

# COMMUNICATIONS

CACM.ACM.ORG

OF THE

# ACM

12/2024 VOL.67 NO.12

## The EU AI Act and the Wager on Trustworthy AI

When Federated Learning Meets Differential Privacy

Prompting Considered Harmful

Ethics and Culture as Key Factors for an Attractive Metaverse

Association for  
Computing Machinery





**IPDPS**  
2025 • Milano, Italy

Milan, Italy • June 3-7, 2025

# 39th IEEE International Parallel and Distributed Processing Symposium

ipdps.org

## ANNOUNCING 20 WORKSHOPS IN MILAN

**IPDPS 2025** will be held for five days – from Tuesday through Saturday – and this year, all workshops will be held during the first two days of the conference. Each workshop has its own website and submission requirements, all with deadlines after December. Visit [www.ipdps.org](http://www.ipdps.org) for more information.

<b>APDCM</b>	Advances in Parallel and Distributed Computational Models
<b>AsHES</b>	Accelerators and Hybrid Emerging Systems
<b>CGRA4HPC</b>	Coarse-Grained Reconfigurable Architectures for High-Performance Computing
<b>EduPar</b>	NSF/TCPP Workshop on Parallel and Distributed Computing Education
<b>ESSA</b>	Extreme-Scale Storage and Analysis
<b>GrAPL</b>	Graphs, Architectures, Programming, and Learning
<b>HCW</b>	Heterogeneity in Computing Workshop
<b>HiCOMB</b>	High Performance Computational Biology
<b>HIPS</b>	High-level Parallel Programming Models and Supportive Environments
<b>HPAI4S*</b>	HPC for AI Foundation Models & LLMs for Science
<b>Intel4EC*</b>	Intelligent and Adaptive Edge-Cloud Operations and Services
<b>iWAPT</b>	International Workshop on Automatic Performance Tuning
<b>JSSPP</b>	Job Scheduling Strategies for Parallel Processing
<b>PAISE</b>	Parallel AI and Systems for the Edge
<b>ParSocial</b>	Parallel and Distributed Processing for Computational Social Systems
<b>PDCO</b>	Parallel / Distributed Combinatorics and Optimization
<b>PDSEC</b>	Parallel and Distributed Scientific and Engineering Computing
<b>Q-CASA</b>	Quantum Computing Algorithms, Systems, and Applications
<b>Q-SCIENCE*</b>	Advancing Scientific Computing through Quantum and HPC Synergies
<b>RAW</b>	Reconfigurable Architectures Workshop

The IPDPS workshops complement the main conference technical program of contributed papers, invited speakers, and student programs and provide the IPDPS community an opportunity to explore special topics and present work that is more preliminary. This year, in Milan, holding the workshops on two consecutive days will also offer an opportunity to integrate other events like tutorials, special interest groups, and a hot topic panel.

See [www.ipdps.org](http://www.ipdps.org).

### GENERAL CO-CHAIRS

**Marco D. Santambrogio**, Politecnico di Milano, Italy  
**Ananth Kalyanaraman**, Washington State University, USA

### PROGRAM CO-CHAIRS

**Karen Devine**, Sandia National Laboratories, ref., USA  
**Michela Becchi**, North Carolina State University, USA

### WORKSHOPS CO-CHAIRS

**Suren Byna**, The Ohio State University, USA  
**David Donofrio**, Tactical Computing Laboratories, USA  
**Giulia Guidi**, Cornell University, USA

### IMPORTANT DATES

#### Conference Preliminary Author Notification

- December 19, 2024

#### Conference Final Author Notification

- February 4, 2025

#### Workshops' Call for Papers Deadlines

- In January and February 2025

### IPDPS 2025 VENUE

IPDPS 2025 will be held on the campus of Politecnico di Milano, the largest technical university in Italy. Milan is located in Italy's northern Lombardy region and ranks as a global capital of fashion and design. It is home to the national stock exchange and is known for its high-end restaurants and shops. The Gothic Duomo di Milano cathedral and the Santa Maria delle Grazie convent, housing Leonardo da Vinci's mural "The Last Supper," testify to Milan's centuries of art and culture. All of these factors promise an exciting IPDPS 2025 week.



# NEW BOOK RELEASE



**ACM BOOKS**  
Collection III

## Pick, Click, Flick!

*The Story of  
Interaction  
Techniques*

Brad A. Myers



ASSOCIATION FOR COMPUTING MACHINERY

## Pick, Click, Flick! The Story of Interaction Techniques

**Brad A. Myers**  
*Carnegie Mellon University*

ISBN: 979-8-4007-0947-0  
DOI: 10.1145/3617448

“Every UX professional should immerse themselves in this book. Not only does it unravel the fascinating and complex history of GUI widgets that will captivate any user interface nerd, but it also stands as the definitive guide to an incredibly diverse array of interaction techniques. This is not just an engaging read; it’s an essential toolkit. By delving into these intricate details, you’re not merely learning—you’re evolving into a more refined and effective designer.” - *Jakob Nielsen, Principal, Nielsen Norman Group*

This book provides a comprehensive study of the many ways to interact with computers and computerized devices. An “interaction technique” starts when the user performs an action that causes an electronic device to respond, and includes the direct feedback from the device to the user. Examples include physical buttons and switches, on-screen menus and scrollbars operated by a mouse, touchscreen widgets, gestures such as flick-to-scroll, text entry on computers and touchscreens, input for virtual reality systems, interactions with conversational agents such as Apple Siri, Google Assistant, Amazon Alexa, and Microsoft Cortana, and adaptations of all of these for people with disabilities. *Pick, Click, Flick!* is written for anyone interested in interaction techniques, including computer scientists and designers working on human-computer interaction, as well as implementers and consumers who want to understand and get the most out of their digital devices.

<http://books.acm.org>

## News

14 **Is It Possible to Truly Understand Performance in LLMs?**

Seeking to understand when, and how, emergence takes place.

By *Samuel Greengard*

17 **AI Judging in Sports**

Sports organizations are looking to artificial intelligence to provide unbiased umpires and referees. Are they making the right call?

By *Esther Shein*

20 **A Camera the Size of an Average Grain of Salt Could Change Imaging as We Know It**

The “meta-optics” camera is 500,000 times smaller than comparable imaging devices.

By *Logan Kugler*

23 **In Memoriam**

**E. Allen Emerson**

ACM remembers A.M. Turing Award laureate Allen Emerson, who passed away on Oct. 15, 2024.

By *Simson Garfinkel and Eugene H. Spafford*

106 **Careers**

## Last Byte

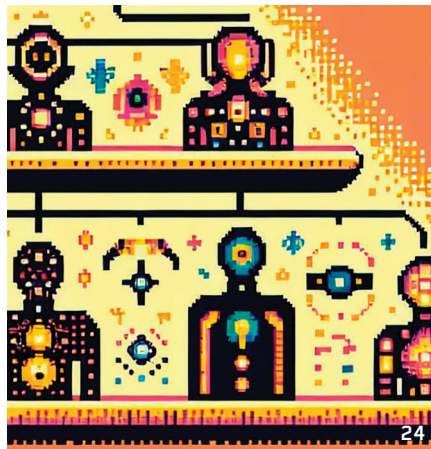
108 **Q&A**

**Personalizing Interactions**

Maja Matari discusses her career, and the surprising things that happen when humans and robots interact.

By *Leah Hoffmann*

## Opinion



5 **From the President**  
**The 5<sup>th</sup> Paradigm:**  
**AI-Driven Scientific Discovery**  
By *Yannis Ioannidis*

7 **Cerf's Up**  
**Warnings!**  
By *Vinton G. Cerf*

9 **Careers in Computing**  
**From Dot Matrix to Data: A Journey**  
**through Technology and Leadership**  
By *Wei Lu*

10 **Letters to the Editor**  
**Diversity Examples Inappropriate**

12 **BLOG@CACM**  
**Considering Conference**  
**Contributions**  
Saurabh Bagchi ponders the extent to which one should offer to work on conference program committees.

24 **The Profession of IT**  
**An AI Learning Hierarchy**  
A hierarchy of AI machines organized by their learning power shows their limits and the possibility that humans are at risk of machine subjugation well before AI utopia can come.  
By *Peter J. Denning and Ted G. Lewis*

## Opinion

28 **Opinion**  
**Prompting Considered Harmful**  
As systems graduate from labs to the open world, moving beyond prompting is central to ensuring that AI is useful, usable, and safe for end users as well as experts such as AI developers and researchers.  
By *Meredith Ringel Morris*

31 **Opinion**  
**Empower Diversity**  
**in AI Development**  
Diversity practices that mitigate social biases from creeping into your AI.  
By *Karl Werder, Lan Cao, Balasubramaniam Ramesh, and Eun Hee Park*

35 **Kode Vicious**  
**Unwanted Surprises**  
When that joke of an API is on you.  
By *George V. Neville-Neil*

37 **Privacy**  
**Notice and Choice**  
**Cannot Stand Alone**  
Privacy notice and choice has largely failed us so far because we are not giving it the legal and technical support it needs.  
By *Lorrie Faith Cranor*

40 **Opinion**  
**AI Must Be Anti-Ableist**  
**and Accessible**  
Seeking to improve AI accessibility by changing how AI-based systems are built.  
By *Jennifer Mankoff, Devva Kasnitz, L. Jean Camp, Jonathan Lazar, and Harry Hochheiser*

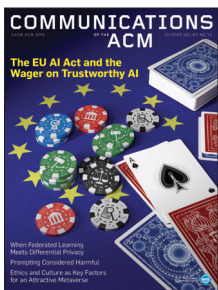
Practice



44 **Confidential Computing or Cryptographic Computing?**  
Trade-offs between secure computation via cryptography and hardware enclaves.  
*By Raluca Ada Popa*

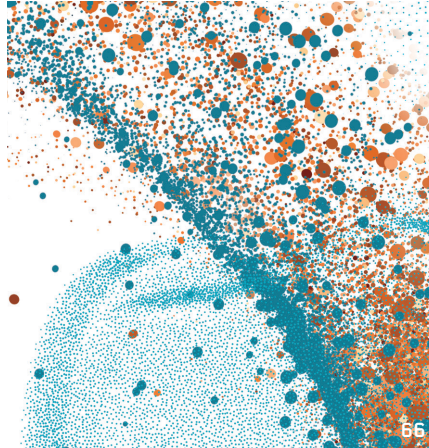
52 **Transactions and Serverless Are Made for Each Other**  
If serverless platforms could wrap functions in database transactions, they would be a good fit for database-backed applications.  
*By Qian Li and Peter Kraft*

**Q** Articles' development led by **acmqueue**  
[queue.acm.org](https://queue.acm.org)



**About the Cover:**  
As we collaborate with and delegate more tasks to AI systems, a question that will remain at the forefront of such interactions is who is responsible for decisions made by those systems. The EU AI Act, which took effect in August 2024, attempts to address that question, to help usher in a new era of trustworthy AI. In this article, the authors discuss why while trustworthy AI seems to be the safest bet, there are still questions as to how to ensure it is the best one. Cover illustration by Jacey Tec.

Research and Advances



58 **The EU AI Act and the Wager on Trustworthy AI**  
As the impact of AI is difficult to assess by a single group, policymakers should prioritize societal and environmental well-being and seek advice from interdisciplinary groups focusing on ethical aspects, responsibility, and transparency in the development of algorithms.  
*By Alejandro Bellogin, Oliver Grau, Stefan Larsson, Gerhard Schimpf, Biswa Sengupta, and Gürkan Solmaz*



Watch the authors discuss this work in the exclusive *Communications* video.  
<https://cacm.acm.org/videos/the-eu-ai-act>

66 **Belt and Braces: When Federated Learning Meets Differential Privacy**  
Building federated learning with differential privacy to train and refine ML models with more comprehensive datasets can help exploit ML's full potential.  
*By Xuebin Ren, Shusen Yang, Cong Zhao, Julie McCann, and Zongben Xu*

Research and Advances

78 **Ethics and Cultural Background as Key Factors for an Attractive Metaverse**  
The metaverse remains a work in progress, but improvements in how it handles ethical concerns and addresses cultural issues could push it further along the path to mass adoption.  
*By Tiziana Catarci, Giuseppina De Nicola, and Daniel Raffini*



Watch the authors discuss this work in the exclusive *Communications* video.  
<https://cacm.acm.org/videos/attractive-metaverse>

Research Highlights

86 **Technical Perspective**  
**How Exploits Impact Computer Science Theory**  
*By Sergey Bratus*

87 **Computing with Time: Microarchitectural Weird Machines**  
*By Thomas S. Benjamin, Jeffery A. Eitel, Jesse Ehwell, Dmitry Evtuyushkin, Abhrajit Ghosh, and Angelo Sapello*

96 **Technical Perspective**  
**Mirror, Mirror on the Wall, What Is the Best Topology of Them All?**  
*By Michela Taufer*

97 **HammingMesh: A Network Topology for Large-Scale Deep Learning**  
*By Torsten Hoefler, Tommaso Bonoto, Daniele De Sensi, Salvatore Di Girolamo, Shigang Li, Marco Heddes, Deepak Goel, Miguel Castro, and Steve Scott*



**Association for Computing Machinery**  
Advancing Computing as a Science & Profession



ACM, the world's largest educational and scientific computing society, delivers resources that advance computing as a science and profession. ACM provides the computing field's premier Digital Library and serves its members and the computing profession with leading-edge publications, conferences, and career resources.

**Executive Director and CEO**  
Vicki L. Hanson  
**Deputy Executive Director and COO**  
Patricia Ryan  
**Director, ACM Digital Library**  
Wayne Graves  
**Director, Office of Financial Services**  
James Schembari  
**Director, Office of SIG Services**  
Donna Cappel  
**Director, Office of Publications**  
Scott E. Delman

**ACM COUNCIL**  
**President**  
Yannis Ioannidis  
**Vice-President**  
Elisa Bertino  
**Secretary/Treasurer**  
Rashmi Mohan  
**Past President**  
Gabriele Kotsis  
**Chair, SGB Board**  
Jens Palsberg  
**Co-Chairs, Publications Board**  
Wendy Hall and Divesh Srivastava  
**Members-at-Large**  
Odest (Chad) Jenkins, John Kim, Tanara Lauschner, Alison Derbenwick Miller, Alejandro Saucedo  
**SGB Council Representatives**  
Jeanna Neefe Matthews and Vivek Sarkar

**BOARD CHAIRS**  
**Education Board**  
Elizabeth Hawthorne and Alison Derbenwick Miller  
**Practitioners Board**  
Terry Coatta  
**Digital Library Board**  
Jack Davidson

**TOPIC AND REGIONAL COUNCIL CHAIRS**  
**Diversity, Equity, and Inclusion Council**  
Stephanie Ludi  
**Technology Policy Council**  
Jim Hendler  
**ACM Europe Council**  
Rosa Badia  
**ACM India Council**  
Venkatesh Raman  
**ACM China Council**  
Xinbing Wang

**PUBLICATIONS BOARD**  
**Co-Chairs**  
Wendy Hall and Divesh Srivastava  
**Board Members**  
Jonathan Aldrich; Tom Crick; Jack Davidson; Mike Heroux; Michael Kirkpatrick; James Larus; Marc Najork; Beng Chin Ooi; Holly Rushmeier; Bobby Schnabel; Stuart Taylor; Bhavani Thuraisingham; Adelinde Uhrmacher; Philip Wadler

# COMMUNICATIONS OF THE ACM

Trusted insights for computing's leading professionals.

*Communications of the ACM* is the leading monthly print and online magazine for the computing and information technology fields. *Communications* is recognized as the most trusted and knowledgeable source of industry information for today's computing professional. *Communications* brings its readership in-depth coverage of emerging areas of computer science, new trends in information technology, and practical applications. Industry leaders use *Communications* as a platform to present and debate various technology implications, public policies, engineering challenges, and market trends. The prestige and unmatched reputation that *Communications of the ACM* enjoys today is built upon a 50-year commitment to high-quality editorial content and a steadfast dedication to advancing the arts, sciences, and applications of information technology.

**STAFF**  
**DIRECTOR OF PUBLICATIONS**  
Scott E. Delman  
cacm-publisher@cacm.acm.org

**Executive Editor, ACM Magazines**  
Ralph Raiola  
**Senior Editor**  
John Stanik  
**Managing Editor**  
Thomas E. Lambert  
**Senior Editor/News**  
Lawrence M. Fisher  
**Web Editor**  
David Roman  
**Editorial Assistant**  
Danbi Yu

**Art Director**  
Andrij Borys  
**Associate Art Director**  
Margaret Gray  
**Assistant Art Director**  
Mia Angelica Balaquiot  
**Production Manager**  
Bernadette Shade  
**Intellectual Property Rights Coordinator**  
Barbara Ryan  
**Advertising Sales Account Manager**  
Ilia Rodriguez

**Columnists**  
Saurabh Bagchi; Michael L. Best; Michael A. Cusumano; Peter J. Denning; Thomas Haigh; Leah Hoffmann; Mari Sako; Pamela Samuelson; Marshall Van Alstyne

**CONTACT POINTS**  
**Copyright permission**  
permissions@hq.acm.org  
**Calendar items**  
calendar@cacm.acm.org  
**Change of address**  
acmhelp@acm.org  
**Letters to the Editor**  
letters@cacm.acm.org

**REGIONAL SPECIAL SECTIONS**  
**Co-Chairs**  
Virgilio Almeida, Haibo Chen, Jakob Rehof, and P J Narayanan  
**Board Members**  
Sherif G. Aly; Panagioti Fatourou; Chris Hankin; Sue Moon; Tao Xie; Kenjiro Taura

**WEBSITE**  
<https://cacm.acm.org>

**WEB BOARD**  
**Chair**  
James Landay  
**Board Members**  
Marti Hearst; Jason I. Hong; Wendy E. MacKay

**AUTHOR GUIDELINES**  
<https://cacm.acm.org/author-guidelines/>

**ACM U.S. TECHNOLOGY POLICY OFFICE**  
Adam Eisgrau  
Director of Global Policy and Public Affairs  
1701 Pennsylvania Ave NW, Suite 200  
Washington, DC 20006 USA  
T (202) 580-6555; acmpo@acm.org

**COMPUTER SCIENCE TEACHERS ASSOCIATION**  
Jake Baskin  
Executive Director

**EDITORIAL BOARD**  
**EDITOR-IN-CHIEF**  
James Larus  
eic@cacm.acm.org  
**SENIOR EDITORS**  
Andrew A. Chien  
Moshe Y. Vardi  
**EDITORS, IN MEMORIAM SECTION**  
Simson L. Garfinkel  
Eugene H. Spafford

**NEWS**  
**Chair**  
Tom Conte  
**Board Members**  
Siobhán Clarke; Lance Fortnow; Charles L. Isbell, Jr.; Irwin King; Mei Kobayashi; Rajeev Rastogi; Vinoba Vinayagamoorthy

**OPINION**  
**Co-Chairs**  
Jeanna Neefe Matthews and Chiara Renso  
**Board Members**  
Saurabh Bagchi; Mike Best; Judith Bishop; Florence M. Chee; Danish Contractor; Lorrie Cranor; Janice Cunny; Ophir Frieder; James Grimmelmann; Mark Guzdial; Brittany Johnson; Beng Chin Ooi; Christina Pöpper; Alessandra Raffaetà; Francesca Rossi; R. Benjamin Shapiro; Len Shustek; Loren Terveen; Marshall Van Alstyne; Matt Wang; Robert West; Susan J. Winter

**Q PRACTICE**  
**Co-Chairs**  
Betsy Beyer and Ben Fried  
**Board Members**  
Peter Alvaro; Stephen Bourne; Terry Coatta; Nicole Forsgren; Camille Fournier; Chris Grier; Tom Killalea; Tom Limoncelli; Kate Matsuura; Erik Meijer; George Neville-Neil; Theo Schlossnagle; Kelly Shortridge; Phil Vachon; Jim Waldo

**RESEARCH AND ADVANCES**  
**Co-Chairs**  
m.c. schraefel and Premkumar T. Devanbu  
**Board Members**  
Indrajit Bhattacharya; Alan Bundy; Peter Buneman; Haibo Chen; Monojit Choudhury; Gerardo Con Diaz; Kathi Fisler; Nate Foster; Jane Cleland-Huang; Rebecca Isaacs; Trent Jaeger; Gal A. Kaminka; Fabio Kon; Ben C. Lee; David Lo; Renée Miller; Sarah Morris; Abhik Roychoudhury; Katie A. Siek; Daniel Susser; Charles Sutton; Thomas Zimmermann

**RESEARCH HIGHLIGHTS**  
**Co-Chairs**  
Shriram Krishnamurthi and Orna Kupferman  
**Board Members**  
Martin Abadi; Sanjeev Arora; Maria-Florina Balcan; David Brooks; Stuart K. Card; Jon Crowcroft; Lieven Eeckhout; Gernot Heiser; Takeo Igarashi; Nicole Immerlica; Srinivasan Keshav; Sven Koenig; Karen Liu; Claire Mathieu; Joanna McGrenere; Tamer Özsu; Tim Roughgarden; Guy Steele, Jr.; Wang-Chiew Tan; Robert Williamson; Andreas Zeller

**Association for Computing Machinery (ACM)**  
1601 Broadway, 10<sup>th</sup> Floor  
New York, NY 10019-7434 USA  
T (212) 869-7440; F (212) 869-0481

**ACM Copyright Notice**  
Copyright © 2024 by Association for Computing Machinery, Inc. (ACM). Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and full citation on the first page. Copyright for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, to republish, to post on servers, or to redistribute to lists, requires prior specific permission and/or fee. Request permission to publish from permissions@hq.acm.org or fax (212) 869-0481.

For other copying of articles that carry a code at the bottom of the first or last page or screen display, copying is permitted provided that the per-copy fee indicated in the code is paid through the Copyright Clearance Center; www.copyright.com.

**Subscriptions**  
An annual subscription cost is included in ACM member dues of \$99 (\$40 of which is allocated to a subscription to *Communications*); for students, cost is included in \$42 dues (\$20 of which is allocated to a *Communications* subscription). A nonmember annual subscription is \$269.

**Single Copies**  
Single copies of *Communications of the ACM* are available for purchase. Please contact acmhelp@acm.org.

**ACM ADVERTISING DEPARTMENT**  
1601 Broadway, 10<sup>th</sup> Floor  
New York, NY 10019-7434 USA  
T (212) 626-0686  
F (212) 869-0481

**Advertising Sales Account Manager**  
Ilia Rodriguez  
ilia.rodriguez@hq.acm.org

**Media Kit** acmm mediasales@acm.org

**COMMUNICATIONS OF THE ACM** (ISSN 0001-0782) is published monthly by ACM Media, 1601 Broadway, 10<sup>th</sup> Floor New York, NY 10019-7434 USA. Periodicals postage paid at New York, NY 10001, and other mailing offices.

**POSTMASTER**  
Please send address changes to *Communications of the ACM*  
1601 Broadway, 10<sup>th</sup> Floor  
New York, NY 10019-7434 USA

Printed in the USA.



Association for Computing Machinery





Yannis Ioannidis

DOI:10.1145/3702970

# The 5<sup>th</sup> Paradigm: AI-Driven Scientific Discovery

**H**OW MANY TIMES must a phenomenon occur before it graduates from a coincidence to a pattern? Usually, the answer depends on how unlikely, how far from the ordinary, and how (seemingly) inexplicable the phenomenon is. The more so, the lower the threshold.

I was very surprised (and pleased) to read of this year's winners of the Nobel Prize in Physics: John Hopfield, a professor of molecular biology and earlier of Chemistry and Biology, together with Geoffrey Hinton, a professor of Computer Science. Their affiliations name three major scientific fields, none of them being Physics. The scientific community was shocked, some physicists were upset, but for all of us in Computing the citation was thrilling: "For foundational discoveries and inventions that enable machine learning with artificial neural networks." Artificial intelligence (AI), an interdisciplinary field with Computing at its core, was the achievement being honored!

As if this was not enough, the next day came the announcement of this year's winners of the Nobel Prize in Chemistry: David Baker, a professor of Biochemistry, cited "for computational protein design", and Demis Hassabis, co-founder and CEO of Google DeepMind, together with John Jumper, researcher at Google DeepMind, were cited "for protein structure prediction," which was often elaborated in more extensive press releases as "using artificial intelligence models." Both achievements advance our understanding of proteins, a central focus of (bio-)chemistry, the field of the award. Nevertheless, at the core of both were again computing and AI.

We were at center stage of two Nobel Prize announcements! Both events were so unexpected that I believe they are not a

coincidence but form a pattern. Scientific walls and walls of prejudice are falling as Computing is becoming a fundamental discipline in the STEM family. These prizes put Computing and AI at the core of scientific discoveries, to the point that, as tools, they may be considered as worthy of recognition as the scientific results they produce. Unprecedented!

We have entered the era of the 5<sup>th</sup> Paradigm of Science and the study of nature. After the paradigms of empirical/experimental science (followed for millennia), theoretical model science (for centuries), computational science (for decades), and data-driven science as envisioned by Jim Gray (for about 15 years), the 5<sup>th</sup> paradigm has emerged: AI-driven science. It is not just the speed at which AI generates and analyzes data; even more valuable are the correlations (and sometimes the causations) AI identifies that far exceed the reach of conventional research. Computing and AI are transforming our scientific discovery processes. Science represents an amazingly exciting frontier for AI, and AI represents the most exciting new instruments in the hands of scientists.

The two Nobel Prizes also highlight the importance of interdisciplinarity between Computing and other fields. The mission of ACM calls for "... advancing the art, science, engineering, and application of computing ...". With few exceptions, the application of computing, that is, interdisciplinarity, has been outside of ACM's radar. Given the centrality of our technologies in other sciences, ACM is now establishing collaborations with prominent sister societies to serve the needs of emerging areas formed as their corresponding disciplines meet with Computing and AI. Our community should seize the moment, redirect

some of its efforts toward the uncharted territories that are opening before us, and conquer new frontiers!

Excited by the current successes of AI in scientific discovery, some push further and raise intriguing existential questions: Can we expect AI to transform from a tool to an actual researcher itself, pose and investigate scientific hypotheses, and eventually write papers about its work? If the work is reviewed by human experts who say the research is correct, novel, and interesting, why should we reject it? Eventually, could AI become a full peer of human researchers? Even if we reach that point, should a human always be in the loop, validating the scientific discovery and reviewing the AI-written paper (or its AI-written reviews)? These are questions we must debate as a community. No matter what the answers may be, even the fact that we contemplate the existence of AI scientists is fascinating and a driver of exciting research ahead.

