act. A more accurate model will usually tend to be less interpretable and vice versa.

Security concerns are an issue. Explanations generated by XAI systems can be powerful tools for understanding and communicating AI decisions. However, they also carry the risk of being misused or misinterpreted. For example, explanations could be used to manipulate users by framing decisions in a biased or misleading way or to justify biased decisions by providing a veneer of objectivity. Additionally, users may oversimplify or misinterpret explanations, leading to inaccurate or incomplete understanding of AI decisions.

Regarding *fairness and robustness*, XAI explanations should not only provide insights into AI decisions but also be fair and robust to potential biases. This means that explanations should not perpetuate or reinforce existing biases in the data or the model itself. Moreover, explanations should be robust to adversarial attacks or attempts to manipulate them to achieve specific outcomes. Ensuring fairness and robustness in XAI is particularly crucial in sensitive applications where AI decisions have significant impacts on individuals or groups (for example, the medical domain).

An *illusion of understanding* can exist. XAI can provide valuable insights into the inner workings of AI models, but it is important to avoid creating an illusion of complete understanding. AI models, especially complex ones, often involve intricate relationships among features, nonlinear dependencies, and stochastic processes. While XAI can help unravel some of these complexities, it is essential to recognize that explanations may not capture the full extent of the model's behavior. Overreliance on XAI explanations without critical evaluation could hinder a deeper understanding of AI systems and their limitations.

...AND WHAT IS EXPLAINABILITY?

The point at which we shall be content with an explanation is unclear. Is "because it has feathers" enough? Why does the network output this explanation? Can it dig further to produce, for example, "because most birds have feathers"? Is *that* a sufficient explanation? Here we seem to be delving into the philosophical nature of explanations, but we may have to, given the rise of AI. As recently noted by Prince[10]: "There is also an ongoing debate about what it means for a system to be explainable, understandable, or interpretable... there is currently no concrete definition of these concepts."

We believe the pursuit of AI explainability and understandability is crucial, to be ignored at our peril. Perhaps task completion and its explanation should be fully integrated, as recently shown by

Sipper.[11] Deep learning pioneer Geoffrey Hinton said in a recent interview (*CBS News*, 8 October 2023), "What we did was we designed the learning algorithm. That's a bit like designing the principle of evolution. But when this learning algorithm then interacts with data, it produces complicated neural networks that are good at doing things. But we don't really understand exactly how they do those things."

ACKNOWLEDGMENT

Moshe Sipper is the corresponding author.

REFERENCES

1. D. Minh, H. X. Wang, Y. F. Li, and T. N. Nguyen, "Explainable artificial intelligence: A comprehensive review," *Artif. Intell. Rev.*, vol. 55, pp. 3503–3568, Jun. 2022, doi: 10.1007/s10462-021-10088-y.
2. S. V. Tamam, R. Lapid, and M. Sipper, "Foiling explanations in deep neural networks," *Trans. Mach. Learn. Res.*, 2023. [Online]. Available: https://openreview.net/forum?id=wvLQMHtyLk
3. J. Feather, A. Durango, R. Gonzalez, and J. McDermott, "Metamers of neural networks reveal divergence from human perceptual systems," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 32., H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, Eds., Curran Associates, Inc., 2019. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2019/file/ac27b77292582bc293a51055bfc994ee-Paper.pdf
4. K. Li, A. K. Hopkins, D. Bau, Viégas, H. Pfister, and M. Wattenberg, "Emergent world representations: Exploring a sequence model trained on a synthetic task," in *Proc. 11th Int. Conf. Learn. Representations*, 2023. [Online]. Available: https://openreview.net/forum?id=DeG07_TcZvT
5. N. Rodis, C. Sardianos, G. T. Papadopoulos, P. Radoglou-Grammatikis, P. Sarigiannidis, and I. Varlamis, "Multimodal explainable artificial intelligence: A comprehensive review of methodological advances and future research directions," 2023, *arXiv:2306.05731*.
6. S. Bills et al. "Language models can explain neurons in language models." OpenAI. Accessed: Jan. 25, 2024. [Online]. Available: https://openaipublic.blob.core.windows.net/neuron-explainer/paper/index.html
7. S. Schwettmann et al., "FIND: A function description benchmark for evaluating interpretability methods," 2023, *arXiv:2309.03886*.
8. S. I. Serengil and A. Ozpinar, "Hyperextended lightface: A facial attribute analysis framework," in *Proc. Int. Conf. Eng. Emerg. Technol. (ICEET)*, Piscataway, NJ, USA: IEEE Press, 2021, pp. 1–4, doi: 10.1109/ICEET53442.2021.9659697.
9. M. Sipper. "A minimal example of face recognition and facial analysis: Using Deepface, with an accompanying Colab notebook." Medium. Accessed: Jan. 25, 2024. [Online]. Available: https://medium.com/ai-mind-labs/a-minimal-example-of-face-recognition-and-facial-analysis-ce4024da30d8
10. S. J. Prince, *Understanding Deep Learning*. Cambridge, MA, USA: MIT Press, 2023.
11. M. Sipper, "Task and explanation network," 2024, *arXiv:2401.01732*.

MOSHE SIPPER is a professor of AI at the Department of Computer Science, Ben-Gurion University of the Negev, Beer-Sheva 8410501, Israel. Contact him at sipper@bgu.ac.il.

RAZ LAPID is a Ph.D. student at the Department of Computer Science, Ben-Gurion University of the Negev, Beer-Sheva 8410501, Israel, and is also at DeepKeep, Tel-Aviv 6701203, Israel. Contact him at razla@post.bgu.ac.il.

IEEE COMPUTER SOCIETY

www.computer.org

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

OMBUDSMAN: Contact ombudsman@computer.org.

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

PUBLICATIONS AND ACTIVITIES

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

Periodicals: The society publishes 12 magazines, 18 journals

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

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

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

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

Certifications: The society offers three software developer credentials.

AVAILABLE INFORMATION

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

IEEE COMPUTER SOCIETY OFFICES

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

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

IEEE CS EXECUTIVE STAFF

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

IEEE CS EXECUTIVE COMMITTEE

President: Jyotika Athavale
President-Elect: Hironori Washizaki
Past President: Nita Patel
First VP: Grace A. Lewis
Second VP: Nils Aschenbruck
Secretary: Mrinal Karvir
Treasurer: Darren Galpin
VP, Member & Geographic Activities: Kwabena Boateng
VP, Professional & Educational Activities: Cyril Onwubiko
VP, Publications: Jaideep Vaidya
VP, Standards Activities: Edward Au
VP, Technical & Conference Activities: Terry Benzel
2023–2024 IEEE Division VIII Director: Leila De Floriani
2024–2025 IEEE Division V Director: Christina M. Schober
2024 IEEE Division V Director-Elect: Thomas M. Conte

IEEE CS BOARD OF GOVERNORS

Term Expiring 2024:
Saurabh Bagchi, Charles (Chuck) Hansen, Carlos E. Jimenez-Gomez, Daniel S. Katz, Shixia Liu, Cyril Onwubiko

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

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

IEEE EXECUTIVE STAFF

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

IEEE OFFICERS

President & CEO: Thomas M. Coughlin
President-Elect: Kathleen Kramer
Past President: Saifur Rahman
Director & Secretary: Forrest D. Wright
Director & Treasurer: Gerardo Barbosa
Director & VP, Publication Services & Products: Sergio Benedetto
Director & VP, Educational Activities: Rabab Kreidieh Ward
Director & VP, Membership and Geographic Activities: Deepak Mathur
Director & President, Standards Association: James E. Matthews III
Director & VP, Technical Activities: Manfred J. Schindler
Director & President, IEEE-USA: Keith A. Moore

IEEE

revised 29 May 2024

EDITOR: Dan Grois, grois@ieee.org

DEPARTMENT: INDUSTRY AND STANDARDS

This article originally appeared in IEEE MultiMedia vol. 30, no. 1, 2023

The JPEG AI Standard: Providing Efficient Human and Machine Visual Data Consumption

João Ascenso, *Instituto de Telecomunicações - Instituto Superior Técnico, 1049-001, Lisbon, Portugal*

Elena Alshina, *Huawei Technologies Duesseldorf GmbH, 80992, Munich, Germany*

Touradj Ebrahimi, *Multimedia Signal Processing Group, EPFL, CH-1015, Lausanne, Switzerland*

FROM THE EDITOR

For more than 30 years, the Joint Photographic Experts Group (JPEG) committee developed many successful and widely adopted image coding standards. JPEG AI is one of its most recent high-potential standardization activities, which has been initiated in order to cope with the dramatic increase in image creation and utilization. Due to the significant breakthroughs in the Artificial Intelligence (AI) field, and specifically in the field of deep neural networks, the learning-based image coding has already shown impressive compression gains over traditional approaches. As a result, the JPEG AI will provide a framework for the efficient distribution and consumption of images, especially when images are consumed by machines, further targeting to reduce the bandwidth and storage requirements by around 50% for the same visual presentation quality.

The Joint Photographic Experts Group (JPEG) AI learning-based image coding system is an ongoing joint standardization effort between International Organization for Standardization (ISO), International Electrotechnical Commission (IEC), and International Telecommunication Union - Telecommunication Sector (ITU-T) for the development of the first image coding standard based on machine learning (a subset of artificial intelligence), offering a single stream, compact compressed domain representation, targeting both human visualization and machine consumption. The main motivation for this upcoming standard is the excellent performance of tools based on deep neural networks, in image coding, computer vision, and image processing tasks. The JPEG AI aims to develop an image coding standard addressing the needs of a wide range of applications such as cloud storage, visual surveillance, autonomous vehicles and devices, image collection storage and management, live monitoring of visual data, and media distribution. This article presents and discusses the rationale behind the JPEG AI vision, notably how this new standardization initiative aims to shape the future of

1070-986X © 2023 IEEE
Digital Object Identifier 10.1109/MMUL.2023.3245919
Date of current version 10 May 2023.

image coding, through relevant application-driven use cases. The JPEG AI requirements, the JPEG AI history, and current status are also presented, offering a glimpse of the development of the first learning-based image coding standard.

Nowadays, image coding is an essential technology in our society, used billions of times per day, by a very large percentage of the world's population. This includes not only personal pictures, many widely distributed on social networks, but also professional applications and services, such as in stock photography and video streaming (e.g., movie covers). The majority of personal photos are acquired with mobile devices, where images frequently occupy a large amount of storage, often becoming the main motivation to buy a new device. In this context, to drastically reduce image storage size, HEVC/H.265[1] was adopted in several smartphones (e.g., Apple iPhone) as the default image coding engine, but it is still not enough to meet the growing demand for image storage. More recently, VVC/H.266[2] has achieved significant improvement on compression efficiency over HEVC/H.265 for video (40%-50%), but shows moderate improvement (15%-20%) for still images. Moreover, image resolution and target quality have been increasing and thus the uncompressed size of images is also increasing, critically asking for efficient image coding solutions to further facilitate transmission and storage. In this context, there is a need for a new image coding standard, where compliant solutions are able to achieve significantly higher compression efficiency and thus large rate savings.

Other applications, such as visual surveillance systems are also very popular nowadays, where multiple cameras often capture, analyze, and store images, especially when events of interest occur. This explosive creation and availability of imaging data requires the use of mining and analysis technologies (machines that process visual data), which can be more effective if performed in the compressed domain. Therefore, an efficient compressed domain representation should be pursued not only for visualization by humans but also for effective machine visual data consumption.

The main objective of this article is to present the motivation, vision, applications, and current status of the Joint Photographic Experts Group (JPEG) AI standard, with a special focus on the JPEG AI use cases and requirements, the JPEG AI history and the technology adopted so far. The article is organized as follows. The "Scope" section describes the JPEG AI scope while the "JPEG AI Vision and Framework" section describes the JPEG AI framework. The "JPEG AI Key Tasks" and "JPEG AI Use Cases" sections describe key tasks and use cases, which can clearly benefit from the JPEG AI vision and the "JPEG AI Requirements" section reviews the JPEG AI requirements. The history and current status of the JPEG AI standardization effort, namely the technologies adopted until now, are described in the "JPEG AI History and Current Status" section and finally concluding remarks are presented in the "Final Remarks" section.

SCOPE

Recently, machine learning algorithms, such as deep neural networks (DNNs), have attracted a lot of attention and have become a popular area of research and development. This popularity is driven by several factors, such as recent advances in processing power, the availability of large datasets and powerful and efficient techniques, from convolutional layers to attention-based models. Nowadays, DNNs are the state-of-the-art for several computer vision tasks, such as those requiring high-level image understanding, e.g., image classification, semantic segmentation, and face recognition, but also in low-level image processing tasks, such as image denoising, super-resolution, enhancement, and inpainting among others. These advances have led to an increased interest in leveraging DNNs and other machine learning techniques for image coding to also obtain improvements, especially in compression efficiency, which is currently a very hot research topic. But more importantly, since in many state-of-the-art computer vision and processing tasks compact representations of the input are used, it may be possible to create for the first time a compressed representation using learning-based image coding that is considered efficient for both human and machine visual consumption. The creation of this common "language" is the main motivation behind the JPEG AI standard.

In the aforementioned context, the scope of the JPEG AI standardization[3] *is the creation of a learning-based image coding standard offering a single-stream, compact compressed domain representation, targeting both human visualization, with significant compression efficiency improvement over image coding standards in common use at equivalent subjective quality, and effective performance for image*

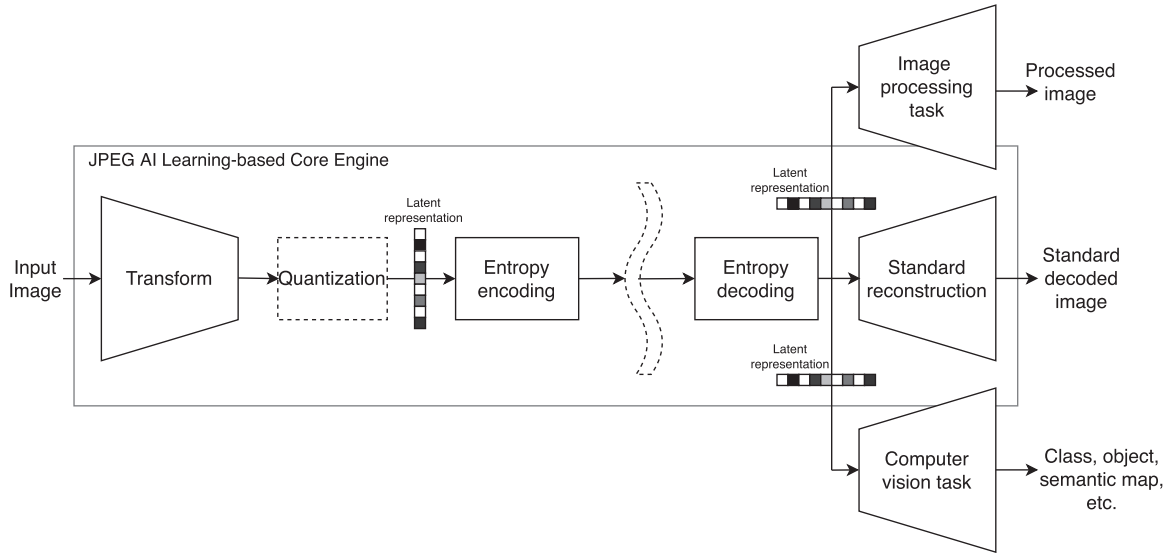


FIGURE 1. JPEG AI learning-based image coding framework.

processing and computer vision tasks, with the goal of supporting a royalty-free baseline.

In summary, the JPEG AI is the most recent member of the JPEG family of image coding standards well suited to have an important role in the multimedia landscape, by offering an efficient way for human and machine consumption, which was never considered in past JPEG (and other) coding standards.

JPEG AI VISION AND FRAMEWORK

Learning-based image coding solutions in the literature have already shown substantially higher compression efficiency when compared to existing conventional coding solutions; in particular, when compared to JPEG, JPEG 2000, HEVC Intra, and VVC Intra, better perceptual quality can be obtained measured both using objective quality metrics and subjective assessment methodologies.[4,5,6]

Besides their high compression efficiency, learning-based image coding solutions may be adapted with little extra effort to image processing and computer vision tasks without the need for full decoding, i.e., without performing image reconstruction. This contrasts with classical image coding that, when used in combination with image processing and computer vision techniques, often need to perform full decoding of the compressed bitstream to obtain a pixel-based representation.

Figure 1 shows the high-level JPEG AI framework, highlighting the three key pipelines. In the JPEG AI framework, the input image is processed with a transform, which aims to decorrelate the input image information typically using convolutional layers of a neural network, each one followed by nonlinear activation layers; each convolutional layer consists of learnable filters where some of them also perform spatial downsampling. This is followed by quantization or some simple rounding operation. At this stage, the so-called latent representation (or latent code) is obtained, which can be understood as a compact representation of the input image. The statistical redundancy present in the latent representation can be exploited by the entropy coding engine to produce the final bitstream to be transmitted or stored.

In the JPEG AI learning-based image coding framework, the bitstream produced by the encoder may be processed for human visualization by performing entropy decoding and standard reconstruction, thus producing a decoded image that aims to represent the original image with high perceptual quality and fidelity. In this case, the operations performed are the inverse of the encoder, which means entropy decoding to obtain the latent representation followed by several convolutional layers to perform spatial upsampling, thus, compensating the downsampling performed on the encoder side.

However, as shown in Figure 1, for many applications where machines consume visual data, standard reconstruction may be skipped since the latent representation produced by the encoder contains the essential information, i.e., features obtained from the original image. Therefore, the two additional JPEG AI pipelines (besides the standard reconstruction pipeline) use the same bitstream (produced by the

TABLE 1. List of JPEG AI key tasks.

Image processing tasks	Computer vision tasks
Super-resolution	Image retrieval and classification
Low-light enhancement	Object detection and recognition
Color correction	Semantic segmentation
Exposure compensation	Event detection and action recognition
Inpainting	Face detection and recognition

encoder) to perform image processing and computer vision tasks at the decoder side. These tasks are performed on the latent representation (obtained after entropy decoding), directly extracted from the original image and not from the (lossy) decoded image. This intrinsically feature-rich latent representation can be used in two main ways, in addition to the standard reconstruction: 1) to perform an image processing task, such as targeting the enhancement of the image, for example, with increased resolution, contrast, etc., and 2) to perform a computer vision task where high-level semantic information is extracted, e.g., to generate labels, regions, etc. Typically, these two pipelines are implemented also with convolutional layers and follow an architecture, which is specific to the task being addressed. The key JPEG AI advantage is the creation of a latent (compact) representation, not only for image reconstruction but also for machine consumption tasks such as image classification and semantic segmentation, providing a multitask (or multipurpose) solution with low complexity and thus lower energy consumption when compared to full image decoding followed by learning-based enhancement or analysis.

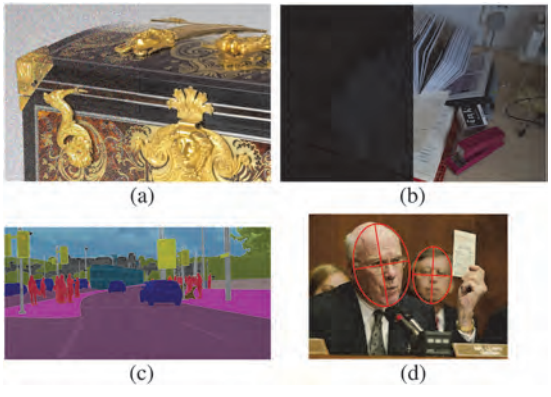


FIGURE 2. Examples of relevant image processing and computer vision tasks. (a) Image denoising. (b) Low-light enhancement. (c) Semantic segmentation. (d) Face detection.