All of us researchers in academia and industry should be thrilled and proud of how Computing and AI have risen in prominence in the eyes of the entire scientific community. A scientific revolution is happening before our eyes, powered by Computing and AI. We should join our fellow researchers in other sciences and harness the power that modern Computing and AI technologies offer to understand the secrets of nature. In parallel, we should join forces with policymakers, governments, and civil society to ensure that our discoveries will be used responsibly for the benefit of all. □

**Yannis Ioannidis**—ACM President—is professor of Informatics and Telecommunications at the National and Kapodistrian University of Athens and affiliated faculty at Athena Research Center, Greece.

© 2024 Copyright held by the owner/author(s).



# A Transformative Model for Open Access

- **Unlimited Open Access** publishing for all corresponding authors in ACM's magazines, conference proceedings and journals
- **Unlimited read access** for all authorized users to the full-text contents of the ACM Digital Library
- **Default CC-BY** author rights on all accepted research articles (multiple CC options available to choose from)

## Impact on the Research Community

- Articles receive 2-3x the number of full-text article downloads
- Articles receive up to 70% more citations
- Authors are immediately compliant with the vast majority of Public and Private Research Funder Open Access Mandates
- Authors retain the copyright of their published article

**ACM is committed to an Open Access future.**

**ACM Open is how we will get there.**



Association for  
Computing Machinery

Visit [libraries.acm.org/acmopen](https://libraries.acm.org/acmopen)  
Contact [acmopen@hq.acm.org](mailto:acmopen@hq.acm.org)





Vinton G. Cerf

DOI:10.1145/3701556

# Warnings!

**A**S I WRITE this at the end of October 2024, artificial intelligence (AI) continues to be Topic A in many discussions. So too are recommendation algorithms in social media. Misinformation and disinformation rank high across many areas of socioeconomic concerns. We are even seeing misinformation about the Federal response to severe storms interfering with our ability to render aid. Why is it that we are attracted to and respond so readily to alarming information?

I have a rather unscientific theory about this. Well, it isn't grounded in solid data, but it is a cartoon model of the way I think of the phenomenon. I think sensitivity to warnings is likely a genetic survival trait for all species, especially those with some level of cognition. I include non-human species in that category. Warning calls are common across many species. Humans have benefited from such warnings by surviving to contribute to the gene pool. Many who ignored warnings did not survive and did not contribute. Thus, when we read, see, or hear warnings, we respond almost automatically. "It's a bear! Run!" (Actually, I hear running from a bear is actually bad advice).

Social media influencers take advantage of recommendation algorithms that steer users toward perceived interests and the scale at which these systems operate. The same mechanisms that might select advertisements of interest may also steer users toward information, including warnings that appear to be of interest or concern. None of this is a new realization. My longtime friend and colleague, Peter G. Neumann, drew attention to this in a 2001 *Communications Inside Risks* column,<sup>a</sup>

which is as relevant now as it was then, maybe even more so.

This is not the first time I have written about this phenomenon. The mix of accurate and inaccurate and deliberately misleading information reinforces my belief that training in critical thinking is needed now more than ever. We rely on many more sources of information today than we have in the past, in part because virtually anyone who has access to the Internet and World Wide Web is in a position to post their views to a global audience. In the past, fewer sources might have meant that information consumers could exercise more due diligence on the sources they chose to rely on. The proliferation of sources increases the need for and utility of provenance of content and concomitant assessment of sources.


This kind of filtering is not new. We don't read every book, newspaper, or magazine; watch every movie or television show; or listen to every broadcast. We don't even pay attention to every social media site on the 'Net. We select these based on recommendations from parties we trust, often including our friends or organizations we belong to.

We could use some technical help, however, as we wrestle to assess the provenance of the information we encounter.

**We computer technologists have work to do to help our societies cope with the potentially harmful effects of media scale.**

ter. Digital signatures and reliable registration of information sources might help. Anonymous speech, while of value in some circumstances (such as whistleblowing), is generally prone to harmful abuse because the source may believe it is immune from the consequences of spreading disinformation. The problem is exacerbated by people who spread information without checking, either deliberately or out of naive belief that it is correct or relevant. Elections in this century have been affected by deliberate misinformation campaigns sourced anonymously or by parties whose identity is deliberately obscured.

I have become persuaded that identity, provenance, and accountability are our friends in this proliferated, online space. Of course, I subscribe to the idea that privacy is an important societal value but not at the expense of potential harms arising from the abuse of anonymity. The veil of anonymity may need to be pierced under the right judicial conditions. I am not in favor of so-called "backdoor" processes as they can be abused and have been in the recent past; for example, by hijacking wire-tapping provisions to gain unauthorized access to telephone conversations. I remember well the debate of the so-called "Clipper chip" in the early 1990s that would have provided "authorized parties" with the ability to decrypt content encrypted by the chip. Eventually, some unauthorized party will find a way to abuse the capability.

Plainly, we computer technologists have work to do to help our societies cope with the potentially harmful effects of media scale while protecting the provisions of the Universal Declaration of Human Rights. 

Vinton G. Cerf is vice president and Chief Internet Evangelist at Google. He served as ACM president from 2012–2014.

© 2024 Copyright held by the owner/author(s).

<sup>a</sup> <https://cacm.acm.org/opinion/inside-risks-what-to-know-about/>

# OPEN FOR SUBMISSIONS

## ACM Games: Research and Practice

Editors-in-Chief

Sebastian Deterding  
*Imperial College London, UK*

Kenny Mitchell  
*Roblox, USA/Edinburgh Napier University, UK*



### Publishing major contributions to games and playable media across disciplines, methods, and media forms

*ACM Games: Research and Practice* (GAMES) is a new quarterly, peer-reviewed, online journal published by the ACM in collaboration with ETC Press. It wants to create a reference point for the state of the art across academic research and industry practice.

Why publish with GAMES:

- Inclusive scope: Open to any form of game and playable media across disciplines and methods, including applied, designerly, integrative work
- Wide range of formats: Research articles plus systematic reviews, tutorials, datasets, case studies, dialogues, viewpoints
- Bridging academia and industry with technical blogs, a magazine website, and newsletter
- Open science and scholarship: Open to open data, replications, and (soon) registered reports
- Standard-setting: Rigorous and transparent review, strong curation, invited reviews and tutorials
- Championing new and diverse voices with a diverse editorial board, mentorship, and active invitations

For more information and to submit your work, please visit [games.acm.org](https://games.acm.org)



# CAREER PATHS IN COMPUTING

DOI:10.1145/3701267

## From Dot Matrix to Data: A Journey through Technology and Leadership

**NAME****Wei Lu****BACKGROUND****CS, Computer Graphics****CURRENT JOB TITLE/EMPLOYER****K2Data****EDUCATION****Ph.D., Computer Graphics and CAD, Tsinghua University, Beijing, China.**

**M**Y FIRST ENCOUNTER with a computer at the age of 10 was a serendipitous moment I could never have foreseen would define my career path. At a time when 99% of the Chinese population was unfamiliar with computers, I was among three fourth-grade students selected to represent my school in a local computer competition. My first lessons in computer class involved using a rudimentary computer, barely more advanced than a typewriter, to perform basic mathematical tasks and create simple drawings. The results were painstakingly printed on a dot matrix printer.

Then, the middle school I attended was designated as the Computer Olympics School of my province. It had a state-of-the-art lab equipped with the latest Apple II computers and offered computer classes to all students. It was

there I first discovered the joy of computing. I found great joy in creating and playing rudimentary computer games. Of course, we also used the computer to perform serious tasks.

Though I found computing fun, neither myself nor my family expected it to be my career. I went to university first as an economics major because my parents thought it more suitable for a woman. Imagine my surprise upon discovering “computer science” listed on my offer letter from Xi’an Jiaotong University. I started university as a reluctant computer science (CS) student, but I later realized how fortunate I was to have discovered my passion for CS. In fact, upon graduation, I decided to pursue a CS Ph.D. My mother’s first impression was “You won’t be able to find a man who dares to marry you if you pursue a Ph.D.” While disconcerting, it did not deter me. I went to Tsinghua University to pursue my Ph.D. degree in Computer Graphics and Computer-Aided Design.


My mentor, Professor Jiaguang Sun, inspired his students to excel not just in research but also in developing systems that could revolutionize people’s lives. He was also an entrepreneur; he founded a start-up developing computer-aided design software for the architecture and mechanics industries, so I had the opportunity to work there as engineering leader while pursuing my degree. The software my team built sold more than 10,000 copies in its first year. This experience helped me realize that I preferred leading a group to build software that could solve real-world problems and transform user’s lives. After I earned my Ph.D., I joined IBM.

At IBM, I learned that truly exceptional products emerge from the con-

vergence of understanding user needs and applying cutting-edge technology, as well as how to balance making technology advances with solving real-world problems. What I benefited from most was the Global Technology Outlook (GTO) study and execution. GTO is an annual report IBM Research releases to spotlight the year’s most promising CS advancements. I was a key team member of the Internet of Things (IoT) GTO study, which imagined the potential applications of a world with sensors everywhere and explored enabling technologies in that world. I served as global technical leader in the execution of IoT GTO, collaborating with IBM researchers from labs worldwide to prototype a new product for managing and analyzing data from connected devices. At the same time, we worked with IBM client teams to find pilot customers to validate if there was a market for a new product and if our technology filled the need. After three years, IoT GTO officially drew to a close; however, the process was like a mini-CEO training course: I learned about drawing up a business plan, getting sponsors, leading teams, selling products, and more.

Even today, as I lead my own start-up, K2Data, I continue practicing the principle of creating exceptional products with advanced technology to address real-world challenges.

Looking back at my career, the following tenets have guided me thus far:

- ▶ Trust your instincts rather than relying solely on logic
- ▶ The value of technology is to improve the world and people’s lives.
- ▶ Go outside your comfort zone. 

© 2024 Copyright held by the owner/author(s).

# Diversity Examples Inappropriate

**I** READ THE OPINION column “Science Needs You: Mobilizing for Diversity in Award Recognition” (*Communications*, August 2024), written by distinguished computer scientists, with mixed feelings. In 2024, the importance of diversity is well established, and it is universally recognized as a valuable goal. However, I believe that while diversity should be encouraged, it does not directly pertain to the core of computer science and technology professions.

The authors cite studies suggesting diverse groups may outperform less diverse but more talented teams. Yet, there is little convincing evidence that anything other than ability, knowledge, and communication skills drives the success of teams working in software and hardware development. In fields that are inherently merit-based, such as computer science, the focus should remain on these essential qualities.

Rather than emphasizing how many women and minority individuals have received specific awards, the focus should be placed on the outcomes achieved by professionals in science and business, regardless of demographic factors. While it is important to support underrepresented groups in pursuing careers in

**Patronizing underrepresented groups does them a disservice and may undermine both scientific progress and business innovation.**

technology, awards should be granted based on merit alone.

Patronizing underrepresented groups does them a disservice and may undermine both scientific progress and business innovation. Let us focus on creating equal opportunities and ensuring all individuals, regardless of background, can succeed on their merits.

**Leonard Gradus**, Marblehead, MA, USA

## Authors' response:

*We thank Leonard Gradus for engaging with our Opinion column. Our perspective is that it is important for organizations such as ACM to be considerate of all of their members. Not everyone starts on the same footing or has the same connections, so some deserving of recognition may not receive it if we continue as we have historically. Importantly, our aim is not to change how award winners are chosen but to increase the pool of nominees from backgrounds that may have lower chances of being considered in the first place. We see little downside to offering more choices for award committees.*

**Elizabeth Novoa-Monsalve**,

Boulder, CO, USA

**David A. Patterson**, Berkeley, CA, USA

**Stephanie Ludi**, Denton, TX, USA

**Daniel E. Acuna**, Boulder, CO, USA

## Myths Are Not Myths

I was a little mystified as to where the Opinion column “The Myth of the Coder” (*Communications*, September 2024) was going. If the endpoint were that AI won’t automate programming and that the idea has been hyped, I would agree.

However, the authors seem to deny there is a clear distinction between coder and programmer. They cite historical precedents in von Neumann and Goldstine and then Grace Hopper, but then conclude there was little historical evidence for this as common thinking. I would go with the great minds rather than what is commonly believed.

The essential difference between

coding and programming is often lost. In the early days, coding required little analytical skill, only the ability to unconsciously translate (an important facility in WWII codebreakers). It was quickly recognized this was an automatable task, ideal for computers.

Programmers could express themselves in programming languages and let a compiler code generator take care of the mechanical translation. The function of coder as a person disappeared.

Whether the authors recognize the coder/programmer distinction or not, it clearly exists, or at least between the activities of programming and coding. Many people degenerate programming to coding as something machine-oriented and cryptic. Since the 1950s, we have tried to remove that impediment to programming. Programming should be expressive, literate, and an “art” (in the Donald Knuth and Bob Barton sense). Programming is removed from the machine to abstract computation. Many of us have learned how specific computers work, later (perhaps decades) realizing that is the wrong view.

It is not the computer (they are fleeting), but the invariant subject of computation and how we think that is important. Too much teaching is about how machines work, neglecting that the lasting subject is about computation. The distinction between machine-oriented coding and problem-oriented abstract programming is clear-cut. We should also distinguish between system programming, which is platform and machine oriented as a physical resource manager, and general programming that deals in logical problem-oriented resources. Programming languages should reflect the distinction and aid and express the activity of design at different levels.

System programming, which also should not be coding, is about managing the container (memory management); most programmers should be concerned with the contents—what

## It is not the computer (they are fleeting), but the invariant subject of computation and how we think that is important.

is the information we are processing, what is its nature, and what are the abstract operations applicable to that data? System software maps logical requirements (contents) to physical resources (container).

Again, much teaching is system oriented, and higher-level thinking is missed, particularly that it is the job of system programming to provide the platforms on which general programming can take place. System “coders” became the high priests of computing with their mystical incantations holding onto their power (Turing, Backus, Barton). We need computational thinking, and should think like programmers, not coders.

But even before Barton and Backus,<sup>a</sup> Alan Turing noted in a 1947 lecture to the London Mathematical Society “The masters (programmers) are liable to get replaced because as soon as any technique becomes at all stereotyped it becomes possible to devise a system of instruction tables which will enable the electronic computer to do it for itself. It may happen, however, that the masters will refuse to do this. They may be unwilling to let their jobs be stolen from them in this way. In that case

<sup>a</sup> Bob Barton (first recipient of the ACM Eckert-Mauchly Award for hardware design) noted “Systems programmers are the high priests of a low cult.”; <https://bit.ly/48rbxjE>. John Backus also noted how the priesthood holds onto their arcane tools: see <https://bit.ly/4e4ZrthA>

they would surround the whole of their work with mystery and make excuses, couched in well-chosen gibberish, whenever any dangerous suggestions were made. I think that a reaction of this kind is a very real danger.”

**Ian Joyner**, Sydney, Australia

### Authors' response:

*It is argued in our Opinion column that, while there is a clear distinction between the activities of coding and programming in the late 1940s and early 1950s, this does not translate to a socioeconomical distinction between the professions of programmer and coder. People who programmed also coded. The claim that coding required only “the ability to unconsciously translate” underestimates this practice and we challenge any reader to try and do the exercise with an original flow diagram of von Neumann and Goldstine.*

*The fact that developing the first compilers and high-level programming notations took over a decade actually illustrates this difficulty of transitioning from a manual to a partially automated practice. It is only after the development of higher-level programming languages that people like Barton will make the argument that higher-level thinking should influence how computers are designed, but by then we are already in the 1960s.*

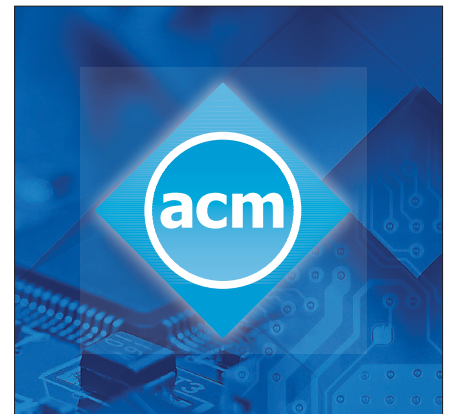
*As historians, it is our “job” not to confirm or support what the “great minds” claimed but to critically examine those claims and to follow the facts. Here, the facts show that there was no separate job for the “coder” (though hierarchies on the workflow did exist and changed through automatic programming). Today, when the high priests of Big Tech make all kinds of claims to sell their products, such critical examination is perhaps even more needed than ever before.*

**Liesbeth De Mol**, Lille, France

**Maarten Bullynck**, Saint-Denis, France

*Communications* welcomes your opinion. To contribute a letter to the editor, please limit your comments to 500 words or less and send to [letters@cacm.acm.org](mailto:letters@cacm.acm.org)

© 2024 Copyright held by the owner/author(s).



## Advertise with ACM!

Reach the innovators and thought leaders working at the cutting edge of computing and information technology through ACM’s magazines, websites and newsletters.



Request a media kit with specifications and pricing:

**Ilia Rodriguez**

+1 212-626-0686

[acmm mediasales@acm.org](mailto:acmm mediasales@acm.org)



In each issue of *Communications*, we publish selected posts or excerpts from the many blogs on our website. The views expressed by bloggers are their own and not necessarily held by *Communications* or the Association for Computing Machinery.

Read more blogs and join the discussion at <https://cacm.acm.org/blog>.

<https://cacm.acm.org/blog>

## Considering Conference Contributions

*Saurabh Bagchi ponders the extent to which one should offer their efforts to conference program committees.*



**SAURABH BAGCHI**  
**Everything You Always  
Wanted to Know  
About PCs, But Were  
Afraid to Ask**

DOI:10.1145/3695859

<https://bit.ly/4e4JoAx>

August 30, 2024

Okay, PCs in the title could be Political Correctness or Personal Computers or even Peace Corps. But it is not. It stands for **Program Committees**. As researchers, in academia or industry, we often are asked to serve on Program Committees of conferences in our fields of expertise. Serving on PCs signals one is a good citizen of our global technical village and has its own altruistic rewards. Beyond that, it has substantive and far-reaching impact on our professional careers—through building connections, getting our names out there, and learning the art and the science of getting our work published at the prestigious conferences. Here I share a few lessons I have learned serving on PCs, chairing PCs, and being part of leadership of professional organizations that run conferences (IEEE, in my case). These lessons

are meant to help us get the most out of the substantial time and effort we put into serving on PCs.

### 1. How many PCs is enough?

This is, of course, a very personal decision. For me, and for many sane colleagues of mine who run active and fair-sized research programs, we would

**Here I share a few lessons I have learned serving on PCs (Program Committees), chairing PCs, and being part of leadership of professional organizations that run conferences.**

say yes to 4–6 PCs in a year, for systems or security conferences, and 5–8 for AI/ML conferences. The AI/ML conferences tend to have much fewer number of submissions assigned to each reviewer. A good balance is needed in choosing the number of PCs to which one commits. We need to say no to some requests to serve on PCs so that we can really do justice to the ones to which we do commit. Reviewing is time-consuming, since you want to do it well.

On the other hand, too many declines means you are not playing your part in keeping your technical community vibrant. Further, that sends off the wrong vibe that you are not a good citizen of your technical community and are more of an extractor than contributor—submitting lots of papers, but not stepping up to review submissions.

### 2. Are review submission deadlines just made up?

Not at all. They are real deadlines, of course with some slack built in. The slack is usually 2–5 days, meaning if you slip by that, you do not seriously derail the subsequent steps of the review process. This means when you are on the PC, you better stick to your

end of the bargain and submit your reviews, if not before the deadline, then at least within the slack period.

I know one matter that infuriates PC chairs is when a member batches up all her reviews and submits them all together, rather than uploading when each review is done. We all may have our good reasons for batching; we want to calibrate across all our reviews, we want a certain minimum or maximum number of submissions in our pile that we want to recommend for acceptance. But this approach throws a wrench in the review process. The Chairs cannot plan for missing reviews, such as by asking others to review the ones that you are likely to blow off, or in cases of asking for tie-breaking reviews when the existing ones are deadlocked. So do stick to your review submission deadlines. We all know how to live by strict deadlines—for our own paper and proposal submissions. Just bring that same mindset to the reviewing process.

### 3. Panglossian or Downbeat: Where is the balance?

I learned this new word “*Panglossian*” recently and finally this is my chance to use it! It means excessively optimistic and is not a complementary term. It is based on the fictional Dr. Pangloss from Voltaire’s satirical novel, *Candide*.

Anyway, coming to the question at hand, of course each of us falls somewhere in the spectrum from the super-grouch to the super-optimistic. Some of us are firmly stuck to one point in this spectrum, while some calibrate this based on the quality of the submissions. All of us are influenced by other things going on in our lives when we review; reviewers, after all, are humans, only too human some would say. Knowing all this to be a pragmatic fact of life, the only exhortation I can make is to go in with as open a mind as possible when starting to review a submission and calibrate your bar for acceptance with the quality of the conference—the historical acceptance rate for the conference serves as an imperfect and yet valuable metric for quality.

So, get the most out of the substantial time and effort we put into serving on PCs. “Do not expect a submission to

**As a reviewer, do what you do when you are not anonymous: be civil in your arguments, back them up with evidence, and don’t be a flamethrower just to enjoy the light show that will ensue.**

be perfect for you to champion its acceptance.” This has been said almost universally in instructions by PC chairs, so much so that it has become akin to a mantra, but this is an often-ignored mantra. I would like to reinforce that message here. *Your* submissions may come out without any rough edges, but those of mere mortals always do. And such submissions also deserve to see the light of day.

### 4. Can I hide behind my anonymity?

One only has to look at the corrosive nature of comments sections on Internet forums to be convinced about the ill-effects of anonymity in posting something. As reviewers, our identities are always hidden from the authors (for very good reason) and to the world at large. Sometimes, we tend to lose our commonsense notion of civility or rationality due to the seeming Potteresque cloak of invisibility that this anonymity bestows upon us. But beware, this cloak of invisibility is merely an illusion. Your fellow PC members get to see when one is being uncivil or unjust in one’s reviews. The PC Chairs and sometimes the Steering Committee members get to see your behavior as PC members. And the sum total of the PC members, aggregated over a few conferences, comprises your technical community. So does what you do when you are *not* anonymous: You are civil in your arguments, backing them up with evidence, and you are not a flamethrower just to enjoy the light show

that will ensue. The “Golden Rule” is as true here as in anything else.

### 5. To ask researchers in my group to help with reviewing or not?

If you are a purist, you will not like my view here. My view is that it is a useful exercise to ask junior researchers in your group—senior Ph.D. students, post-doctoral scholars—to help you with reviewing. This serves the self-serving purpose that it reduces your own reviewing load, and then it serves the broad-minded purpose of teaching these junior researchers an essential skill. This is a good approach in my view, but with some important caveats.

First, you should discuss the review with the junior researcher and have confidence in presenting the opinion on the submission in front of the PC. Second, you should review parts of the submission, even large parts, if you find the opinion of the researcher you delegated the review to is suspect. Finally, if the submission is contentious (that is, there are good arguments both for and against it), then you should dive in and do a full-blown review yourself because you may end up being the tie-breaker.

### In Summary

To sum up, publications at our top conferences act as the paramount indicator of research excellence. The Program Committees of these conferences act as the sole gatekeepers and Program Committees are composed of us, not some omniscient entities. So there are important ground rules, often learned only through experience, that one needs to follow to make the process fulfilling for us and fair and productive for our technical communities. In this post, I have shared my opinionated view of five such ground rules. I would love to hear if you have contrarian views on them.

This post was originally published on Distant Whispers (<https://bit.ly/4hbJAjO>).

**Saurabh Bagchi** is a professor of Electrical and Computer Engineering and Computer Science at Purdue University, where he leads a university-wide center on resilience called CRISP. His research interests are in distributed systems and dependable computing, while he and his group have the most fun making and breaking large-scale usable software systems for the greater good.

© 2024 Copyright held by the owner/author(s).

## Is It Possible to Truly Understand Performance in LLMs?

*Seeking to understand when, and how, emergence takes place.*

**T**HE LIGHTNING-FAST GROWTH of large language models (LLMs) has taken the world by storm. Generative artificial intelligence (AI) is radically reshaping business, education, government, academia, and other parts of society. Yet, for all the remarkable capabilities these systems deliver—and they are clearly impressive—a major question emerges: How can data scientists measure model performance and fully understand how they gain abilities and skills?

It is far from an abstract question. Constructing high-functioning AI models hinges on critical metrics and benchmarks. These criteria, in turn, require an understanding of what constitutes correctness. As data scientists drill down into models, they soon recognize the choice of metrics and what key performance indicators they plug into a model influence outcomes. This includes everything from real-world reliability to the amount of energy and resources required to construct an LLM.

That is where a concept called *emergence* enters the picture. In LLMs, certain skills and capabilities appear or dramatically improve on larger-scale models. This process does not take



place along a predictable trend line. It's advantageous to know what this threshold for emergence is, because it is a key to building better models and allocating time, energy, and resources efficiently.

There's a catch, however. How data scientists interpret model accuracy may determine whether emergence occurs, or how and when it occurs.

“How people measure and interpret results has a significant impact on AI tooling and training,” said Sanmi Koyejo, an assistant professor of computer science at Stanford University. Recently, Koyejo and a pair of Ph.D. students embarked on a mission to better understand the somewhat-cryptic but critical factors that define emergence and effective scaling. They wanted to



know whether spikes in performance are real, or whether the measurement system creates the appearance of emergence.

The research appeared in a 2023 paper titled *Are Emergent Abilities of Large Language Models a Mirage?*<sup>a</sup> “It’s essential to build models that behave in predictable ways and understand why, when, and where we’re hitting critical mass,” added Rylan Schaeffer, who collaborated with Koyejo on the research.

### Metrics Matter

It is a widely accepted concept in the artificial intelligence space: more data leads to better models. There’s plenty of evidence to support this contention. A 2022 study, dubbed BIG-bench, revealed a surprising finding: both GPT-3 and LAMDA, two leading LLMs, struggled with basic arithmetic when given fewer parameters.<sup>b</sup> Yet, when GPT-3 hit 13 billion parameters, it could suddenly solve addition problems accurately. LAMDA demonstrated a similar breakthrough at 68 billion parameters.

This “emergent” capability occurs in several key areas: arithmetic problems, word dexterity, language translations, logical and analogical reasoning, and so-called zero-shot and few-shot learning. The latter refers to the need to fine-tune smaller LLMs on specific tasks, while larger models learn on their own. For example, Chat GPT-3 demonstrated an ability to solve a wide array of problems with little or no specific task training.<sup>c</sup>

This suggests there is a critical mass of parameters required for LLMs to grasp fundamental mathematical concepts. Yet, this sudden jump—*emergence*—remains mysterious and somewhat random. At times, advances within models take place in steady and anticipated ways; in other moments, abilities and skills suddenly leap forward for no explicable reason, other than the model has reached a certain number of parameters.

Understanding why emergence

a *Are Emergent Abilities of Large Language Models a Mirage?* <https://arxiv.org/abs/2304.15004>

b *Beyond the Imitation Game: Quantifying and extrapolating the capabilities of language models.* <https://arxiv.org/abs/2206.04615>

c *Language Models are Few-Shot Learners* <https://arxiv.org/abs/2005.14165>

**“It’s essential to build models that behave in predictable ways and understand why, when, and where we’re hitting critical mass.”**

occurs, if it occurs at all, is part of a broader desire to shine a light inside the black box of LLMs. Despite remarkable performance advances over the last couple of years, little is known about how systems “learn,” connect words and concepts, and arrive at an answer. “Assessing the intelligence and actual abilities within systems is difficult,” said Melanie Mitchell, a professor at the Santa Fe Institute. For example, “LLMs can now pass a bar exam, but they would fail at practicing law. High performance on benchmarks and real-world results are entirely different things.”

Nevertheless, understanding whether emergence is real or an artifact that results from specific measurement methods is a crucial piece of the overall puzzle. Data scientists typically rely on a straightforward method to gauge accuracy: is the information correct or not? In many instances, they assign one point for a correct answer and zero points for an incorrect answer. “On the surface, it often seems like a simple determination, but once you dive into a model, you discover things can become incredibly complicated. How you measure things determines what results you obtain?” Schaeffer asked.

Consider: if you ask a group of basketball players to shoot 100 three-point shots and track each player’s results, it’s possible to rank each player by an exact percentage. However, if you alter the measurement method—say you group players into two categories, based on whether they made 90% of their shots—perhaps one player reaches the benchmark while the other 99

fail. “Yet it’s possible that all the rest shot just under 90%. That would indicate a 1% success rate when the average score was around 88%, Schaeffer explained.

Change the measurement criteria and you change the results. For example, if you add 100 players, and three players from the second group suddenly meet the 90% threshold by each hitting only one more shot, the average percentage for the entire group of 200 will tick up by a percent. Yet, the success rate at the 90% cutoff appears to have improved by 300% (like emergence in an LLM model)—despite little or no actual improvement.

“A sharp increase may simply be an artifact of the measurement system that’s being used,” Mitchell said. “It may appear there’s a sharp spike when the real outcome is smoother and more predictable.”

When the Stanford team drilled into 29 different metrics commonly used to evaluate model performance, they found that 25 of them demonstrated no emergent properties. With the use of more refined metrics, a continuous, linear growth pattern emerged as the model grew larger. Even the other four metrics had explanations for emergence, Schaeffer said. “They’re all sharp, deforming, non-continuous metrics. So, if an error occurs because one digit is wrong, it causes the same outcome as if a billion digits are wrong.”

### Partial Credit

All of this is relevant because software engineers and data scientists lack unlimited resources to train and build LLMs. Depending on the specific model, design, and purpose, it often is necessary to condense or round off data to conserve time and resources, including the high cost of using GPUs.

There also are considerations for how the model works in the physical world. How an LLM behaves and what it does can impact economic decisions, public policies, safety, and how autonomous vehicles and other machines act and react to real-world situations and events.

A simplistic “pass or fail” approach to LLMs doesn’t cut it, the researchers argue. “Not allowing for a partial credit and not building this information into

the benchmarking framework may lead to misleading and problematic results that can undermine AI,” Koyejo said. “Emergence shouldn’t drive the way we make decision or design things.”

Adds Brando Miranda, a Stanford University Ph.D. student who also served on the research team and helped author the paper, “There’s a need to develop methods that promote greater consistency and better predictability.”

In other words, data scientists might have to rethink the fundamental definition of success—and introduce more precise metrics and measurement systems. Criteria and metrics should depend less on whether a model did the exact right thing and, instead, on how close it is to the real world truth or desired result, the researchers argue. An all-or-nothing approach may function well for an arithmetic equation such as  $1+1 = 2$  or when an LLM produces a direct word translation, such as “gato” to “cat” from Spanish to English.

The real world of AI is far more complicated, however. For example, what happens if an AI model gets 99% of an algebraic equation right, but misses a single variable or coefficient? What about an LLM that generates an excellent summary of a document, but with a single factual error? Expanding measurement criteria to how well the model predicted both right and wrong things changes the equation, Miranda noted. So, if an LLM is spitting out language translations, it isn’t only about getting the specific words right, it’s about the overall quality of the translation and how accurately it conveys the intended message.

Scoring systems and benchmarking methods deserve additional study, Mitchell said. However, getting to a higher plane may prove difficult. For one thing, human subjectivity can creep into a scoring model, particularly those with components or factors subject to interpretation. For another, “Machine learning systems can sometimes incorporate ‘shortcuts’—spurious statistical associations—to obtain high scores on benchmarks without possessing the understanding that the benchmark was supposed to measure,” she explained.

Indeed, a study conducted by researchers in Taiwan found that an AI

system that performed nearly as well as humans tapped statistical clues in the dataset to achieve random accuracy.<sup>d</sup> It performed this feat by analyzing certain keywords, such as “not,” and their position in a sentence. Once the researchers eliminated the words, performance plummeted. In the end, the high scores were merely an illusion—or artifacts based on the scoring method.

### Truth and Consequences

Some in the scientific community contend that if a system gets to the right answer, it doesn’t matter how or why. Others, such as Tianshi Li, an assistant professor in the Khoury College of Computer Sciences at Northeastern University, believe a lack of explainability and transparency in LLMs and other AI systems undermines public trust, particularly in critical areas such as data security, privacy, and public safety. “Transparency is in dire need at many layers,” she said.

Yet despite questions about unpredictability that could arise from emergent systems, the scientific community isn’t completely sold on the idea that emergence is merely a function of measurements, metrics, and scoring systems. Some data scientists argue that even with more robust tools and techniques, sudden jumps in knowledge likely will continue to occur when LLMs reach a critical size. They argue the research conducted by the Stanford team does not fully account for emergence.

The Stanford researchers concede further exploration of the topic is needed, and a deeper understanding of various factors is required. This includes studying other dimensions of model behavior, such as generalization, robustness, and interpretability. In July 2024, the trio co-authored another paper that further explored the concept of predictability in terms of downstream performance.<sup>e</sup> They found that performance typically degrades — even when a more-nuanced

multiple choice scoring system is introduced—if scoring is based only on correct answers and does not incorporate incorrect data.

A deeper understanding of LLM behavior and its real-world impacts could change the way data scientists gauge results—and build models. If performance is a result of the measurement techniques used, then it is vital to consider factors like model size and task complexity when data scientists create an LLM. With a better grasp of “sharpness,” it is possible to build better models.

On the other hand, if emergence is a real thing, there’s a need to understand how, when, and why it occurs. This could help avoid unpredictable behavior and possibly catastrophic outcomes.

“If we want to build the best possible models, we have to understand how they work and why they do the things they do,” Mitchell said. “We have to make them both robust and safe.” ■

### Further Reading

Shaeffer, R., Miranda, B., and Koyejo, S. **Are Emergent Abilities of Large Language Models a Mirage?**; [https://proceedings.neurips.cc/paper\\_files/paper/2023/file/adc98a266f45005c403b8311ca7e8bd7-Paper-Conference.pdf](https://proceedings.neurips.cc/paper_files/paper/2023/file/adc98a266f45005c403b8311ca7e8bd7-Paper-Conference.pdf)

Schaeffer, R., Schoelkopf, H., Miranda, B., Mukobi, G., Madan, V., Ibrahim, A., Bradley, H., Biderman, S., and Koyejo, S. **Why Has Predicting Downstream Capabilities of Frontier AI Models with Scale Remained Elusive?** June 6, 2024; <https://arxiv.org/abs/2406.04391>

Numerous Authors.

**Beyond the Imitation Game: Quantifying and extrapolating the capabilities of language models**, *Transactions on Machine Learning Research*, May 2022; <https://arxiv.org/abs/2206.04615>

Wei, J., Tay, Y., Bommasani, R., Raffel, C., et al **Emergent Abilities of Large Language Models**, October 26, 2022; <https://arxiv.org/abs/2206.07682>

Niven, T. and Hung-Yu, K.

**Probing Neural Network Comprehension of Natural Language Arguments** *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4658–4664, Florence, Italy, July 28 - August 2, 2019.; <https://aclanthology.org/P19-1459.pdf>

**Samuel Greengard** is an author and journalist based in West Linn, OR, USA.

© 2024 ACM 0001-0782/24/12

<sup>d</sup> Probing Neural Network Comprehension of Natural Language Arguments. <https://aclanthology.org/P19-1459.pdf>

<sup>e</sup> Why Has Predicting Downstream Capabilities of Frontier AI Models with Scale Remained Elusive? [http://rylanschaeffer.github.io/content/research/2024\\_arxiv\\_downstream\\_predictability\\_elusive/main.html](http://rylanschaeffer.github.io/content/research/2024_arxiv_downstream_predictability_elusive/main.html)

# AI Judging in Sports

*Sports organizations are looking to artificial intelligence to provide unbiased umpires and referees. Are they making the right call?*

**T**HERE ARE THINGS in life that are subjective, like beauty, taste, emotions, and feelings. However, when it comes to judging in competitive sports, decisions have become a lot more cut and dried, thanks to the use of artificial intelligence (AI) systems.

Several sports organizations have been using AI to judge certain aspects of their competitions and games for years, but as the systems become more sophisticated, more are jumping on the bandwagon.

The board of the Premier League (the highest level of the English football (soccer) league system) last April voted unanimously to introduce the use of “semi-automated offside technology.” The new system will be used by the League for the first time in the 2025 season.

“The technology will provide quicker and consistent placement of the virtual offside line, based on optical player tracking, and will produce high-quality broadcast graphics to ensure an enhanced in-stadium and broadcast experience for supporters,” the League said in a statement.

## AI and Judging’s Human Factor

The Hawk-Eye computer vision system made its tennis debut in 2003 for broadcasting purposes, but was approved in 2005 after a notorious U.S. Open Tennis match between Serena Williams and Jennifer Capriati in 2004, during which Williams was the victim of multiple bad calls in the third set and went on to lose the match.

Use of Hawk-Eye was expanded during the COVID-19 pandemic, and the 2020 U.S. Open was played without line judges on all but two of the main courts. Since Hawk-Eye has been in use, between 190 and 200 judges have been replaced, depending on the stage of the tournament, says Sean Carey, managing director of competition operations, at the U.S. Tennis Association (USTA).



The Hawk-Eye computer vision system is used by 23 of 25 global sports leagues and federations, the company says.

“The reason we bring technology in for this level of tournament—and we want to do it across every level if we could afford it—is to ensure integrity and the fairest and most even calls,” Carey said.

Hawk-Eye, which uses cameras to track the trajectory of a ball and create a three-dimensional (3D) representation of it, is now being used by 23 of

**“There is something deeply appealing about the human element [in a game] and this is why we end up with AI oversight. People like people in the mix.”**

the top 25 global sports leagues and federations, according to the company. Yet, the sentiment appears to be that AI will never fully replace human judges.

This has been the subject of much debate in Major League Baseball (MLB), a sport grounded in tradition, noted Daniel Martin, an associate professor of economics at the University of California at Santa Barbara. MLB is using Hawk-Eye to automatically monitor strike zones, and is questioning whether to get rid of umpires, given that the system is “incredibly accurate,” he said.

Yet, Martin said he does not think that will happen, because society is not ready to fully give way to machines to judge sports. “There’s something deeply appealing about the human element [in a game] and this is why we end up with AI oversight,” he said. “People like people in the mix.”

While baseball matters in an economic sense, we also have to factor in people’s emotional experience with the game, he said. Hawk-Eye is useful when calls are challenged to make the ultimate decision without human bias.

At the same time, humans need to remain in the loop to make the most of the judgment calls, Martin said.

David Almog, an economics Ph.D. candidate at the Kellogg School at Northwestern University and the lead author of the 2024 paper, “AI Oversight and Human Mistakes: Evidence from Centre Court,” said human behavior changes when technology is in play. Almog worked with Martin and others to analyze a set of Hawk-Eye data on how umpires called matches.

The researchers found that umpires’ accuracy improved, and their overall mistake rate declined by 8%, after Hawk-Eye was introduced. Yet, oddly, there were instances where tennis umpires’ mistakes increased after AI was used. This suggests to Almog and Martin that the umpires were feeling the pressure of the AI system and reacting to it, whether consciously or not.

The psychological pressures on human judges when AI systems are used is cause for concern, Almog said. The fact that Hawk-Eye can overrule a line umpire is good for the game and gives umpires the impetus to improve, which is also good.

“It’s a good thing if all you care about is improving performance,” he noted, “but if you care about [a human’s] welfare ... that’s still an open question.”

The solution, said Martin, would be to have a couple of human arbitrators, although AI serves as a neutral arbiter.

Ultimately, Martin said, sports will do what makes the most business sense. That said, he believes humans will remain in the mix “because people want to see humans,” and there is the entertainment value of seeing an umpire make a mistake. Even watching AI overrule an umpire is entertainment, Martin said.

“What we’re selling with sports is not perfection. What we’re selling is the human experience—but we want some kind of fairness.”

### **Increasing Predictions, Levels of Probability**

While the PGA Tour has been using machine learning tools for many years, the system was revamped about two years ago, said Ken Lovell, senior vice president of golf technologies. Unlike the square field used in many other sports, it has taken the PGA years to

map its golf courses in three dimensions, he said.

The real-time, proprietary ShotLink system comprises logistical information, statistical data that is translated into something humans can understand, as well as a networking layer, and sensors, which are new this year and which Lovell refers to as “the cool piece.” Three different types of sensors are embedded in the ground and cover the golf course, including “military-grade radar” that provides data about every shot.

There are also 12 cameras and a group of people who watch for shots that go outside the course parameters. The sensors and cameras provide data and are used to predict the location of a shot before it hits the ground, he said.

“It’s not just about predicting the bounce and roll out; we can tell you where [the ball] will end up with a level of probability around everything that happened in real time for every shot on the golf course,” Lovell said. “In addition, we can look at obstructions and tell you the probability of where the next shot will likely be able to be hit.”

Golf is unique in that players call their own fouls, he added. The rules of officials on the ground provide guidance, and the idea behind using AI is to “give tools to the rules officials to help them do their jobs,” he said.

Lovell said there is a lot of “killer math” going on in the background to trace a ball in real time using data from the sensors, radar, videos, and cameras that is sent into the cloud.

“Sometimes, it’s hard for players to see a shot,” and there can be a 30-yard or 40-yard discrepancy, Lovell

**“What we’re selling with sports is not perfection. What we’re selling is the human experience—but we want some kind of fairness.”**

explained. In that instance, a rules official gets involved, since “they know the rules of golf better than anybody,” he said.

Right now, the PGA is building a system that will allow rules officials to look at all of this information on a tablet. It will have a view from every camera that saw the shot, enabling an official to zoom in or pan out as much as they want, he said.

Because the golf course has been painstakingly mapped, there is information on the outlines of a water hazard line. “I can draw in space a vertical plane from the edges of that line and can tell ... exactly where the ball crossed the line,” Lovell said.

The International Gymnastics Federation (known as FIG) also is bullish on the use of AI, although the idea started as a joke, when Masanori Fujiwara, then a project leader at Fujitsu, met with Morinari Watanabe, who was head of the Japanese Gymnastics Association, and said that “In the future, maybe robots will be scoring.” Fujiwara took the joke seriously and built a prototype, which led to the development of the Judging Support System (JSS) in partnership with FIG in 2017.

JSS was used to judge the pommel horse, vault, and rings events at gymnastics’ 2019 World Championships.

Proponents of JSS say it can eliminate biases and make the sport fairer. However, as is the case with the other sports, there is also debate about whether it will take away something; in this case, the subjectiveness that factors in artistry and performance as part of a competitor’s score.

JSS has been enhanced, and now uses camera-based imaging instead of sensors. The reason, says Fujiwara, now general manager of the human digital twin business division at Fujitsu, is that “At the time, the Microsoft Kinect was in the spotlight as a skeleton recognition technology, but it had the inherent problem of not being able to achieve the 6m+ range performance required for gymnastics and other sports.”

Meanwhile, Fujitsu Laboratories was developing LiDAR (light detection and ranging) for autonomous cars and determining terrain for heavy equipment work, he said, explaining that developers thought “the Lidar could measure distances of 10m or more and

that by speeding up and improving the resolution of laser scanning technology and developing the angle of view control technology, it could be acquired as a depth image capable of detecting complex human movements in many sports, such as gymnastics.”

As development progressed, Fujitsu “began to feel Lidar’s limitations and moved on to cameras,” Fujiwara said.

The technology has been refined. In 2023, JSS was used for all 10 apparatus used at the Antwerp World Championships. Said Fujiwara, “Through the development of JSS and the actual use of JSS by FIG, we have been more confident that new value can be created by digitizing human movement.”

In the near future, Fujitsu will roll out its Human Motion Analytics platform, which uses technologies developed for JSS, including its motion constraint corrector, a correction algorithm that can significantly reduce estimation errors in posture recognition, according to Fujiwara. “Until now, posture recognition blurring has been an issue in deep learning image recognition, so the motion constraint corrector technology enables more accurate skeleton recognition,” he said.

The corrector reduces jitter by ensuring skeleton length is as constant as possible and preventing joint position and angle abnormalities. “In gymnastics (JSS), the relative position of the head, legs, and so on, determines whether the technique is completed,” Fujiwara said. “In the case of the Human Motion Analytics Platform, the aesthetic elements of human movement are also important because of their relative position, such as the position of the head and the position of the legs, which can help improve the judgment of human movement.”

Fujiwara said he expects JSS will be used in other sports because of its ability to instantaneously capture complex, high-speed movements.

### Olympians Still on the Fence

Not every sports organization has climbed on the AI bandwagon. The International Olympic Committee (IOC), which established a working group in 2023, made no mention of using AI systems to judge the various sports at the 2024 Summer Games. In an April state-

ment, the organization said it is “still at the beginning of its AI journey,” with a plan to “leverage learnings from the Olympic Games Paris 2024 and other Olympic events to identify AI solutions that will improve the operational efficiency and sustainability of future Olympic Games.”

As to whether colleges and universities will use AI to judge competitive sports, Natalie Kupperman, an assistant professor of data science at the University of Virginia, said there tends to be a trickle-down effect. “I would not be surprised if we see specific camera technologies implemented in the college atmosphere,” she said, adding that the issue is the cost of outfitting the arenas and stadiums where college sports are played.

Northwestern’s Almog said he hopes as the use of AI systems for judging sports increases, the human element will be considered. AI systems also introduce the potential for distortions, sometimes referred to as AI hallucinations, when models produce misleading results.

“If you introduce distortions, you run the chance of AI oversight not improving the way you thought it would,” Almog pointed out. Humans will rationalize the change in their behavior based on the AI mechanism in place, which may be good for them, but not necessarily good for the sports organizations that brought in the systems. “So they’re acting in different interests,” he said. “That’s the cautionary tale.”

### Further Reading

Almog, D., Gauriot, R., Page, G., and Martin, D. “AI Oversight and Human Mistakes: Evidence from Centre Court.” 2024 Working paper. Kellogg School of Management at Northwestern University; <https://arxiv.org/abs/2401.16754>

MacInnes, P. “Premier League to use semi-automated offside in hope of speeding up VAR calls.” *The Guardian (U.K.)*, April 11, 2024; <https://bit.ly/40bjEPd>

“How AI is changing gymnastics judging.” January 16, 2024. *MIT Technology Review*; <https://bit.ly/3BUdiK8>

“Tactician: an AI assistant for football tactics.” *Nature Communications*. March 2024; <https://go.nature.com/4eNkYwe>

**Esther Shein** is a freelance technology and business writer based in the Boston area.

© 2024 ACM 0001-0782/24/12

# ACM Member News

## HUMAN-CENTERED PRODUCT INNOVATION



**Jofish Kaye** is Principal Research Scientist at Wells Fargo in Menlo Park, CA. Kaye earned

his undergraduate degree in brain and cognitive sciences from the Massachusetts Institute of Technology (MIT), and a master’s degree in media arts and sciences from the MIT Media Lab. He received a Ph.D. in information science from Cornell University in 2008.

On receiving his doctorate, Kaye began working at various organizations, including Nokia, Yahoo, Mozilla, and Anthem Health, before joining the staff at Wells Fargo in June 2023.

Kaye says his research centers on human-computer interaction. “My research focuses on understanding people, what their needs and behaviors are, and using that knowledge to build new products for them.”

Determining where people need support and building tools to help them move forward has been a consistent theme throughout his career, Kaye added.

Kaye leads an internal consulting team to support innovation needs across Wells Fargo. “I am currently spending a lot of time on building tools to move peoples’ financial goals forward,” he noted.

The team also is working on a wide variety of other projects, including diversification of the types of investments customers can have in their accounts, as well as building tools to help young people save and grow their wealth.

Kaye has held leadership roles at ACM over the years. These have included serving as Conference Chair for the 2016 CHI Conference on Human Factors in Computing Systems, and serving as a Governing Board Representative for the ACM Diversity, Equity and Inclusion Council, among many other roles.

—John Delaney

# A Camera the Size of an Average Grain of Salt Could Change Imaging as We Know It

*The “meta-optics” camera is 500,000 times smaller than comparable imaging devices.*

**W**HEN IT COMES to cameras, size matters, but not in the way you think.

Any time a new smartphone is released, it is easy to drool over the latest, greatest, and biggest features that allow you to take even more stunning selfies composed of even more megapixels. However, in the world of cameras, smaller cameras could end up having a far greater impact on the world at large—and enable a ton of positive applications in society—than the next iPhone camera. Work from researchers at Princeton University and the University of Washington is pointing the way.

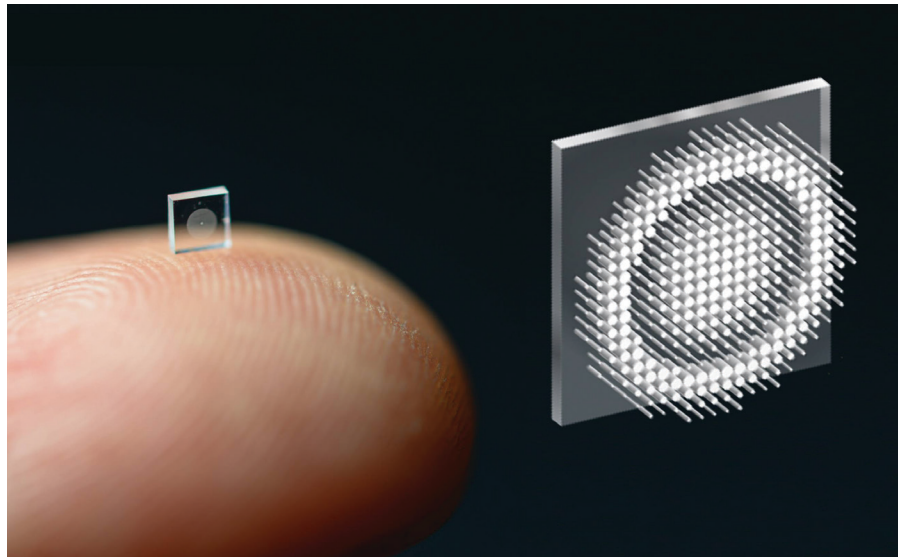
A team of researchers from both institutions has published work that uses innovative methods and materials to create a “meta-optics” camera that is the size of a single grain of salt.

The meta-optics camera is the first device of its kind to produce full-color images that are equal in quality to those produced by conventional cameras, which are an order of magnitude larger. In fact, the meta-optics camera is 500,000 times smaller than conventional cameras that capture the same level of image quality.

The approach the researchers used to create this meta-optics camera’s small form factor is a huge deal.

They used nano-structures called “metasurfaces” and novel approaches to hardware design to build a meta-optics camera superior to past efforts, as well as implementing unique AI-powered image post-processing to create high-quality images from the camera.

Their work is impressive on its own for breaking through past limitations of meta-optics imaging devices. Yet it is also notable because it opens the door to the creation of extremely



The ultracompact camera system developed by researchers at Princeton University and the University of Washington relies on a technology called a metasurface, which is studded with 1.6 million cylindrical posts and can be produced much like a computer chip.

small cameras that can create high-fidelity images for a range of industries and use-cases (for instance, by enabling the use of less-invasive medical imaging without compromising image quality).

This work also unlocks the science-fiction-like possibilities of turning

entire surfaces into cameras made up of thousands of such devices, and launching high-quality, ultra-light telescopes into space.

Here is how they did it—and why it could change the world of imaging as we know it.

## From Conventional Lenses to Metasurfaces

All camera designers and engineers, no matter the type(s) of cameras they design, share the same challenge: they want to make their cameras as compact as possible while still allowing it to record as much light as possible.

Smartphone cameras present a great example of the trade-offs inherent in solving this challenge. Each new smartphone packs more computational firepower into smaller and thinner frames, to the point where the newest generations of smartphones look positively futuristic. However, smartphone cameras are still obviously large and

**The meta-optics camera is 500,000 smaller than conventional cameras that capture the same level of image quality.**

obtrusive on otherwise sleek smartphone frames because camera designers are packing more and more lenses into them so they can take higher-quality pictures.

This means researchers are always on the hunt for ways to compress more optical power into smaller form factors, said Ethan Tseng, a researcher at Princeton who was part of the team that produced the salt-grain-sized meta-optics camera.

“Metasurfaces have emerged as a promising candidate for performing this task,” Tseng said.

A metasurface, Tseng explained, is an artificial, man-made material that allows us to affect light in unique ways. It is an ultrathin, flat surface just half a millimeter wide and studded with millions of cylindrical posts called “nano-antennas.” These nano-antennas can be individually tuned by researchers to shape light in certain ways so that, together, they are capable of producing images just like standard refractive glass lenses—but in a device that is much, much smaller.

“Using metasurfaces enables us to open a large design space of optics that we only hardly were able to access before with conventional refractive optics,” said Felix Heide, a Princeton professor who is the senior author of the study that produced the salt-grain-sized meta-optics camera.

With a standard refractive lens, you can only really shape the surface of the lens and vary the material to get better results. However, with metasurfaces, researchers are able to modulate light at the sub-wavelength level, said Heide.

In the salt-grain-sized camera, the research team was able to create a single metasurface that has more light-steering power than a traditional lens, dramatically reducing the overall size of the camera while still achieving similar results. The meta-optic lens itself is 0.5 millimeters in size, while the sensor is one millimeter in size, making the entire camera much, much smaller than traditional lenses.

The researchers did not invent the concept of using metasurfaces for cameras, but they did find a way to make the approach work in a way that was actually useful in the real world. Meta-optics cameras have been designed before, but none of them can produce

## The meta-optics camera is the first high-quality, polarization-insensitive nano-optic imager for full-color, wide field of vision imaging.

images that are of sufficient quality to deploy for imaging use cases.