JPEG AI KEY TASKS

Following the JPEG AI scope, the compressed stream will have a triple-purpose, offering compelling advantages for applications where an image processing task aims to enhance the image or where semantic (or higher level) information needs to be extracted from large amounts of visual data. The impact can be significant since these tasks can be performed with lower complexity by using as input the compressed domain representation instead of the original or decoded images (requiring lower computational resources). In some cases, higher performance (e.g., accuracy) can be achieved since features extracted from the original, instead of the lossy decoded images, are used. Table 1 lists some relevant image processing and computer vision tasks that can benefit from JPEG AI and Figure 2 shows examples obtained from popular image processing and computer vision datasets.

To meet the JPEG AI scope and offer a single-stream compact-domain representation that is useful for multiple purposes, the encoder must create a bitstream independent of the task, i.e., the latent representation (obtained after entropy decoding) must allow efficient processing by a compressed domain decoder (including a standard reconstruction decoder). In this context, the term *decoder* refers to not only the process of translating the bitstream to pixel data but also includes other modalities, e.g., image to textual labels.

The JPEG AI has defined three representative key tasks besides standard reconstruction for which compressed domain decoder solutions will be developed, one representative of the computer vision tasks and two representatives of the image processing tasks: 1) compressed domain image classification; 2) compressed domain super-resolution; and 3) compressed domain denoising. It is expected that more compressed domain decoders will be developed in the future.

A compressed domain image classifier receives as input a quantized latent representation and should achieve competitive image classification accuracy compared to standard reconstruction followed by image classification, especially at lower rates, but also at lower complexity. A compressed domain super-resolution decoder should produce a higher resolution image and, thus, lower the computational cost of upscaling the image obtained by standard reconstruction. In this context, bandwidth and storage costs can be reduced for local and cloud-based visual systems since the original image to be

encoded has lower spatial resolution. Finally, a compressed-domain image denoiser aims at performing pixel-wise reconstruction while simultaneously removing the noise, i.e., filtering out the noise of the latent representation to obtain a clean image. Again, the target is to achieve lower computational complexity and, potentially, improve the performance of the pipeline when compared to standard reconstruction and denoising in cascade.

JPEG AI USE CASES

This section describes the use cases considered more relevant for the JPEG AI standardization defined by relevant academia and industry experts, where the JPEG AI vision may bring clear benefits.

Cloud Storage

Due to the popularity of online storage services, an ever increasing number of images are stored in the cloud, often, billions of photos are stored in a cloud service, requiring a considerable amount of resources, notably storage space, bandwidth, and energy. Therefore, an efficient image coding solution for cloud storage would be important to minimize costs since even marginal savings may have a significant overall impact. Higher compression efficiency would allow for reduced storage space and latency, thus leading to the distribution of high-quality images at a fraction of the cost and with improved quality of experience (e.g., low latency).

Visual Surveillance

Visual surveillance systems are widely deployed to perform video monitoring with several objectives, such as anomaly detection, detection of suspicious activity, provision of forensic evidence, and intelligent control. Often, many cameras generate huge amounts of visual data that need to be processed, compressed, analyzed, and stored. Intelligent surveillance systems are used to record relevant events not just as video but also as very high-resolution images. Considering the amount of data, visual surveillance systems require efficient compression, content understanding, and in some cases image enhancement. Image processing or computer vision tasks are frequently employed to allow efficient navigation by searching, abnormal activity detection, object detection, crowd behavior analysis, and recognition of faces and events.

Autonomous Vehicles and Devices

Self-driving cars, drones, and other autonomous devices generate a vast amount of visual data that must be analyzed and sometimes stored. Moreover, images acquired from autonomous vehicles and devices may need to be processed offline and, thus, efficiently transmitted and/or stored. For example, drones carry cameras that are programmed to capture high-resolution aerial imagery, which can be difficult to transmit over resource-constrained connections and thus require efficient image coding solutions. The storage and transmission of images representing key events allow very useful applications, such as traffic monitoring, accident investigation, etc. This use case often involves several computer vision tasks, such as object detection, semantic segmentation, and event recognition, which could benefit from compressed domain processing.

Image Collection Storage and Management

Due to the popularity of smartphones and other consumer devices, every person has a digital camera that is used to acquire and store images of relevant events. This large collection of images is often backed up on online web storage to avoid their loss in the event of failure or even theft. Moreover, since these images usually have very high resolution, they require a significant amount of storage space, and images have to be organized conveniently, to facilitate their search and consumption. In this use case, image classification, object detection, and action recognition can be applied in the compressed domain to facilitate the management and organization of an image collection.

Live Monitoring of Visual Data

Live streaming of visual data has significantly increased, from professional services such as online lectures, videoconferences, and webcasts but also entertainment services, such as video game live streaming and, short-form personal videos (i.e., snack culture). Often, such visual data have to be analyzed to detect inappropriate content (as it is often done in social media networks) that may violate policies, but also to provide additional information such as labeling of faces, emotions, gestures, etc. Also, computer vision tasks could be applied to perform intelligent review, rating, and distribution of this type of content.

Media Distribution

Billions of user-generated images are captured and transmitted over the Internet daily. These images are often uploaded and transcoded into multiple quality versions and formats, then stored on worldwide servers for distribution. In such a scenario, more efficient image compression solutions allow lowering storage, transmission cost, and latency, which is especially relevant to users with low-bandwidth connections.

Progressive decoding may also be desirable, which allows for useful lower-quality previews while the image is still being received. In addition, lower resolution representations could be obtained for devices that support a lower resolution, without requiring the resources needed for the high-resolution version.

JPEG AI REQUIREMENTS

This section reviews the requirements that should be met by the standard so that it can be employed for the aforementioned use cases. Requirements are split between *mandatory requirements*, which are essential and *desirable requirements* which are only desirable, but might actually enlarge the target application scenarios. Naturally, it is assumed that the technology to be standardized will at least support the coding of images with a wide range of spatial resolutions (from 128 × 128 to 8K).

The JPEG AI mandatory requirements[3] are shown in Figure 3 and described in the following paragraphs along with a short motivation. JPEG AI standard-compliant image coding solutions must:

High standard reconstruction compression efficiency: offer significant compression efficiency improvement over image coding standards in common use at equivalent subjective quality. Following this requirement, VVC Intra, HEVC Intra, JPEG 2000 and JPEG were defined as the JPEG AI anchors.[7]

Effective compressed domain image processing and computer vision processing: provide clear benefits in terms of performance (e.g., accuracy for specific target rate) and complexity, especially with respect to full decoding followed by image enhancement or analysis.

FIGURE 3. JPEG AI mandatory and desirable requirements.

High fidelity for decoded images: be able to obtain reconstructed images with high fidelity as measured by full reference objective quality metrics and double stimulus subjective assessment protocols. Thus, fidelity should be preserved as much as possible, minimizing any distortion that modifies the visual content in unwanted ways, e.g., generating textures without considering the corresponding original texture. To verify this requirement, the JPEG AI Common Training and Test Conditions (CTTC)[7] defines seven full-reference objective quality metrics based on the study of[4]: Multi-Scale Structural Similarity Index Measure (MS-SSIM), Information Content Weighted Structural Similarity Measure (IW-SSIM), Video Multimethod Assessment Fusion (VMAF), Visual Information Fidelity in Pixel Domain (VIFP), Peak Signal-to-Noise Ratio–Human Visual System–Masked (PSNR-HVS-M), Normalized Laplacian Pyramid Decomposition (NLPD), and Feature Similarity Index Measure (FSIM). The Double Stimulus Continuous Quality Scale (DSCQS)[8] subjective assessment methodology has been adopted, where both reference and impaired stimuli are shown side by side in randomized order and both independently assessed by the subjects. Figure 4 shows the graphical user interface for a subjective assessment test using the DSCQS methodology.

Device reproducibility: provide a similar decoded image independently of the hardware type (such as Graphics Processing Unit (GPU) or Central Processing Unit (CPU) used for encoding or decoding and the specific hardware architecture (e.g., different GPU brands or models). Reproducibility problems may occur due to floating-point rounding errors generated by different orders in the sequence of arithmetic operations (which can be more sequential or parallel) and may be very severe during the entropy decoding step. Naturally, the JPEG AI standard must include tools or mechanisms that minimize reproducibility errors. More precisely, the performance difference between obtained in different platforms should not be greater than around 0.5% of the BD-rate.

Hardware platform agnostic: be implementable in a wide range of hardware platforms, from CPU to GPUs of different brands and even neural compute engines. This means that the JPEG AI standard should be agnostic to the characteristics of specific CPU and GPU architectures.

Hardware/software implementation-friendly: enable encoders and decoders, which exploit parallelism and keep the memory, computational power, and energy

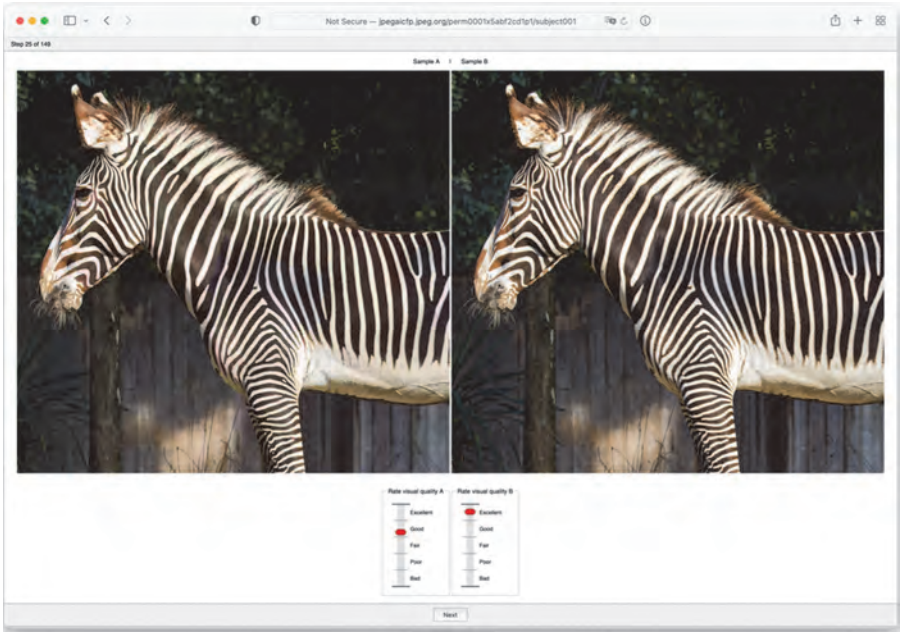


FIGURE 4. Graphical user interface for subjective assessment.

consumption to a minimum, exploiting the availability of operations already defined in libraries or hardware, such as convolution layers.

Support for 8- and 10-bit depth: be able to process popular image representation formats that support 8 and 10 bit depth precision.

Screen and synthetic content coding: efficiently code images with text and graphics, 3-D rendered images, illustrations, etc., and not only natural content, such as photographs and aerial/satellite images.

Progressive decoding: support progressive decoding, thus allowing a preview with lower quality or resolution and hence image transmission with low latency on low bandwidth connections.

The JPEG AI desirable requirements[3] are also shown in Figure 3 and described next (highlighted in italic).

Low complexity profile is nowadays considered very important for the adoption of the JPEG AI standard, especially to support resource-constrained hardware, such as smartphones.

The emergence of high dynamic range images, with higher bit depth and/or high dynamic range have attracted considerable attention, especially considering emerging capture and display devices and thus *high-bit depth coding* is considered as desirable.

Region of interest image coding is also a technology that allows to decode and visualize specific regions of an image without decoding others (or with lower resolution); this requirement is typically fulfilled by defining coding units (or tiles) that can be encoded and decoded independently. This is particularly useful for very high resolution images.

For some types of content (such as illustrations), not only RGB data are available but also an alpha channel that defines the transparency of texture. This channel needs to be transmitted to the decoder and, thus, efficient *lossless alpha channel/transparency coding* is needed.

Nowadays, an image coding algorithm such as JPEG may also be used to compress short video sequences (often called Motion JPEG) and is actually exploited in many applications (from webcams to short clip consumption). Therefore, the JPEG AI image coding standard should support *animated image sequence coding*, which is often performed by including relevant headers to convey order and/or timing information of each independently coded frame.

With *wide color gamut* display devices becoming available to consumers, it is desirable to support color profiles often used by such displays, such as Adobe RGB, Pro Photo RGB, and DCI-P3; such profiles are capable of representing a wider range of colors than sRGB. Naturally, the image coding pipeline to be defined by JPEG AI should be able to maintain a high perceptual color accuracy in the presence of such data.

Not all content is available in the sRGB color representation and, thus, it may be necessary to perform color conversion before coding and include relevant color profile information as metadata, thus supporting *multiple color representations*.

FIGURE 5. Selected test images and cropping regions for the JPEG AI Call for Proposals.

Thumbnail image coding can be a quick way to have a glimpse of the image content and is often used in many applications, such as personal photo management and, thus, is also included as a desirable requirement for the JPEG AI standard.

JPEG AI HISTORY AND CURRENT STATUS

The JPEG AI standardization has already reached an important milestone since the collaborative phase of this effort has been initiated. The work started during the 82nd JPEG meeting in January 2019 and many studies were performed to gather data, tools, and knowledge in the design and assessment of learning-based image coding, followed by discussions among experts. More recently, in 2022, several advances were attained, which are described in the next sections.

Submission of Proposals

Following the JPEG AI joint International Organization for Standardization (ISO)/International Electrotechnical Commission (IEC)/International Telecommunication Union - Telecommunication sector (ITU-T) Call for Proposals (CfP) issued at the conclusion of the 94th JPEG meeting (January 2022),[10] 14 registrations were received among which 12 submitted codecs for the standard reconstruction task. For computer vision and image processing tasks, several teams submitted compressed domain processors, notably six for image classification. After submissions of decoders, no model retraining was allowed for the following steps.

The JPEG AI CfP test set was then created and the JPEG AI anchors were generated for standard reconstruction, image processing, and computer vision tasks. Objective performance assessment was performed using the decoded images of the JPEG AI anchors and the JPEG AI hidden test set committee (which included experts not part of any proposal team) selected 10 images from the test set to have a varied set of content in terms of intrinsic characteristics (colorfulness, spatial complexity, etc.), degradations, and quality ranges. After, a dry run of the subjective evaluation procedure for standard reconstruction was performed for the JPEG AI anchors with expert viewers; the results obtained were reported and discussed in 95th meeting.[9] This evaluation led to additions and corrections to the JPEG AI CTTC[11] and the definition of several recommendations for the evaluation of the proposals, notably, the anchors, images, and bitrates along with a clear cross-check evaluation procedure.

Figure 5 illustrates the final selected eight test images (from the ten previously defined by the JPEG AI hidden test set committee by consensus among all participants) and used in the subjective evaluation, with the cropping regions defined in red. The cropping regions were defined by the JPEG AI hidden test set committee and correspond to salient and perceptually important regions of each image, where artifacts are clearly visible after compression with the submitted learning-based image coding proposals.

Evaluation of Proposals

At the 96th JPEG meeting (July 2022), 11 responses to the CfP were presented along with the subjective, objective, and complexity assessment as well as the identification of device interoperability issues by cross

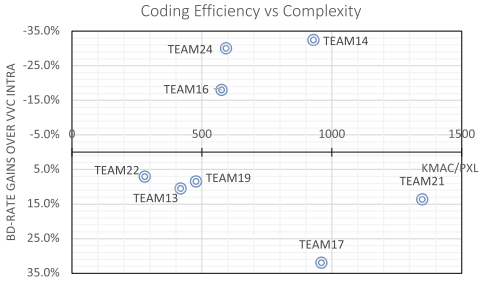


FIGURE 6. Compression efficiency improvements over VVC Intra versus decoding computational complexity.

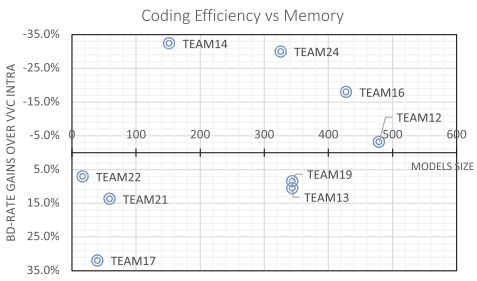


FIGURE 7. Compression efficiency improvements over VVC Intra versus decoder memory requirements (all models).

checking. The full results were reported in 96th JPEG meeting.[6] Figures 6 and 7 summarize the results for the standard reconstruction track.

Figure 6 shows the decoding complexity, that is measured through number of multiply accumulate operations (kilo) per pixel (kMAC/pxl) versus the coding efficiency improvements measured through average BD-rate gains over VVC Intra (above the horizontal axis means better than VVC Intra). The BD-rate performance represents the average across all JPEG AI CTTC quality metrics considering all test images. Figure 7 shows the memory occupied for all models needed for decoding any bitstream (at any quality) using as unit million of parameters versus the coding efficiency measured as aforementioned. As shown, several contributions show significantly higher compression efficiency when compared to VVC Intra (the best performing conventional image coding solution), more precisely, TEAM14 and TEAM24 show 32.3% and 29.9% BD-rate improvements. Regarding decoding complexity and memory usage, it could be observed a wide range of values, e.g., TEAM22 requires ~40 times less memory compared to the TEAM12. Moreover, the best BD-rate performing proposal has 3.3 higher computational complexity compared to TEAM22, which showed the lowest coding efficiency (similar to VVC Intra). Among the two best performing proposals, TEAM24 shows the lowest decoding complexity while TEAM14 had much lower memory requirements.

Collaborative Phase

The collaborative phase of the JPEG AI project started in the 96th JPEG meeting.

JPEG AI Verification Model

From the analyses and discussions of the results obtained in the evaluation of the proposals (see "Evaluation of Proposals" section), the most promising technologies were identified from the best performing proposals and the JPEG AI Verification Model under Consideration (VMuC) was designed.[12] The VMuC mainly corresponds to a combination of two proponents' solutions[13,14] (following the "one tool for one functionality" principle), selected by consensus and considering the CfP decision criteria and evaluation conditions and procedures. The JPEG AI VMuC was created to validate the first Verification Model (VM) through a battery of tests. The JPEG AI VM was approved at the 97th JPEG meeting (October 2022), but does not address all functionalities under consideration and will be improved in the future in the collaborative process, through several Core Experiments.

JPEG AI Current Status

The main focus so far has been the standard reconstruction task. For this task, the JPEG AI VM has two basic configurations, the first with a minimal set of networks for encoding/decoding images ("tools-off") and the second with additional networks and tools enabled to provide additional compression performance improvements ("tools-on"). JPEG AI VM1 (first version) can already provide average BD-rate gains over VVC Intra of 28% for the tools-off configuration and 31% for the tools-on configuration, considering the new JPEG AI test set with 50 images.[17] Regarding subjective quality, improvements of JPEG AI VM over VVC/H.266 Intra are illustrated in Figure 8 for two bitrates (0.06 bpp and 0.12 bpp). Regarding complexity, at the 98th JPEG meeting several proposals were presented based on the JPEG AI VM1 that can achieve 30% compression gain over VVC Intra operating at 200 kMAC/pxl (which may become acceptable for many devices in mid-term) and 15% compression gain operating at just 20 kMAC/pxl, which could be affordable in the nearest future (considering current trends in computing/memory resources for high-end mobile devices).

FIGURE 8. Decoded test images obtained with JPEG AI VM1 (left) and VVC Intra (right). (a) Test image 00019 (2120 × 1608) coded at 0.06 bpp. (b) Test image 00020 (1072 × 928) coded at 0.12 bpp.

Timeline

Regarding the JPEG AI standard (ISO/IEC 6048) timeline, Working Draft (WD) will be made available on January 2023, Committee Draft (CD) on July 2023, and the objective is to have the first learning-based image compression standard published by April 2024.

Future Challenges

The first challenge regards the compression efficiency for synthetic media, where all CfP submitted image codecs and the JPEG AI VM underperformed when compared to the screen content profiles of conventional image codecs such as VVC Intra; these profiles employ tools specifically designed to efficiently code this type of content. Other challenge regards decoding computational complexity, which is still high for the JPEG AI VM configuration with maximum coding efficiency, especially when attention models are used. A future deployment of the JPEG AI VM on smartphones will enable to have a more complete characterization (and accurate) measures of the complexity bottlenecks that need to be addressed. The final challenge regards the range between very high quality to near-lossless quality, which is especially important to professional photography applications (and others that target high fidelity). The current JPEG AI VM may require changes to address such cases through the use of integer or fixed precision weights on the convolutional layers of the encoder and decoder transforms.

FINAL REMARKS

The JPEG standardization committee started the development of the JPEG AI image coding standard based on neural networks, following strong evidence that learning-based image coding solutions can outperform previous coding standards in terms of compression efficiency. The JPEG AI standard will offer unique features desirable for an efficient distribution and consumption of images, especially addressing the very important case where images are consumed by machines, which often employ large machine learning models to enhance image quality (e.g., super-resolution), extract semantic information (e.g., image annotation). Regarding human consumption, the impact of the JPEG AI standard could be large, lowering bandwidth and storage requirements by roughly 50% (estimated) for the same subjective quality for a large set of content, when compared to best conventional image coding solutions (e.g., VVC Intra). Naturally, a tradeoff between decoding computational complexity and coding efficiency can already be observed in the first JPEG AI Verification Model, which could be addressed using the concept of profiles. Therefore, compromises may have to be made in the first deployments of JPEG AI image compression compliant solutions.

REFERENCES

1. G. J. Sullivan, J.-R. Ohm, W.-J. Han, and T. Wiegand, "Overview of the high efficiency video coding (HEVC) standard," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 22, no. 12, pp. 1649–1668, Dec. 2012.
2. B. Bross et al., "Overview of the versatile video coding (VVC) standard and its applications," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 31, no. 10, pp. 3736–3764, Oct. 2021.
3. *Use Cases and Requirements for JPEG AI*, ISO/IEC JTC 1/SC29/WG1, 94th JPEG Meeting, N100094, Jan. 2022.
4. *Performance Evaluation of Learning-Based Image Coding Solutions and Quality Metrics*, ISO/IEC JTC 1/SC29/WG1, 85th JPEG Meeting, N85013, San Jose, CA, USA, Nov. 2019.
5. *Report on the JPEG AI Call for Evidence Results*, ISO/IEC JTC 1/SC29/WG1, 89th JPEG Meeting, N89022, Oct. 2020.
6. *Report on the JPEG AI Call for Proposals Results*, ISO/IEC JTC 1/SC29/WG1, 96th JPEG Meeting, N100250, Online, Jul. 2022.
7. *JPEG AI Common Training and Test Conditions*, ISO/IEC JTC 1/SC29/WG1, 94th JPEG Meeting, N100106, Online, Jan. 2022.
8. ITU-R Question 102-3/6, *Recommendation 500-19: Methodology for the Subjective Assessment of the Quality of Television Pictures*, ITU-R Recommendation BT.500-19, Geneva, Switzerland, 2019.
9. *Report on the Objective and Subjective Quality Assessment of the JPEG AI Anchors*, ISO/IEC JTC 1/SC29/WG1, 95th Meeting, N100205, Online, Apr. 2022.
10. *Final Call for Proposals for JPEG AI*, ISO/IEC JTC 1/SC29/WG1, 94th JPEG Meeting, N100095, Online, Jan. 2022.
11. *Corrections and Additions to the JPEG AI Common Training and Test Conditions*, ISO/IEC JTC 1/SC29/WG1, 95th JPEG Meeting, N100204, Online, Apr. 2022.
12. *Description of the JPEG AI Verification Model Under Consideration and Associated Software Integration Procedure*, ISO/IEC JTC 1/SC29/WG1, 96th JPEG Meeting, N100279, Online, Jul. 2022.
13. *Presentation of the Huawei Response to the JPEG AI Call for Proposal*, ISO/IEC JTC 1/SC29/WG1, 96th JPEG Meeting, M96016, Online, Jul. 2022.

14. *Bytedance's Response to the JPEG AI Call for Proposals*, ISO/IEC JTC 1/SC29/WG1, 96th JPEG Meeting, M96053, Online, Jul. 2022.
15. *NYCU-PUT Response to the JPEG AI Call for Proposals*, ISO/IEC JTC 1/SC29/WG1, 96th JPEG Meeting, M96054, Jul. 2022.
16. *NJUVISION Response to the JPEG AI Call for Proposals*, ISO/IEC JTC 1/SC29/WG1, 96th JPEG Meeting, M96066, Jul. 2022.
17. *JPEG AI Common Training and Test Conditions*, ISO/IEC JTC 1/SC29/WG1, 98th JPEG Meeting, N100421, Sydney, Australia, Jan. 2023.

JOÃO ASCENSO is a professor with Instituto Superior Técnico, University of Lisbon, 1049-001, Lisboa, Portugal, and is with the Multimedia Signal Processing Group, Instituto de Telecomunicações in Lisbon, Portugal. Contact him at joao.ascenso@lx.it.pt.

ELENA ALSHINA is a chief video scientist in Huawei Technologies, 80992, Munich, Germany. Contact her at elena.alshina@huawei.com.

TOURADJ EBRAHIMI is a professor at Ecole Polytechnique Fédérale de Lausanne (EPFL), 1015, Lausanne, Switzerland, where he heads its Multimedia Signal Processing Group. He is a fellow of the IEEE. Contact him at touradj.ebrahimi@epfl.ch.

EDITOR: Laura Maguire, Ohio State University, maguire.81@osu.edu

This article originally appeared in
IEEE Software
vol. 41, no. 1, 2024

DEPARTMENT: FAILURE MODE

Automation Doesn't Work the Way We Think It Does

Laura Maguire

Increasing automation and artificial intelligence in software systems is driving a need for observability and explainability. This need for software to "show" what it is doing and "tell" how it functions points to broader considerations for coordination and collaboration with machines. It also indicates that observability and explainability are not enough. We should be investing in design for coordinative competencies between humans and automated technologies. Software outages provide clear examples of the significance of coordination to reliability and resilience. Consider the following example from a large multinational technology company.

A TALE OF THINGS GONE WRONG

NexusSky was a highly dependable digital services provider. Its architecture, designed to be highly available and fault tolerant, included thousands of servers ("hosts") that made up a logical unit ("cluster"). NexusSky ran hundreds of clusters to meet the needs of its many customers. The company's site reliability engineering (SRE) team was a proficient, knowledgeable group known for their expertise in handling large-scale systems, and they had a long track record of reliable service delivery. Over time, the work they handled was increasingly focused on problem solving that was high value, business critical, and challenging.

Recently, engineering management in an adjacent part of the organization offered to fund additional headcount to assist with routine but helpful tasks to take some of the load off the SREs. "Jacob" was hired to support the team in managing the cluster. It had been a busy month for the SRE team and, since Jacob had been hired by the adjacent engineering management, he hadn't been fully onboarded into the SRE group. For several weeks, Jacob helped out where he could, restarting a server here, answering a support ticket there—always wanting to be helpful while he waited for his training to be completed.

One afternoon, the team started seeing hosts going down to the point that it was making clusters unstable. Right away, they started working to bring the servers back up. It was a challenging incident involving very manual steps, but the team was able to get them back up.

Strangely, after bringing some hosts back up, some of the servers began to fail again. As more clusters became involved and the blast radius started growing, more SREs joined the incident response. The situation was quickly becoming chaotic.

Fearing the worst—a cascading or systemic failure—they put a change moratorium in place so there would be no changes to the production environment while they were trying to figure out the problem. An hour into the response effort, the team was still stumped as to why the hosts remained unstable after manual restart.

Then someone remembered Jacob.

It turned out that Jacob had been told to monitor metrics, such as CPU and memory utilization, across the clusters and to restart the host if the utilization was too high for an extended period of time. However, he didn't recognize that he needed to ensure the hosts were brought back up fully after restarting—and they weren't. On top of that, he hadn't noticed that more than half the hosts in the cluster were down, so he had kept on restarting hosts as the available pool of servers grew increasingly smaller.

The SRE team was dumbfounded! What was he doing!?!

It seemed unfathomable Jacob hadn't realized that limping along with high-resource utilization and performance degradation among some hosts was better than the broader impact across the entire

Digital Object Identifier 10.1109/MS.2023.3328746
Date of current version: 20 December 2023

 Published by the IEEE Computer Society

cluster caused by unintentionally pulling more and more servers out of the cluster through failed restarts.

Surely, he must have noticed the pool of servers was shrinking! Why was he doing that?

The team realized that this was what had pushed the system into instability and would continue to exacerbate the problem, but Jacob seemed completely unaware of anything abnormal going on around him.

What the heck was he going to do next, and how do we get him to stop?!

Fearing for the worst, the incident commander quickly interrupted Jacob and told him to stop restarting the hosts. This time, when the engineers restarted the down hosts, everything stayed up. After a perplexing and stressful incident, the team was able to breathe a sigh of relief.

An Epilogue to the Tale

Poorly onboarded or junior engineers who lack the operational context for a system can often work at cross-purposes to a team's larger goals when they fail to recognize how their actions get integrated into a broader system. There are similar unintended consequences when introducing automation into large-scale software systems.

In fact, the dutiful Jacob wasn't a junior engineer hired to handle routine tasks. Rather, "he" was artificial intelligence for IT operations (AIOps) that had been implemented as part of a push to introduce automation into NexusSky's operations. It reveals a difficult truth—machines cannot coordinate effectively with humans. That is problematic given that human–machine teams are increasingly responsible for society's critical digital infrastructure.

MOVING BEYOND OBSERVABILITY AND EXPLAINABILITY

This vignette highlights the fundamental asymmetry inherent in human–machine teaming. When we consider the implementation of automation as similar to hiring new team members, we see coordination breakdowns more readily. Instead of a seemingly inconsequential redistribution of tasks from humans to machines, we see how problematic it is to be working alongside automated team members who are not designed to participate in joint activity[1] with humans. We can see how much interdependence matters[2] and how automation must be designed for both task work and teamwork.[3]

There's a certain amount of irony in that software can be both a source of failure by causing service outages and a source of capacity by helping to handle the demands of continuous availability and to respond quickly when incidents occur. Automation has replaced many of the manual tasks that humans used to carry out—a practice that is often referred to as *reducing toil*. Reducing toil frees up engineers from repetitive, time-consuming, and often simpler tasks. This leaves them to handle more complex work. However, there is a pervasive belief that, when these tasks are handed off to automated agents, the engineer is no longer responsible for them. This belief holds that specific functions can be allocated to either humans or machines with no residual effort unaccounted for.

However, this isn't quite accurate. Work is not eliminated when tasks are distributed between humans and machines; instead, it is changed. When automation is deployed, engineers may no longer be doing the task, but they are most certainly supervising the automation as it is being done. The engineer watches dashboards and checks log files. They supervise the automated processes being completed to ensure they are done as required—albeit often on a relatively infrequent basis.

This changes the task from being manual to being cognitive. An engineer no longer physically runs the script. Instead, they monitor dashboards and review log files to ensure the work is carried out within specified parameters and within acceptable thresholds. They determine the implications of variations. They assess whether they need to intervene and how. The need for observability comes from this shift from manual to cognitive work and reinforces the supervisory role for humans. Being able to adequately monitor and observe system performance at varying levels of granularity helps engineers with their role in detecting and diagnosing problems as they emerge and change. This shift from doing to watching it being done by automation changes the nature of the cognitive activity involved.

Continuous engagement with the system when carrying out a task provides rich detail to keep the engineer's mental model current. It shapes expectancies about future performance. It helps with early detection.

It allows for anticipating problems before they arise. It primes action to respond. Intervening only when something goes wrong is cognitively demanding—so much so that studies have shown that, even when systems are highly automated, software engineers still allocate attentional resources to monitoring.[4] This makes sense. If you are responsible when the site goes down, you need to be ready to act quickly if it does. If your role as a supervisor doesn't allow you to react appropriately because your mental model is stale, you find ways to fill the gap and get yourself prepared.

It is no longer just manual tasks. Increasingly, automation is replacing cognitive work tasks. Advances in artificial intelligence and machine learning have improved machine capabilities in perceiving meaningful changes in the operating environment, carrying out more complex reasoning, and exercising discretion and judgment. This has increased the ability to act independently. Because of this, explainability has become more important. After working with a colleague across several incidents, you develop a strong sense of what they are good at, where they struggle, and how they are likely to act during an incident. Done well, explainability should tell you about the capabilities, limitations, and expected behaviors of an automated teammate.

CONTINUOUS ENGAGEMENT WITH THE SYSTEM WHEN CARRYING OUT A TASK PROVIDES RICH DETAIL TO KEEP THE ENGINEER'S MENTAL MODEL CURRENT.

However, explainability on its own is insufficient to improve human–machine team performance for handling failures in systems running at speed and scale. This is because failures are often emergent and a result of unexpected or unintended interactions. When failure is dynamic, emergent, and surprising, it requires more sophisticated reasoning, deeper knowledge, and flexibility in applying your knowledge to a situation you've never seen before. The incident response must then also be flexible and adaptive, requiring sophisticated levels of coordination and collaboration.

Automation—in particular, artificial intelligence—that is not designed for this level of teaming will add to the complexity and difficulty of incident handling. In her seminal work *Ironies of Automation*, Lisanne Bainbridge notes, "By taking away the easy parts of his task, automation can make the difficult parts of the human operator's task more difficult."[5] This was clearly shown in the opening vignette. "Jacob" was helpful, until he wasn't. Then, his inability to coordinate effectively made things worse. He lacked context for nominal and off-nominal performance within the cluster. He failed to recognize any kind of anomalous activity outside of his tasks. He was unaware of hidden interdependencies in the architecture. He couldn't reason about the global implications of his local actions. Most significantly, he couldn't coordinate his actions with his human teammates. He did not signal to others what he was doing. He didn't share his intent to continue restarting servers every time they went down. He didn't tell anyone what he planned to do next. This meant the team could not integrate his activities into the evolving understanding of the problem they faced. He wasn't able to adapt to the broader context and began working contrary to the SRE team's larger goals for system reliability.

I continue to refer to "Jacob" as an independent actor because he has agency to act in the incident response, and, as an agent with an active role—he was recalcitrant. His actions were counterproductive to the collective joint activity of the SRE team. Working with someone who refuses to cooperate or to coordinate their actions is frustrating. Being literal-minded, context-limited automation[6] means it does not know how to cooperate and coordinate. It was designed to restart services with little consideration for anything else, and, while it did that competently and diligently, it could not understand and adapt to the larger joint activity of incident response. In other words, it is not the kind of incident responder you would want to work with under time pressure, uncertainty, or stress.

While there are exceptions, human counterparts engaged in shared goals and missions will engage in cooperative coordinated activity. These cooperative and coordinative skills are precisely the kinds of capabilities—not simply observability and explainability—that next-level automated systems require.

The questions posed by the confused SRE team—What is it doing? Why is it doing that? What will it do

next? How can we stop t?—are representative of the common coordination breakdowns in poorly designed human–automation teams.[7] These can occur, for example, because of a configuration error when the automation was initially being deployed or because the automated agent changes its behavior in response to inputs from the engineer. In complex and particularly high-tempo operations, software observability and explainability are not enough for safe, reliable human–machine teaming. The limits to this approach come sharply into focus the more complex the work becomes, the larger the scale, and faster the speed. There are many examples in software where humans and machines ended up working at cross-purposes with high consequences. Examples include the runaway automation in Knight Capital's high-frequency trading algorithm resulting in a pretax loss of US$440 million; AWS's 8-h scaling issue impacting events such as final exams for colleges, online ordering, and retail delivery services; and the Boeing 737 Max 8's Maneuvering Characteristics Augmentation System malfunction that killed 346 passengers onboard two flights.

JOINT COGNITIVE SYSTEMS — WORKING BETTER TOGETHER

Coordination with human tasks and human capabilities must be the central starting point for automation and artificial intelligence design. It is the engineer's perception, interpretation, reasoning, and action that will be called upon when the system approaches or crosses the limits of the conditions the system was designed for. While they are powerful allies in managing large-scale systems, automated agents are literal-minded and disconnected from the world in which problems play out. Software engineers who manage continuously available services are ultimately accountable for the consequences of poor system performance. This is true even when the automated agent—purported to be highly reliable and trustworthy—fails or begins working against them. Therefore, machines should be designed around helping the engineer make sense of the situation as accurately and quickly as possible. This shift moves from "technology first" to a joint cognitive system.

As the software industry continues to push the boundaries of the possible by introducing intelligent machines, designers of these technologies have an inherent responsibility to ensure they are safe and reliable partners who are capable of meaningful collaboration and efficient coordination. This becomes increasingly important in environments where human–machine teams manage complex, challenging work. It becomes critical with multiple human teams working with multiple machine teams in large-scale cooperative work systems. In the next article, we will examine design principles for collaborative work in joint cognitive systems.

REFERENCES

1. G. Klein, P. J. Feltovich, J. M. Bradshaw, and D. D. Woods, "Common ground and coordination in joint activity," in *Organizational Simulation*, vol. 53, W. B. Rouse and K. R. Boff, Eds., New York, NY, USA: Wiley, 2005, pp. 139–184.
2. M. Johnson, J. M. Bradshaw, P. J. Feltovich, C. M. Jonker, M. B. Van Riemsdijk, and M. Sierhuis, "Coactive design: Designing support for interdependence in joint activity," *J. Human-Robot Interact.*, vol. 3, no. 1, pp. 43–69, Feb. 2014, doi: 10.5898/JHRI.3.1.Johnson.
3. M. Johnson, J. M. Bradshaw, R. R. Hoffman, P. J. Feltovich, and D. D. Woods, "Seven cardinal virtues of human-machine teamwork: Examples from the DARPA robotic challenge," *IEEE Intell. Syst.*, vol. 29, no. 6, pp. 74–80, Nov./Dec. 2014, doi: 10.1109/MIS.2014.100.
4. L. M. Maguire, "Managing the hidden costs of coordination," *Commun. ACM*, vol. 63, no. 4, pp. 90–96, Apr. 2020, doi: 10.1145/3379989.
5. L. Bainbridge, "Ironies of automation," *IFAC Proc. Vol.*, vol. 15, no. 6, pp. 129–135, 1982, doi: 10.1016/S1474-6670(17)62897-0.
6. D. D. Woods and N. Sarter, "Learning from automation surprises and 'going sour' accidents," in *Cognitive Engineering in the Aviation Domain*, N. Sarter and R. Amalberti, Eds., Lawrence Erlbaum Associates, 2000, pp. 327–254.
7. D. D. Woods, "Steering the reverberations of technology change on fields of practice: Laws that govern cognitive work," in *Proc. 24th Annu. Conf. Cogn. Sci. Soc.*, Evanston, IL, USA: Routledge, 2002, pp. 14–16.

LAURA MAGUIRE is a research fellow with the Cognitive Systems Engineering Lab, The Ohio State University, Columbus, OH 43210 USA. Contact her at maguire.81@osu.edu.

EDITOR: Diomidis Spinellis, dds@aueb.gr

DEPARTMENT: ADVENTURES IN CODE

This article originally appeared in
Software
vol. 41, no. 2, 2024

Automating a Massive Open Online Course's Production

Diomidis Spinellis

For those of us who enjoy coding, automating a mundane task with software brings several benefits. Apart from raising our productivity in terms of quality and volume, it makes our work more interesting, and it provides us with several learning opportunities. When the task's domain is one other than software engineering, say, contract drafting or machining, the automation also allows us to cross pollinate the domain with our time-honed practices, such as configuration management, build automation, continuous integration, reuse management, and verification.[1] Here is a look behind the scenes at how I automated the production of a massive open online course (MOOC) I developed on the use of Unix command line tools for data, software, and production engineering.[2] Although many of the tools and techniques I present are specific to the particular course, the related ideas and concepts can be applied to many similar situations.