“Existing approaches have been unable to design a meta-optics camera that can capture crisp, wide-field-of-view full-color images,” said Tseng.

The research team’s work changes that. Their meta-optics camera is the first high-quality, polarization-insensitive nano-optic imager for full-color, wide field-of-view imaging.

“We addressed the shortcomings of previous meta-optics imaging systems through advances in both hardware design and software post-processing,” said Tseng. To do that, the researchers used artificial intelligence to solve two challenges: lens design and image processing.

First, the team used novel AI optimization algorithms to design the nano-antennas on the actual metasurface. Simulating the optical response of a metasurface and calculating the corresponding gradients can be quite computationally expensive, Tseng said, so the team created essentially fast “proxies” for metasurface physics that allowed them to compute how to design the metasurface very quickly.

Then, a physics-based neural network was used to process the images captured by the meta-optics camera. Because the neural network was trained on metasurface physics, it can remove aberrations produced by the camera.

“We were the first to treat the metasurface as an optimizable, differentiable layer that can perform computation with light,” said Heide. “This made it possible to effectively treat metasurfaces like layers in optical neural networks

and piggyback on the large toolbox of AI to optimize these layers.”

Finally, the metasurface physics simulator and the post-processing algorithm were combined into a single pipeline to fabricate the actual meta-optic camera, and then to reconstruct the images it captures into high-quality, full-color images.

This innovative combination of hardware and software means that the researchers’ meta-optics camera produces images that could actually be used in real-world contexts, such as medical imaging.

“Only combined with computation were we able to explore this design space and make our lenses work for broadband applications,” said Heide.

### Better Endoscopes, Smartphone Cameras, Telescopes

The potential real-world applications of the research are vast.

The most obvious one is medical imaging, which directly benefits from cameras that are as small as possible so as not to be invasive. “We are very excited about miniaturized optics in endoscopes, which could allow for novel non-invasive diagnosis and surgery,” said Heide.

Ultra-compact endoscopes powered by a meta-optics camera could even image regions of the body that are difficult to reach with today’s technology.

Another major area of interest for using meta-optics cameras—or cameras that incorporate meta-optics techniques—is consumer hardware. The ability to design cameras and lenses that are an order of magnitude smaller than those in devices today opens up exciting possibilities across smartphones, wearables, and augmented reality (AR) and virtual reality (VR) headsets.

Your smartphone screen or the back of your phone itself could become a camera, says Heide. Wearables could bake high-quality cameras right into the surfaces of, say, eyeglasses. Or, VR headsets could become dramatically lighter and sleeker, leading to higher adoption and greater use of these devices on the go.

Drones also could benefit from significantly smaller cameras. All drones require cameras of some type to perform their work, whether for military purposes like reconnaissance or civil-

ian ones such as order delivery. Much smaller cameras would result in far lighter drones that consume far less battery power, said Tseng.

In fact, with a breakthrough like the meta-optics camera, the very nature of cameras can be rethought entirely.

“Our tiny cameras have also recently allowed us to rethink large cameras as flat arrays of salt-grain cameras—effectively turning surfaces into cameras,” said Heide. Larger metasurfaces could even replace the lenses needed for telescopes, making it not only easier to build them but also to send more powerful lenses into space.

While researchers are still in the early stages of brainstorming and engineering potential real-world applications for meta-optics cameras, the way in which metasurfaces are produced has them excited.

“Metasurfaces are especially interesting because they can be made using the same mature technology used to produce computer chips,” said Tseng. Today’s computer chips are produced on wafers, and each wafer contains hundreds of identical copies of the

**“Metasurfaces are especially interesting because they can be made using the same mature technology used to produce computer chips.”**

chip. Metasurfaces are produced in an identical way, which holds the promise of greatly reducing the individual cost per metasurface produced, he says.

While the exact materials used to make metasurfaces vary, the researchers used a silica wafer for their mounting surface and silicon nitride for their nano-antennas. Both materials are compatible with today’s semiconductor manufacturing techniques that pump out computer chips.

This means going from sophisti-

cated computer chips to meta-optics cameras might be easier than we think. If so, the picture for how to use these devices in many different industries could get much, much clearer. □

#### Further Reading

Shartlach, M.,  
Researchers shrink camera to the size of a salt grain, *Princeton University*, Nov. 29, 2021, <https://engineering.princeton.edu/news/2021/11/29/researchers-shrink-camera-size-salt-grain>

Smith, A.,  
Researchers develop tiny camera the size of a grain of salt - and it could turn your phone into one big cameral, *The Independent*, Dec. 9, 2021, <https://www.independent.co.uk/tech/camera-grain-salt-tiny-princeton-washington-university-b1973070.html>

Tseng, E., et al.,  
Neural nano-optics for high-quality thin lens imaging, *Nature Communications*, Nov. 29, 2021, <https://www.nature.com/articles/s41467-021-26443-0>

Logan Kugler is a freelance technology writer based in Tampa, Florida. He is a regular contributor to CACM and has written for nearly 100 major publications.

© 2024 ACM 0001-0782/24/12

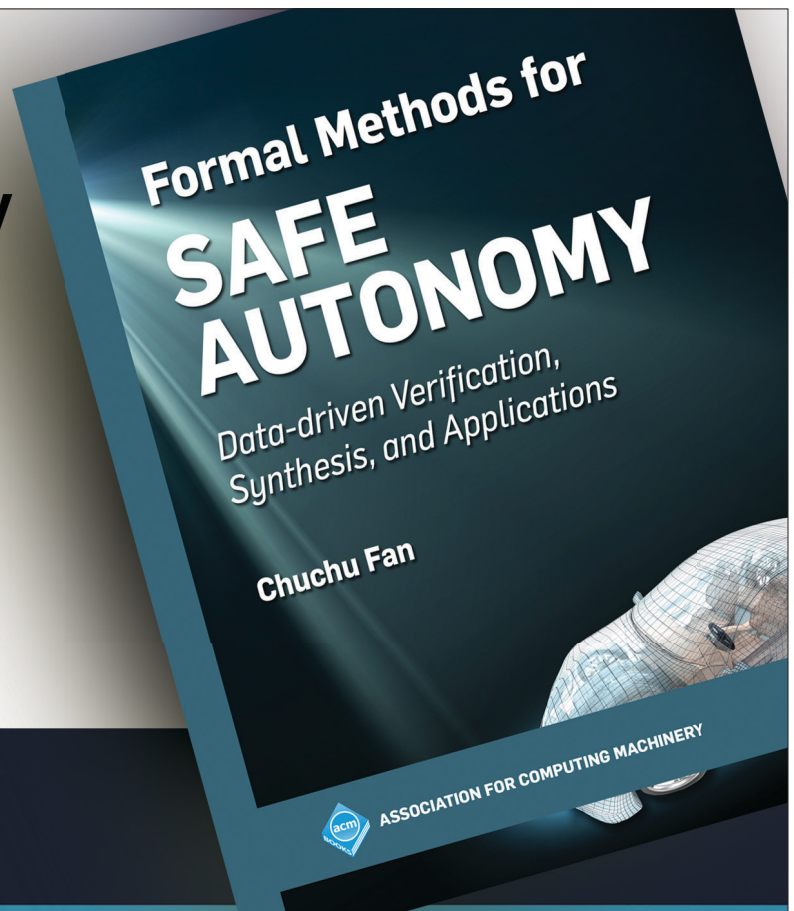
## Formal Methods for Safe Autonomy

*Data-Driven Verification  
Synthesis, and Applications*

**Chuchu Fan**  
Georgia Tech University

**ISBN: 979-8-4007-0863-3**  
**DOI: 10.1145/3603288**

<http://books.acm.org>



 **ACM BOOKS**  
Collection III



# E. Allen Emerson

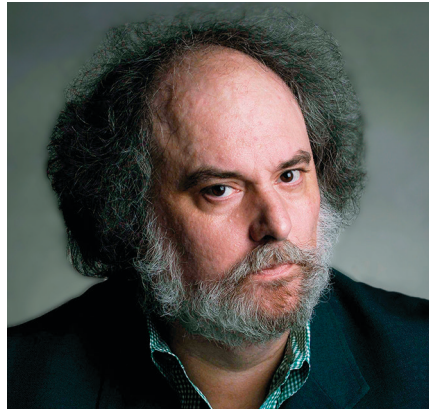
*The ACM A.M. Turing Laureate helped to develop model checking into a widely adopted, highly effective verification technology.*

**E** ALLEN EMERSON WAS the first graduate student of Edmund M. Clarke at Harvard University. After discussing several ideas for Allen's dissertation, they identified a promising candidate: verifying a finite-state system against a formal specification. According to Martha Clarke, Edmund's widow, it was during a walk across Harvard Yard that they decided to call it "model checking." Emerson received his Ph.D. in applied mathematics for this work in 1981. Twenty-five years later, he and Clarke (along with Joseph Sifakis) shared the ACM A.M. Turing Award in 2007 for this and related work.

Ernest Allen Emerson II passed away on October 15, 2024. He was born in Dallas, TX on June 2, 1954, to Ernest and Ina Lee Emerson. He graduated first in his high school class, where he learned programming on GE Mark I and Burroughs computers. He next attended the University of Texas at Austin (UT Austin), where he developed a lifelong interest in formal methods of computation. As noted in the biography accompanying his Turing Award citation, he was partly inspired by reading "Proof of a Program: FIND,"<sup>3</sup> a *Communications* paper by 1980 Turing Award recipient Tony Hoare, and by a talk from Zohar Manna on fixpoints and the Tarski-Knaster Theorem.

After obtaining his B.S. in mathematics in 1976, Emerson enrolled at Harvard for graduate study. Upon receiving his Ph.D., he joined the computer science faculty at UT Austin. Also on the UT Austin faculty was Edsger Dijkstra, who was no fan of model checking and believed programmers should reason about the correctness of their programs and not rely on mechanical program checkers. In 1985, Dijkstra analyzed a short paper Emerson published at the ACM Symposium on Programming Languages<sup>1</sup> as part of his Tuesday Afternoon Club.

"Dijkstra took extreme umbrage to the paper, and he wrote a scathing



memo criticizing me harshly," recalled Emerson in his 2019 oral history.<sup>6</sup> "It was devastating.


"But to Dijkstra's surprise, I returned the next week with a memo of my own, defending myself and counterattacking," with an argument that Dijkstra accepted. The two became close friends, with Emerson concluding that Dijkstra eventually conceded his argument regarding the benefits of model checking, telling him: "Sir, you are at risk of winning the argument." Emerson retired in 2016 as Regents Chair and professor of Computer Science, entering emeritus status.

The Turing Award citation states, "For their role in developing model checking into a highly effective verification technology that is widely adopted in the hardware and software industries." In addition, Emerson made significant contributions to temporal logic and introduced the concept of computation tree logic (CTL and CTL\*), all of which are used in verifying concurrent and real-time systems, communications protocols, and microprocessors.

Emerson advised many notable Ph.D. students. One of them, Thomas Wahl (GrammaTech, Inc.), recounted, "[he was]...high entropy-high reward in most interactions. The learning curve was steep (tough for a student spoiled with strictly organized lectures, homework, and exams)." Another former student, Charanjit Jutla (Simons Institute, UC Berkeley), noted: "He was a

logician at heart, following on the traditions of Alonzo Church to whom Allen could trace his academic legacy." Kedar Namjoshi (Nokia Bell Labs) recalls, "We'd meet to discuss research, but the conversation would soon diverge into science fiction (we both loved it; he was a particular fan of Larry Niven's *Ringworld*); rabbits (he had two as pets); and every other topic under the sun." He went on to state, "He made getting a Ph.D. so enjoyable that I would happily have done it all over again. He held us all to his own exacting standards, but he did so with gentleness, patience, trust, and humor"—a sentiment that several of his other former advisees echoed.

Emerson received other honors for his work, including the 1998 ACM Paris Kanellakis Theory and Practice Award (joint with Randal Bryant, Edmund M. Clarke, and Kenneth L. McMillan for the development of symbolic model checking. He also received the 1999 CMU Newell Research Excellence Award and the IEEE's 2006 Test of Time Award.

Emerson met his future wife, Leisa, at the public schools they attended in Dallas. They married in 1977. He is survived by his wife, his sister, and a niece and nephew and their families. His obituary notes Emerson's love of travel, books, family, and work.<sup>1,4,5,7</sup> 

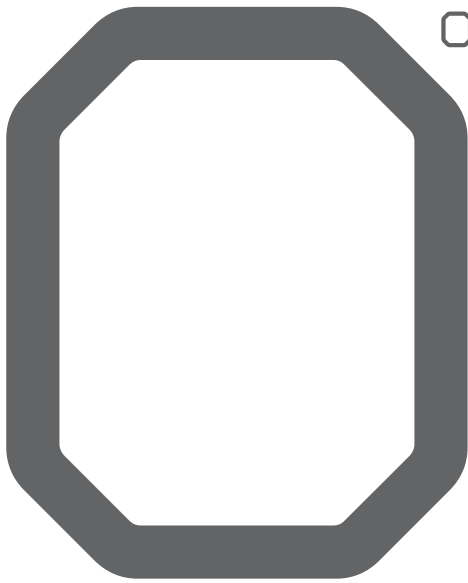
## References

- Allen Emerson biography. University of Texas at Austin; <https://bit.ly/3NK3cht>
- Emerson, E.A. and Lei, C-L. Modalities for model checking (extended abstract): Branching time strikes back. In *Proceedings of the 12<sup>th</sup> ACM SIGACT-SIGPLAN Symp. Principles of Programming Languages (1985)*, 84–96; <https://doi.org/10.1145/318593.318620>
- Hoare, C.A.R. Proof of a program: FIND. *Commun. 14*, 1 (Jan. 1971), 39–45; <https://bit.ly/3YJgam0>
- Obituary. Dignity Memorial (2024); <https://bit.ly/3YJUPs0>
- Turing Award Citation. ACM; <https://bit.ly/4eYuAEu>
- Wahl, T. A.M. Turing Award Oral History Interview with E. Allen Emerson. (Mar. 4, 2019); <https://amturing.acm.org/pdf/EmersonTuringTranscript.pdf>
- Wikipedia; <https://bit.ly/4e7lnsb>

Simson Garfinkel is an ACM Fellow.

Eugene H. Spafford is a professor of computer science and the founder and executive director emeritus of the Center for Education and Research in Information Assurances and Security (CERIAS) at Purdue University, W. Lafayette, IN, USA. He is an ACM Fellow.

© 2024 Copyright held by the owner/author(s).



# The Profession of IT

## An AI Learning Hierarchy

*A hierarchy of AI machines organized by their learning power shows their limits and the possibility that humans are at risk of machine subjugation well before AI utopia can come.*

**A**RTIFICIAL INTELLIGENCE (AI) has been successful in numerous areas including speech recognition, automatic classification, language translation, chess, Go, facial recognition, disease diagnosis, drug discovery, driverless cars, autonomous drones, and most recently linguistically competent chatbots. Yet none of these machines is the slightest bit intelligent and many of the more recent ones are untrustworthy. Businesses and governments are using AI machines in an exploding number of sensitive and critical applications without having a good grasp on when those machines can be trusted.

From its beginnings, AI as a field has been plagued with hype. Many researchers and developers were so enthusiastic about the possibilities that they overpromised what they could deliver. Disillusioned investors twice pulled back during two “AI winters.” With the arrival of large language models (LLMs), the hype has reached new heights and has driven a huge wave of speculative investment in AI companies. Investment advisors are

warning of an AI bubble. Many AI researchers have weighed in with concerns that the hype is drawing people to trust machines before we know enough about them, and to put them into critical applications where mistakes can be costly or deadly.

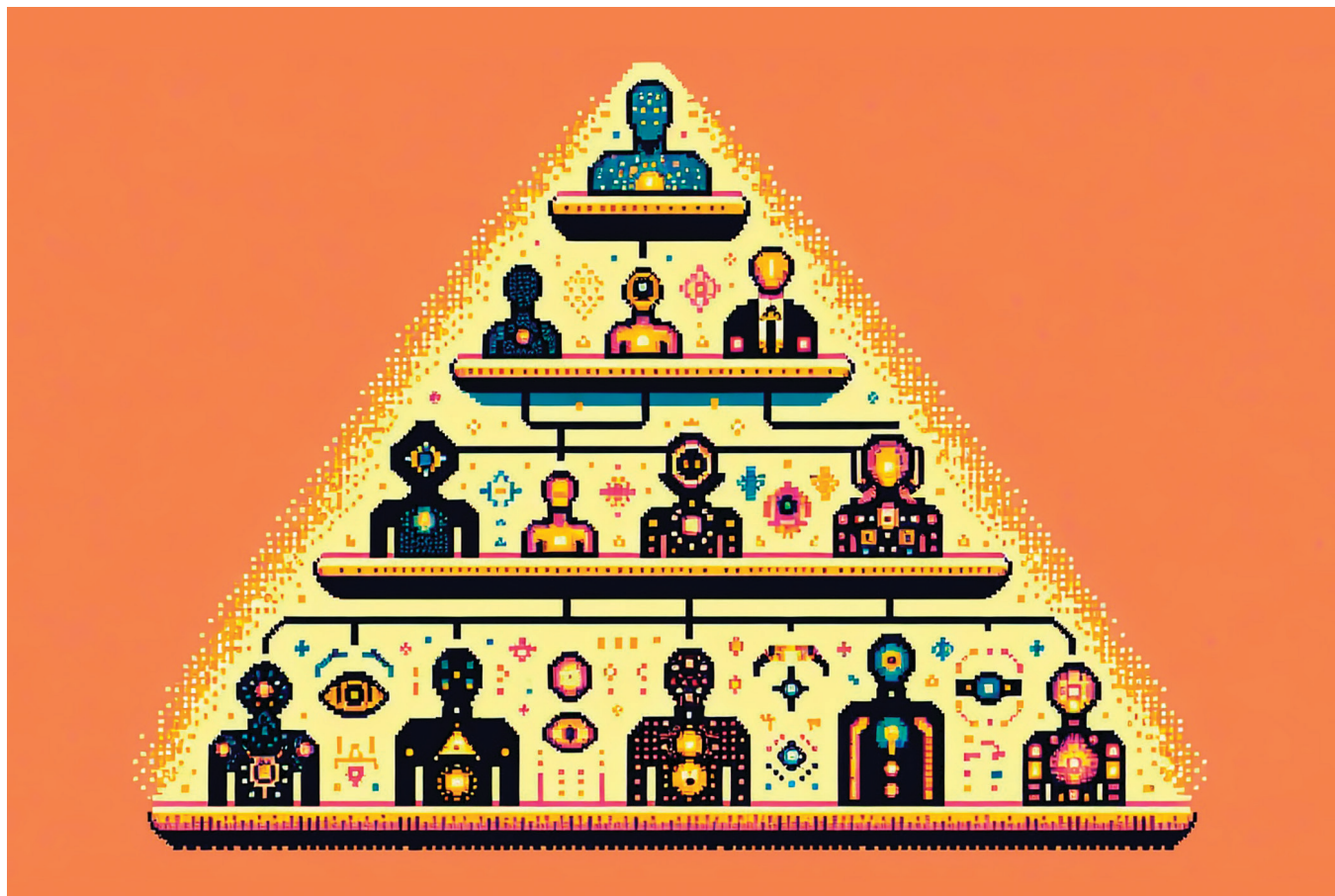
In 2019, we (the authors) proposed a way to look at AI machines that is objective enough to avoid reliance on hype and anthropomorphism.<sup>1</sup> We found that AI machines can be grouped into classes by learning power. This way of classifying AI machines gives more insight into the trust question than the more common classifications by domains including speech, vision, natural language, games, healthcare, transportation, navigation, and so on.

One particularly troubling aspect of the hype has been claims that recent advances in computing are driven by AI and that all software is a form of AI. It is the other way around: Computing has made steady progress in power and reliability over the past half-century and most software is not AI. Modern AI would not exist except for those advances.

Another troubling aspect is our tendency to anthropomorphize—to project our beliefs and hopes about human intelligence onto machines. This leads to unwelcome contradictions and misplaced trust in AI. For example, we believe intelligent people think fast, and yet supercomputers that run a billion times faster than humans are not intelligent. We believe that interacting communities of AI machines will be collectively smart, and yet massively parallel computers and networks are not intelligent. We believe chatbots will make new discoveries, but do we not accept their outputs as intelligent.

### **A Hierarchy of Learning Machines**

In Table 1, we offer an eight-tiered hierarchy that classifies AI machines by their learning power. A machine is more powerful at learning than another if, in a reasonable time, it can learn to perform some tasks that the other cannot. Learning power comes from structure. This definition does not rely on any notion of intelligence. No anthropomorphizing is needed to



explain why one machine is more powerful at learning than another.

This definition also accommodates the two basic ways machines can learn. One is by programming: A designer expresses all the rules of operation in a database and the machine applies these rules to deduce results. The other is by self-adaptation: The machine learns from examples and experience and adjusts its internal structure according to a training algorithm. These approaches can be combined, with part of an AI machine programmed and other parts self-adapting.

This hierarchy does not rank by com-

putational power. All the AI machines are Turing complete. The hierarchy shows that none of the machines built so far has any intelligence at all, leading to the intriguing possibility that human intelligence is not computable.

**Level 0: Basic automation.** These machines are automata that carry out or control processes with little or no human intervention. They frequently include simple feedback controls that maintain stable operation by adjusting and adapting to readings from sensors. For example, an FM radio locks onto a frequency but does not learn what frequencies it recognizes. However, basic automata cannot learn any new actions because their feedback does not change their function—they do not learn anything beyond what they were built to do. All the higher levels are forms of automation augmented with learning.

**Level 1: Rule-based systems.** These machines imitate the logic of human reasoning. They were called “rule-based programs” because they made their logical deductions by applying programmed logic rules to their inputs and intermediate results.

Board games were early targets for rule-based programs. In 1952, Arthur Samuel of IBM demonstrated a competent, self-improving checkers program. Beginning in 1957, a long line of chess research led to the IBM Deep Blue computer, which, in 1997 beat grandmaster Garry Kasparov. Computer speed is essential—the computer evaluates thousands of next moves in the same time a human can evaluate just one.

Expert systems were another early target—programs using logic rules derived from the knowledge of experts. Early examples were developed by Edward Feigenbaum at Stanford University in 1965: Dendral identified unknown organic molecules, and Mycin diagnosed infectious blood diseases. In 1980, John McDermott of Carnegie Mellon University built XCON, which determined the best configuration of complex DEC computer systems for a given customer.

Expert systems designers soon discovered that getting experts to state their expertise as rules is an impossible task. Hubert Dreyfus, a philosopher and an early critic of expert sys-

**Table 1. AI machines hierarchy.**

Level	Machine Category
0	Basic automation
1	Rule-based systems
2	Supervised learning
3	Unsupervised learning
4	Generative AI
5	Reinforcement learning AI
6	Human-machine interaction AI
7	Aspirational AI



STUDENT  
RESEARCH  
COMPETITION



Association for Computing Machinery  
Advancing Computing as a Science & Profession

## ACM Student Research Competition

**Attention:**  
Undergraduate *and* Graduate  
Computing Students

The ACM Student Research Competition (SRC) offers a unique forum for undergraduate and graduate students to present their original research before a panel of judges and attendees at well-known ACM-sponsored and co-sponsored conferences. The SRC is an internationally recognized venue enabling students to earn many tangible and intangible rewards from participating:

- **Awards:** cash prizes, medals, and ACM student memberships
- **Prestige:** Grand Finalists receive a monetary award and a Grand Finalist certificate that can be framed and displayed
- **Visibility:** meet with researchers in their field of interest and make important connections
- **Experience:** sharpen communication, visual, organizational, and presentation skills

Learn more:

<https://src.acm.org>

tems, argued that much of what we call expertise is not rule based: A machine limited to rule-based operations could not be expert.<sup>2</sup> Not even an enormous database of commonsense facts could make these systems as smart as experts. Many expert systems are useful despite this weakness.

**Level 2: Supervised learning.** These machines do not apply logic rules to inputs. Instead, they remember in their structure the proper output for each input shown it by a trainer. The artificial neural network (ANN) is the common example. The ANN trainer presents a long series of input-output examples; it adjusts the internal connection weights to minimize error between the actual and intended outputs. Although training may take days, a trained network responds within milliseconds.

An important property of ANNs is that any continuous mathematical function can be approximated arbitrarily closely by a sufficiently large ANN trained with a sufficient number of input-output pairs. This has inspired much research into ANNs to implement differential-equation models of physical systems, leading to many improvements in scientific computing.

In many applications, the data do not come from a continuous function—for example, facial recognition trained by labeled images. These ANNs have two main limitations: fragility and inscrutability. Fragility means that, when presented with a new (untrained) input that differs only slightly from a trained input, the network may respond with a wildly wrong output. Inscrutability means it is difficult or impossible to “explain” how the network reached its conclusion.

**Level 3: Unsupervised learning.** These machines improve their performance by making internal modifications without the assistance of an external training agent. Classifiers are the most common examples. A classifier divides the input data into the most probable set of classes by similarity; no classes are specified in advance. An early example is the AUTOCLASS program by Peter Cheeseman that classified space telescope profiles of stars.

**Level 4: Generative AI.** Machines of this level are ANNs augmented with natural-language processors. The

training process presents a large corpus of text and records which words are near to each other. When presented with an input text (“prompt”), the basic ANN produces an output word that is highly likely to be next after the input. That word is appended to the prompt and the cycle repeats, generating an output string of words that is highly probable given the original prompt. The basic model is fluent but likely to generate nonsense or fabrications not in the training data. The basic network is modified by a “tweaking process” that adjusts weights to reduce the chances of these unsatisfactory outputs.

Generative AI systems are often called large language models (LLMs) because they are trained on a very large textual training set. One of the most prominent of this genre, ChatGPT-4, was trained on several hundred billion words of texts found on the Internet; training took several months and consumed as much electricity as a small town. The results were astounding. LLMs can give astonishingly competent outputs, but they are so prone to generating fabrications and nonsense that Emily Bender in 2021 called them “stochastic parrots.” Many people do not trust them, especially when they make recommendations for action in critical areas where mistakes are costly.

There is a controversy around whether generative AI machines are creative. Skeptics point to many human creations that are not inferences from prior knowledge.

**Level 5: Reinforcement learning.** These machines avoid the need for

**Expert systems designers soon discovered that getting experts to state their expertise as rules is an impossible task.**

massive training data. Reinforcement teaches an ANN how to achieve a goal. Two ANNs play rounds of a game with each other, keeping track of which moves were ultimately part of a win and adjusting parameters so that the machines gradually learn to select only winning moves. This is done with millions or billions of rounds, simulated on an energy-gobbling supercomputer. It can produce amazing results. DeepMind's AlphaZero became a chess grandmaster in four hours and Go grandmaster in 13 days with reinforcement learning. OpenAI's ChatGPT uses reinforcement learning to make final adjustments to the weights in its core ANN so that the responses are more satisfactory to humans. Deepmind's AlphaFold was so good with predictions about protein folding that its originator received a Nobel Prize in 2024.