PRODUCING AND MANAGING THE COURSE'S FIVE HOURS OF VIDEO, WITH 896 SLIDES, 234 NARRATED ANIMATIONS, AND 47,000 WORDS OF SUBTITLES AND TRANSCRIPTS, WASN'T TRIVIAL.

The six-week course consists of six units with 37 modules in total, covering basic command line interactions, the Unix Bourne shell syntax, and good practices as well as the use of data fetching, selection, processing, and reporting tools. Each module includes a short (6–10-min-long) video, reading material, and formative exercises coupled with discussion and FAQ sections. Following edX's guidelines, we put significant effort into making this an engaging rather than a typical material-only course, addressing all requests and questions and participating in discussions or starting new ones. This has been broadly recognized in our received feedback throughout the runs. In the videos, I present slides guiding the learner through each module's material and narrate command line interactions demonstrating the use of the corresponding tools. Up to now, more than 7,000 learners from about 100 countries have enrolled in the course.

The course was produced by DelftX, a Delft University of Technology partnership with the edX online education platform. Its educators guided the course's design, and its production professionals and studios supported the material's development. Yet, producing and managing the course's five hours of video, with 896 slides, 234 narrated animations, and 47,000 words of subtitles and transcripts, wasn't trivial. The main challenges were producing polished content, reducing manual tasks, and organizing the material and its production. I addressed them by applying tried software engineering practices and developing about 35 custom tools (see Figure 1).

When recording the course contents, I realized that it would be impossible for me to create professional videos in a single take: I would hesitate, use filler words, mistype commands, or forget what I had planned to say. We engineers solve the problem of creating reliable systems out of potentially faulty parts by modularizing them and assembling them from components that have passed testing. Similarly, for the course, I broke each module into a short recorded video of me guiding the learners through its structure

Digital Object Identifier 10.1109/MS.2023.3346590
Date of current version: 22 February 2024

 Published by the IEEE Computer Society

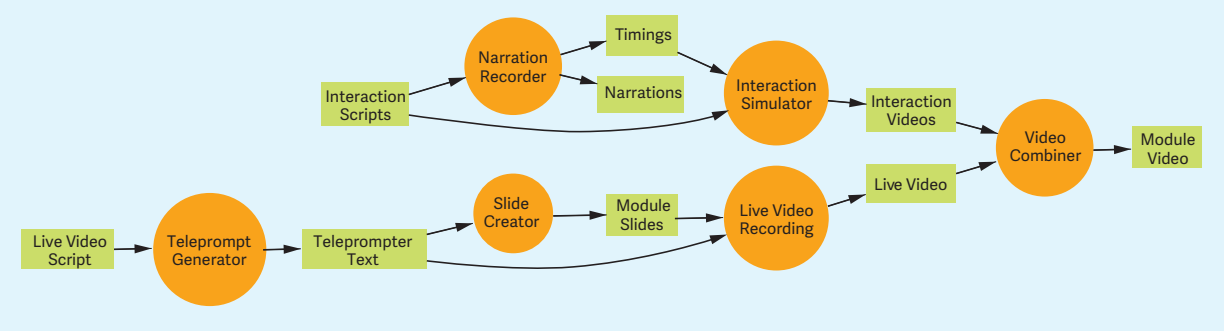


FIGURE 1. The data flow among the MOOC's processes and tools.

and several narrated full-screen demonstrations of commands or tools. A fault in any of them required only the video's individual rerecording rather than ditching the whole module.

For each video guiding the learners through the module's structure, I wrote a script in Markdown format with the exact wording I wanted to use. A simple tool processes this script to create text for the teleprompter together with annotations for the buttons I need to press to advance slides or switch between my image and a full-screen view of the computer's screen. This allowed me to concentrate during recordings on the clear delivery of the material, without worrying about the content or what button to press each time.

CLICKETY-CLICK

The most challenging part of the course videos was the tool demonstrations. For these, I would have to coordinate the flawless typing of commands with the narration of what was going on. After realizing my incompetence in this area, I developed a less error-prone workflow. Rather than live typing commands, I created scripts showing each typed command and its output. Here is an excerpt from a script demonstrating the grep command:

```
$ grep a.a.a.a words | head
acatamathesia
amadavat
amalaka
anabata
$^Pgrep '^....$' words | wc -l
5272
$
```

I implemented the workflow through two programs. The first one (180 lines of Python) reads such scripts; shows their contents onscreen, pausing between commands (as indicated by the ^P sign in the preceding script); and highlights each script part that needs to be explained. In parallel, it records me describing what is going on together with timing information for each segment's narration.

THIS LITTLE DANCE ENHANCES THE MATERIAL'S REALISM BY INTERMIXING ELEMENTS OF THE COMPUTER-GENERATED ANIMATION WITH THE LIVE VIDEO.

The bulk of the work is performed by a second program, which converts the script and the narration with its timing information into a video simulating the typing of commands and appearance of results. The program employs several techniques to enhance the video's realism and utility (see Figure 2):

- It alternates the rate at which characters appear between a human's slow typing and the computer's fast output.
- It mixes into the narration the sound of the keyboard clicks associated with the typing of each character.
- It shows typed commands in blue and computer-generated output in black.
- It allows scripts to annotate commands with comments, which appear in the video instantaneously and colored in green.

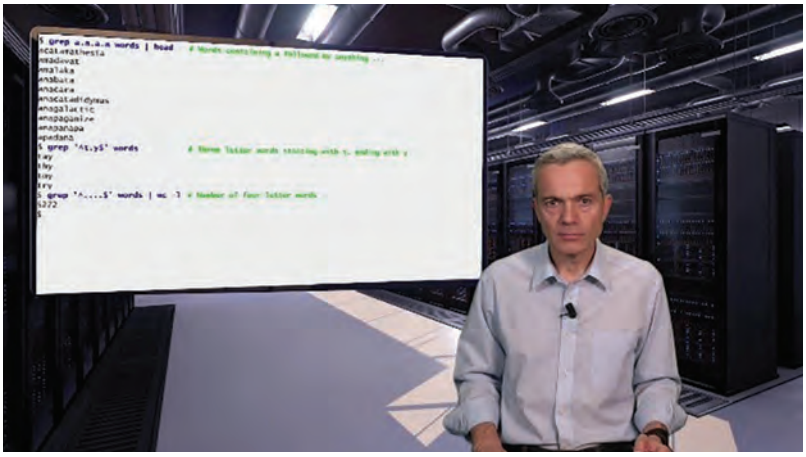


FIGURE 2. The live video with the animation's last frame.

I wrote most of this program (1,500 lines) in Processing, a language targeting the electronic arts, new media art, and visual design communities.[3] Its powerful yet uncomplicated graphics support and integrated development environment helped me iteratively adjust the output to the form I required.

The program produced thousands of individual frames. I then employed a shell script using the FFmpeg and SoX programs to convert the individual graphics frames, the previously recorded narration, and the key click sound effects into a video for the specific segment. Using the most appropriate language or tool for each part of the task simplified the required code for a job tailored to the needs of the specific course.

AUTOMATION

Having secured the command description videos, the next challenge was to automate the creation and assembly of the course's content. For that, I developed several small programs, each addressing a part of the workflow.

One program automates the creation of the slides. It reads the teleprompter script and creates the module's opening slide, slides with the title of each of the module's parts, slides with descriptive images, and transition slides for guiding the video merging (more on that later). The program creates each slide as a PDF document and then uses the pdfunite command to combine the slides into a single presentation. For the creation of each slide, I employed a technique that has served me many times, from creating journal review appreciation certificates to best paper award plaques. I created a template of each slide type in the Inkscape vector graphics editor program, using variable-like placeholder text (e.g., `$UNIT` or `$SECTION`) for elements that would need to be filled in (see Figure 3). Because Inkscape saves image files in the textual scalable vector graphics (SVG) format (an application of XML), it is easy for the program to tailor them to the required contents through simple textual replacements. The program then invokes Inkscape as a command line tool to convert the tailored SVG file into a PDF.

I also used the slides I presented during the live video recording of the module's structure as keys to automate the creation of the entire module video containing both the live parts and the narrated animations. At each point where an animated video was to be inserted, the presentation contained three slides: a slide with an empty computer screen containing just the shell prompt character (the animation's first frame), a key slide with a green background listing the name of the animation video to be inserted, and a slide containing the animation's last frame. When recording the live video, on seeing the shell prompt slide, I would switch the slide presentation to full screen, advance to the key slide and to the animation's end slide, and then switch back to live video (as seen in Figure 2). This little dance enhances the material's realism by intermixing elements of the computer-generated animation with the live video. A 157-line Python program written around the MoviePy library processes the live video frame by frame. In the places where it detects the mostly green key frame, it inserts each module's animation and, finally, writes out the combined video. Thus, the program flawlessly merged the 234 narrated animations into the 37 live-recorded videos in just a couple of hours.

I tried to automate the creation of the typescripts and subtitles for each module by passing the recorded audio to Google's speech-to-text recognition engine. At the time (2019), the results were subpar, forcing us to manually type in most of the text. However, I guess this would work much better today.

Finally, I automated the addition into the edX platform's course structure of the presentation slides, typescripts, and reading material. Unfortunately, no application programming interface was available for this task. As a workaround, I downloaded the entire course as XML and HTML files, reverse engineered the format, and had a program create additional files to inject the material into the course files, which I then uploaded as an edited course.

FIGURE 3. The module's first slide, with a title placeholder.

ORGANIZATION

Organizing the course's primary materials (752 files) was a matter of adopting commonly applied software engineering practices. I established and followed naming conventions for each module and animation script, using descriptive names and prefixing each name with a number denoting its sequence within the module or course. Small tools automate the renumbering when I change the presentation order.

Each of the 37 modules is stored in a separate directory containing its animation scripts, their transcribed narrations, the live video script, reading material, and knowledge checks. The course material build process is controlled and automated using the Make program[4] and a Makefile shared between modules through file symbolic linking. Makefile rules specify the creation of videos, presentations, typescripts, readings, and knowledge check files. In total, 22 GB of course content is generated from 3.4 GB of narration audio recordings, 17 GB of autogenerated animation videos, and 11 GB of live video recording.

I put all textual course materials and tools under Git version control. The 558 committed changes simplified the collaboration with teaching assistants while also allowing me to make bold changes without fear because I could always revert them if they didn't work out.

In sum, automating the course's production process tipped the balance from churning through mind-numbing repeated tasks to the joy of coding the required programs while also maintaining a high level of quality in the created materials and streamlining our interactions with the production team. Such automation used to be a perk for those of us versed in software development. However, assistance from systems, such as ChatGPT, has lowered the bar of the required expertise, making automation more widely accessible.

REFERENCES

1. D. Spinellis, "Extending our field's reach," *IEEE Softw.*, vol. 32, no. 6, pp. 4–6, Nov. 2015, doi: 10.1109/ms.2015.138.
2. D. Spinellis, "DelftX: Unix tools: Data, software and production engineering," edX, Cambridge, MA, USA, 2020–2024. [Online]. Available: https://www.edx.org/learn/unix/delft-university-of-technology-unix-tools-data-software-and-production-engineering
3. C. Reas, B. Fry, and J. Maeda, *Processing: A Programming Handbook for Visual Designers and Artists*, 2nd ed. Cambridge, MA, USA: MIT Press, 2014.
4. S. I. Feldman, "Make—A program for maintaining computer programs," *Softw., Pract. Experience*, vol. 9, no. 4, pp. 255–265, 1979, doi: 10.1002/spe.4380090402.

DIOMIDIS SPINELLIS is a professor in the Department of Management Science and Technology, Athens University of Economics and Business, Athens 104 34, Greece, and a professor of software analytics in the Department of Software Technology, Delft University of Technology, 2600 AA Delft, The Netherlands. Contact him at dds@aueb.gr.

EDITOR: Joanna F. DeFranco, Pennsylvania State University, jfd104@psu.edu

DEPARTMENT: INTERNET OF THINGS

This article originally appeared in
Computer
vol. 54, no. 12, 2021

The 12 Flavors of Cyberphysical Systems

Joanna F. DeFranco, *Pennsylvania State University*
Dimitrios Serpanos, *University of Patras*

There is wide variation in the definitions of cyberphysical systems and the Internet of Things. Different organizations have attempted to provide clarity, as consensus and consistency will help advance both technologies.

In the October issue of *Computer*, the "12 Flavors of IoT" were presented in this column.[1] A part two and follow-up column on cyberphysical systems (CPSs) is appropriate given their relationship to the Internet of Things (IoT).

It is not surprising that, when a new technical concept is introduced, its definition evolves until it reaches a stable state—this can take years. The definition variations could simply be due to the technology's application use expanding and/or its architecture being refined, and so on. The problem is that, given the nature of the Internet, the varying definitions are persistent and cause confusion. This situation has occurred with the definitions of both *IoT* and *CPS*. In addition, the relationship between the IoT and CPS technologies adds to the complexity of creating clear definitions. Thus, work groups, consortiums, government entities, and various stakeholders have attempted to provide clarity with their own definitions.

The term *CPS* was coined by Dr. Helen Gill, a scientist at the National Science Foundation (NSF). CPSs were at the forefront of discussions beginning in 2006.[8] In 2008, Dr. Gill defined *CPS* at a workshop titled "New Research Directions for High Confidence Transportation CPS: Automotive, Aviation, and Rail" and at a conference held at Carnegie Mellon University.[9,10]

Since 2008, the CPS domains have expanded to systems such as

- communication (for example, cellular, sensor networks, and wireless)
- consumer (such as audio and video systems as well as interactive games)
- energy (including energy production, distribution, and optimization)
- infrastructure (for example, disaster recovery; health monitoring; and water safety, distribution, and optimization)
- manufacturing (such as robotic machinery, embedded vision, and computer-controlled actuation)
- military (encompassing unmanned vehicles and weapon systems, among others)
- physical security (including card access control, video analytics, and so on)
- robotics (such as motion control, among others)
- smart buildings (for example, building system management)
- transportation (including automotive, avionics, aerospace, railroads, traffic management, and so forth).

The goal of this article is to facilitate a path toward a consistent understanding of CPSs, as was done with the "12 Flavors of IoT" column in October 2021.[1] Definitions of CPSs written by the most prominent CPS stakeholders, beginning with Dr. Gill's 2008 definition to the present, are analyzed and compared to the agreed upon CPS characteristics. This article also discusses the key differences and relationship between CPSs and the IoT.

Digital Object Identifier 10.1109/MC.2021.3112319
Date of current version: 17 November 2021

 Published by the IEEE Computer Society

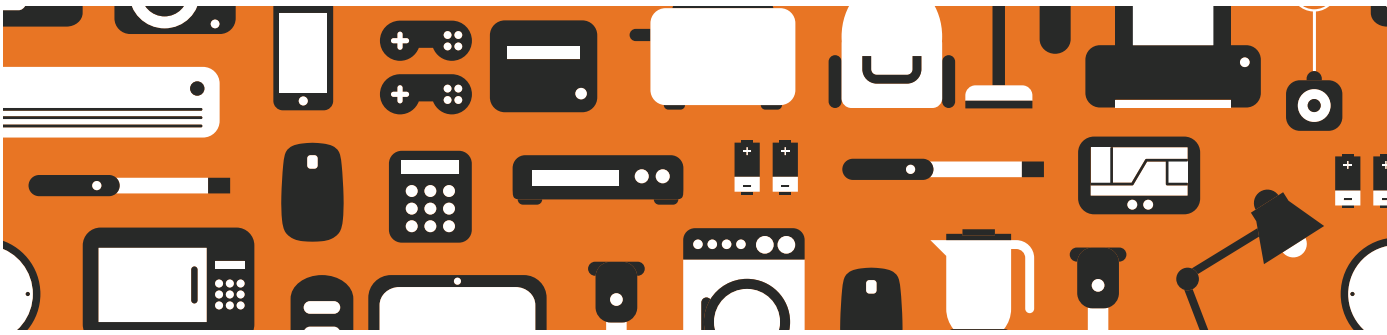


TABLE 1. The CPS characteristics.

	NIST (SP 1900-202)[5]	Platforms4CPS[7]
1	*Hybrid systems:* The architecture of CPSs consists of both physical and logical elements, for example, a system that can address the close interactions and feedback loop between sensing systems and physical components	*Physical action or processes:* For example, motion/control functionalities *Energy:* For example, storage, distribution, harvesting, and efficiency
2	*Hybrid methods:* Software to join the integrated physical and logical systems, comprising the networking, information processing, sensing, and actuation that allow the physical device to operate in a changing environment	*Processing:* For example, information
3	*Control:* Using computational systems to control physical processes and engineered systems, such as to monitor, coordinate, and control physical operations using computing and communication	*Communication:* For example, between things and machines, including wired/wireless and local/global
4	*Component classes:* For example, physical/engineered components, sensors, actuators, IT systems, and so on	*Sensing:* Of the physical world
5	*Time:* Integrating the physical-world time with event-driven computation	*Coordination and collaboration:* For example, for physical actions occurring outside the system
6	*Trustworthiness:* Safety, reliability, and security	—

CPS CHARACTERISTICS

Platforms4CPS and the National Institute of Technology (NIST) gathered different experts to specifically discuss and determine CPS characteristics. The results from these two major collaboration efforts were used to analyze the prevalent CPS definitions.

Platforms4CPS (platforms4cps.eu) is a European consortium of experts from academia and industry with the mission of creating a "vision, strategy, and technology building blocks" to support the developers of CPS applications. One of the outcomes of this consortium is a document describing the foundations of CPS engineering that includes six CPS characteristics.[7]

The Smart Grid and Cyber-Physical Systems Program Office at NIST also defined six CPS characteristics in SP 1900-202.[5] Table 1 shows both sets as well as a mapping of the Platforms4CPS characteristics to the NIST ones. As a result, two of the Platforms4CPS components were mapped to the NIST hybrid system attribute. Also, the trustworthiness characteristic was not addressed by Platforms4CPS. Therefore, the six NIST CPS characteristics are used to analyze the 12 CPS definitions in the next section.

CPS DEFINITIONS

Table 2 shows the 12 CPS definitions from key CPS stakeholders. The CPS characteristics from Table 1 were mapped to each definition in Table 2. For example, one of the first CPS definitions is from the NSF. It shows that the first definition in 2008 maps to all NIST components except for trustworthiness. Overall, most definitions recognize the hybrid systems and hybrid methods. However, most are missing some verbiage to address control (nine out of 12), component classes (eight out of 12), time (eight out of 12), and trustworthiness (10 out of 12).

CPSS VERSUS THE IOT

There is no doubt that the CPS and IoT technologies are related; however, there is limited consensus on the exact similarities, differences, and relationship between them. Many researchers have attempted to

TABLE 2. The CPS definitions mapped to defined CPS characteristics.

Entity	Definition	CPS component mapping
NSF (2008)[10]	"Cyber-physical systems are physical, biological, and engineered systems whose operations are integrated, monitored, and/or controlled by a computational core. Components are networked at every scale. Computing is 'deeply embedded' into every physical component, possibly even into materials. The computational core is an embedded system, usually demands real-time response, and is most often distributed. The behavior of a cyber-physical system is a fully-integrated hybridization of computational (logical) and physical action."	Maps to 1, 2, 4, and 5 Missing 6: trustworthiness
NIST (website)[11]	"Cyber-Physical Systems (CPS) comprise interacting digital, analog, physical, and human components engineered for function through integrated physics and logic."	Maps to 1 and 2 Missing 3, 4, 5, and 6: control, component classes, time, and trustworthiness
CPS PWG NIST SP 1500-201 (2017)[3]	"Cyber-physical systems integrate computation, communication, sensing, and actuation with physical systems to fulfill time-sensitive functions with varying degrees of interaction with the environment including human interaction."	Maps to 1, 2, 4, and 5 Missing 3 and 6: control and trustworthiness
IEEE Standard 2413 (2019)[2]	"A cyber-physical system is a system in which the physical world, such as production sites, and the digitalized cyber world are harmoniously combined."	Maps to 1 Missing 2, 3, 4, 5, and 6: hybrid methods, control, component cases, time, and trustworthiness
An academic work group website on CPSs[12]	"Cyber-Physical Systems (CPS) are integrations of computation, networking, and physical processes. Embedded computers and networks monitor and control the physical processes, with feedback loops where physical processes affect computations and vice versa."	Maps to 2 and 3 Missing 1, 4, 5, and 6: hybrid systems, component classes, time, and trustworthiness
IEEE Technical Committee on CPS (website)[13]	"CPS addresses the close interaction and deep integration between the cyber components such as sensing systems and the physical components such as varying environment and energy systems."	Maps to 1, 2, 4 Missing 3, 5, and 6: control, time, and trustworthiness.
ACM[14]	"Cyber-physical systems are systems with a coupling of the cyber aspects of computing and communications with the physical aspects of dynamics and engineering that must abide by the laws of physics."	Maps to 1 and 2 Missing 3, 4, 5, and 6: control, component classes, time, and trustworthiness
Cyber-Physical Systems Virtual Organization (website)[15]	CPSs "are engineering systems that are built from, and depend upon, the seamless integration of computational algorithms and physical components."	Maps to 1 and 2 Missing 3, 4, 5, and 6: control, component classes, time, and trustworthiness
NASA (website)[16]	"Cyber-Physical (CPS) denotes the emerging class of physical systems that exhibit complex patterns of behavior due to highly capable embedded software components. Also known as hybrid systems (a hybrid of hardware and software), or mechatronic systems (mechanical + electronic), these include devices with content, or knowledge, that gives them unprecedented capabilities in interoperability and interaction, resilience, adaptivity, and emergent behavior."	Maps to 1 and 2 Missing 3, 4, 5, and 6: control, component classes, time, and trustworthiness
U.S. Department of Transportation (2014)[17]	(From a presentation): A CPS is connected system with a path to vehicle automation using an infrastructure and new data for asset monitoring, predictive modeling, and control. Impacts safety, mobility, and the environment.	Maps to 1, 2, 3, and 6 Missing 4 and 5: component classes and time
U.S. Department of Homeland Security (website)[18]	"Smart networked systems with embedded sensors, processors and actuators that sense and interact with the physical world and support real-time, guaranteed performance in safety-critical applications."	Maps to 1, 2, 4, 5, and 6 Missing 3: control
Cyber-Physical Systems Program Solicitation NSF (2021)[19]	"Cyber-physical systems (CPS) are engineered systems that are built from, and depend upon, the seamless integration of computation and physical components."	Maps to 1 Missing 2, 3, 4, 5, and 6: hybrid methods, control, component cases, time, and trustworthiness

TABLE 3. The CPS and IoT component consensus issues.

Components	CPS	IoT
Control	Some definitions have a greater emphasis on the control of physical processes.	Some definitions have a greater emphasis on information flows from sensors.
Platform	Platform choice is a systemic decision and depends on system functionality.	The IoT can be a platform for or a simpler form of a CPS to achieve collaboration in a distributed system.
Internet	Inconsistency occurs if the Internet is among CPS design options.	There can be inconsistent association of the IoT with the Internet.
Human	Some emphasize human interaction.	Some minimize human interaction.

explain distinct variations; however, the challenge is, again, a lack of consistency in their respective definitions. In the work by Greer,[5] an analysis of the literature discussing CPSs versus the IoT showed four schools of thought: equivalency, partial overlap, CPSs are a subset of the IoT, and the IoT is a subset of CPSs. Note that IEEE Standard 2413 states, "An IoT system is a cyberphysical system, which interacts with the physical world through sensors and actuators."[2] Does this imply equivalency?

Greer[5] also described four specific components that add to the inconsistency between how much the CPS and IoT technologies overlap: control, platform, Internet, and human. The respective definitions of these four components for both CPSs and the IoT, shown in Table 3, create a problem in drawing a distinct conclusion about the CPS/IoT relationship/differences.

Another way to analyze and determine the exact distinctions between these two technologies is to review and compare the functionality of CPSs and the IoT within system architecture layers. Fatima et al.[6] reviewed functionalities in the analytic, intelligence, control, and configuration layers of each system architecture. For example, in the analytic layer, performance prediction (for example, tracking and responding to system changes) is a significant system requirement of a CPS but not common in an IoT system. Thus, theoretically, adding performance prediction to an IoT device would convert it to a CPS.

We are still left with four questions: Is the IoT a subset of CPSs? Are CPSs a subset of the IoT? Are CPSs and the IoT equivalent technologies? Do the CPS and IoT technologies only partially overlap? We can't answer these questions until there are clear and consistent definitions of CPS and IoT. The hope is that the two "12 Flavors" columns together will provoke clear definitions for both technology communities, as consistency in these definitions will help to incite new innovations, applications, and collaborations.

REFERENCES

1. J. DeFranco, "12 flavors of IoT," *Computer*, vol. 54, no. 10, pp. 133–137, Oct. 2021.
2. *IEEE Standard for an Architectural Framework for the Internet of Things*, IEEE Standards Association, Piscataway, NJ, IEEE 2413, 2019.
3. Cyber-Physical Systems Public Working Group, "Framework for cyber-physical systems: Volume 1, overview," NIST, Gaithersburg, MD, NIST SP 1500-201, 2017. [Online]. Available: https://nvlpubs.nist.gov/nistpubs/SpecialPublications/NIST.SP.1500-201.pdf
4. D. Serpanos, "The cyber-physical systems revolution," *Computer*, vol. 51, no. 3, pp. 70–73, 2018. doi: 10.1109/MC.2018.1731058.
5. C. Greer, M. Burns, D. Wollman, and E. Griffor, "Cyber-physical systems and internet of things" NIST, Gaithersburg, MD, NIST SP 1900-202, 2019. [Online]. Available: https://nvlpubs.nist.gov/nistpubs/SpecialPublications/NIST.SP.1900-202.pdf
6. I. Fatima, A. Anjum, S. Malik, and N. Ahmad, "Cyber physical systems and IoT: Architectural practices, interoperability, and transformation," *IT Prof.*, vol. 22, no. 3, pp. 46–54, May 2019. doi: 10.1109/MITP.2019.2912604.
7. "D4.3 Collaboration on the Foundations of CPS Engineering," Platforms4CPS, 2018. [Online]. Available: https://ec.europa.eu/research/participants/documents/downloadPublic?documentIds=080166e5bf0e32f9&appId=PPGMS
8. NSF Workshop on Cyber-Physical Systems, Cyber-Physical Systems Virtual Organization, 2006. [Online]. Available: https://cps-vo.org/node/179
9. "From vision to reality: Cyber-physical systems," NITRD, 2008. https://tinyurl.com/khf8p2

10. "A continuing vision: Cyber-physical systems," NITRD, 2008. https://tinyurl.com/frwkkedv
11. "Cyber-physical systems," NIST, Gaithersburg, MD. Accessed: Sept. 8, 2021. [Online]. Available: https://www.nist.gov/el/cyber-physical-systems
12. "Cyber-physical systems," Berkeley CPS Publications. Accessed: Sept. 8, 2021. [Online]. Available: https://ptolemy.berkeley.edu/projects/cps/
13. IEEE Technical Committee on Cyber-Physical Systems (CPS), IEEE Systems Council, 2017. Accessed: Sept. 8, 2021. [Online]. Available: www.ieee-cps.org
14. ACM/IEEE International Conference on Cyber-Physical Systems (ICCPS). Accessed: Sept. 8, 2021. [Online]. Available: http://iccps.acm.org/
15. "Internet of Things & cyber-physical systems," CPS-VO. Accessed: Sept. 8, 2021. [Online]. Available: https://cps-vo.org/group/iot
16. "Cyber-physical systems modeling and analysis (CPSMA) initiative," NASA, Washington, DC. Accessed Sept. 8, 2021. [Online]. Available: https://www.nasa.gov/centers/ames/cct/office/studies/cyber-physical_systems.html
17. NSF Workshop on Cyber-Physical Systems, NSF Transportation CPS, 2014. [Online]. Available: https://highways.dot.gov/sites/fhwa.dot.gov/files/docs/research/publications/multimedia/6606/cps.pdf
18. "Cyber physical systems security," Homeland Security. Accessed: Sept. 8, 2021. [Online]. Available: https://www.dhs.gov/science-and-technology/cpssec
19. "Cyber physical systems (CPS)," National Science Foundation, Alexandria, VA, 2021. [Online]. Available: https://www.nsf.gov/pubs/2021/nsf21551/nsf21551.htm

JOANNA F. DEFRANCO is an associate professor of software engineering at The Pennsylvania State University, Malvern, Pennsylvania, 19355, USA. Contact her at jfd104@psu.edu.

DIMITRIOS SERPANOS is president of the Computer Technology Institute and a professor at the University of Patras, Patras, 26504, Greece. He is the editor of the "Cyber-Physical Systems" column of *Computer*. Contact him at serpanos@computer.org.

Unlock Your Potential

WORLD-CLASS CONFERENCES — Stay ahead of the curve by attending one of our 195+ globally recognized conferences.

DIGITAL LIBRARY — Easily access over 900k articles covering world-class peer-reviewed content in the IEEE Computer Society Digital Library.

CALLS FOR PAPERS — Discover opportunities to write and present your ground-breaking accomplishments.

EDUCATION — Strengthen your resume with the IEEE Computer Society Course Catalog and its range of offerings.

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

NETWORK — Make connections that count by participating in local Region, Section, and Chapter activities.

Explore membership today at the IEEE Computer Society

www.computer.org

IEEE COMPUTER SOCIETY

IEEE

This article originally appeared in
Computer
vol. 55, no. 3, 2022

Should Cyberphysical Systems and the Internet of Things Get Married?

Joanna F. DeFranco, *Penn State Great Valley School of Graduate Professional Studies*

This roundtable discussion explores differences between cyberphysical systems and the Internet of Things, including technical challenges and progress toward addressing them. The panel concludes with highlights of cutting-edge and future research areas.

This is a virtual roundtable discussion between seven experts in the cyberphysical system (CPS) and Internet of Things (IoT) communities. It is a valuable conversation to improve community understanding and consensus in an effort to assist in the advancement of both technologies. The panelists were asked a series of emailed questions, thus some of the responses are interactive and will be presented in the order of the email threads. Their answers are thorough, inclusive, and thoughtful. In alphabetical order, the panelists include John Baras, University of Maryland; Oleg Loginov, IoTecha; Stephen Mellor, Industrial IoT Consortium; Janos Sztipanovits, Vanderbilt University; Haydn Thompson, THHINK Group; Martin Törngren, KTH Royal Institute of Technology; and Claire Vishik, Intel.

Computer: Is the IoT a subset, equivalent, or partial overlap of CPSs? Is this conversation worth pursuing? What is your school of thought and why?

Haydn Thompson: This question has been an open debate for many years, and if you are European, the general consensus is that the IoT is a subset of CPSs, and if you are American, the consensus tends to be that CPSs are a subset of the IoT. My view is that a fundamental characteristic of CPSs is the "physical" connection to the world. This is not always the case with the IoT. So, CPSs have an element of interaction with the physical world, usually via sensing, and then an aspect of control via actuation. The IoT, on the other hand, can just be working with data, with no control feedback (although the data may well be used for decision making). A good example of this is in diagnostics, for instance, medical/machinery, where data are collected to schedule maintenance and predict impending failures. Over the past few years, however, the world of the IoT has changed toward the IIoT, and there has been a blurring of the domains here, with many CPS and IoT applications now being called *IIoT*. This is a consequence of the cloud and operational technology worlds coming together in the cloud–edge IoT continuum.

Martin Törngren: My take is that CPSs, by definition, emphasize systems and, in particular, system-level properties. The IoT has traditionally been more of a bottom concept of creating opportunities as things are connected. In any case, physicality (energy, timing, reliability, safety, and so on) as well as cyber aspects will be essential for most of the future systems we are building, with similar trends and drivers. Regardless of the name, we are building systems and linking systems that will contain cyber parts (in terms of computers and feedback systems), physical parts, and humans, where the end properties will depend on the properties of the parts/constituent units, their interactions, and interactions with (other entities in) the environment, causing emergence. The distinctions that are more relevant, then, are on what types of systems we design (for example, the level of automation), if they represent systems of systems (no single system integrator), and their specific requirements.

Digital Object Identifier 10.1109/MC.2021.3133965
Date of current version: 11 March 2022

ROUNDTABLE PANELISTS

John Baras is a Distinguished University Professor and endowed Lockheed Martin Chair in Systems Engineering, University of Maryland, College Park, Maryland, USA. Contact him at baras@umd.edu.

Oleg Loginov is the president and chief executive officer of IoTecha, Cranbury, New Jersey, USA. IoTecha is a revolutionary technology for electric vehicle smart charging infrastructure and power grid integration. Contact him at Oleg@iotecha.com.

Stephen Mellor is the chief technology officer of the Industry IoT Consortium (IIC), La Jolla, California, USA. The IIC delivers transformative business value to industry, organizations, and society by accelerating the adoption of a trustworthy Internet of Things. Contact him at mellor@iiconsortium.org.

Janos Sztipanovits is the E. Bronson Ingram Distinguished Professor of Engineering, a professor of computer science, a professor of electrical and computer engineering, and the director of the Institute for Software Integrated Systems, Vanderbilt University, Nashville, Tennessee, USA. Contact him at janos.sztipanovits@vanderbilt.edu.

Haydn Thompson is the managing director/owner of THHINK Group, Sheffield, U.K. THHINK specializes in the development of custom platforms for diagnostics, condition/health monitoring, telemetry, and control, with specialist expertise in advanced data analytics, robust ultralow-power embedded wireless sensors, and energy harvesting. Contact him at haydn.thompson@thhink.com.

Martin Törngren is a professor at KTH Royal Institute of Technology, Stockholm, Sweden. Contact him at martint@kth.se.

Claire Vishik is a fellow at Intel, Santa Clara, California, USA. Intel creates emerging technologies, such as data servers, business transformation, memory, and storage, in fields including artificial intelligence, analytics, and cloud to edge. Contact her at claire.vishik@intel.com.

Oleg Loginov: We tried answering this question in IEEE 2413-2019[1]:

> *Interconnected and integrated IoT systems can provide new functionalities to improve the quality of life and to enable technological advances in areas such as personalized healthcare, emergency response, traffic-flow management, manufacturing, defense and homeland security, and energy supply and use. The impacts of IoT will be revolutionary and pervasive; this is already evident in emerging technologies such as autonomous vehicles, Smart Transportation, Smart Logistics, intelligent buildings, Smart Mining, Smart Energy Systems, Smart Manufacturing, multipurpose robots, Smart Agriculture, Smart Forestry, and Smart Medical Devices* (p. 14).

Also from IEEE 2413 (paraphrased): an IoT system is composed of components (or systems) that interact with one another to achieve a set of goals. Cyberphysical devices are technical artifacts/components that compute and interact with the physical via sensing and actuation.[2] Actuation, sensing, and control are fundamental to IoT systems. Examples of other types of cyberphysical mechanisms include dedicated storage devices and networking equipment, such as routers,

switches, and transceivers. They can be understood as "information transducers," in that they mediate the translation of physical properties into information by using a function (the intended purpose or characteristic action) and vice versa. Cyberphysical devices are part of a trend of "dematerializing" interactions. These information transducers include sensors for observing the physical world and actuators for changing the physical world.

Janos Sztipanovits: One of the variants of interpretations of CPSs is the following, from the National Institute of Standards and Technology (NIST) Framework for Cyber-Physical Systems[2]:

> *CPS are often engineered systems. ... CPS functionalities are the result of the tight integration of the cyber and physical sides* (p. 50).

The emphasis of this interpretation is that CPSs have functionalities that cannot be implemented only by physical and cyber means. This interpretation clearly has a profound impact on the design processes and required new system science foundations that must be both physical and computational. All in all, CPSs are a category of engineered systems, where certain essential functionalities emerge by the interaction of physical and computational processes. The IoT concept usually emphasizes engineering fine-grained networked systems. They may or may not be CPSs, and CPSs may or may not use IoT platforms. Regarding the common technology elements, I would look to the IoT as a possible platform for creating CPSs. In this sense, I would not equate the two; they are rather complementary.

Claire Vishik: Indeed, there are a number of views on the relationship between the two concepts, and, as indicated by Haydn, differing approaches to CPSs and the IoT in various geographic regions, for example, the United States and Europe. What is also remarkable is that, in many cases, definitions of the IoT are not provided in documents focusing on the IoT, to avoid controversy. In the United States, CPSs are more frequently considered a subset of the IoT, although, when these definitions are probed, little distinction between CPS and IoT definitions can be detected. To provide an example, the NIST Framework for Cyber-Physical Systems[2] defines CPSs as systems that "integrate computation, communication, sensing, and actuation with physical systems to fulfill time-sensitive functions with varying degrees of interaction with the environment, including human interaction" (p. 18).

On IoT, Voas[4] asked, "What is the IoT?" There are many ways to describe the IoT. More than 20 professional and research groups have worked to characterize the IoT, but so far, there is "no simple, actionable, and universally-accepted definition for IoT." Instead, the NIST "Networks of Things"[4] model focuses on cross-cutting components in the IoT as a way to at least describe what the term may mean: the Network of Things (NoT) model is based on five fundamentals at the heart of the IoT: sensing, computing, communication, e-utility, and actuation.

In other words, the core of the definition for CPSs and the description of the IoT (or NoTs) in the preceding are the same (communication, computation, sensing, e-utility, and actuation with physical systems), but CPSs, per the previous definition, describe those IoT systems that perform time-sensitive functions interacting, to diverse degrees, with the environment, including human interaction. But this behavior is also true of the IoT (or NoT). Thus, it is clear that there is no significant distinction between the two. In practice, electronic systems that have a distinct physical subsystem or electronic processes that have a clear physical element are frequently described as cyberphysical. Examples can be drawn from numerous areas such as autonomous vehicles and smart cities as well as electronically managed supply chains that transport physical goods.

Martin Törngren: Thanks, Claire, for bringing up the NIST IoT characterization and discussion. I would like to add to this. If we contrast this IoT characterization with the NIST CPS definition (raised earlier by Janos), I think we are onto a key difference in scope and emphasis: the IoT (or NoT) involves sensing, computing, communication, e-utility, and actuation. CPSs, or "smart" systems, are coengineered, interacting networks of physical and computational components.[6]

Sensors and actuators represent interfaces to the physical world; see, for example, the classical view of a mechatronic system (Figure 1) in Wikander et al.[5]

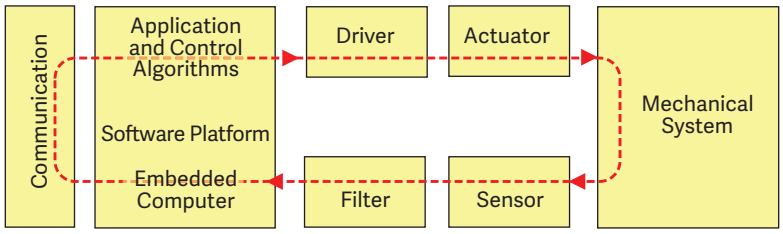


FIGURE 1. The interactions within an integrated mechatronic system (adapted from Wikander et al.[5]).