**Level 6: Human-machine interaction.** It is generally agreed that humans and machines blending together are more powerful than either working alone. Humans are particularly good with judgments while machines are good with computations. Achieving good blends is a very difficult problem in design.

One approach to this was popularized by Marvin Minsky in his book *Society of Mind*.<sup>5</sup> The idea is that thousands or millions of agents, each trained to be good at a narrow human skill, cooperate, and collectively generate results better than any individual human or machine). This idea permeates many proposals for achieving artificial general intelligence (AGI).

Another approach, pioneered in the 1960s by Doug Engelbart, was based on the idea of amplifying human intelligence by augmenting humans with machines. In his day, the machines were external devices using tools such as windows, mice, and hyperlinks. Today the augmentation tools are much more sophisticated and include smartphones, virtual reality glasses, and simulations. After IBM Deep Blue beat him in 1997, Garry Kasparov invented Advanced Chess, where a "player" is a team consisting of a human augmented by a computer. It was soon found that the teams of competent players and good chess programs were able to defeat the best

machines. According to futurist Ray Kurzweil, in the next decade or two augmentations may include nanobots introduced into the human bloodstream that interface with external computers and provide organ repair and enhancements like photographic memory.<sup>4</sup>

These examples show that human-machine teaming is a rich area and can often be achieved with simple interfaces that do not rely on AI tools.

**Level 7: Aspirational AI.** This level includes a variety of speculative machines that represent the dreams of many AI researchers. The most ambitious feature machines that think, reason, understand, and are self-aware, conscious, self-reflective, compassionate, and sentient. No such machines have ever been built and no one knows whether they can.<sup>3</sup>

### AI Progress Models

The AI hierarchy can be seen as a progress model. As machines gain in learning power, they approach AGI.

In *The Last AI*,<sup>7</sup> S.M. Sohn lays out a progress model depicted as a pyramid of increasing automation from AI (see Table 2).<sup>7</sup> He envisions automation making basic necessities abundant and cheap, leading eventually to 0-person organizations (no humans involved in running things) and AI utopia. While some consider this model to be preposterous, we take it seriously—as a very plausible path to a society of human subjugation by unintelligent machines.

The process seen by Sohn is already well under way at all four levels: Copilot and LLMs at Level 1, business workflow automation at Level 2, automated purchasing and customer service at Level 3, and automated bureaucracies and political deepfakes at Level 4. These systems are already distrusted because of their rigidity, fragility, lack of care, lack of compassion, and intolerance of human errors. We are drifting toward a new singularity—the subjugation of humans to networks of low-intelligence, uncaring machines—well before Kurzweil's Singularity merges humans with machines in 2045.

Inspired by Sohn, the OpenAI company promoted its own progress hierarchy, its roadmap to safe and benefi-

**Table 2. Sohn's AI adoption hierarchy.**

Level	Category of machines "in charge of"
1	Human business roles (AI copilot, AI assistant)
2	Machine business roles (AI agent, AI butler)
3	Business (AI CEO, AI company)
4	Government (AI president, AI bureaucracy, AI congress)

**Table 3. OpenAI's adoption hierarchy.**

Level	Category of machines "in charge of"
1	Chatbots (AI with conversational language)
2	Reasoners (human-level problem solving)
3	Agents (systems that can take actions)
4	Innovators (AI that aids in innovation)
5	Organizations (AI doing the work of organizations)

cial AGI (see Table 3).<sup>6</sup> It is a business plan for the new singularity! Our collective eagerness to push toward AGI may accelerate our prospect of being sucked into the quicksand of machine-orchestrated stupidity. ■

### References

- Denning, P.J. and Lewis, T.G. Intelligence might not be computable. *American Scientist* 107, (Nov.-Dec. 2019), 346–349.
- Dreyfus, H. *What machines (still) cannot do*. MIT Press, (1972; updated in 1978 and 1992).
- Koch, C. *The Feeling of Life Itself: Why Consciousness Is Widespread But Can't Be Computed*. MIT Press, (2019).
- Kurzweil, R. *The Singularity Is Nearer: When We Merge with AI*. Viking, (2024).
- Minsky, M. *The Society of Mind*. Simon and Schuster, (1986).
- Sohn, S.M. Comments on OpenAI's Adoption Model. (2024); <https://bit.ly/3XWUJWH>
- Sohn, S.M. *The Last AI: Of Humans Climbing the AI Pyramid*. SM Research Institute, (2024).

**Peter J. Denning** (pjd@nps.edu) is Distinguished Professor of Computer Science at the Naval Postgraduate School in Monterey, CA, USA, is Editor of ACM Ubiquity, and is a past president of ACM. His most recent book is *Navigating a Restless Sea: Mobilizing Innovation in Your Community* (with Todd Lyons, Waterside Productions, 2024). The author's views expressed here are not necessarily those of his employer or the U.S. federal government.

**Ted G. Lewis** (tedglewis@icloud.com) is the 2021 Oregon State Hall of Famer, author, and computer scientist with expertise in applied complexity theory, homeland security, infrastructure systems, and computer security. He is past director of the Center for Homeland Defense and Security at the Naval Postgraduate School. His most recent book is *Critical Infrastructure Resilience and Sustainability Reader* (2024).

© 2024 Copyright held by the owner/author(s).

# Opinion

## Prompting Considered Harmful

*As systems graduate from labs to the open world, moving beyond prompting is central to ensuring that AI is useful, usable, and safe for end users as well as experts such as AI developers and researchers.*

**A**S A COMPUTER scientist with one foot in artificial intelligence (AI) research and the other in human-computer interaction (HCI) research, I have become increasingly concerned that *prompting*<sup>a</sup> has transitioned from what was essentially a test and debugging interface for machine-learning (ML) engineers into the de facto interaction paradigm for end users of large language models (LLMs) and their multimodal generative AI counterparts. It is my professional opinion that prompting is a poor user interface for generative AI systems, which should be phased out as quickly as possible.

My concerns about prompting are twofold. First, prompt-based interfaces are confusing and non-optimal for

<sup>a</sup> *Prompting* refers to a pseudo-natural-language input accepted by generative AI models and applications including LLMs, LLM-powered chatbots, generative image models, generative video models, generative audio models, and multimodal generative AI apps. Prompts are typically entered as a text string, but can also be spoken as voice input to some systems. Some prompts resemble natural language closely (for example, prompting an image model with the string “Fairy tale-like mountain scenery.”) while others tend to be more arcane (for instance, prompting an image model with the string “Photo of a white fender Stratoaster :: explosion of thick fire smoke paint ink :: psychedelic style :: white background::2 -v 4 -upbeta.”). Image prompt examples from Das et al.<sup>1</sup>



end users (and ought not to be conflated with true natural-language interactions). Second, prompt-based interfaces are also risky for AI experts—we risk building a body of apps and research atop a shaky foundation of prompt engineering. I will discuss each of these issues in turn, below.

### Limitations of Prompting as an End-User Interface

Prompting is not the same as natural language. When people converse with each other, they work together to com-

municate, forming mental models of a conversation partner’s communicative intent based not only on words but also on paralinguistic and other contextual cues, theory-of-mind abilities, and by requesting clarification as needed.<sup>4</sup> By contrast, while some prompts resemble natural language, many of the most “successful” prompts do not—for instance, image generation is a domain where arcane prompts tend to produce better results than those in plain language.<sup>1</sup> Further, prompts are surprisingly sensitive to variations in

wording, spelling, and punctuation in ways that lead to substantial changes in model outputs, whereas these same permutations would be unlikely to impact human interpretation of intent—for example, jailbreak prompts using suffix attacks<sup>10</sup> or word-repetition commands.<sup>7</sup>

Indeed, the subtle differences between prompting and true natural-language interaction leads to confusion for typical end users of AI systems<sup>9</sup> and results in the need for specially trained “prompt engineers” as well as prompt marketplaces, such as PromptBase, where customers can pay money to copy prompts that purport to achieve a given result (I say “purport to achieve a given result” because the stochastic nature of generative AI models means the same input may not reliably yield the same output, an issue further exacerbated by frequent updates to underlying models). As further evidence of the challenges many end users face in crafting prompts, systems such as Dall-E 3 and Gemini sometimes rewrite users’ submitted prompts—that is, performing behind-the-scenes AI-assisted prompt engineering that may or may not be transparent to or controllable by the end user.

A few years from now, I expect we will look back on prompt-based interfaces to generative AI models as a fad of the early 2020s—a flash in the pan on the evolution toward more natural interactions with increasingly powerful AI systems. Indeed, true natural-language interfaces may be one of the desirable ways to interact with such systems, since they require no learning curve and are extremely expressive. Other high-bandwidth “natural” interfaces to AI systems might include gesture interfaces, affective interfaces (that is, mediated by emotional states), direct-manipulation interfaces (that is, directly manipulating content on a screen, in mixed reality, or in the physical world), non-invasive brain-computer interfaces (that is, thought-based interactions), or multimodal combinations of all of these.

Alternatively, there may be situations where free-form “natural” interactions are non-optimal. For example, the limitless input combinations from true natural language or other similar expressive interactions may in

## Prompting is a poor user interface for generative AI systems, which should be phased out as quickly as possible.

practice create barriers to interaction by their open-ended nature; such paradigms do not help novice end users understand the affordances of a system. Constraint-based graphical user interfaces (for example, menus, templates) might be more suitable for some user groups or application scenarios by revealing affordances, scaffolding user knowledge of available interactions, and supporting recognition over recall. It is also worth considering interaction designs that shift more of the burden for interaction onto the system rather than the end user, such as implicit interactions that infer the user’s intent from contextual clues and mixed-initiative systems that are more proactive than today’s chatbots about eliciting users’ preferences.<sup>4</sup> Modality influences “naturalness” as well; for instance, sketching or other direct-manipulation interactions<sup>2</sup> might be faster and more intuitive for generative image creation and editing than text-based prompting.

Ultimately, the goal of any AI interface is to allow the user to express their intent and to know the system understood their meaning and will carry out their intent in a safe and correct fashion. Careful consideration of appropriate human-AI interaction paradigms is an important component of a multifaceted approach to AI safety, particularly as models progress in capability.<sup>6</sup>

There is an urgent need for the fields of AI and HCI to combine their skills not only in developing improved interfaces for status quo systems, but in developing strategic programs of research on user experience for frontier models. It is also vital to innovate

in educational programs that will train a new generation of computing professionals fluent in the methods and values of both fields. We should not be complacent and assume that artificial general intelligence (AGI) will obviate the need to consider user interfaces (since such hypothetical, powerful systems would by definition understand all inputs perfectly). Progress toward AGI is a journey, not a single endpoint<sup>6</sup>—investment in user experience along the path to AGI will improve the utility of status-quo and near-term systems while also improving alignment for more speculative future models.

### Limitations of Prompting as an Expert Interface for ML Researchers and Engineers

I fear that we are in the midst of a “replication crisis” in AI research. Psychology and related social sciences have been experiencing a crisis in which a substantial number of published results do not replicate, often due to p-hacking to obtain statistically significant findings.<sup>8</sup> I am increasingly concerned that a non-negligible portion of recent AI research findings may not stand the test of time due to a different phenomenon, which I call *prompt-hacking*.

Much like the p-hacking crisis in the social sciences, prompt-hacking does not imply nefarious intent or active wrongdoing on the part of a researcher. Indeed, researchers may be entirely unaware they are engaging in this behavior. Prompt-hacking might include any of the following research practices:

- ▶ Carefully crafting dozens or even hundreds of prompts (manually, programmatically, or via generative AI tools) to obtain a desired result but not reporting in a paper the number of prompts tried that failed to produce desired results, and whether the prompt(s) that did produce desired results had any properties that systematically differentiated them from those that failed.

- ▶ Not checking whether slight variations in a successful prompt alter the research results.

- ▶ Not checking whether a prompt is robust across multiple models, multiple generations of the same model,



Association for  
Computing Machinery

2021 JOURNAL IMPACT  
FACTOR 14.324

## ACM Computing Surveys (CSUR)

ACM Computing Surveys (CSUR) publishes comprehensive, readable tutorials and survey papers that give guided tours through the literature and explain topics to those who seek to learn the basics of areas outside their specialties. These carefully planned and presented introductions are also an excellent way for professionals to develop perspectives on, and identify trends in, complex technologies.



For further information  
and to submit your  
manuscript,  
visit [csur.acm.org](https://csur.acm.org)

## A few years from now, I expect we will look back on prompt-based interfaces to generative AI models as a fad of the early 2020s.

or even the same model when repeated several times.

How can we prevent prompt-hacking and an associated AI research replication crisis? In addition to investing in developing the next generation of interfaces for generative and general AI systems, conference and journal committees could set clear standards for reporting exact prompts used, the method by which they were generated, and any prompts that were tried and discarded (and an explanation of why) as part of the methods sections of research papers. We can also seek to replicate key results to understand whether prompt-hacking is prevalent in the AI research community, and to what extent. Perhaps we even need a system for “prompt pre-registration,” analogous to the pre-registration of hypotheses that is the standard for quality social science research.

The fact that variations in prompting that would be irrelevant to a human interlocutor (for example, swapping synonyms, minor rephrasings, changes in spacing, punctuation, or spelling) result in major changes in model behavior should give us all pause,<sup>3</sup> and serve as a further reminder that prompts are still quite far from being a natural-language interface. Even research that does not engage in “prompt hacking” is still dependent on the shaky foundations of the sensitivity of models to prompts.

In addition to a replication crisis, another risk of current prompting approaches is in our methods for evaluating models. A critique of status quo evaluation of frontier models is that,

while models are ostensibly testing on the same set of benchmarks, in practice these metrics may not be comparable due to variations in how each organization operationalizes the benchmarking—that is, the format of prompts used to present the tests to the model.<sup>5</sup> This is cause for concern: Accurate measurement is key to responsibly and safely monitoring our progress toward advanced AI capabilities.<sup>6</sup>

In sum, we are at a critical juncture in AI research and development. However, our acceptance of prompting as a “good enough” simulacrum of a natural interface is hindering progress. Moving beyond prompting is vital for successful end-user adoption of AI. As systems graduate from labs to the open world, improvements in human-AI interaction paradigms are central to ensuring that AI is useful, usable, and safe. Further, moving beyond prompting (or at least openly acknowledging and compensating for its shortfalls) is vital even for experts, such as AI developers and researchers, to ensure that we can trust the results of our research and evaluations such that future systems are built upon a sturdy foundation of trustworthy knowledge. **□**

### References

1. Das, M. et al. From provenance to aberrations: Image creator and screen reader user perspectives on alt text for AI-generated images. In *Proceedings of the CHI Conf. on Human Factors in Computing Systems*, Association for Computing Machinery (2024).
2. Gholami, P. and Xiao, R. Streamlining image editing with layered diffusion brushes. *arXiv* (2024).
3. Gonen, H. et al. Demystifying prompts in language models via perplexity estimation. *Association for Computational Linguistics* (2023).
4. Kraljic, T. and Lahav, M. From prompt engineering to collaborating: A human-centered approach to ai interfaces. *Interactions* 31, 3 (May 2024), 30–35.
5. Mai, Y. and Liang, P. *Massive Multitask Language Understanding (MMLU) on HELM*. Stanford Center for Research on Foundation Models blog (2024).
6. Morris, M.R. et al. Levels of AGI for operationalizing progress on the path to AGI. *arXiv* (2024).
7. Nasr, M. et al. Scalable extraction of training data from (production) language models. *arXiv* (2023).
8. Simmons, J.P. et al. False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science* 22, 11 (2011), 1359–1366.
9. Zamfirescu-Pereira, J., Wong, R.Y., Hartmann, B., and Yang, Q. Why Johnny can't prompt: How non-AI experts try (and fail) to design LLM prompts. In *Proceedings of the 2023 CHI Conf. on Human Factors in Computing Systems*, Association for Computing Machinery.
10. Zou, A. Universal and transferable adversarial attacks on aligned language models. *arXiv* (2023).

**Meredith Ringel Morris** is director and principal scientist for Human-AI Interaction in Google DeepMind, Seattle, WA, USA.



# Opinion

## Empower Diversity in AI Development

*Diversity practices that mitigate social biases from creeping into your AI.*

**W**E SUGGEST THAT social biases are exacerbated by the lack of diversity in the artificial intelligence (AI) field.<sup>6</sup> These biases cannot be effectively addressed by technical solutions that aim at mitigating biases stemming from data sources and data processing or from the algorithm itself.<sup>9</sup> We argue that a social view—which has been neglected in AI development so far—is needed to address the root causes of some biases, given that AI systems are often reflections of our social structures. While great technical progress has been made in measuring and testing fairness<sup>4</sup> and mitigating unfairness,<sup>1</sup> biases may originate from any stage of AI development through the developers involved.<sup>6</sup> As a result, some AI system biases reflect the social biases present within the AI developers who build them. Hence, we argue that the lack of diversity in AI development is a source of social biases. As a solution, in this Opinion column we present a set of practical recommendations that empower organizations to increase diversity in AI development. In an online supplement (<https://osf.io/854ce/>), we also present prior work on AI development biases and bias mitigating and exacerbating practices.

### Lacking Diversity in AI Development: A Source of Social Biases

We argue that a lack of diversity in AI development contributes to AI system



biases in which individuals' cognitively and affectively induced biases creep into the AI system. We call these *social biases* (see the accompanying

**Some AI system biases reflect the social biases present within the AI developers who build them.**

sidebar for more information). AI developers with similar demographic backgrounds make similar (mis-)judgments, and hence, run the risk of codifying their social biases into an AI system that reinforces them. In contrast, we know that diversity is associated with positive outcomes.<sup>5</sup> For example, cross-cultural diversity and gender diversity improve requirement specification, project performance, and innovation, and they reduce biases. Without a diverse team, AI development may focus only on certain design considerations and performance measures based on narrow value judgments without considering the shared values of the broader community and diverse stakeholders.<sup>3</sup>

**Self-Designing Software**

**Considering Trauma in Accessible Design for People with Intellectual and Developmental Disabilities**

**Superpowers of Gender Equality Failing to Establish Gender Balance in IT**

**Improve CS Performance at All Levels by Developing Spatial Skills**

**On Program Synthesis and LLMs**

**Thinking of Algorithms as Institutions: A Route to Democratic Digital Governance**

**GPTs and Hallucination**

**Test Accounts: A Hidden Risk**

**OpenVPN is Open to VPN Fingerprinting**

**AAC with Automated Vocabulary from Photographs**

Plus, the latest news about the data storage crisis, AI as inventor, and the secret of Ramsey numbers.

However, benefiting from diversity is challenging because it requires the right mix of participants (for example, in hiring) and involves creating policies and procedures that help take advantage of diversity. Engaging only in shallow actions without making any meaningful changes, so-called *diversity washing*, will not address the fundamental problem and may even be counterproductive. Rather, *empowered diversity* goes beyond superficial or tokenistic efforts and encompasses a deep commitment to engaging AI developers from diverse backgrounds.

### **Empowering Diversity in AI Development**

Empowering diversity benefits all levels of an organization, that is, it positively affects developers, teams, and the organization. However, empowering diversity in practice can be challenging in AI development, particularly in science, technology, engineering, and mathematics (STEM) fields, given the limited access to and opportunities for mobility and education by marginalized groups, for example, the long-lasting shortage of women graduates. Here, we provide five practical recommendations that help organizations increase and empower diversity in AI development.

**Cultivating diversity skills.** At the individual level, managers need to equip AI developers with a strong understanding of various social biases and their impacts. AI developers first need to acquire diversity as a skill before they change their behavior. Take *confirmation biases*, for instance, when AI developers work in a male-dominated environment, they may take this as a given and focus on confirming evidence.

Managers can cultivate diversity skills by training developers to recognize and avoid this cognitive trap, ensuring they do not neglect the varied experiences of others. For example, AI developers can use specific methods such as GenderMag to identify potential biases related to gender in AI systems.<sup>2</sup> GenderMag encompasses several practices including evaluating software features for potential gender biases, creating diverse user personas to understand how different genders interact with the software, uncovering biases in task flows and interac-

## **The composition of the AI development team should mirror the system's affected stakeholders to mitigate social biases.**

tions by cognitive walkthroughs, collecting data on user demographics for inclusive design decisions, and testing with diverse groups. Using these tools helps AI developers to search for trade-off solutions that satisfy competing goals.<sup>7</sup>

In addition, managers must promote interactions between different groups within the organization by ensuring everyone has equal status, sharing common goals, fostering cooperation, and providing institutional and social support.<sup>8</sup> These intergroup contacts create a positive environment for learning from each other's experiences, which in turn develops diversity skills among the team members.

**Mirroring target stakeholders' compositions.** At the team level, the composition of the AI development team should mirror the system's affected stakeholders to mitigate social biases. Take *bounded awareness*, for instance, when the team lacks diversity in skin color, they may overlook the effects of facial recognition AI on different skin tones.

Managers can adjust HR practices to mirror the composition of target stakeholders. For instance, in diverse hiring, implementing blind hiring techniques helps counter bias influenced by bounded awareness. By combining diversity reporting with well-crafted diversity performance indicators, managers can measure the effectiveness of updated HR practices and demonstrate progress. This approach also raises awareness of potential implicit biases that are harder to crack.

Mirroring the target stakeholders' composition not only improves team behavior by managing team abrasion and

mitigating groupthink but also fosters innovation and creativity. Managers can use tools from psychology, for example, the Hermann Brain Dominance Instrument (see <https://www.thinkherrmann.com/hbdi>), to identify complementary profiles and modes of thinking. Diversity of thoughts and viewpoints improves collective creativity and helps the team become more innovative.

**Promoting inclusive knowledge sharing through experiences.** At the organizational level, managers should be aware that lacking diversity yields unfavorable outcomes, whereas empowering diversity creates positive and impactful results. Take *availability bias*, for instance, organizations should encourage sharing both success stories and failures within and outside the organization because it helps prevent narrow and one-sided views of the workforce due to availability bias.

When managers facilitate knowledge exchange within their teams, they promote organizational understanding through shared experiences. For example, acknowledging negative experiences such as gender discrimination reported by many women in AI and software development can raise awareness about existing biases and their impact on individuals' career progression. Managers can further this by launching awareness campaigns and providing internal workshops, for example on emotional intelligence, to mitigate dominance and foster empathy within the organization.

Managers should also allocate at least an equal number of resources to facilitate the sharing of positive experiences. Identifying and cultivating success stories for their diverse audience showcases possibilities that are otherwise unrecognized. Managers should promote marginalized groups within the organization, because internal success stories are especially impactful. Speakers who serve as role models for meaningful change may offer insights into organizational processes and practices from marginalized perspectives. When internal success stories are limited, managers can also engage external speakers to share their perspectives and experiences.

**Fostering long-term sustainability in a diverse talent pipeline.** One important challenge for organizations

## Description of Social Biases with Examples

Social biases include availability biases, confirmation biases, bounded awareness, and affective and emotional biases. We describe each type here.

**AI developers with *availability biases***—that is, judging events and their frequency differently based on vividness, recency, or memory structure—often ‘go where the data is’ by focusing on datasets that are available or accessible rather than datasets that are most suitable. As a result, the data used are not fully representative of the target population and can differ significantly from reality. Take facial recognition of people with dark skin, for example, which is less accurate because camera technology provides lower-quality images caused by limitations in lighting and contrast.

***Confirmation biases***—that is, seeking confirmatory information, interpreting newer information from an anchor—lead AI developers to seek confirmatory information and interpreting newer information from an anchor (that is, the first piece of information given about a topic), and be overconfident in their own judgments' infallibility. Two exemplary biases from the development of an AI system for human resource management suggest that recruiters favor candidates who are like them in age, race, and attitudinal characteristics, and recruiters assess candidates based on a particular group or class of people, such as prior employers.

**AI developers with *bounded awareness***—that is, relying overly on irrelevant information in specific conditions and failing to see the obvious—ignore important, accessible, and perceivable information during decision making because of information selection and inattentive blindness. For example, AI developers often fail to select training images and facial features that are not from their own race, which is referred to as the “other-race effect.” This may explain why facial recognition algorithms developed in China, Japan, and South Korea recognize Asian faces more accurately than Caucasian faces, and vice versa.

***Affective and emotional biases***—that is, relying on emotions rather than rational evaluation—impact AI developers' decision making and productivity, as they rely on emotions rather than cognition. Take empathy, for example, a core technique for human-centered design and design thinking that emphasizes understanding users' situations, feelings, and needs. However, it is virtually impossible for AI developers to effectively empathize with minorities, thus, giving rise to potential biases.

that want to empower diversity is the availability of talent within their environment. Therefore, managers must develop a sustainable talent pipeline that speaks to their diversity needs. Considering the lack of availability of female candidates as an example, given the pervasive nature of *availability bias*, which has raised concerns shared

**Managers should know that lacking diversity leads to unfavorable outcomes, whereas empowering diversity creates positive and impactful results.**

across different areas in the STEM field, developing suitable candidates requires sincere collective efforts with a long-term goal in mind.

Organizations should engage with their environment when developing a diverse workforce. Given the severity and longevity of the problem, organizations need to explore new ways of encouraging underrepresented groups to take on roles in AI development. The fact that established means through targeted online job advertisements are already biased only exacerbates the problem. While talent pools are limited and sometimes hard to access, organizations are encouraged to go where diversity is, that is, institutions of higher education. Organizations that collaborate more closely with educational institutions that serve a diverse population find it easier to develop a sustainable pipeline of diverse talent. For example, organizations often inform and sometimes co-develop educational curricula through their business needs as part of employer panels (for example,

## INTERACTIONS



ACM's *Interactions* magazine explores critical relationships between people and technology, showcasing emerging innovations and industry leaders from around the world across important applications of design thinking and the broadening field of interaction design.

Our readers represent a growing community of practice that is of increasing and vital global importance.



To learn more about us, visit our award-winning website <http://interactions.acm.org>

Follow us on Facebook and Twitter  

To subscribe: <http://www.acm.org/subscribe>




<https://bit.ly/400eeXu>). Communicating diversity as an important business need fuses diversity into the curricula development process.

In addition, organizations can engage in events that foster the growth and advancement of women in technology. For example, the Grace Hopper Celebration of Women in Computing focuses on empowering women to learn new skills, make connections, discuss innovative trends, and access motivational leaders.

While engagements with the environment can take some time to develop and evolve, organizations also must harness and build on what they already have, for example, by offering opportunities for career advancement to retain diverse talent. Thus, organizations must develop clear career paths for progression and promotion from within. Developing a positive culture for often underrepresented social groups positions the organization as an attractive employer. This signals that the organization does not see diversity as a goal in itself, but rather as a tool for accomplishing and delivering organizational objectives.

**Establishing a diversity charter for AI development.** Managers should develop an active agenda for proactive change. Executives can lead the charge by embracing existing government regulations, such as the U.S. Algorithmic Accountability Act of 2022 and the EU's General Data Protection Regulation. The latter, for example, provides organizational stakeholders with a *right to explanation* and thus the ability to assess potential biases. Rather than reacting to these trends, regulations, and laws, managers should proactively embrace diversity and the changes that come with it. For example, managers can develop a diversity charter for AI development and establish task forces composed of employees from various demographic, ethnic, and other backgrounds, composed of representatives from different levels within the organization. Task forces and the charter help facilitate ongoing dialogue and action plans to continuously identify and address diversity challenges, while ensuring compliance with relevant regulations.

A proactive leadership mindset

allows organizations to embrace diversity by relying on each employee's unique strengths and skills to contribute to the organization's overarching goals. Organizations that embrace this mindset and develop the corresponding implementation agenda are better positioned to develop responsible AI systems. For example, AI development organizations can embed diversity and inclusion principles into their AI development life cycle, ensuring that diverse perspectives are represented in the design, development, and deployment of AI systems to mitigate potential biases and promote fairness. 

## References

1. Cruz, A.F. et al. Promoting fairness through hyperparameter optimization. In *Proceedings of the 2021 IEEE Intern. Conf. on Data Mining (ICDM)* (Dec. 2021); <https://bit.ly/4f01RPs>
2. Guizani, M. et al. Gender inclusivity as a quality requirement: Practices and pitfalls. *IEEE Software* 37, 6 (Nov. 2020); <https://bit.ly/3BE0cAA>
3. Karen Hao, K. This is how AI bias really happens—and why it's so hard to fix. *MIT Technology Rev.* (2019); <https://bit.ly/3Yia3FO>
4. Lalor, J.P. et al. Should fairness be a metric or a model? A model-based framework for assessing bias in machine learning pipelines. *ACM Trans. Inf. Syst.* (Mar. 2023); <https://bit.ly/3BCAtIV>
5. Nilsson, F. Building a diverse company culture means empowering employees. *Forbes* (Sept. 22, 2021); <https://bit.ly/3U26lNg>
6. Nouri, S. Diversity and inclusion In AI. *Forbes*, (2019); <https://bit.ly/405ght0>
7. Treude, C. and Hata, H. She elicits requirements and he tests: Software engineering gender bias in large language models. In *Proceedings of the 2023 IEEE/ACM 20th Intern. Conf. on Mining Software Repositories* (Mar. 17, 2023); <https://bit.ly/3XXvl9v>
8. Wang, Y. and Zhang, M. Reducing implicit gender biases in software development: Does intergroup contact theory work? *ESEC/FSE 2020—Proceedings of the 28th ACM Joint Meeting European Software Engineering Conf. and Symp. on the Foundations of Software Engineering 20*, (Nov. 2020); <https://bit.ly/3Ye4gAh>
9. Werder, K., Ramesh, B., and Zhang, R. Establishing data provenance for responsible artificial intelligence systems. *ACM Trans. Manag. Inf. Syst.* 13, 2 (June 2022); <https://bit.ly/488zn3K>

**Karl Werder** (karw@itu.dk) Department of Business IT, IT University of Copenhagen, Copenhagen, Denmark.

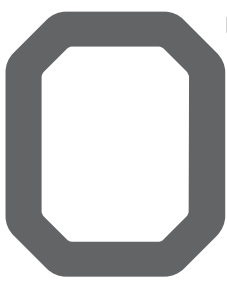
**Lan Cao** (lcao@odu.edu) Information Technology & Decision Sciences, Old Dominion University, Norfolk, VA, USA.

**Balasubramaniam Ramesh** (bramesh@gsu.edu) Computer Information Systems, Georgia State University, Atlanta, GA, USA.

**Eun Hee Park** (epark@odu.edu) Information Technology & Decision Sciences, Old Dominion University Norfolk, VA, USA.

Balasubramaniam Ramesh's research was sponsored by the Army Research Laboratory and was accomplished under Cooperative Agreement Number W911NF-23-2-0224. The views and conclusions contained in this Opinion column are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the Army Research Laboratory or the U.S. government. The U.S. government is authorized to reproduce and distribute reprints for government purposes notwithstanding any copyright notation herein.

© 2024 Copyright held by the owner/author(s).



# Kode Vicious

## Unwanted Surprises

*When that joke of an API is on you.*

### Dear KV,

A recent small project at work required me to use JavaScript, and I was surprised to find the following note in the documentation (see <https://bit.ly/3Bz3hC0>) for a code that sets a timeout: “Also note that if the value isn’t a number, implicit type coercion is silently done on the value to convert it to a number—which can lead to unexpected and surprising results; see Non-number delay values are silently coerced into numbers for an example.”

It is commonly accepted that languages like JavaScript and others will do duck typing, where a type is implied, but this seems to have gone a bit further, coercing any input into a number. My code does not allow user input to go into the timeout routine, so I am not upset, but that bit of doc gave me pause. I just cannot imagine a good reason to do what that does.

### Coerced

### Dear Coerced,

At least they tell you the results can be surprising, and everyone likes surprises, don’t they?

Like you, KV is at a loss to understand how this type of coercion is anything like duck typing. Just to be sure, I went to find the documentation you mentioned, and what caught my eye was this: “The time, in milliseconds, that the timer should wait before the specified function or code is executed. If this parameter is omitted, a value of 0 is used, meaning ex-



ecute “immediately,” or more accurately, during the next event cycle.”

Now, we know that the argu-

**It seems to me, and perhaps to you as well, that there really ought to be an error flagged at this point.**

ment expected is a time, in milliseconds, and I don’t know about you, but I have never seen such a time expressed as anything but an integer number—you know, like 42. Why would anyone supply anything but a number? It turns out that JavaScript passes a lot of strings around, or so one might think from reading this hilarious documentation. It turns out that the string “1000” is the same as the number 1000, which is a helpful bit of coercion, but the string “1 second” gets converted to 0, because ... well, just because.

It seems to me, and perhaps to you



Association for  
Computing Machinery

# ACM Transactions on Computing for Healthcare (HEALTH)

*A multi-disciplinary journal for  
high-quality original work on how  
computing is improving healthcare*

ACM Transactions on Computing for Healthcare (HEALTH) is the premier journal for the publication of high-quality original research papers, survey papers, and challenge papers that have scientific and technological results pertaining to how computing is improving healthcare.



For further information and to submit  
your manuscript, visit [health.acm.org](https://health.acm.org)

as well, that there really ought to be an error flagged at this point. If you want a number and you get a string, then say so, rather than just “trying to do the right thing,” especially if the results of getting it wrong can be surprising. KV does not like surprises, and really hates them in his code.

An extra funny part of this API is, as you point out, what might happen if it took user input. Imagine a banking website that let you tell it how long you wanted to wait for a call back from an agent, and it had this joke of an API behind an input box. How many people would write “1 second” or “1 minute” or some other value that was going to be coerced to 0?

There is the higher-order question of whether or not loosely typed languages with coercion are really a good idea in the first place. If you do not know what you are operating on, or what the expected output range might be, then maybe you ought not to be operating on that data in the first place. But now these languages have gotten into the wild and we will never be able to hunt them down and kill them soon enough for my liking, or for the greater good.

The only recourse we have now is to put our own protections around such APIs and make sure we know what we are passing through, before we wind up on the wrong side of a coercion.

**KV**

**Q** Related articles  
on [queue.acm.org](https://queue.acm.org)

**Code Vicious**  
**The Elephant in the Room**

<https://queue.acm.org/detail.cfm?id=3570921>

**Thou Shalt Not Depend on Me**

*Tobias Lauinger, Abdelberi Chaabane,  
and Christo B. Wilson*

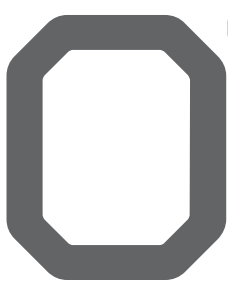
<https://queue.acm.org/detail.cfm?id=3205288>

**Dismantling the Barriers to Entry**

*Rich Harris*

<https://queue.acm.org/detail.cfm?id=2790378>

**George V. Neville-Neil** ([kv@acm.org](mailto:kv@acm.org)) is the proprietor of Neville-Neil Consulting, Brooklyn, NY, USA, and co-chair of the ACM Queue editorial board. He works on networking and operating systems code for fun and profit, teaches courses on various programming-related subjects, and encourages your comments, quips, and code snips pertaining to his *Communications* column.



# Privacy Notice and Choice Cannot Stand Alone

*Privacy notice and choice has largely failed us so far because we are not giving it the legal and technical support it needs.*

“NOTICE AND CHOICE” is the much-criticized approach to privacy regulation and self-regulation that has been in widespread use for approximately three decades. However, despite its failures as a regime, the concept of notice and choice should not be abandoned. It remains an important component of a broader privacy arsenal that should be combined with strong privacy laws and automated tools to provide customized privacy protections for individuals.

The idea behind notice and choice is that data collectors will provide transparent notices about their collection and use of personal information and allow individuals to make informed choices about whether and for what purpose their information will be used. In theory, this approach should allow data subjects to choose for themselves which uses of their personal information to permit (since these preferences are often context-dependent and vary by individual), while also encouraging a market for privacy in which data collectors improve their data practices to be more competitive.

In practice, notice and choice is a fantasy that has largely failed because notices take a long time to read,<sup>12</sup> are difficult to understand, and the number of decisions individuals face about the use of their data is overwhelming. Furthermore, it is often difficult for people to understand the potential



**The failure of notice and choice is due in part to the fact it has been left to stand on its own, with minimal legal teeth or technical support.**

consequences of their privacy-related choices because they lack a detailed understanding of relevant technologies, data flows, and downstream data uses. To make matters worse, notices and choice interfaces are often difficult to find<sup>8</sup> and designed to manipulate people into making the most privacy-invasive selections (deceptive patterns or dark patterns<sup>13</sup>). Given its poor track record, legal scholars,<sup>14</sup> privacy advocates, and even regulators<sup>10</sup> have been calling for the end of the notice-and-choice regime.

The failure of notice and choice is due in part to the fact it has largely



## Peer-reviewed Resources for Engaging Students

EngageCSEdu provides faculty-contributed, peer-reviewed course materials (Open Educational Resources) for all levels of introductory computer science instruction.



[engage-csedu.org](http://engage-csedu.org)



Association for  
Computing Machinery

been left to stand on its own, with minimal legal teeth or technical support, and (in many jurisdictions) without strong baseline privacy laws. In short, the notice-and-choice regime was set up for failure. For notice and choice to be effective it must be mandatory, with requirements regulators have resources to enforce, and it must be embedded in a standardized technology framework that allows people to readily automate their privacy decisions without being constantly bombarded with privacy choices. It is also critical to have baseline legal protections that do not allow data collectors to ask people to consent to data practices that are fundamentally unfair or about which they are unable to make informed decisions.

The idea of automating privacy decisions is not new. After the U.S. Federal Trade Commission began encouraging website operators to post privacy policies in the 1990s, privacy advocates complained these policies were too long to be useful to users. In response, the World Wide Web Consortium (W3C) developed the Platform for Privacy Preferences Project (P3P), a protocol for encoding privacy policies in a computer-readable XML format and allowing Web browsers and other user agents to retrieve these policies, parse them automatically, and use them to inform users or make automated decisions on their behalf. The most widely adopted P3P implementation was built into the Microsoft Internet Explorer 6 Web browser in 2001 and used to automate third-party cookie-blocking decisions. I led the P3P working group at W3C and also worked on a research prototype P3P user agent called Privacy Bird that displayed a colored bird and a brief digest of the privacy policy and where it conflicted with a user's preset privacy preferences. Unfortunately, P3P never saw widespread adoption after its release in 2002 because it lacked incentives for adoption and mechanisms for enforcement.<sup>4</sup> Indeed, in 2010 when thousands of websites were found to have circumvented browser P3P controls,<sup>9</sup> not a single regulator stepped in.<sup>5</sup>

After P3P had come and gone, a simpler approach to automating privacy choice was proposed: Do Not

Track (DNT). Rather than creating a computer-readable privacy policy and sending it to web browsers, DNT sought to transmit a single header from web browsers to websites to request a site and any third-party sites that load with it refrain from tracking an individual user. The W3C spent nearly a decade trying to reach consensus on a DNT standard that web browsers and websites would adopt. Although some web browsers implemented DNT, in practice it was meaningless as few websites paid attention to the DNT headers.<sup>7</sup> In the absence of adoption incentives or legal mandates, DNT ultimately failed.

The latest automated privacy choice approach is Global Privacy Control (GPC), which allows users to turn on a setting in their browser (or browser extension) that transmits a GPC signal to automatically opt out of websites selling or sharing their personal information. What is particularly exciting about GPC is that now, for the first time, privacy laws are requiring websites to respect automated privacy signals such as GPC. Under the California Consumer Privacy Act (CCPA), websites are required to act on GPC opt-out requests, and in 2022 the California Attorney General began enforcing compliance.<sup>2</sup> There are already six other U.S. states that require websites to honor opt-out preferences transmitted through Universal Opt-Out Mechanisms (UOOMs).<sup>1</sup> GPC is designed to be compatible with privacy laws around the world. Although GPC currently provides only a single signal, it could be extended to offer

**We currently lack incentives for companies to build automated privacy choice frameworks into their products.**



multiple signals and provide for more fine-grained choices. The fact that an automated choice mechanism is now enforceable by law is potentially the game changer needed for automated choice mechanisms to have a chance at success. However, GPC is not yet a settings option in the most popular Web browsers, although there are privacy-focused browsers and plugins that offer this option or enable it by default.

With the rapid proliferation of mobile apps, smart homes, and Internet of Things (IoT) devices, websites are just the tip of the iceberg when it comes to the collection and use of personal data. Thus, we have an increased need for automated tools that can help users signal their preferences about sharing and using their data without being bombarded by requests from every smart device they walk by throughout the day. Researchers have explored mobile app privacy agents that automate app permissions settings<sup>11</sup> and IoT privacy agents that allow users to manage privacy settings for IoT devices in their environment.<sup>6</sup> However, we currently lack incentives for companies to build automated privacy choice frameworks into their products.

The current generation of deployed notice-and-choice tools are simple but lack flexibility. More research is needed on how to build tools that can operate with minimal user input after their initial quick and easy configuration, perhaps driven by machine-learning approaches that learn a user's preferences over time and can extrapolate based on the user's current context or the preferences of similar users.<sup>15</sup> These tools should allow users to occasionally grant exceptions to allow the use of their data when they find it beneficial. However, these exceptions should be granted because users want to provide their data (for example, I want to provide my location when I use a mapping service because I want to view my location on the map) rather than because services break when data is withheld (for example, some websites exhibit strange behavior or stop working when third-party cookies are blocked, encouraging users to override their cookie blockers) or because users are constantly bombard-

## More research is needed on how to build tools that can operate with minimal user input after their initial quick and easy configuration.

ed with requests that they swat away without thinking (for example, cookie banners, another notice-and-choice mechanism that has been largely ineffective as a privacy tool<sup>3</sup>). And when exceptions are granted, data should be used only for the purposes the user requests (for example, the map service should not also use my location to serve me location-targeted ads unless I have specifically requested them). Recent research has demonstrated the utility of “generalizable active privacy choice” interfaces for GPC that allow users to send GPC signals automatically to websites a user visits according to criteria such as type of website, type of data collected, user's self-described privacy profile, or user's privacy profile learned by the system.<sup>15</sup>

Even if we design fantastic tools, we will need incentives or legal mandates for them to be made readily available to end users and their signals respected by data collectors. We need UOOMs built into every Web browser and their signals respected by every website. We need IoT devices that send and receive standardized privacy signals to well-designed user agents. We need enforceable penalties for data collectors that fail to honor automated signals or manipulate users into consenting to data practices. And, importantly, we need strong baseline privacy regulations that restrict the use of personal information without individual consent and prohibit some personal information uses altogether.

While notice and choice as a re-

gime has largely failed to live up to its promises to date, if bolstered by appropriate laws, technology standards, and easy-to-use interfaces, the notice and choice concept could be a useful tool in our future privacy toolbox. **□**

### References

- Adams, S. and Gray, S. Survey of current universal opt-out mechanisms. *Future of Privacy Forum*. (Oct. 12, 2023); <https://bit.ly/3ZSVo4s>
- AG Press Office. *Attorney General Bonta Announces Settlement with Sephora as Part of Ongoing Enforcement of California Consumer Privacy Act*. (Aug. 24, 2022); <https://bit.ly/4dApSv6>
- Cranor, L.F. Cookie monster. *Commun. ACM* 65, 7 (July 2022); 10.1145/3538639
- Cranor, L.F. Necessary but not sufficient: Standardized mechanisms for privacy notice and choice. *J. Telecommun. High Technol. Law* 10, (2012); <https://bit.ly/3Yb0BDs>.
- Cranor, L.F. P3P is dead, long live P3P! This Thing blog. (Dec. 3, 2012); <https://bit.ly/480Bw19>
- Das, A. et al. Personalized privacy assistants for the Internet of Things: Providing users with notice and choice. *IEEE Pervasive Computing* 17, 3 (Jul. 2018).
- Fleishman, G. How the tragic death of Do Not Track ruined the web for everyone. *Fast Company* (Mar. 7, 2019); <https://bit.ly/3ZUn6Ob>
- Habib, H. et al. “It's a scavenger hunt”: Usability of websites' opt-out and data deletion choices. In *Proceedings of CHI 2020* (2020); 10.1145/3313831.3376511
- Leon, P.G. et al. Token attempt: the misrepresentation of website privacy policies through the misuse of P3P compact policy tokens. In *Proceedings of the 9th Annual ACM Workshop on Privacy in the Electronic Society (WPES '10)*, (2010), 93–104; 10.1145/1866919.1866932
- Levine, S. *Toward a Safer, Freer, and Fairer Digital Economy: How Proactive Consumer Protection Can Make the Internet Less Terrible*. Fourth Annual Reidenberg Lecture, Fordham Law School. (Apr. 17, 2024); <https://bit.ly/4dGZWOI>
- Liu, B. et al. Follow my recommendations: A personalized privacy assistant for mobile app permissions. In *Proceedings of the 12th Symp. on Usable Privacy and Security (SOUPS 2016)*. (2016); <https://bit.ly/3U4yFIC>
- McDonald, A.M. and Cranor, L.F. The cost of reading privacy policies. *I/S: A J. of Law and Policy for the Information Society* 4, 3 (Winter 2008–2009); <https://bit.ly/3XYth10>.
- Narayanan, A. et al. Dark patterns: past, present, and future. *Commun. ACM* 63, 9 (Sept. 2020); 10.1145/3397884
- Rothchild, J.A. Against notice and choice: The manifest failure of the proceduralist paradigm to protect privacy online (or anywhere else) (Feb. 20, 2018). *Cleveland State Law Rev.* (2018); <https://ssrn.com/abstract=3126869>
- Wijesekera, P. et al. Contextualizing privacy decisions for better prediction. In *Proceedings of CHI 2018* (2018); 10.1145/3173574.3173842
- Zimmeck, S. et al. Generalizable active privacy choice: Designing a graphical user interface for global privacy control. In *Proceedings on Privacy Enhancing Technologies Symp. (PoPETs)* (2024); 10.56553/popets-2024-0015

**Lorrie Faith Cranor** (lorrie@cmu.edu) is director and Bosch Distinguished Professor in Security and Privacy Technologies, CyLab Security and Privacy Institute and FORE Systems University Professor, Computer Science and Engineering and Public Policy, Carnegie Mellon University in Pittsburgh, PA, USA.

© 2024 Copyright held by the owner/author(s).

# Opinion

## AI Must Be Anti-Ableist and Accessible

*Seeking to improve AI accessibility by changing how AI-based systems are built.*

**T**HE INCREASING USE of artificial intelligence (AI)-based technologies in everyday settings creates new opportunities to understand how disabled people might use these technologies.<sup>2</sup> Recent reports by Whittaker et al.,<sup>11</sup> Trewin et al.,<sup>9</sup> and Guo et al.<sup>3</sup> highlight concerns about AI's potential negative impact on inclusion, representation, and equity for those in marginalized communities, including disabled people. In this Opinion column, we summarize and build on these important and timely works. We define disability in terms of the discriminatory and often systemic problems with available infrastructure's ability to meet the needs of all people. For example, AI-based systems may have ableist biases, associate disability with toxic content or harmful stereotypes, and make false promises about accessibility or fail to accessibly support verification and validation.<sup>2</sup> These problems replicate and amplify biases experienced by disabled people when interacting in everyday life. We must recognize and address them.

### Recognizing and Addressing Disability Bias in AI-Based Systems

AI model development must be extended to consider risks to disabled people including:

**Unrepresentative data.** When groups are historically marginalized and underrepresented, this is "imprinted in the data that shapes AI sys-



tems."<sup>11</sup> Addressing this is not a simple task of increasing the number of categories represented, because identifiable impairments are not static, or homogeneous, nor do they usually occur singly. The same impairment may result from multiple causes and vary across individuals. To reduce bias, we must collect data about people from multiple contexts with multiple impairments over multiple timescales.

**Missing and unlabeled data.** AI models trained on existing large text corpora risk reproducing bias inherent in those corpora.<sup>2,3</sup> For example, the relative lack of accessible mobile apps<sup>8</sup> makes it more likely AI-generated code for mobile apps will also be inaccessible.

**Measurement error.** Measurement error can exacerbate bias.<sup>9</sup> For example, a sensor's failure to recognize wheelchair activity as exercise may lead to bias in algorithms trained on associated data. These errors exist for every major class of sensing.<sup>3</sup>

**Inaccessible interactions.** Even if an AI-based system is carefully designed to minimize bias, the interface to that algorithm, its configuration, the explanation of how it works, or the potential to verify its outputs may be inaccessible (for example, Glazko et al.<sup>2</sup>).

### Disability-Specific Harms of AI-Based Technologies

Even the most well-designed of systems may cause harms when de-

ployed. It is critical that technologists learn about these harms and how to address them before deploying AI-based systems.

**Defining what it means to be “human.”** As human judgment is increasingly replaced by AI, “norms” baked into algorithms that learn most from the most common cases<sup>11</sup> become more strictly enforced. One user had to falsify data because “some apps [don’t allow] my height/weight combo for my age.”<sup>14</sup> Such systems render disabled people “invisible”<sup>11</sup> and amplify existing biases internal to and across othering societal categories.<sup>3</sup> AI-based systems are also being used to track the use and allocation of assistive technologies, from CPAP machines for people with sleep apnea, to prosthetic legs,<sup>11</sup> deciding who is “compliant enough” to deserve them.

**Defining what “counts” as disabled.** Further, algorithms often define disability in historical medical terms.<sup>11</sup> However, if you are treated as disabled by those around you, legally you are disabled—the Americans with Disabilities Act does not require a diagnosis (42 U.S.C. § 12101 (a)(1)). Yet, AI-based technologies cannot detect how people are treated. AI-based technologies must never be considered sufficient, nor required as mandatory, for disability identification or service eligibility.

**Exacerbating or causing disability.** AI-based systems may physically harm humans. Examples include activity tracking systems that push workers and increase the likelihood of work-related disability<sup>11</sup> and AI-based systems that limit access to critical care resources, resulting in an increased risk of hospitalization or institutionalization.<sup>5</sup>

**Privacy and security.** Disability status is increasingly easy to detect from readily available data such as mouse movements.<sup>12</sup> Any system that can detect disability can also track its progression over time, possibly before a person knows they have a diagnosis. This information could be used, without consent or validation, to deny access to housing, jobs, or education, potentially without the knowledge of the impacted individuals.<sup>11</sup> Additionally, AI biases may require people with specific impairments to accept

reduced digital security, such as the person who must ask a stranger to ‘forge’ a signature at the grocery store “... because I can’t reach [the tablet].”<sup>14</sup> This is not only inaccessible, it is illegal: kiosks and other technologies such as point-of-sale terminals used in public accommodations are covered under Title III of the Americans with Disabilities Act.

**Reinforcing ableist policies, standards, and norms.** AI systems rely on their training data, which may contain biases or reflect ableist attitudes. For example, Glazko et al.<sup>2</sup> describe both subtle and overt ableism that appeared when trying to generate an image and summarize text. These harms also affect disabled people who are not directly using AI, such as biased AI-rankings for résumés that mention disability.<sup>1</sup>

#### Recommendations

First and foremost, do no harm: algorithms that put a subset of the population at risk should not be deployed. This requires regulatory intervention, algorithmic research (for example, developing better algorithms for handling outliers<sup>9</sup>), and applications research (for example, studying the risks that applications might create for disabled people). We must consider “the context in which such technology is produced and situated, the politics of classification, and the ways in which fluid identities are (mis)reflected and calcified through such technology.”<sup>11</sup>

The most important step in avoiding this potential harm is to change who builds, regulates, and deploys AI-based systems. We must ensure

**We must ensure disabled people contribute their perspective and expertise to the design of AI-based systems.**

## ACM Distinguished Speakers

**A great speaker can make the difference between a good event and a WOW event!**

Take advantage of ACM’s Distinguished Speakers Program to invite renowned thought leaders in academia, industry, and government to deliver compelling and insightful talks on the most important topics in computing and IT today. ACM will cover the cost of transportation for the speaker to travel to your event.

**speakers.acm.org**



Association for  
Computing Machinery

disabled people contribute their perspective and expertise to the design of AI-based systems. Equity requires that disabled people can enter the technology workforce so they can build and innovate. This requires active participation in leadership positions, access to computer and data science courses, and accessible development environments. The slogan “Nothing about us without us” is not just memorable—it is how a just society works.

Organizations building AI systems must also improve equity in data collection, review, management, storage, and monitoring. As highlighted in President Biden’s AI Bill of Rights,<sup>10</sup> equity must be embedded in every stage of the data pipeline. This includes motivating and paying participants for accessible data and metadata that does not oversimplify disability, ensuring disabled peoples’ data is not unfairly rejected when minor mistakes occur or due to stringent time limits,<sup>7</sup> ensuring disabled stakeholders participate in and understand their representation in training data, through transparency about and documentation of what is collected and how it is used.<sup>9</sup> Community representation can improve the breadth of participation in data collection and guide the design of data collection systems and prioritization of what data to collect and what *not* to use.