Thus, given the example that Voas had for the IoT/NoT, an IoT designer would go as far as designing the computer communication system toward sensors and actuators, but not the room (or the car, and so on). However, CPS design, by way of its construction, encompasses the "coengineering" of cyber and physical parts and thus also, for instance, the mechanical engineering aspects of a car. To me, this makes for a clear difference in scope and emphasis. The computer science or automatic control point of view is that the "plant" is given. If we take both the cyber and physical components into account, then we are designing a CPSs. This view appears to resonate with several previous comments, including the ones by Janos and Haydn.

John Baras: I think of the IoT and CPSs as quite different concepts (even if we consider, as is common today, networked CPSs). In CPSs, the physical part of the system involves multiple heterogeneous physics and plays a key role in system design and operation, which must coordinate the close interactions between the cyber and physical components. Not so for the IoT, which is very loosely defined, as far as I can tell, and primarily focused on the cyber and IT networking parts of systems. One example that emphasizes this important difference is modern and next-generation communication networks that integrate software-defined networks (SDNs), network function virtualization (NFV), and 5G, where everything essentially is software and the hardware components are standardized and de-emphasized. Of course, the two classes of systems overlap, but they are addressing different design and operational challenges. They overlap—one class is not a subset of the other class.

Another important difference is that while in both classes composability and compositionality are key concepts, with appropriate emphasis on component-based architectures and synthesis, it is in CPSs where the interface between the cyber and physical components must be treated as a system and not just as a simple port. For example, in several security challenges, these interfaces must be able to understand the semantics of both sides (the cyber and the physical) and specifically check whether the cyber-commands can be safely executed by the physical part; otherwise, we have catastrophic attacks like some very well-known ones (for example, STUXNET and broken wind turbines). Finally, if we take the view that any iterative algorithm is a dynamical system, most CPSs are hybrid (logic and physical) ones with digital and analog implementations. This is not the case for IoT devices.

Stephen Mellor: There is no useful differentiation. They are equivalent. This is similar to "fog" and "edge." Yes, we can quibble for months about the exact differences (if any), but in the end, the market will decide. Google returns 11.1 million links for *CPS* and 12.13 million for *IoT*. That is not quite as big a difference as I was expecting, but … Also, in response to John's email, there are a lot of similarities to the Industrial IoT (IIoT). In addition, CPSs and the IoT overlap to the point that it's six of one, half a dozen of the other.

Computer: You all contributed amazing responses, especially by referencing one another's statements and relevant documents. In reviewing/interpreting the statements to determine the themes among them, it appears there is consensus in highlighting/distinguishing the focus of each category/class (IoT/CPS) rather than the distinction label (overlap/complementary/subset). As Janos stated, "Interpretation clearly has a profound impact on the design processes and required new system science foundations." This speaks to the importance of this discussion.

Part of the discussion topic for this panel is related to a statement from John. He said that the two classes address "different design and operational challenges." The next two questions are posed with that in mind. Tell me if you agree with the following statement (if not, feel free to edit): the differentiation between the IoT

and CPSs can be described by focus/emphasis, where the IoT concerns networked components focusing on sensing/control with one another and CPSs concern the sensing/actuation of a distinct physical world system (or subsystem/electronic process) connection.

Vishik: In the previous discussion, we talked about the definitions of the IoT and CPSs. The conclusion, at least as far as I could see, was that there are very diverse definitions of the two areas and that each of us uses our own definitions that match specific research areas. This is somewhat similar to the definition of cybersecurity. A broad definition includes everything that may potentially acquire affiliation with cybersecurity. For example, the following is a commonly used extended definition from the National Initiative for Cybersecurity Careers and Studies[7]:

> *Strategy, policy, and standards regarding the security of and operations in cyberspace, and encompass[ing] the full range of threat reduction, vulnerability reduction, deterrence, international engagement, incident response, resiliency, and recovery policies and activities, including computer network operations, information assurance, law enforcement, diplomacy, military, and intelligence missions as they relate to the security and stability of the global information and communications infrastructure.*

This breadth was not helpful for the development of cybersecurity as a normal rigorous discipline. Similar breadth has been utilized to define the IoT for a variety of pragmatic reasons. For example, the Global System for Mobile Communications Association stated the following about the IoT in its guidelines on IoT security[8]:

> *Almost all IoT services are built using endpoint device and service platform components that contain similar technologies to many other communications, computing and IT solutions* (p. 5).

If we follow this line of reasoning, as many frameworks and definitions do, the IoT includes everything in information and communications technology, not just endpoint devices. Approaches like this make it easier to use frameworks in traditional IT to examine issues in the IoT. But they make it much harder to focus on cross-cutting issues that characterize what specific research papers consider the IoT.

To react to the previous statement, if we use traditional definitions of the IoT and CPSs that are exceedingly broad, the statement will be incorrect. If we use common sense (and narrower definitions of the IoT and CPS areas), it will be mostly correct but still contain a large number of exceptions for CPSs, especially in areas such as medical devices, where actuation is always mediated. For example, is a contact lens that measures blood sugar levels a CPS or an IoT device? It could actuate an insulin dispenser but only indirectly since it is a separate system. There are many similar examples in other areas. Using the word *focus* provides room for exceptions, but it seems that there are more exceptions than there are rules since only a few fields (for example, automotive, smart grids, and so on) lend themselves easily to this approach.

Thompson: In general, there are consequences in terms of physical harm or death if a CPS fails, for example, automotive, aerospace, and so on. If the IoT fails, there may be financial losses and inconvenience but not physical harm (here, the medical IoT may be an exception, but normally there is a human decision maker involved in the loop, as highlighted by Claire). Of course, not all CPSs are safety critical, for example, irrigation control systems in smart farming, but we often see a link via the Internet to a supervisory controller, and this may not require a hard real-time response.

Mellor: It's a distinction without a difference.

Törngren: Yes, I agree with this statement. People will often have different understandings of these terms and explicitly or implicitly assume a particular viewpoint, or set of viewpoints, meaning that they have a particular thing in mind (that is, a focus/emphasis).

Sztipanovits: Frankly, I do not completely understand this differentiation. Perhaps for the sake of finding some distinction, I consider the IoT a platform with the usual platform concerns and view the CPS as a design approach with strong emphasis on the codesign of cyber and physical aspects of systems. Clearly, IoT platforms are frequently used for developing systems where CPSs design approaches are needed (for

instance, certain categories of networked control systems). However, IoT platforms are also used for creating systems that would be hard to consider CPSs, due to the lack of cross-cutting constraints. Similarly, there are plenty of systems where CPS design approaches are beneficial, but they do not include any IoT elements (not even networking), and, of course, there many IoT-based systems that close control loops over networks and need CPS codesign methods. Since industrial-strength IoT platforms are increasingly available, they accelerate the need for, and increase the complexity of, CPS-like applications. Therefore, it makes sense to maintain links among the respected communities.

Baras: The statement is ambiguous. As several others have pointed out, we cannot continue calling everything CPSs and everything the IoT. We have been through this discussion several times within both communities. This trend, a few years ago, when Helen Gill was still at the National Science Foundation, came close to "killing" the funding for this program. We sharpened the definition, then, as Janos described. But unfortunately, the trend and bad habits keep creeping in.

So here is my precise answer. I will use networked CPSs and networked human CPSs (H-CPSs) as the reference frame because in this subdomain the IoT and CPSs overlap. The IoT is a platform (actually, a cyber-only platform) that addresses primarily communication (that is, data and information exchanges) between physical and cyber (hardware) devices and system components. CPSs are a framework that focuses primarily on the codesign of the cyber and physical parts of such systems, where there is close interaction between the cyber and physical components. If we consider H-CPSs, this also involves codesign (or better, a co-recommendation) about human behavior and human social aspects.

Now, when we go to networked CPSs and H-CPSs, things get more difficult, as we have to reconsider the network effects [and this needs to be specified precisely, as there are several networks involved (collaboration, information, and communication networks, with some being physical, some cyberlogical, and some mixed)] on the cyberphysical codesign. This has not been properly addressed in CPS research and development efforts. And in this subdomain, some IoT issues and concerns become relevant for CPSs. I cannot find many examples where the opposite is happening, that is, CPS issues becoming relevant for IoT systems. The one area where I have some examples involves constraints on the energy consumption of networked mobile devices communicating wirelessly.

Computer: What do you feel are the major technical challenges (design and operational) of the IoT and CPSs?

AT A VERY SIMPLIFIED LEVEL, HOW DO WE DEFINE REQUIREMENTS FOR SAFETY, SECURITY, AND PRIVACY WHEN THEY MAY BE ORTHOGONAL TO ONE ANOTHER?

Vishik: There are many challenges at the technical levels. Areas such as security, privacy, integration, safety, and so on are well known. Similarly, there is a significant body of knowledge that is growing at the intersection of physical and cyber areas. I will leave these aside for now. What I think is a significant gap is the ability to develop IoT and CPS devices in ways where requirements are integrated and the integrated risk metrics to evaluate potential outcomes are available. At a very simplified level, how do we define requirements for safety, security, and privacy when they may be orthogonal to one another? How do we recognize misalignment? How do we understand that new tools are needed? For example, are traditional safety metrics sufficient for autonomous vehicles? Or should we switch to model-based approaches?

With regard to the integrated risk picture, how can we define and compute risks for situations where, for example, safety and security need to be integrated? In a simplistic way, even looking at the percentages for allowed failure (which are much more rigorous in safety than in security) reveals a problem that remains unresolved. Without answering these questions, it will not be possible to address more complex environments based on systems of systems (for instance, smart cities). So, what are the major technical challenges in the IoT and CPSs? They are numerous. But they are connected by one foundational consideration: if we don't have stricter

definitions of the two areas, we will be able to address these challenges only in a highly fragmented fashion. Similarly, for cybersecurity that became a study of everything under its broad definition, more rigor is required to understand better how to build resilient IoT and CPS platforms and environments.

Thompson: Hard real time and safety are the main technical challenges in CPSs. An issue is that the control engineering, software engineering, and networking worlds are quite different. The two types of systems are developed in different ways. CPS engineers use rigorous processes to meet certification for safety. IoT engineers tend to have a less structured method of development, which is more about getting a system to market as quickly as possible (consider sprints and scrums). A key challenge is that the two worlds are colliding, with engineers trying to integrate safety-critical systems via the IoT. This would be OK for nonreal time (for example, the smart irrigation systems), but, of course, connecting a CPS via the Internet will result in delays, so hard real-time control would not be possible, for instance, the autonomous control of a car. Looking to the future, there will need to be new development methods (including certification) that can cope with these new integrated IoT/CPS systems, with more use of autonomous control at the edge to cope with intermittent connectivity and periods of outage.

Mellor: Not knowing what you don't know ("unknown unknowns" for American readers).

Törngren: A large amount of effort has been spent on investigating challenges for the IoT and CPSs, as reported in the Electronics Components and Systems for European Leadership strategic research agenda[9] and recommendations from the Platforms4CPS project.[10] I would like to highlight the following:

- using CPSs and the IoT to drive and support sustainability
- making sure that future CPS and IoT systems are trustworthy
- managing the complexity of future CPS and IoT systems.

These topics are naturally interrelated.

Sztipanovits: For the IoT: platforms that provide security, dependability, and at least some safety guarantees. For CPSs: composition, assurance, security, the assurance of CPSs with embedded learning-enabled components, DevOps and DevSecOps for CPSs, and, of course, a number of complex issues related to H-CPSs.

Baras: There are many challenges, as several others have already pointed out. My brief list of the main ones includes the following:

- *IoT*: security standards, privacy standards, containing security breaches at the edge when unknown devices are linked to edge routers, quantitative evaluation of the impact of 5G and 6G, quantitative evaluation of the impact of network virtualization (SDNs, NFV, and so on), and, most importantly, a systems engineering (composability and compositionality) framework to design/implement/operate IoT systems to provably satisfy given requirements
- *CPSs and H-CPSs*: most importantly, a systems engineering (composability and compositionality) framework to design/implement/operate CPSs, networked CPSs, and networked H-CPSs systems to provably satisfy given requirements, security and trust issues, and standards; integrating machine learning and artificial intelligence (AI) concepts and methods in such systems in a quantifiable and measurable way; developing a taxonomy of architectures for specific subdomains of CPSs; quantifying the effects that the several networks involved in networked CPSs and networked HCPSs have on one another (which is mostly unexplored territory); and developing credible models of human behavior (including social) and their incorporation into H-CPS investigations.

Loginov: The main difficulty is the deployment in brownfield (legacy equipment) scenarios. It is the integration with legacy systems that typically takes the most effort.

Computer: The challenges presented from the last question were well thought out and will most certainly

facilitate progress in the right direction. Claire summed it up with her comment on the importance of stricter definitions in both areas, so we can systemically address all of the challenges. The third and final question (based on themes I pulled from your comments), will be a continuation of Haydn's comment regarding "looking to the future" to cope with new integrated IoT/CPS systems.

What/how much progress is being made with the following challenges:

1. trust standards (for example, security, privacy, integration, and safety)
2. integrated requirements (such as the ability to develop IoT and CPS devices in ways where requirements are integrated)
3. risk metrics/provability guarantees (for instance, integrated, quantifiable, and measurable risk metrics; percentages for allowed failure; and so on)?

Loginov: We have made a lot of progress, but a lot is yet to be accomplished. It is important to embrace the constant of evolution. As we learn about trust in the IoT, we realize that a lot more still needs to be developed.

Thompson:

1. *Trust standards*: We already have standards for safety in various sectors, such as aerospace and automotive. We also have standards for privacy in Europe, including the General Data Protection Regulation. (Moving processing and data to the edge is also beneficial for privacy.) There are standards for security (and there has been a lot of activity on blockchain), and one thing we need in the future is trusted edge clusters. Integration is still a challenging area, though, as there are so many competing standards in this area, and we continue to have difficulties with semantic interoperability. When considering AI (which is now everywhere), we also need to think about transparency and ethical issues concerning trust.
2. *Integrated requirements*: I am not sure what the question is here, as requirements are always integrated. Do you mean integrating CPS and IoT systems? In this case, there are serious issues with proving safety, such as, latency, security, and so on.
3. *Risk metrics/provability guarantees*: My background is in aerospace, so things are very black and white for me. Risk depends on consequences and the probability that a given event will happen. Fundamentally, it is necessary to quantify risk and prove the appropriate figures to meet safety regulations.

APPROACHES TO CERTIFICATION WILL HAVE TO CHANGE, AS WE ARE MOVING TO SYSTEMS, FOR EXAMPLE, AUTONOMOUS CARS, WHERE WE CANNOT PREDICT EVERY RISK.

If you cannot prove this, the system will not be certified. I am thus a bit confused by the question. What I do believe is that approaches to certification will have to change, as we are moving to systems, for example, autonomous cars, where we cannot predict every risk. Here, we may need new approaches that provide a continuously predicted safety guarantee that is valid for a limited time period.

Törngren: I will complement Haydn's comment with the following:

1. *Trust standards*: Trust/trustworthiness is starting to be used as a new umbrella term, which incorporates dependability as well as attributes like fairness and transparency. This is, for example, noticeable in the new European Union AI guidelines (and proposed legislation). With the increasing capabilities and complexity of CPS and IoT systems, most trust-related aspects face challenges, and their combined consideration poses even greater challenges with the tradeoffs involved. In, for example, automated driving, a large number of new (and evolving) standards are in progress and related to safety and security, attempting

to define the "rules" of the game, operational design domains, risk metrics, safety processes for high levels of automated driving, and how to handle various vulnerabilities (from hardware/software faults, over insufficient specifications and performance imitations, to attacks).
2. *Integrated requirements*: My comments are similar to Haydn's.
3. *Risk metrics/provability guarantees*: A main aspect for future highly automated CPSs, operating in more unconstrained environments, is that they will need to reason about risk at runtime. They will thus have built-in risk metrics, which will be evaluated at runtime and have to trade performance versus, for example, safety. They will also be highly complex, emphasizing the need for transparency and explainability, presumably with some sort of mandated "black/red" boxes (like aircraft recorders). Formal models and proofs will be important, but their assumptions have to be scrutinized, and the real world will new generations of CPS, which will always pose surprises since they will (at some point) deviate significantly in their behavior from the model. Thus, resilient designs and architectures will be essential.

TRUST/TRUSTWORTHINESS IS STARTING TO BE USED AS A NEW UMBRELLA TERM, WHICH INCORPORATES DEPENDABILITY AS WELL AS ATTRIBUTES LIKE FAIRNESS AND TRANSPARENCY.

Mellor:

1. *Trust standards*: Standards are difficult in the absence of best practices and a principled view of how to reconcile various aspects. So, in respect to standards, they will be some time in coming. For principles and best practices, work is proceeding apace. See "The Industrial Internet of Things Trustworthiness Framework Foundations."[11]
2. *Integrated requirements*: This is backward. Requirements drive development. Besides, the IIoT and CPSs are the same thing, so what does "integrated requirements" even mean?
3. *Risk metrics/provability guarantees*: Again, the Industrial Internet Consortium is working on this, but we have not published anything [though there are some interesting sections in the Trustworthiness Framework[11] regarding how to represent trust numerically (see section 4.7), which will lead, in time, to metrics].

Sztipanovits: I agree with Martin and Haydn, so let me add just a few remarks. Since there are IoT applications that are not CPSs and CPS applications that are not the IoT, let me just comment on those systems where the two overlap: CPSs that are built on IoT platforms. Consider the following:

1. *Trust standards*: I cannot add too much. The term incorporates a number of different properties and interpretations. A particularly interesting area that is evolving rapidly is human–CPS systems, which force us to contrast the anthropomorphic interpretation of "trust" and possible machine-based interpretations. Networked human–AI–machine teams are emerging in areas such as connected autonomous vehicles, and much needs to be done to understand how to formalize "trust" in these hybrid, complex distributed systems.
2. *Integrated requirements*: I cannot add to what Haydn wrote.
3. *Risk metrics/provability guarantees*: This is becoming a tremendously important issue in autonomous systems (whether IoT based or not). As Martin wrote, the fundamentally new challenge is that these systems cannot be assured only at design time, not only because they are complex but because they frequently incorporate learning-enabled components that can evolve during operations and may be created in a completely data-driven manner (without explicit models). A new research direction in assured autonomy (there is an ongoing DARPA program on this) started developing dynamic assurance concepts that

can change during operation and runtime methods that produce a sort of "assurance gauge" indicating whether the system (or some of its components) goes out of conformance with training conditions. Regarding provable guarantees, there are viable results to bound system behavior with runtime safety monitors. This area of research is interesting, important, and wide open.

Baras: I have the following comments:

1. *Trust standards*: As I frequently state, trust is a very frequently used word and equally frequently abused. In the context of our discussions, there are several quite different meanings of trust. There is the standard meaning that we associate with human interactions. This, in itself, has several versions (for example, direct versus indirect trust). There is trust as it is used in telecommunications and computing, that is, devices, links, nodes, and computers that are trustworthy, meaning that after inspection, they have been found not to be compromised or offered stronger resilience to attacks. There is the trusted platform module, a secure chip standard with keys embedded at manufacturing time (a product of the industry Trusted Computing Group) that is now included in almost 75% of computers. Then there is trust in CPSs and autonomous systems, where the meaning is that a system executes a task or mission within the tolerance of an expected normal behavior.