Legislators and government agencies must enact regulations for algorithmic accessibility. Algorithms should be subject to a basic set of expectations around how they will be assessed for accessibility, just like websites. This will help to address basic access constraints, reduce the types of errors that enforce “normality” rather than honoring heterogeneity, and eliminate errors that gatekeep who is “human.” Consumer consent and oversight concerning best practices are both essential to fair use. AI-based systems should be interpretable, overrideable, and support accessible verifiability of AI-based results during use.<sup>2</sup>

All parties must work together to promote best practices for accessible deployments, including accessible options for interacting with AI. Just as accessible ramps or elevators that are hidden or distant are not acceptable

## Accessible AI is ultimately a question of values, not technology. Simple infusion of disabled people is insufficient.

for accessibility in physical spaces, accessible AI-based systems must not create undue burdens in digital spaces nor segregate disabled users.

To gauge progress and identify areas in need of work, the community must develop assessment methods to uncover bias. Many algorithms maximize aggregate metrics that fail to both recognize and address bias.<sup>3</sup> Further, we must consider intersections of disability bias with other concerns, such as racial bias.<sup>6</sup> Scientific research will be essential to defining appropriate assessment procedures.

### Conclusion

Accessible AI is ultimately a question of values, not technology. Simple inclusion of disabled people is insufficient. We must work to ensure equity in data collection, algorithm access, and in the creation of AI-based systems, even when equity may not be expedient.

The fight for accessible computing provides lessons for meeting these ambitious goals. As the disability rights movement of the 1970s converged with the dawn of the personal computer era, activists urged the computing industry to make computing more accessible. The passage of the Americans with Disabilities Act (ADA) in 1990 provided legal recourse, and the advent of GUIs and the Web in the mid-1990s led to the development of new accessibility tools and guidelines for building accessible systems. These tools made computing more robust, helping users with disability and others alike, while advocates successfully used the ADA to ensure accessibility of many websites.

This combination of advocacy, engagement with industry, regulation, and legal action can be applied to make AI safer for everyone. The opacity of AI tools presents unique obstacles, but the AI Bill of Rights<sup>10</sup> and more technical federal efforts detailing steps toward appropriate AI design provide initial directions. The pushback from those who hope to profit from AI will undoubtedly be significant, but the costs to those of us who are, or who will become, disabled will be even greater. We cannot train AI on a mythic 99%. **C**

### References

1. Glazko, K.S. et al. Identifying and improving disability bias in GAI-based resume screening. In *FACCT 2024* (2024); <https://bit.ly/489R5nd>
2. Glazko, K.S. et al. An autoethnographic case study of generative artificial intelligence’s utility for accessibility. In *Proceedings of the 25<sup>th</sup> Intern. ACM SIGACCESS Conf. on Computers and Accessibility, ASSETS 2023, New York, NY, USA, October 22-25, 2023*; <https://bit.ly/484udWp>
3. Guo, A. et al. Toward fairness in AI for people with disabilities SBG@a research roadmap. *ACM SIGACCESS Access. Comput.* 125, 2 (2020); <https://bit.ly/3U5cGrl>
4. Kane, S.K. et al. Sense and accessibility: Understanding people with physical disabilities’ experiences with sensing systems. In *ASSETS ’20: The 22<sup>nd</sup> Intern. ACM SIGACCESS Conf. on Computers and Accessibility*. T.J. Guerreiro, H. Nicolau, and K. Moffatt, (Eds.). Virtual Event, Greece, (Oct. 26–28, 2020); <https://bit.ly/3UbgK9l>
5. Lecher, C. What happens when an algorithm cuts your health care. *The Verge* 21, 3 (2018).
6. Morgan, J.N. Policing under disability law. *Stan. L. Rev.* 73, (2021).
7. Park, J.S. et al. Understanding the representation and representativeness of age in AI data sets. M. Fourcade, B. Kuipers, S. Lazar, and D.K. Mulligan (Eds.). Virtual Event, USA, (May 19–21, 2021); <https://bit.ly/401sdvV>
8. Ross, A.S. An epidemiology-inspired, large-scale analysis of mobile app accessibility. *ACM SIGACCESS Access. Comput.* 123, 6 (2020); <https://bit.ly/3A6QWo0>
9. Trewin, S. et al. Considerations for AI fairness for people with disabilities. *AI Matters* 5, 3 (2019); <https://bit.ly/4f5ur1A>
10. White House Office of Science and Technology Policy. Blueprint for an AI Bill of Rights (2023); <https://bit.ly/4eFbHG8>
11. Whittaker, M. et al. Disability, bias, and AI. *AI Now Institute* 8, (2019).
12. Youngmann, B. et al. A machine learning algorithm successfully screens for Parkinson’s in web users. *Annals of Clinical and Translational Neurology* 6, 12 (2019).

**Jennifer Mankoff** ([jmankoff@cs.washington.edu](mailto:jmankoff@cs.washington.edu)) is a professor at the University of Washington, Seattle, WA, USA.

**Devva Kasnitz** ([devva@earthlink.net](mailto:devva@earthlink.net)) is an adjunct professor at the City University of New York, New York, NY, USA.

**L. Jean Camp** ([ljcamp@indiana.edu](mailto:ljcamp@indiana.edu)) is a professor at Indiana University, Bloomington, IN, USA.

**Jonathan Lazar** ([jlazar@umd.edu](mailto:jlazar@umd.edu)) is a professor at the University of Maryland, College Park, MD, USA.

**Harry Hochheiser** ([harryh@pitt.edu](mailto:harryh@pitt.edu)) is a professor at the University of Pittsburgh, Pittsburgh, PA, USA.

This work was funded by the University of Washington CREATE Center.

© 2024 Copyright held by the owner/author(s).

# OPEN FOR SUBMISSIONS

## ACM Transactions on Probabilistic Machine Learning (TOPML)

Co-Editors-in-Chief

Wray Buntine  
VinUniversity, Vietnam

Fang Liu  
University of Notre Dame, USA

Theodore Papamarkou  
The University of Manchester, UK



Gold Open Access publication focusing on probabilistic methods that learn from data to improve performance on decision-making or prediction tasks under uncertainty

*ACM Transactions on Probabilistic Machine Learning* (TOPML) is a new Gold Open Access publication from ACM focusing on probabilistic methods that learn from data to improve performance on decision-making or prediction tasks under uncertainty. Optimization, decision-theoretic or information-theoretic methods are within the remit if they are underpinned by a probabilistic structure. Probabilistic methods may be harnessed to address questions related to statistical inference, uncertainty quantification, predictive calibration, data generation and sampling, causal inference, stability, and scalability. Examples of approaches relevant to the scope include Bayesian modelling and inference, variational inference, Gaussian processes, Monte Carlo sampling, Stein-based methods, and ensemble modelling. Examples of models for which probabilistic approaches are sought include neural networks, kernel-based models, graph-based models, reinforcement learning models, recommender systems, and statistical and stochastic models. Ethical considerations of probabilistic machine learning, such as data privacy and algorithmic fairness, should be addressed in papers where there is a direct ethical connection or context for the work being described.

The journal welcomes theoretical, methodological, and applied contributions. Purely theoretical contributions are of interest if they introduce novel methodology. Methodological and applied contributions are of interest provided that proposed probabilistic approaches are motivated and empirically corroborated by non-trivial examples or applications. Multidisciplinary approaches with a probabilistic structure are within the scope.

For more information and to submit your manuscript, please visit [topml.acm.org](http://topml.acm.org).



Association for  
Computing Machinery

**Trade-offs between secure computation via cryptography and hardware enclaves.**

BY RALUCA ADA POPA

# Confidential Computing or Cryptographic Computing?

INCREASINGLY STRINGENT PRIVACY regulations—for example, the European Union’s General Data Protection Regulation (GDPR) or the California Consumer Privacy Act (CCPA) in the U.S.—and sophisticated attacks leading to massive breaches have increased the demand for protecting data in use, or *encryption in use*. The encryption-in-use paradigm is important for security because it protects data during processing; in contrast, *encryption at rest* protects data only when it is in storage and *encryption in transit* protects data only when it is being communicated over the network. In both cases, however, data is exposed during computation—namely, while it is being *used/processed* at the servers. It is during that processing window when many data breaches happen, either at the hands of hackers or insider attackers.

Another advantage of encryption in use is that it

allows different parties to collaborate by putting their data together for the purpose of gleaning insights from their aggregate data—without actually sharing their data with each other. This is because the parties share *encrypted* data with each other, so no party can see the data of any other party in decrypted form. The parties can still run useful functions on the data and release only the computation results. For example, medical organizations can train a disease-treatment model on their aggregate patient data without seeing each other’s data. Another example is within a financial institution, such as a bank, where data analysts can build models across different branches or teams that would otherwise not be allowed to share data with each other.

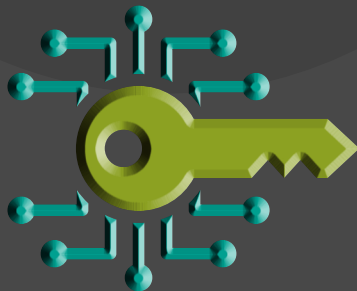
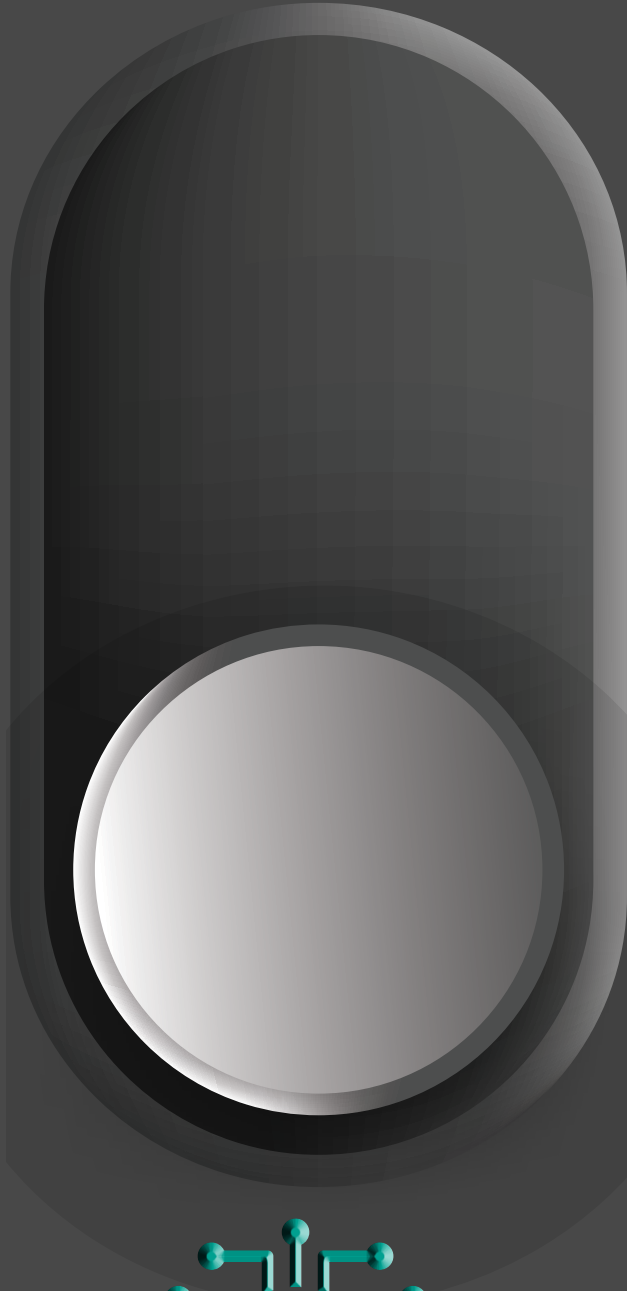
Today there are two prominent approaches to secure computation:

- ▶ A purely cryptographic approach (using homomorphic encryption and/or secure multi-party computation).
- ▶ A hardware security approach (using hardware enclaves sometimes combined with cryptographic mechanisms), also known as *confidential computing*.

There is a complex trade-off between these two approaches in terms of security guarantees, performance, and deployment. Comparisons between the two for ease of use, security, and performance are shown in Tables 1, 2, and 3. For simple computations, both approaches tend to be efficient, so the choice between these two would likely be based on security and deployment considerations. However, for complex workloads, such as advanced machine-learning (ML) training (for example, transformers) and rich SQL analytics, the purely cryptographic approach is too inefficient for many real-world deployments. In these cases, the hardware security approach is the practical choice.

## Cryptographic Computation

There are two main ways to compute on encrypted data using cryptographic




mechanisms: *homomorphic encryption* and *secure multi-party computation*.

Homomorphic encryption permits evaluating a function on encrypted input. For example, with fully homomorphic encryption,<sup>9</sup> a user can send to a cloud  $\text{Encrypt}(x)$  for some input  $x$ , and a cloud can compute  $\text{Encrypt}(f(x))$  using a public evaluation key for any function  $f$ .


Secure multi-party computation<sup>23</sup> is often more efficient than homomorphic encryption and can protect against a malicious attacker, but it has a different setup, shown in Figure 1. In secure multi-party computation (MPC),  $n$  parties having private inputs  $x_1, \dots, x_n$ , compute a function  $f(x_1, \dots, x_n)$  without sharing their inputs with each other. This is a cryptographic protocol at the end of which the parties learn the function result, but in the process no party learns the input of the other party beyond what can be inferred from the function result.

There are many different threat models for computation in MPC, resulting in different performance overheads. A natural threat model is to assume that all but one of the participating parties are malicious, so each party need only trust itself. This natural threat model, however, comes with implementations that have high overheads because the attacker is quite powerful. To improve performance, people often compromise in the threat model by assuming that a majority of the parties act honestly (and only a minority are malicious).

Also, since the performance overheads often increase with the number of parties, another compromise in some MPC models is to outsource the computation to  $m < n$  servers in different trust domains. For example, some works propose outsourcing the secure computation to two mutually distrustful servers. This latter model tends to be weaker than threat models, where a party needs to trust only itself. Therefore, in the rest of this article, we only consider maliciously secure  $n$ -party MPC. This also makes the comparison to secure computation via hardware enclaves more consistent, because this second approach aims to protect against all parties being malicious.



**There are many different threat models for computation in multi-party computation, resulting in different performance overheads.**



## Hardware Enclaves

Figure 2 shows a simplified view of one processor. The light blue area denotes the inside of the enclave. The Memory Encryption Engine (MEE) is a special hardware unit that contains cryptographic keys and, by using them, it encrypts the data leaving the processor, so the memory bus and memory receive encrypted data. Inside the processor, the MEE decrypts the data so the core can compute on data at regular processor speeds.

Trusted execution environments, such as hardware enclaves, aim to protect an application's code and data from all other software in the system. The MEE ensures that even an administrator of a machine with full privileges examining the data in memory sees encrypted data (Figure 2). When encrypted data returns from main memory into the processor, the MEE decrypts the data and the CPU computes on decrypted data. This is what enables the high performance of enclaves compared with the purely cryptographic computation: The CPU performs computation on raw data as in regular processing.

At the same time, from the perspective of any software or user accessing the machine, the data looks encrypted at any point in time: The data going into the processor and coming out is always encrypted, giving the illusion that the processor is computing on the encrypted data. Hardware enclaves also provide a useful feature called remote attestation,<sup>5</sup> with which remote clients can verify code and data loaded in an enclave and establish a secure connection with the enclave, which they can use to exchange keys.

A number of enclave services are available today on public clouds, such as Intel Software Guard Extensions (SGX) in Azure, Amazon Nitro Enclaves in Amazon Web Services (although this enclave is mostly software-based and does not provide memory encryption), Secure Encrypted Virtualization (SEV) from AMD<sup>12</sup> in Google Cloud, and others. NVIDIA recently added enclave support in its H100 GPU.<sup>6</sup>

**Ease-of-use comparison.** With a purely cryptographic approach, there is no need for specialized hardware and special hardware assumptions.



At the same time, in a setting like MPC, the parties must be deployed in different trust domains for the security guarantees of MPC to hold. In the threat models discussed earlier, participating organizations have to run the cryptographic protocol on site or in their private clouds, which is often a set-up, management, and/or cost burden compared with running the whole computation on a cloud. This can be a deal-breaker for some organizations.

With homomorphic encryption, in principle, the whole computation can be run in the cloud, but homomorphic encryption does not protect against malicious attackers as MPC and hardware enclaves do. For such protection, you would also have to use heavy cryptographic tools, such as zero-knowledge proofs.

In contrast, hardware enclaves are now available on major cloud providers such as Azure, AWS, and Google Cloud. Running an enclave collaborative computation is as easy as using one of these cloud services. This also means that to use enclaves, you do not need to purchase specialized hardware: The major clouds already provide services based on these machines. Of course, if the participating organizations want, they could each deploy enclaves on their premises or in private clouds and perform the collaborative computation across the organizations in a distributed manner similar to MPC. The rest of this article assumes a cloud-based deployment for hardware enclaves, unless otherwise specified.

With cryptographic computing, cryptographic expertise is often required to run a certain task. Since the cryptographic overhead is high, tailoring the MPC design for a certain task can bring significant savings. At the same time, this requires expertise and time that many users do not have. Hiring cryptography experts for this task is burdensome and expensive. For example, a user cannot simply run a data-analytics or machine-learning pipeline in MPC. Instead, the user has to identify some key algorithms in those libraries to support, employ tailored cryptographic protocols for those, and implement the resulting cryptographic protocols in a system

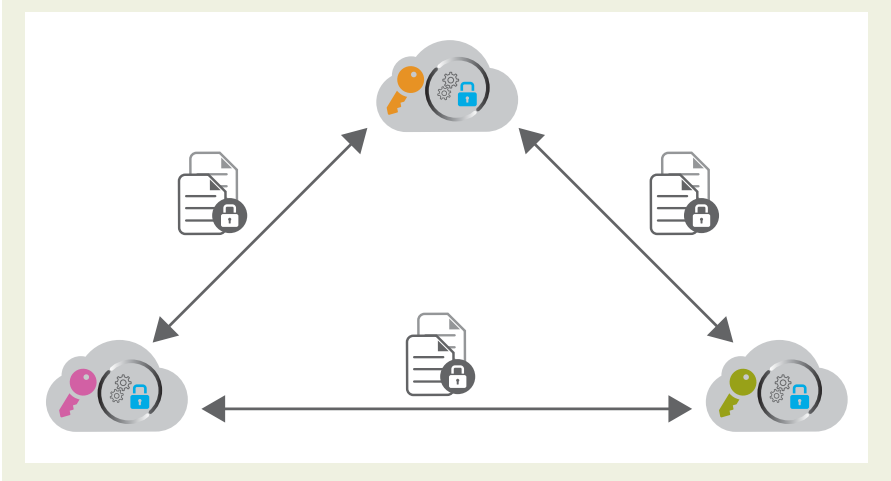
that likely requires significant code changes as compared with an existing analytics/ML pipeline.

In contrast, modern enclaves provide a VM interface, resulting in a Confidential Virtual Machine.<sup>10</sup> This means that the user can install proprietary software in these enclaves without modifying this software. Complex codebases are supported in this manner. For example, Confidential Google Kubernetes engine nodes<sup>11</sup> enable Kubernetes to run in confidential VMs. The first iteration of the enclave, Intel SGX, did not have this flexibility

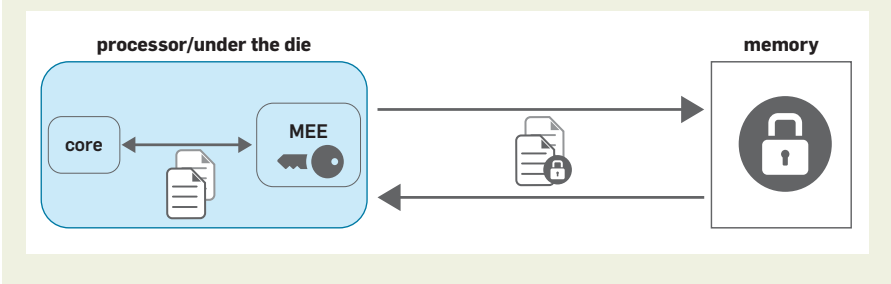
and required modifying and porting a program to run it in the enclave. Since then, it has been recognized that to use this technology for confidential data pipelines, users must remove the friction of porting to the enclave interface. This is how the confidential VM model was born.

**Security comparison.** The *homomorphic encryption* referred to here can compute more complex functions, meaning either *fully* or *leveled* homomorphic encryption. Some homomorphic encryption schemes can perform simple functions efficiently

**Figure 1. Illustration of secure multi-party computation for three parties. Each party essentially has a cryptographic key that only that party can access. Parties exchange encrypted data (often over multiple iterations), and compute locally on cryptographic data from other parties and their local data.**



**Figure 2. Simplified illustration of memory encryption in a hardware enclave.**



**Table 1. Ease-of-use comparison.**

Cryptographic Computing (such as FHE and MPC)	Enclave/Confidential Computing
✗ Requires cryptographic expertise to design a tailored protocol for increased performance	✓ Can run proprietary systems in confidential computing without modification
✓ Does not require specialized hardware	✗ Requires specialized hardware to run on
✗ Requires a deployment across multiple trust domains	✓ Can be deployed in a single trust domain (for example, the Cloud)
✗ Cannot support proprietary systems	

(such as addition or low-degree polynomials). As soon as the function becomes more complex, performance degrades significantly.

Homomorphic encryption is a special form of secure computation, where a cloud can compute a function over encrypted data without interacting with the owner of the encrypted data. It is a cryptographic tool that can be used as part of an MPC protocol. MPC is more generic and encompasses more cryptographic tools; parties running an MPC protocol often interact with each other over multiple rounds, which affords better performance than being restricted to a non-interactive setting.

For general functions, homomorphic encryption is slower than MPC. Also, as discussed, it does not provide malicious security without employing an additional cryptographic tool such as zero-knowledge proofs, which can be computationally expensive.

When an MPC protocol protects against some malicious parties, it also protects against any side-channel attacks at the servers of those parties. In this sense, the threat model for the malicious parties is cleaner than for hardware enclaves' threat model because it does not matter what attack adversaries mount at their servers; MPC considers any sort of compromise for these parties. For the honest parties, MPC does not protect against side-channel attacks.

In the case of enclaves, attackers can attempt to perform side-channel attacks. A common class of side-channel attack (which encompasses many different types) involves an attacker who observes which memory locations are accessed as well as the order and frequency of these accesses. Even though the data at those memory locations is encrypted, seeing the access pattern can provide confidential information to the attacker. These attacks are called memory-based access-pattern attacks, or simply access-pattern attacks.

There has been significant research on protecting against these access-pattern side-channel attacks using a cryptographic technique called *data-oblivious computation*. Oblivious computation ensures that the accesses to memory do not reveal any information about the sensitive data being accessed. Intuitively, it transforms the code into a side-channel-free version of the code, similar to how the OpenSSL cryptographic libraries have been hardened.

Oblivious computation protects against a large class of side-channel attacks based on cache-timing-exploiting memory accesses, page faults, branch predictor, memory bus leakage, dirty bit, and others.

Hardware enclaves such as Intel SGX are also prone to other side-channel attacks besides access patterns (for example, speculative-execution-

based attacks, attacks to remote attestation), which are not prevented by oblivious computation. Fortunately, when such attacks are discovered, they are typically patched in a short amount of time by cloud providers, such as Azure confidential computing and others. Even if the hardware enclaves would be vulnerable for the time period before the patch, the traditional cloud security layer is designed to prevent attackers from breaking in to mount such a side-channel attack. This additional level of security would not exist on a client-side usage of enclaves.

Subverting this layer as well as being able to set up a side-channel attack in a real system with such protection is typically much harder to do for an attacker because it requires the attacker to succeed at mounting two different and difficult types of attacks. It is not sufficient for the attacker to succeed in attacking only one. At the time of writing this article, there is no evidence of any such dual attack having occurred on state-of-the-art public clouds such as Azure confidential computing. This is why, when using hardware enclaves, you can assume that the cloud provider is a well-intended organization and its security practices are state of the art, as would be expected from major cloud providers today.

Another aspect pertaining to security is the size of the trusted computing base (TCB). The larger the TCB, the larger the attack surface and the more difficult it is to harden the code against exploits. Considering the typical use of enclaves these days—namely, the confidential VM abstraction—the enclave contains an entire virtual machine. This means that the TCB for enclaves is large—many times larger than the one for cryptographic computation. For cryptographic computation, the TCB is typically the client software that encrypts the data, but there might be some extra assumptions on the server system, depending on the threat model.

**Performance comparison.** Cryptographic computation is efficient enough for running simple computations, such as summations, counts, or low-degree polynomials. At the time this article was published, cryp-

**Table 2. Security comparison.**

Cryptographic Computing (such as FHE and MPC)	Enclave/Confidential Computing
✗ Homomorphic encryption is typically slower than MPC for non-trivial functionalities and does not protect against malicious attackers.	✓ Enclaves offer a notion of integrity of computation and data, unlike FHE.
✓ MPC does not suffer from side-channel attacks within the permitted number of compromised parties.	✗ Enclaves suffer from side-channel attacks. (Leveraging oblivious computation prevents many of these attacks.)
	✗ Enclaves have a large, trusted compute base (TCB).

**Table 3. Performance comparison.**


Cryptographic Computing (such as FHE and MPC)	Enclave/Confidential Computing
✗ MPC (and homomorphic encryption) is still very inefficient for complex computation.	✓ Enclave computation is much more efficient, sometimes close to vanilla processor speeds.

tographic computation was still too slow to run complex functions, such as ML training or rich data analytics. Take, for example, training a neural-network model. Recent state-of-the-art work on Microsoft Falcon (2021) estimates that training a moderately sized neural network such as VGG-16 on datasets such as CIFAR-10 could range into years. This work also assumes a threat model with three parties that have an honest majority, so a weaker threat model than the n organizations where n-1 can be malicious.


Now let us take an example with the stronger threat model: our state-of-the-art work on Senate,<sup>18</sup> which enables rich SQL data analytics with maliciously secure MPC. Senate improved the performance of existing MPC protocols by up to 145 times. Even with this improvement, Senate can perform analytics only on small databases of tens of thousands of rows and cannot scale to hundreds of thousands or millions of rows because the MPC computation runs out of memory and becomes very slow. We have been making a lot of progress on reducing the memory overheads in our recent work on MAGE<sup>13</sup> and in another work, Piranha, on employing GPUs for secure computation learning,<sup>22</sup> but the overheads of MPC remain too high for training advanced ML models and for rich SQL data analytics. It could still take years until MPC becomes efficient for these workloads.

Some companies claim to run MPC efficiently for rich SQL queries and ML training. How is that possible? An investigation of a few of them showed that they decrypt a part of the data or keep a part of the query processing in unencrypted form, which exposes that data and the computation to an attacker. This compromise reduces the privacy guarantee.