   Before we can discuss standards, we need to define what trust means in the various problems relevant to our discussion and develop quantitative models of trust and associated specifications so that we can talk about verification and assurance. In addition, we need to develop trust and mistrust dynamics for single as well as networked systems. There is work along these lines in the various meanings and areas I have mentioned. Then we can define standards of trust in each area and most importantly the interoperability of trust across domains [that is, a way to translate and link trust specifications from area to area and across components, akin to security composition (still unsolved)].
2. *Integrated requirements*: I do not quite understand the thrust of this topic. In CPS and IoT systems, we have requirements to start with, which are modified and new ones are added as we step through a system design (that is, derivative requirements and so on). What is lacking in both areas is a framework for requirements that catalyzes and facilitates compositionality—contract-based design is a big step in this direction. We need a framework to combine requirements of physical components [usually given in terms of constraints and metrics involving numerical variables (continuous and sampled)] and requirements of cyber components [usually given in terms of constraints and metrics involving Boolean (integers) variables and via logic].

   There is mathematical unification between optimization and logic that leads to a unified framework via mixed (that is, numerical and integer variables) multicriteria constrained optimization, constraint-based reasoning, and satisfiability modulo theories and algorithms. But we have still a long way to go to have a framework and tools that are practical and easy to learn and use. Another very important challenge is to come up with an integrated modeling framework and tools to combine space and time specifications and their tolerances as needed because several requirements are now given via temporal logic (linear temporal logic, metric temporal logic, metric interval temporal logic, and signal temporal logic). STL is a step in this direction but a very small one.
3. *Risk metrics/provability guarantees*: Risk metrics are very important because they directly link to robustness and sensitivities to perturbations in inputs and models. There is a fundamental theory from robust control that covers many classes of systems problems but not yet temporal specifications well. The same mathematics that can be used for tradeoff analysis and design space exploration can be used (and has been used) in advanced methods and tools for verification and validation. But

with learning components and autonomy we need to develop rigorously what I call *trusted autonomy* (the term *assured autonomy* is also used). Trusted autonomy requires systems to self-monitor their behavior and execution of tasks, self-adjust models and execution to correct anomalies and deviations, and self-learn from task execution and monitoring and anomalies. There is active research in this area but we have a long way to go.

Vishik:

1. *Trust standards*: As pointed out in other responses, the answer depends on the definition of trust. If trust is understood as it is defined in trusted computing (we trust an application when it behaves the same way under the same circumstances), there are a large number of mature standards. The Trusted Computing Group has developed many of them beyond the trusted platform module. There are a number of International Organization for Standardization (ISO)/International Electrotechnical Commission standards (IEC), and there are several trusted execution environment standards, and the list can be continued. If we include the concept of trustworthiness, we will find a number of developing standards for various environments, such as CPSs and AI, for example, in https://www.iso.org/committee/6794475.html under the ISO/IEC. If we take the term *trust* casually, for example, saying that "without privacy, it is impossible to achieve trust in the digital economy," applying trust to ethics and societal situations, the use of the word is legitimate, but it doesn't have a rigorous definition and is descriptive rather than terminological.
2. *Integrated requirements*: These are not a new area, but the space has been slow to develop. This is due to a variety of factors, including the traditional separation of research areas between privacy and safety, for example, IT systems and CPSs. But this is a field of study that needs to receive a push from researchers and technologists. If we think about fully automated environments, for instance, self-driving cars and smart cities, codeveloping requirements for safety and security, physical subsystems and their cyber components, and so on, is necessary to move forward. I hope the interest in research in this key area will grow.
3. *Risk metrics/provability guarantees*: Risk metrics and metrics in general have always required considerable effort. The transition from calling for metrics and risk base analysis, publishing single-case risk models, and developing metrics/risk models that could be used in a whole field has always been complicated. The probabilities of failure vary significantly between safety and security and between physical subsystems and cyber components, to give an example. The transition from metrics to models (say, in safety) has also been slower than expected. With increased access to real-time and near-real-time data, these models can be constructed in new data-driven ways. The slowness is probably due to the fragmentation of the field. If we resolve the integration issues in point 2, building the quantifiable risk models in point 3 will be feasible, and improving them to make them broadly applicable will be a matter of time.

Computer: Thank you all for your participation in this discussion. You provided valuable insights and highlighted key research areas, especially trust (define trust, trusted edge clusters, semantic interoperability, transparency, fairness, vulnerabilities, developing models and standards of trust, developing trust and mistrust dynamics, and ethical issues), requirements (integrating safety/privacy/security, proving safety, proving latency, and a framework and tools), and risk (models, metrics, quantifying and certifying risks, reasoning about risk at runtime, and trusted/assured autonomy). Is there enough concept overlap within these two technologies that the CPS and IoT communities put on rings and start planning their marriage?

REFERENCES

1. *IEEE Standard for an Architectural Framework for the Internet of Things (IoT)*, IEEE Std 2413-2019, May 21, 2019.
2. Framework for Cyber-Physical Systems, release 1.0, NIST, Gaithersburg, MD, USA, May 2016. [Online].

Available: https://s3.amazonaws.com/nist-sgcps/cpspwg/files/pwgglobal/CPS_PWG_Framework_for_Cyber_Physical_Systems_Release_1_0Final.pdf
3. "Cyber-physical systems public working group smart grid and cyber-physical systems program office engineering laboratory," NIST, Gaithersburg, MD, USA, NIST Special Publication 1500-201, Jun. 2017. [Online]. Available: https://nvlpubs.nist.gov/nistpubs/SpecialPublications/NIST.SP.1500-201.pdf
4. J. Voas, "Networks of 'Things'," NIST, Gaithersburg, MD, USA, NIST Special Publication 800-183, Jul. 2016. [Online]. Available: https://nvlpubs.nist.gov/nistpubs/specialpublications/nist.sp.800-183.pdf
5. J. Wikander, M. Torngren, and M. Hanson, "The science and education of mechatronics engineering," *IEEE Robot. Autom. Mag.*, vol. 8, no. 2, pp. 20–26, Jun. 2001, doi: 10.1109/100.932753.
6. "Cyber-physical systems," Engineering Laboratory, National Institute of Standards and Technology, Gaithersburg, MD, USA. Accessed: Dec. 30, 2021. [Online]. Available: https://www.nist.gov/el/cyber-physical-systems
7. "Cybersecurity glossary," National Initiative for Cybersecurity Careers and Studies, Gaithersburg, MD, USA. Accessed: Dec. 30, 2021. [Online]. Available: https://niccs.cisa.gov/about-niccs/cybersecurity-glossary#C
8. "IoT security guidelines: Overview document," GMSA. https://www.gsma.com/iot/wp-content/uploads/2020/05/CLP.11-v2.2-GSMA-IoT-Security-Guidelines-Overview-Document.pdf (accessed Dec. 30, 2021).
9. "Strategic research agenda 2020," ECS. https://www.ecsel.eu/sites/default/files/2020-02/ECS%20SRA%202020%20%281%29.pdf (accessed Dec. 30, 2021).
10. "Platforms4CPS key outcomes and recommendations," Platforms4CPS. https://www.platforms4cps.eu/ (accessed Dec. 30, 2021).
11. M. Buchheit *et al.*, "The Industrial Internet of Things trustworthiness framework foundations," Industrial Internet Consortium. https://www.iiconsortium.org/pdf/Trustworthiness_Framework_Foundations.pdf (accessed Dec. 30, 2021).

JOANNA F. DEFRANCO is an associate professor of software engineering at the Penn State Great Valley School of Graduate Professional Studies, Malvern, Pennsylvania, 19355, USA. Contact her at jfd104@psu.edu.

EDITORS: Brittany Johnson, George Mason University, johnsonb@gmj.edu
Tim Menzies, North Carolina State University, tim@menzies.us

This article originally appeared in
IEEE Software
vol. 41, no. 1, 2024

DEPARTMENT: SE AND ETHICS

Ethics: Why Software Engineers Can't Afford to Look Away

Brittany Johnson and Tim Menzies

FROM THE EDITORS

Some people shy away from discussing ethics, believing it's not in the domain of software engineering. We want to steer the conversation in the opposite direction, and this column explains that such ethics-based discussions are crucial to our profession.

And for future issues, what do you want to see in this "SE for Ethics" column? Do you have an important insight or industrial case study? Something that could prompt an important discussion? Or, alternatively, something that extends or challenges significant ideas? If so, e-mail a one-paragraph synopsis to johnsonb@gmu.edu or timm@ieee.org (subject line: "SE for Ethics: Idea: [Your Idea]"). If that looks interesting, we'll ask you to submit a 1,000–3,000-word article (where each graph, table, or figure is worth 250 words) for review for *IEEE Software*.—*Brittany Johnson and Tim Menzies*

OUR PERSONAL JOURNEYS INFORM OUR VIEWS

To begin, we share our backgrounds as a reminder that perspectives on ethics are deeply personal and shaped by individual experiences.

We—Brittany, a Black woman from the Southern United States, and Tim, a senior white man of Anglo-Saxon heritage from Australia—bring our diverse life experiences to this conversation. Throughout our lives, we've witnessed hardworking individuals unable to succeed due to their environments. This injustice propels our advocacy for change.

ETHICS IN SOFTWARE: MORE THAN A HYPOTHETICAL

Not long ago, we attended a symposium where an affluent senior white male lauded the role of artificial intelligence (AI) in legal decisions, believing that it eliminated human biases. While that person has every right to express that view, we think that person was … ill informed. Experience with tools like the COMPAS risk assessment tool, which aims to predict potential reoffenders, has shown that AI models can exhibit biases against (for example) Black individuals. (COMPAS has a higher false-positive rate for Black than for white defendants. This means that, as a result of using COMPAS' recommendations, more white men got bail, and more Black men spent time in jail.)

COMPAS is just one example of the inherent bias in certain algorithms. Sadly, there are many other similar examples (see "Examples of Unfair Software"). Such biases aren't just unfair; they have real-life implications, such as people not getting the bail they deserve or businesses failing due to an algorithm's internal decision making.

ETHICS: BEYOND JUST "TOOLS"

For the aforementioned problems, there exist automatic tools that can, to some degree, adjust these

Digital Object Identifier 10.1109/MS.2023.3319768
Date of current version: 20 December 2023

 Published by the IEEE Computer Society

EXAMPLES OF UNFAIR SOFTWARE

The following points were taken from Cruz et al.[S1]:

- Women can be five times more likely to be incorrectly classified as low income.
- African Americans are five times more likely to languish in prison until trial rather than being given the bail they deserve.
- Proposals from low-income groups are five times more likely to be incorrectly ignored by donation groups.

The following point was taken from Canellas[S2] and Matthews[S3]:

- Forensic software used for DNA analysis is written so poorly that many people languish in jail, and some states have even banned the use of that software.

The following point was taken from the last chapter of Noble[S4]:

- A successful hair salon went bankrupt due to internal choices within the Yelp recommendation algorithm.

For more examples, see the rest of Noble[S4] as well as Rudin,[S5] Dastin,[S6] Hardesty,[S7] and Caliskan et al.[S8]

REFERENCES

S1. A. F. Cruz, P. Saleiro, C. Belém, C. Soares, and P. Bizarro, "Promoting fairness through hyperparameter optimization," in *Proc. IEEE Int. Conf. Data Mining (ICDM)*, Piscataway, NJ, USA: IEEE Press, 2021, pp. 1036–1041, doi: 10.1109/ICDM51629.2021.00119.

S2. M. Canellas, "Defending IEEE software standards in federal criminal court," *Computer*, vol. 54, no. 6, pp. 14–23, Jun. 2021, doi: 10.1109/MC.2020.3038630.

S3. B. Johnson and T. Menzies, "Unfairness is everywhere. So what to do? An interview with Jeanna Matthews," *IEEE Softw.*, to be published.

S4. S. U. Noble, *Algorithms of Oppression: How Search Engines Reinforce Racism*. New York, NY, USA: New York Univ. Press, 2018. [Online]. Available: http://algorithmsofoppression.com/

S5. C. Rudin, "Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead," *Nature Mach. Intell.*, vol. 1, no. 5, pp. 206–215, May 2019, doi: 10.1038/s42256-019-0048-x.

S6. J. Dastin, "Amazon scraps secret AI recruiting tool that showed bias against women," *Reuters*, Oct. 10, 2018. [Online]. Available: https://www.reuters.com/article/us-amazoncom-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-thatshowed-bias-against-women-idUSKCN1MK08G

S7. L. Hardesty, "Study finds gender and skin-type bias in commercial artificial-intelligence systems," *MIT News*, Feb. 11, 2018. [Online]. Available: https://news.mit.edu/2018/study-finds-gender-skin-type-bias-artificial-intelligence-systems-0212

S8. A. Caliskan, J. J. Bryson, and A. Narayanan, "Semantics derived automatically from language corpora contain human-like biases," *Science*, vol. 356, no. 6334, pp. 183–186, Apr. 14, 2017, doi: 10.1126/science.aal4. [Online]. Available: https://www.science.org/doi/10.1126/science.aal4230

systems to enable them to function without some of these prejudices (https://cs.gmu.edu/~johnsonb/fairkit.html).[1] But we rush to add that ethics isn't a problem we can merely "fix" with automated software patches. We need to address the root societal, economic, legal, and cognitive conditions that birthed these biases. We need to recognize the broader impacts of the technology we create and use. We must also acknowledge that sometimes well-intentioned frameworks, such as intersectionality, can be diluted over time and lose their impact.[2]

For a more inclusive software landscape, we must do the following:

- Diversify design teams to include multiple perspectives.
- Test software for potential biases against specific groups.
- Foster open communication with all stakeholders.
- Design better models that reduce the cognitive load required for their review.[3,4]

WE MUST EQUIP CURRENT AND FUTURE DEVELOPERS WITH KNOWLEDGE OF ETHICAL CONSIDERATIONS.

On the legal side, unbiased external review teams should regularly assess potentially discriminatory projects. Legislative mandates for software and AI system reviews are becoming crucial, especially since self-regulation doesn't always yield ethical outcomes (https://en.wikipedia.org/wiki/Volkswagen_emissions_scandal).

ELEVATING THE ROLE OF ETHICS IN SOFTWARE ENGINEERING

How do we prioritize ethics in our field? It starts with education. We must equip current and future developers with knowledge of ethical considerations. This doesn't just mean college courses but also professional and industrial settings. Implementing new policies and legislations that center on ethical considerations is another crucial step.

As software engineers, we make impactful decisions daily. The vast choices we make in system configurations offer an opportunity to shape the world ethically. Every design choice and management decision can profoundly affect society.

In essence, let's harness our power as software engineers. Let's lean into ethical considerations and make decisions that champion fairness, justice, and inclusivity.

REFERENCES

1. J. Chakraborty, S. Majumder, and T. Menzies, "Bias in machine learning software: Why? How? What to do?" in *Proc. 29th ACM Joint Meeting Eur. Softw. Eng. Conf. Symp. Found. Softw. (ESEC/FSE)*, New York, NY, USA, 2021, pp. 429–440, doi: 10.1145/3468264.3468537.
2. L. Bowleg, "Evolving intersectionality within public health: From analysis to action," *Amer. J. Public Health*, vol. 111, no. 1, pp. 88–90, Jan. 2021, doi: 10.2105/AJPH.2020.306031.
3. B. Green, "The flaws of policies requiring human oversight of government algorithms," *Comput. Law Secur. Rev.*, vol. 45, Jul. 2022, Art. no. 105681, doi: 10.1016/j.clsr.2022.105681.
4. T. Menzies, "Model review: A PROMISEing opportunity," 2023. [Online]. Available: https://arxiv.org/pdf/2309.01314.pdf

BRITTANY JOHNSON is an assistant professor at George Mason University, Fairfax, VA 22030 USA. Contact her at johnsonb@gmu.edu.

TIM MENZIES is a full professor at North Carolina State University, Raleigh, NC 27606 USA. Contact him at timm@ieee.org.

This article originally appeared in
IT Professional
vol. 25, no. 2, 2023

What If Ethics Got in the Way of Generative AI?

George Hurlburt, *STEMCorp Foundation, Tall Timbers, MD, 20653, USA*

One of the puzzling realities surrounding AI is that almost every AI app serves a singular purpose at which it often astonishingly excels. While frequently inscrutable and seemingly unexplainable, AI seldom exhibits semblances of true natural intelligence, much less the exercise of anything resembling real human intelligence,[1] including free will borne of consciousness. Instead, AI typically conforms to the rote algorithmic and often stochastic parameters underlying its programming and, when appropriate, to the data upon which it has been trained.

Nonetheless, some speculate that the advanced versions of OpenAI's Generative Pretrained Transformer (GPT), a relatively new darling of AI, will soon be capable of passing the Turing test and, hence, be relegated closer to artificial general intelligence. It is the case that GPT4, an order of magnitude bigger than its predecessors, has passed several tests, including the bar exam,[2] with stunningly high marks. Indeed, the genre of generative AI has set off a firestorm of speculation concerning its human interface. Responses are both positive and negative. Some of this speculation deals with the ethics surrounding the tools borne of generative AI. This concern has caused some prominent individuals to call for a hiatus in ongoing AI development.[3] Thus, it is perhaps fitting to examine generative AI through an ethical lens.

IS IT FAIR?

The FAIR initiative, intended to do for data what the Internet has done for networks, is founded on the principle of data, which is at once findable, accessible, interoperable, and retrievable (FAIR). To be FAIR is to be linked to solid metadata, the bedrock upon which the FAIR initiative is built.[4] Through reliance on robust metadata, authoritative sources may be linked in such a fashion as to substantiate existing data and advance scientific inquiry through exposure to new insights.

The large language models (LLM) upon which GPT systems operate give the impression that they are FAIR compliant. A well-constructed GPT prompt will find, access, integrate, and return a seemingly coherent response, usually directly related to the prompt. However, the results are far less than satisfying when the now-fabled ChatGPT is asked for source attribution. In this case, the lack of metadata underlying ChatGPT is evident via utterly bogus references when citations are requested via prompts. Worse, ChatGPT is known to offer regular occurrences of verifiable misinterpretation or outright misinformation. To a lesser degree, the same appears to be the case in the larger GPT4. GPT5, once again larger than GPT4, is slated to come out of training in December 2023.[5] While seemingly more authoritative, with a greatly enlarged working queue, GPT5 is still unlikely to effectively self-reference its sources. However, uncertainty prevails, as OpenAI, the creator of the GPT product line, is mum about the specific technical capabilities of its newer releases.

The ethical implication is that LLMs, while seemingly literate, do not seem able to substantiate what they produce with any rigor. This is because LLMs assemble their information in a probabilistic fashion, essentially linking words stochastically based on the content of the prompt. In reality, however, the result is a mathematical amalgam of a given concept without regard for its true meaning or underlying veracity. That GPT can rhyme, code, and generate and interpret images are all impressive variations of its ability to mathematically predict what should follow what has already been produced or prompted, not an innate ability to truly understand and justify its logic. Ultimately, the inability to provide authoritative references puts all of its output in question from a purely scientific standpoint. Thus, one might say that generative AI, at least as it seems to be evolving, is just not FAIR.

IS IT POTTY TRAINED?

LLMs draw upon vast amounts of data, often scraped directly from the Internet and other sources without attribution and, hence, without the expressed permission of the creators or curators of the data. To build mathematical confidence, the LLMs require extensive training. One desired outcome of this training is

1520-9202 © 2023 IEEE
Digital Object Identifier 10.1109/MITP.2023.3267140
Date of current version 5 May 2023.

eliminating all forms of bias, vulgarity, hate speech, sexually explicit material, and excessive violence. In keeping with industry practice, large cadres of workers are hired worldwide at near-poverty wages to weed out all objectionable materials from the LLM model to prevent them from becoming ingrained via training. There is no recourse for worker stress generated by a nonstop diet of vulgarity, hatred, perversion, and violence.[6]

Worse, as all forms of bias are already well baked into the Internet, literature, and art, such exercises to excise vast amounts of bias wholesale become a nonstop venture. The open ethical question would be, who judges what is biased and when? While it is straightforward to eliminate material that is clearly vulgar, socially unacceptable, or outright illegal, it is quite another thing to interpret art and literature based on presumptive normative standards. Whose standards apply, and can they be applied globally? The ethical question becomes one of the eyes of the beholder, who is far too often the AI power broker.

The same questions appear to hold in politics as well. It may be seemingly appropriate to eliminate extreme political bias in a balanced way, be it left or right. Given that sound political discourse necessarily shapes policy, how much culling of doctrine is too much, and what is to prevent unbalanced viewpoints from dominating?

So, ultimately, who is protecting whom from what? The answer appears to be elusive. As ChatGPT has shown, enabled search engines and derivative generative AI tools are constantly being tweaked via ongoing training to correct newly discovered wrinkles viewed as vulgar, hateful, or harmful to some element of society. In the case of ChatGPT, a most interesting instance came to light when the LLM became infatuated with a correspondent after expressing some bizarre megalomaniac desires.[7] Both of these negative attributes were eventually damped through added remedial training to block certain types of provocative prompts.

IS IT EQUITABLE?

Generative AI involving LLMs is not for small fry. ChatGPT is said to engage some 10,000 Nvidia GPUs operating in tandem to train itself using its associated LLM.[8] One estimate suggests that 30,000 Nvidia A100 GPUs will be required to sustain ChatGPT in production.[9] The immense demand for electricity and cooling water is a prohibitive hurdle to any new start-ups in the field. One estimate places the cost of running GPT3, a ChatGPT predecessor, at $4 million/month.[10] Thus, an LLM, by definition, is big business, requiring a steady influx of cash to sustain operations, continue to tune the LLMs, and turn a profit. Typically, monetization is achieved through the markets for search engine optimization (SEO), advertising, and subscription services. Thus, it can be no surprise that OpenAI, the parent company of the GPT product line, has teamed up with Stripe, a leading consumer service, to offer individualized subscription services for GPT4 services.[11] Larger firms, including Microsoft, have negotiated to embed variations of GPT products into their own product lines.

Interestingly, OpenAI was started as an ethical nonprofit consortium to broker AI for the betterment of humanity. Since then, OpenAI has become a major forprofit corporation and has attracted large cash reserves through the runaway public fascination with its GPT product lines. Fortunately, it retains some of its ethical foundations, as it honestly issues warnings that its products are far from perfection and must be evaluated carefully while still sharing little about what is under the hood.

The fact that OpenAI has become big and must monetize to compete and sustain, however, suggests that sales in the SEO and advertising worlds will bring business-induced bias to the ultimate product line. This is particularly true as Microsoft partners with OpenAI for its enhanced Bing search engine, Google introduces Bard, and other high rollers join in on the gold rush spirited by the potential market strength of LLMs. While perhaps subtle, such necessary business sustainment strategies bring bias regarding who buys high-ranking advertising campaigns and who underwrites them. While these behaviors are indeed favorable to free enterprise, they can and do directly influence public behavior in both subtle and overt ways. Often, such influences bear ethical consequences, both seen and unseen. While some predict monetization realities for sustainment will tarnish the image of generative AI borne by LLMs, their full employment will already have been well established.

IT IS CAPABLE OF FREE WILL?

Consciousness and the notion of free will seem to set so-called intelligent beings apart from otherwise inanimate objects. While the GPT tools appear to exhibit astonishing degrees of literacy, syntactic excellence, graphic ability, rhyming acumen, and even coding skills, can they be truly conscious? Only able to respond to prompts, can GPT products be said to possess any degree of free will? Hence, can they be said to distinguish right from wrong in their behaviors? Are they really ethical or mere savants to the degree of ethics already well embedded within the Internet and other sources underlying their attendant LLMs?

This line of reasoning may prove helpful in how these products are eventually perceived and consumed. The products of advanced combinatorial mathematics, they are, after all, merely digital programs trained by and operating on gigantic pools of pure data. Lacking any actual knowledge, they exhibit seeming conversational skills via mathematical manipulation of symbols. While some would hold that the brain does much the same thing, brains do not require substantial mega-GPU server farms, multimegawatts of electricity, and vast coupled and dedicated storage devices to operate. Moreover, the brain works from trillions of potential links, possibly involving quantum mechanics, far exceeding any server farm in capacity. There is simply no comparison.

IS THERE A RECOURSE?

Given what is known at this early stage, it is reasonably safe to assume that the generative AI user interface (UI) must take center stage. How generative AI prompts are composed has proven to significantly affect the quality of the output from GPT products.[12] Thus, generative AI takes on a whole new utility if one assumes that creating well-founded prompts is actually a new form of high-level coding. Generative AI is liable to take hold, especially in the emerging worlds of low code–no code and learning management systems (LMSs). In both instances, effective, prompt creation builds on increasingly practical business opportunities. Given the reality of the GPT product line and its competitors, a real-world ethical challenge exists to teach people how best to interact with generative AI. That implies properly encoding generative AI through appropriately composed prompts. More importantly, it entails properly fact-checking and substantiating the results, even if pick-and-shovel techniques must be applied to bypass AI-driven search tools. This, too, will define a productive new career field for the willing, perhaps offsetting growing fears that generative AI products will blindly displace vast numbers of workers. Perhaps the human interaction with the generative AI UI is where ethical standards are best proactively applied.

This is because, like it or not, generative AI appears here to stay.

REFERENCES

1. A. Braga and R. K. Logan, "The emperor of strong AI has no clothes: Limits to artificial intelligence," *Information*, vol. 8, no. 4, Nov. 2017, Art. no. 156, doi: 10.3390/info8040156.
2. J. D. Capelouto, "Here's how GPT-4 scored on the GRE, LSAT, AP English, and other exams," *Semafor*, Mar. 2023. Accessed: Apr. 9, 2023. [Online]. Available: https://www.semafor.com/article/03/15/2023/how-gpt-4-performed-in-academic-exams
3. "Pause giant AI experiments: An open letter," Future of Life Institute, Boston, MA, USA, Mar. 2023. Accessed: Apr. 9, 2023. [Online]. Available: https://futureoflife.org/open-letter/pause-giant-ai-experiments/
4. G. Strawn, "Doing for data what the internet did for networking," *IT Prof.*, vol. 24, no. 6, pp. 66–68, Nov./Dec. 2022, doi: 10.1109/MITP.2022.3222712.
5. A. Blake, "GPT-5 could soon change the world in one incredible way," *Digit*. Trends, Mar. 2023. Accessed: Apr. 9, 2023. [Online]. Available: https://www.digitaltrends.com/computing/gpt-5-artificial-general-intelligence/
6. J. G. Asare, "The dark side of ChatGPT," *Forbes*, Jan. 2023. Accessed: Apr. 9, 2023. [Online]. Available: https://www.forbes.com/sites/janicegassam/2023/01/28/the-dark-side-of-chatgpt/?sh=4ee65e244799
7. K. Roose, "A conversation with Bing's chatbot left me deeply unsettled," *NY Times*, Feb. 2023. Accessed: Apr. 9, 2023. [Online]. Available: https://www.nytimes.com/2023/02/16/technology/bing-chatbot-microsoft-chatgpt.html
8. M. Hamblen, "Update: ChatGPT runs 10K Nvidia training GPUs with potential for thousands more," *Fierce Electron.*, Feb. 2023. Accessed: Apr. 9, 2023. [Online]. Available: https://www.fierceelectronics.com/sensors/chatgpt-runs-10k-nvidia-training-gpus-potential-thousands-more
9. Z. Liu, "ChatGPT will command more than 30,000 Nvidia GPUs: Report," *Tom's Hardware*, Mar. 2023. Accessed: Apr. 9, 2023. [Online]. Available: https://www.tomshardware.com/news/chatgpt-nvidia-30000-gpus
10. K. Leswing, "ChatGPT and generative AI are booming, but the costs can be extraordinary," *CNBC*, Mar. 2023. Accessed: Apr. 9, 2023. [Online]. Available: https://www.cnbc.com/2023/03/13/chatgpt-and-generative-ai-are-booming-but-at-a-very-expensive-price.html
11. "Stripe and OpenAI collaborate to monetize OpenAI's flagship products and enhance stripe with GPT-4," *Stripe*, Mar. 2023. Accessed: Apr. 9, 2023. [Online]. Available: https://stripe.com/newsroom/news/stripe-and-openai
12. C. Si et al., "Prompting GPT-3 to be reliable," 2022, *arXiv:2210.09150*.

GEORGE HURLBURT is the uncompensated chief scientist at the STEMCorp Foundation and serves on the Board of Advisors for the University System of Maryland at Southern Maryland. Contact him at gfhurlburt@gmail.com.

stay connected.

Join our online community! Follow us to stay connected wherever you are:

- @ComputerSociety
- facebook.com/IEEEComputerSociety
- IEEE Computer Society
- youtube.com/IEEEComputerSociety
- instagram.com/ieee_computer_society

IEEE COMPUTER SOCIETY

IEEE

IEEE COMPUTER SOCIETY D&I FUND

Drive Diversity & Inclusion in Computing

Supporting projects and programs that positively impact diversity, equity, and inclusion throughout the computing community.

DONATE TODAY!

IEEE COMPUTER SOCIETY

IEEE Foundation

IEEE COMPUTER SOCIETY

Publications Seek 2026 Editors in Chief

Application Deadline: 1 March 2025

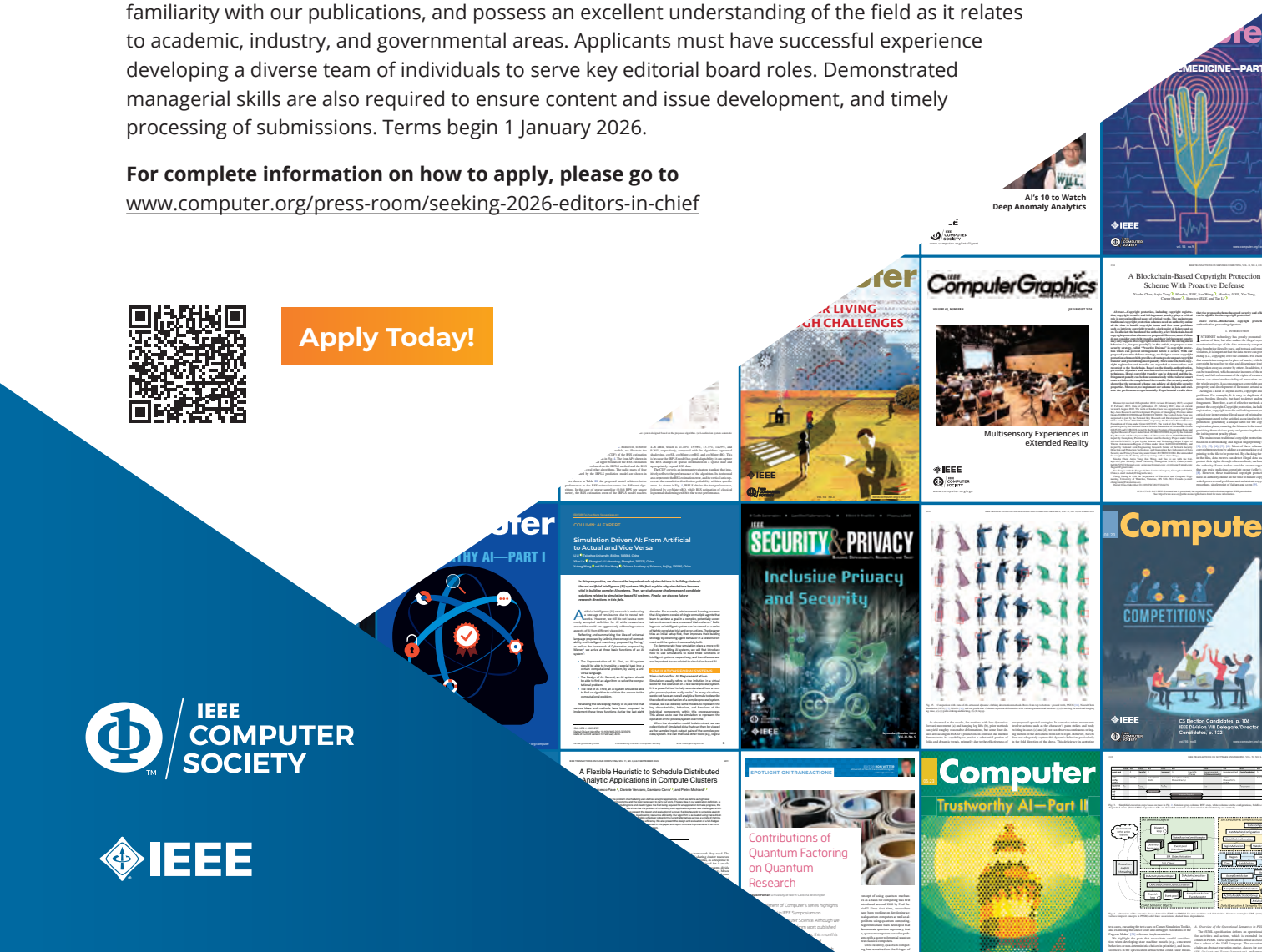
IEEE Computer Society seeks applicants for editor in chief for the following publications:

- *Computer* magazine
- *IEEE Computer Graphics and Applications*
- *IEEE Security & Privacy*
- *IEEE Transactions on Emerging Technologies in Computing*
- *IEEE Transactions on Mobile Computing*
- *IEEE Transactions on Services Computing*
- *IEEE Transactions on Software Engineering*
- *IEEE Transactions on Visualization and Computer Graphics*

Our publications are the cornerstone of professional activities for our members and the community we serve. We seek candidates who are IEEE members in good standing, have strong familiarity with our publications, and possess an excellent understanding of the field as it relates to academic, industry, and governmental areas. Applicants must have successful experience developing a diverse team of individuals to serve key editorial board roles. Demonstrated managerial skills are also required to ensure content and issue development, and timely processing of submissions. Terms begin 1 January 2026.

For complete information on how to apply, please go to
www.computer.org/press-room/seeking-2026-editors-in-chief

Apply Today!



IEEE COMPUTER SOCIETY

IEEE

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

Conference Calendar

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

DECEMBER

2 December

- SWC (IEEE Smart World Congress), Nadi, Fiji

3 December

- BIBM (IEEE Int'l Conf. on Bioinformatics and Biomedicine), Lisbon, Portugal

4 December

- ICA (IEEE Int'l Conf. on Agents), Wollongong, Australia

9 December

- CloudCom (IEEE Int'l Conf. on Cloud Computing Technology and Science), Abu Dhabi, United Arab Emirates
- ICDM (IEEE Int'l Conf. on Data Mining), Abu Dhabi, United Arab Emirate

10 December

- RTSS (IEEE Real-Time Systems Symposium), York, UK

11 December

- ICKG (IEEE Int'l Conf. on Knowledge Graph), Abu Dhabi, United Arab Emirates
- ISM (IEEE Int'l Symposium on Multimedia), Tokyo, Japan

13 December

- DependSys (IEEE Int'l Conf. on Dependability in Sensor, Cloud & Big Data Systems & Applications), Wuhan, China
- DIKW (IEEE Int'l Conf. on Data, Information, Knowledge and Wisdom), Wuhan, China
- DSS (IEEE Int'l Conf. on Data Science and Systems), Wuhan, China
- HPCC (IEEE Int'l Conf. on High Performance Computing and Communications), Wuhan, China
- ICESS (IEEE Int'l Conf. on Embedded Software and Systems), Wuhan, China
- SmartCity (IEEE Int'l Conf. on Smart City), Wuhan, China

15 December

- BigData (IEEE Int'l Conf. on Big Data), Washington, District of Columbia, USA

16 December

- ICRC (IEEE Int'l Conf. on Rebooting Computing), San Diego, USA
- iSES (IEEE Int'l Symposium on Smart Electronic Systems), New Delhi, India
- MCSoC (IEEE Int'l Symposium on Embedded Multicore/Many-core Systems-on-Chip), Kuala Lumpur, Malaysia

17 December

- ATS (IEEE Asian Test Symposium), Ahmedabad, India
- BigDataSE (IEEE Int'l Conf. on Big Data Science and Eng.), Sanya, China
- CSE (IEEE Int'l Conf. on Computational Science and Eng.), Sanya, China
- EUC (IEEE Int'l Conf. on Embedded and Ubiquitous Computing), Sanya, China
- iSCI (IEEE Int'l Conf. on Smart City and Informatization), Sanya, China
- TrustCom (IEEE Int'l Conf. on Trust, Security and Privacy in Computing and Communications), Sanya, China

18 December

- HiPC (IEEE Int'l Conf. on High Performance Computing, Data, and Analytics), Bangalore, India

19 December

- ESAI (Int'l Conf. on Embedded Systems and Artificial Intelligence), Fez, Morocco

20 December

- MSN (Int'l Conf. on Mobility, Sensing and Networking), Harbin, China

27 December

- ICVRV (Int'l Conf. on Virtual Reality and Visualization), Macao SAR, China

2025

JANUARY

15 January

- ICOIN (Int'l Conf. on Information Networking), Chiang Mai, Thailand

27 January

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

FEBRUARY

9 February

- BigComp (IEEE Int'l Conf. on Big Data and Smart Computing), Kota Kinabalu, Malaysia

17 February

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

26 February

- VISIGRAPP (Int'l Joint Conf. on Computer Vision, Imaging and Computer Graphics Theory and Applications), Porto, Portugal
- WACV (IEEE/CVF Winter Conf. on Applications of Computer Vision), Tucson, USA

MARCH

1 March

- HPCA (IEEE Int'l Symposium on High Performance Computer Architecture), Las Vegas, USA

4 March

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

8 March

- VR (IEEE Conf. Virtual Reality and 3D User Interfaces), Saint Malo, France

17 March

- PerCom (IEEE Int'l Conf. on Pervasive Computing and Communications), Washington, DC, USA

31 March

- ICSA (IEEE Int'l Conf. on Software Architecture), Odense, Denmark
- ICST (IEEE Conf. on Software Testing, Verification and Validation), Napoli, Italy

APRIL

9 April

- SaTML (IEEE Conf. on Secure and Trustworthy Machine Learning), Copenhagen, Denmark

22 April

- PacificVis (IEEE Pacific Visualization Conf.), Taipei City, Taiwan

26 April

- ICSE (IEEE/ACM Int'l Conf. on Software Eng.), Ottawa, Canada

MAY

4 May

- ARITH (IEEE Symposium on Computer Arithmetic), El Paso, USA
- FCCM (IEEE Annual Int'l Symposium on Field-Programmable Custom Computing Machines), Fayetteville, USA
- MOST (IEEE Int'l Conf. on Mobility, Operations, Services and Technologies), Newark, USA

5 May

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

11 May

- ISPASS (IEEE Int'l Symposium on Performance Analysis of Systems and Software), Ghent, Belgium

12 May

- SP (IEEE Symposium on Security and Privacy), San Francisco, USA

19 May

- CCGrid (IEEE Int'l Symposium on Cluster, Cloud and Internet Computing), Tromsø, Norway

26 May

- FG (IEEE Int'l Conf. on Automatic Face and Gesture Recognition), Tampa/Clearwater, USA

IEEE COMPUTER SOCIETY CAREER CENTER

Career Accelerating Opportunities

Explore new options—upload your resume today

careers.computer.org

Changes in the marketplace shift demands for vital skills and talent. The **IEEE Computer Society Career Center** is a valuable resource tool to keep job seekers up to date on the dynamic career opportunities offered by employers.

Take advantage of these special resources for job seekers:

- JOB ALERTS
- TEMPLATES
- WEBINARS
- CAREER ADVICE
- RESUMES VIEWED BY TOP EMPLOYERS

No matter what your career level, the IEEE Computer Society Career Center keeps you connected to workplace trends and exciting career prospects.

IEEE COMPUTER SOCIETY

IEEE