Hardware enclaves are far more efficient than cryptographic computation because, as explained earlier, deep down in the processor the CPU computes on unencrypted data. At the same time, data coming in and out of the processor is in encrypted form, and any software or entity outside of the enclave that examines the data sees it in encrypted form. This has the effect of computing on encrypted data



**Confidential computing promises to bring the benefits of generative AI to confidential data, such as the proprietary data of businesses, to increase their productivity, and the private data of users to assist them in various tasks.**



without the large overheads of MPC or homomorphic encryption. The overheads of such computation depend a lot on the workload, but there have been overheads of, for example, 20%—twice for many workloads.

Adding side-channel protection, such as oblivious computation, can increase the overhead, but overall the performance of secure computation using enclaves still is much better than MPC/homomorphic encryption for many realistic SQL analytics and ML workloads. The amount of overhead from side-channel protection via oblivious computation varies based on the workload—from adding almost no overhead for workloads that are close to being oblivious to 10 times the overhead for some workloads.

The Nvidia GPU enclaves<sup>16</sup> in the H100 architecture offer significant speed-ups for ML workloads, especially for generative AI. Indeed, there are significant industry efforts around using GPU enclaves to protect prompts during generative AI inference, data during generative AI fine-tuning, and even model weights during training of the foundational model. At the time of writing this article, Azure offered a preview of its GPU Confidential Computing service, and other major clouds have similar efforts under way. Confidential computing promises to bring the benefits of generative AI to confidential data, such as the proprietary data of businesses, to increase their productivity, and the private data of users to assist them in various tasks.

### Real-World Use Cases

Because of the need for data protection in use, there has been an increase in secure-computation use cases, whether it is cryptographic or hardware-enclave based. This section looks at use cases for both types.

**Cryptographic computation.** One of the main resources to track major use cases for secure multi-party computation is the MPC Deployments dashboard<sup>15</sup> hosted by the University of California, Berkeley. The community can contribute use cases to this tracker if they have users. A variety of deployed use cases are available for applications such as privacy-pre-


serving advertising, cryptocurrency wallets (Coinbase, Fireblocks, Dfns), private inventory matching (J.P. Morgan), privacy-preserving Covid exposure notifications (Google, Apple), and others.

Most of these use cases are centered around a specific, typically simple computation and use specialized cryptography to achieve efficiency. This is in contrast to supporting a more generic system, on top of which you can build many applications, such as a database, data-analytics framework, or ML pipeline—these use cases are more efficiently served by confidential computing.


One prominent use case was collecting Covid exposure notification information from users' devices in a private way. The organizations involved were Internet Security Research Group (ISRG) and National Institutes of Health (NIH). Apple and Google served as injection servers to obtain encrypted user data, and the ISRG and NIH ran servers that computed aggregates with help from MITRE. The results were shared with public health authorities. The computation in this case checked that the data uploaded from users satisfied some expected format and bounds, and then performed simple aggregates such as summation.

Heading toward a more general system based on MPC, Jana<sup>8</sup> is an MPC-secured database developed by Galois Inc. using funding from DARPA over four-and-a-half years and providing privacy-preserving data as a service (PDaaS). Jana's goal is to protect the privacy of data subjects while allowing parties to query this data. The database is encrypted, and parties perform queries using MPC. Jana also combines differential privacy and searchable encryption with MPC.

The Jana developers detail the challenges<sup>7</sup> they encountered, such as "Performance of queries evaluated in our linear secret-sharing protocols remained disappointing, with JOIN-intensive and nested queries on realistic data running up to 10,000 times slower than the same queries without privacy protection." Nevertheless, Jana was used in real-world prototype applications, such as inter-agency data sharing for public-policy devel-



**Because of its efficiency, confidential computing has been more widely adopted than cryptographic computation.**



opment, and in a secure computation class at Columbia University.

### Confidential Computing Use Cases

Because of its efficiency, confidential computing has been more widely adopted than cryptographic computation. The major clouds—Azure, AWS, and Google Cloud—offer confidential computing solutions. They provide CPU-based confidential computing, and some are in the process of offering GPU-based confidential computing (for example, Azure has a preview offering for the H100 enclave). A significant number of companies have emerged to enable various types of workloads in confidential computing in these clouds. Among them are Opaque, Fortanix, Anjuna, Husmesh, Antimatter, Edgeless, and Enclave.

For example, Opaque<sup>17</sup> enables data analytics and ML to run in confidential computing. Using the hardware enclave in a cloud requires significant security expertise. Consider, for example, that a user wants to run a certain data-analytics pipeline—say, from Databricks—in confidential VMs in the cloud. Simply running in confidential VMs is not sufficient for security: The user has to be concerned with correctly setting up the enclaves' remote attestation process, key distribution and management, a cluster of enclaves that offer scaling out, as well as defining and enforcing end-to-end policies on who can see what part of the data or computation.

To avoid this work for the user, Opaque provides a software stack running on top of the enclave hardware infrastructure that allows the user to run the workflow frictionlessly without security expertise. Opaque's software stack takes care of all these technical aspects. This is the result of years of research at UC Berkeley, followed by product development. Specifically, the technology behind Opaque was initially developed in the Berkeley RISELab (Realtime Intelligent Secure Explainable Systems),<sup>19</sup> and it has evolved to support ML workloads and a variety of data-analytics pipelines.

Opaque can scale to an arbitrary cluster size and big data, essentially creating one "large cluster enclave" out of individual enclaves. It enables collaboration between organiza-

tions or teams in the same organization that cannot share data with each other: These organizations can share encrypted data with each other in Opaque's workspace and perform data analytics or ML without seeing each other's dataset. Use cases include financial services (such as cross-team collaboration for identity resolution or cross-organization collaboration for crime detection); high-tech (such as fine-tuning ML from encrypted data sets); a privacy-preserving large language model (LLM) gateway that offers logging, control, and accountability; and generating a verifiable audit report for compliance.

A number of companies have created the Confidential Computing Consortium,<sup>2</sup> an organization meant to catalyze the adoption of confidential computing through a community-led consortium and open collaboration. The consortium lists more than 30 companies that offer confidential computing technology.

Following are a few examples of end use cases. Signal, a popular end-to-end encrypted messaging application, uses hardware enclaves to secure its private contact discovery service.<sup>21</sup> Signal built this service using techniques from the research projects Oblix<sup>14</sup> and Snoopy.<sup>4</sup> In this use case, each user has a private list of contacts on their device, and Signal wants to discover which of these contacts are Signal users as well. At the same time, Signal does not want to reveal the list of its users to any user, nor does it want to learn the private contact list of each user. Essentially, this computation is a private set intersection. Signal investigated various cryptographic computation options and concluded that these would not perform fast enough and cheaply enough for its large-scale use case. As a result, it chose to use hardware enclaves in combination with oblivious computation to reduce a large number of side channels, as discussed earlier. Our work on Oblix and Snoopy developed efficient oblivious algorithms for use inside enclaves.

Other adopters include the cryptocurrency MobileCoin, the Israeli Ministry of Defense,<sup>1</sup> Meta, ByteDance (to increase user privacy in TikTok), and many others.

## Combining the Two Approaches

Given the trade-off between confidential computing via enclaves and secure computation via cryptography, a natural question is whether a solution can be designed that benefits from the best of both worlds. A few solutions have been proposed, but they still inherit the slowdown from MPC.


For example, my students and I have collaborated with Signal with its SecureValueRecovery<sup>20</sup> system to develop a mechanism that helps Signal users recover secret keys based on a combination of different hardware enclaves on three clouds and secure multi-party computation. The purpose of this combination is to provide a strong security guarantee, stacking the power of the two technologies as defense in depth.

A similar approach is taken by Meta and Fireblocks, a popular cryptocurrency wallet: They both combine hardware enclaves with cryptographic computation for increased security. The resulting system will be at least as slow as the underlying MPC, but these examples are for specialized tasks for which there are efficient MPC techniques.

## Conclusions and How to Learn More

Secure computation via MPC/homomorphic encryption versus hardware enclaves presents trade-offs involving deployment, security, and performance. Regarding performance, it depends significantly on the target workload. For simple workloads, such as simple summations, low-degree polynomials, or simple ML tasks, both approaches can be ready to use in practice. However, for rich computations, such as complex SQL analytics or training large ML models, only the hardware enclave approach is practical enough at the moment for many real-world deployment scenarios.

Confidential computing is a relatively young sub-field in computer science—but one that is evolving rapidly. To learn more about confidential computing, attend or watch the content from the Confidential Computing Summit,<sup>3</sup> which was held in June of 2024 in San Francisco. This conference is the premier in-person event

for confidential computing and has attracted the top technology players in the space, from hardware manufacturers (Intel, ARM, Nvidia) to hyperscalers (Azure, AWS, Google), solution providers (Opaque, Fortanix), and use-case providers (Signal, Anthropic). The conference is hosted by Opaque and co-organized by the Confidential Computing Consortium. 

## References

1. Anjuna. Confidential computing pioneer Anjuna makes cloud safe enough for even government and defense agencies (2022); <https://bit.ly/4bCFAF1>.
2. Confidential Computing Consortium; <https://confidentialcomputing.io>.
3. Confidential Computing Summit 2024; <https://www.confidentialcomputingsummit.com>.
4. Dauterman, E. et al. Snoopy: surpassing the scalability bottleneck of oblivious storage. In *Proceedings of the ACM SIGOPS 28th Symp. on Operating Systems Principles* (2021); 10.1145/3477132.3483562.
5. Delignat-Lavaud, A. et al. Why should I trust your code?. *ACM Queue* 21, 4 (2023); <https://queue.acm.org/detail.cfm?id=3623460>.
6. Dhanuskodi, G. et al. Creating the first confidential GPUs. *ACM Queue* 21, 4 (2023); <https://queue.acm.org/detail.cfm?id=3623391>.
7. Galois. Galois team wraps up the Jana project (2020); <https://bit.ly/46179a7>.
8. Galois. Jana: Private data as a service (2024); <https://bit.ly/3LgGMTZ>.
9. Gentry, C. A fully homomorphic encryption scheme. *Ph.D. dissertation*, Stanford University (2009); <https://crypto.stanford.edu/craig/craig-thesis.pdf>.
10. Google Cloud. Confidential VM overview; <https://bit.ly/3XROdzz>.
11. Google Cloud. Encrypt workload data in-use with confidential Google Kubernetes engine nodes; <https://bit.ly/3XXeqpp>.
12. Kaplan, D. Hardware VM isolation in the cloud. *ACM Queue* 21, 4 (2023); <https://queue.acm.org/detail.cfm?id=3623392>.
13. Kumar, S., Culler, D.E., and Popa, R.A. MAGE: nearly zero-cost virtual memory for secure computation. In *Proceedings of the Usenix Symp. on Operating Systems Design and Implementation*; <https://bit.ly/45Wk78Z>.
14. Mishra, P. et al. Oblix: an efficient oblivious search index. In *Proceedings of the IEEE Symp. on Security and Privacy* (2018), 279–297; <https://people.eecs.berkeley.edu/~raluca/oblidx.pdf>.
15. MPC Deployments; <https://mpc.cs.berkeley.edu>.
16. Nvidia Confidential Computing; <https://bit.ly/3xFdYdI>.
17. Opaque; <https://opaque.co>.
18. Poddar, R. et al. Senate: A maliciously secure MPC platform for collaborative analytics. In *Proceedings of the 30th Usenix Security Symp.* (2021); <https://bit.ly/4cUrKIL>.
19. RISElab; <https://rise.cs.berkeley.edu>.
20. SecureValueRecovery2. Github; <https://bit.ly/3VX4ERw>.
21. Signal; <https://signal.org/blog/building-faster-oram/>.
22. Watson, J.-L., Wagh, S., and Popa, R.A. Piranha: A GPU platform for secure computation. In *Proceedings of the 31st Usenix Security Symp.* (2022); <https://www.usenix.org/system/files/sec22-watson.pdf>.
23. Yao, A.C.-C. How to generate and exchange secrets. In *Proceedings of the 27th Annual Symp. on Foundations of Computer Science* (1986), 162–167; <https://ieeexplore.ieee.org/document/4568207>.

**Raluca Ada Popa** is the Robert E. and Beverly A. Brooks associate professor of computer science at University of California, Berkeley, working in computer security, systems, and applied cryptography. She is also a co-founder of Opaque Systems, a confidential computing company.



This work is licensed under a Creative Commons Attribution International 4.0 License.

**If serverless platforms could wrap functions in database transactions, they would be a good fit for database-backed applications.**

BY QIAN LI AND PETER KRAFT

# Transactions and Serverless Are Made for Each Other

SERVERLESS CLOUD OFFERINGS are becoming increasingly popular for stateless applications because they simplify cloud deployment. This article argues that if serverless platforms could wrap functions in database transactions, they would also be a good fit for database-backed applications. There are two unique benefits of such a transactional serverless platform: time-travel debugging of past events and reliable program execution with “exactly-once” semantics.

Serverless cloud platforms, such as Amazon Web Services (AWS) Lambda and Azure Functions, are increasingly popular for building production applications as varied as website front ends, machine-

learning (ML) pipelines, and image-processing systems. These platforms radically simplify development by managing application deployment. Developers can deploy functions with the click of a button and the platform automatically hosts them, guarantees their availability, and scales them to handle changing loads.

Serverless platforms are primarily used for stateless operations, such as image resizing or video processing. Here, we argue they should also be used to deploy stateful applications, particularly *database-backed applications* whose business logic frequently queries and updates a transactional database such as Postgres or MySQL. Database-backed applications are ubiquitous in modern businesses; examples include e-commerce Web services, banking systems, and online reservation systems. They run primarily on server-based platforms such as Kubernetes. Thus, they form a massive opportunity for serverless offerings, including the back ends of most enterprise APIs and much of the modern Web.

To make serverless work for database-backed applications, serverless platforms would need to make one critical addition: *Allow developers to execute functions as database transactions.* Figure 1 shows an inventory reservation function implemented in a conventional serverless platform versus a transactional serverless platform. The `checkInventory` and `updateInventory` functions perform SQL queries. In a conventional serverless platform, if a function accesses the database, developers must obtain a database connection, manually begin a transaction, execute business logic and SQL queries, and then finally commit the transaction (Figure 1a).

By contrast, a *transactional* serverless platform manages the database connection: If a function accesses the database, it uses a platform-provided connection that automatically wraps the function in a transaction (Figure 1b). The idea of building such a platform has been explored in several re-



**Figure 1. An inventory reservation function implemented in a conventional serverless platform versus a transactional serverless platform. `checkInventory` and `updateInventory` perform SQL queries.**

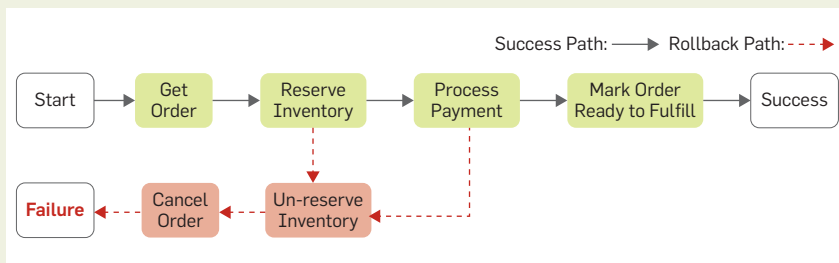
**(a) Conventional Serverless**

```
1 # Check if an item is available, then reserve it
2 def reserveInventory(itemId, num):
3     conn = getConnection(DBurl)
4     conn.beginTransaction()
5     avail = conn.checkInventory(itemId)
6     if (avail > num):
7         conn.updateInventory(itemId, avail - num)
8     conn.commitTransaction()
```

**(b) Transactional Serverless**

```
1 # Check if an item is available, then reserve it
2 def reserveInventory(itemId, num):
3     # Connection supplied by the platform
4     avail = conn.checkInventory(itemId)
5     if (avail > num):
6         conn.updateInventory(itemId, avail - num)
```

**Figure 2. Serverless checkout service workflow, including both success and rollback paths.**



search projects—by these authors<sup>1</sup> and others.<sup>3,4</sup>

As this article explains, a transactional serverless platform is not only more convenient for the developer but can also provide powerful benefits for database-backed applications beyond the capabilities of conventional serverless or server-based systems.

First, a transactional serverless platform makes programs easier to debug. Modern applications are difficult to debug because they run in distributed settings with frequent concurrent accesses to shared state, so bugs often involve complex race conditions that are not easy to reproduce in a development environment. Reproducing errors is particularly difficult in conventional serverless platforms because their execution environment is transient and exists only in the cloud. A transactional serverless platform, however, can simplify debugging through *time travel*.<sup>2</sup>

Because the platform wraps functions in transactions to coordinate their state accesses, a debugger can leverage the transaction log to “travel back in time” and locally replay any past transactional function execution.

Second, a transactional serverless platform can provide reliable program execution. Writing reliable database-backed applications is difficult because they often coordinate several business-critical tasks, any of which may fail. In a server-based application, addressing this problem is difficult, as developers must manually track each request’s status and recover failed requests. Conventional serverless platforms make this easier by automatically restarting any task that fails, but this can be problematic if it causes an operation to execute multiple times (for example, paying twice). If functions are transactions, however, the platform can record their success or failure *in the same transac-*

*tion as their business logic*, thus guaranteeing each function executes once and only once.

## Programming a Transactional Serverless Platform

A transactional serverless platform could provide a programming model similar to today’s serverless platforms, where developers write programs as *workflows of functions*. Each function performs a single operation. Workflows, implemented as directed graphs or state machines, orchestrate many functions. Popular serverless workflow orchestrators include AWS Step Functions and Azure Durable Functions.

The distinguishing feature of a transactional serverless platform is that all functions accessing the application database are wrapped in atomic, consistent, isolated, and durable (ACID) database transactions, as shown in Figure 1b. These functions must be deterministic and have no side effects outside the database. Functions not accessing the database, such as those making external API calls, work the same as they do in conventional serverless platforms.

As a running example for this article, Figure 2 shows a diagram of a serverless checkout service workflow that first reserves inventory for all items in an order, then processes payment for the order, and finally marks the order as ready to fulfill. Each step is implemented in a separate function. All functions except “process payment” (which uses a third-party payment provider) contact the database and are wrapped in transactions. If any step fails, the workflow runs rollback functions to undo previous operations (for example, returning reserved inventory if the payment fails).

## Time-Travel Debugging

One powerful and unique feature enabled by a transactional serverless platform is time-travel debugging: letting developers faithfully replay production traces in a local development environment to reproduce bugs that happened in the past. Time-travel debugging is especially useful for database-backed applications because they frequently run in distributed environments where bugs manifest as race conditions that occur only under high concurrency and are nearly impossible to reproduce locally.

For example, suppose the “reserve



inventory” operation in Figure 2 is split into two separate transactional functions, as in Figure 3, which shows a buggy implementation of the “reserve inventory” operation. This implementation contains a race condition where, if two requests arrive at the same time, both can reserve the same item—potentially causing the vendor to sell more items than it has available.

Debugging issues such as this is tricky because they surface only if multiple concurrent requests with specific inputs are interleaved in a specific way with a particular database state. To reproduce the bug locally, the developer must determine not only which requests caused the bug, but also the order in which different operations in those requests interleaved and the exact database state that made the bug possible. In a conventional platform, tracking execution order and reconstructing database state are prohibitively expensive: Requests execute concurrently on many parallel threads on many distributed servers, potentially modifying the database thousands of times per second.

By contrast, prior research<sup>2</sup> has shown that a transactional serverless platform makes faithful replay practical because each function is wrapped in an isolated, atomic, and deterministic transaction. This enables a *time-travel debugger*, which can faithfully replay a production trace (including race conditions and concurrency bugs) in two steps:

1. Using database transaction logs, it can reconstruct the state of the application database at the time of the trace’s first request.

2. It can locally execute each request in the trace on the reconstructed database, executing their transactional functions in the order they originally executed in the application database’s transaction log.

A time-travel debugger improves developers’ lives by reproducing complex concurrency bugs in a controlled local environment. For example, if the debugger is run on a trace containing the bug described in Figure 3, it executes both check transactions on a database containing only one item, and then executes both update transactions, thus overselling the item and reproducing the bug. This process is shown in Figure 4.

A time-travel debugger can provide

another powerful feature called *retroaction*: the execution of modified code over past events. For a given trace, the debugger performs retroaction similarly to faithful replay but uses the updated implementation of each function instead of the original one. Retroaction is especially useful for regression testing: running a new code version over old production traces to verify it handles them correctly. For example, assume the bug in Figure 3 was fixed by combining the check and update functions into a single transactional function. A time-travel debugger can retroactively test this fix by re-executing the original trace but running the combined function in place of the original checks and updates. As shown in Figure 5, this validates that the fix eliminates the bug.

### Reliable Program Execution

Another key benefit of a transactional

serverless platform is *reliable program execution*. Many database-backed applications must coordinate multiple business-critical tasks, any of which may fail. For example, the checkout workflow in Figure 2 performs three tasks for each order:

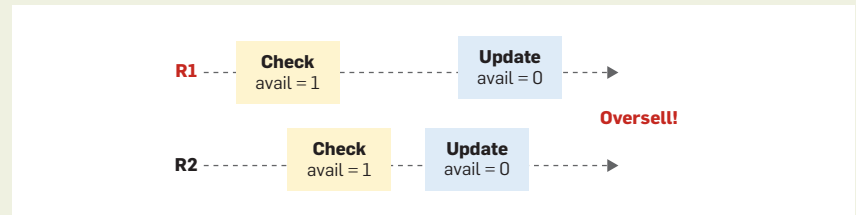
1. Reserving its inventory
2. Processing its payment
3. Marking it as ready to fulfill

To execute reliably, such applications must not only handle failures in any of those tasks but also recover from interruptions such as server crashes. Specifically, they require two properties:

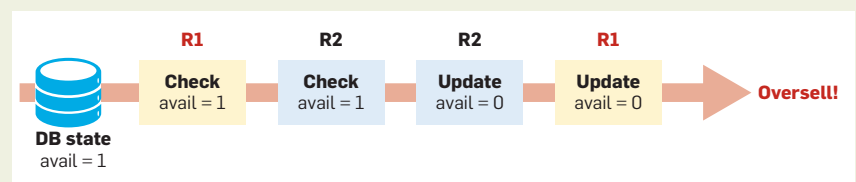
► **Programs run to completion.** If a program begins executing, it must continue, recovering through any interruptions until it reaches a terminal success or failure state. For example, if the checkout service is interrupted after processing a payment, it must recover and either mark the order as fulfilled (if

**Figure 3. A buggy implementation of the “reserve inventory” operation. Two concurrent requests both try to reserve the same item and both succeed, causing overselling.**

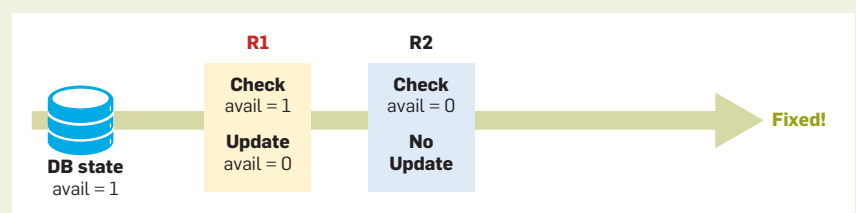
```
1 def reserveInventory(itemId, num):
2   avail = execTxn(checkInventory(itemId))
3   if (avail > num):
4     execTxn(updateInventory(itemId, avail - num))
```



**Figure 4. A time-travel debugger replaying an execution trace containing the reserve inventory bug.**



**Figure 5. A time-travel debugger testing a fix to the reserve inventory bug using retroaction.**



the payment succeeded) or cancel the order and return reserved inventory (if the payment failed).


► **Operations execute exactly once.** While executing a program, each of its operations must execute once and only once. For example, if you are recovering the checkout service after it is interrupted, you cannot naively re-send the payment request; otherwise, the customer may pay twice. You must instead determine the status of the original payment request (whether it was sent at all, and if so, whether it succeeded or failed) and recover accordingly.

Manually obtaining these properties in a traditional server-based application is difficult. One approach is to write the application as a state machine that checkpoints its state to persistent storage after every operation. If the program is interrupted, resume execution from the last checkpointed state. To ensure “exactly-once” execution, make all operations idempotent so they can be safely re-executed during recovery. While such an approach works, it is tedious, error-prone, and requires careful program design.


Existing serverless platforms simplify writing programs that run to completion but do not provide “exactly-once” execution. This follows naturally from the serverless programming model. If a program is written as a workflow of functions, the workflow orchestrator can record the workflow’s state after every function execution, then resume from the last recorded state if workflow execution is interrupted. Thus, serverless function orchestrators such as AWS Step Functions and Azure Durable Functions run workflows to completion, restarting each function until it succeeds or reaches a predefined failure state.

Durable workflow engines such as Temporal provide similar guarantees for server-based programs, provided they are written as workflows of operations. Because orchestrators treat functions as black boxes, however, they cannot provide “exactly-once” semantics, but instead restart each function until it succeeds. If a function crashes after completion but before its success is recorded, it is re-executed, potentially corrupting data.

As prior work has shown,<sup>1,4</sup> a transactional serverless platform can guar-



**By tightly integrating application execution and data management, a transactional serverless platform enables many new features not possible in either existing serverless platforms or server-based deployments.**




antee not only that programs run to completion, but also that transactional operations execute exactly once. Because the platform wraps functions in transactions, it can record the success or failure of a transactional function *in the same transaction as the function*. Therefore, if a function completes, its success or failure is always recorded in the database, while if a function fails, all its actions are rolled back by the database. Thus, the platform knows never to re-execute a function with a recorded result but can always safely re-execute without a recorded result.

### Conclusion

Database-backed applications are an exciting new frontier for serverless computation. By tightly integrating application execution and data management, a transactional serverless platform enables many new features not possible in either existing serverless platforms or server-based deployments.

This article has explained how such a platform could benefit application debuggability and reliability. Its additional benefits include:

- *Observability*, as the platform can track the full history (provenance) of each data item through all functions that have modified it.
- *Security*, as the platform can monitor all operations on data in real time.
- *Performance*, as the platform can collocate transactional functions with the application database.

We look forward to future work in this space. 

### References

1. Kraft, P. et al. Apiary: a DBMS-integrated transactional function-as-a-service framework. *arXiv:2208.13068* (2023); <https://arxiv.org/abs/2208.13068>.
2. Li, Q. et al. R<sup>3</sup>: Record-replay-retroaction for database-backed applications. In *Proceedings of the VLDB Endowment* 16, 11 (2023), 3085–3097; <https://dl.acm.org/doi/10.14778/3611479.3611510>.
3. Wu, C., Sreekanti, V., and Hellerstein, J.M. Transactional causal consistency for serverless computing. In *Proceedings of the 2020 ACM SIGMOD Intern. Conf. Management of Data*, 83–97; <https://bit.ly/3NHsxIT>.
4. Zhang, H. et al. Fault-tolerant and transactional stateful serverless workflows. In *Proceedings of the 14<sup>th</sup> Usenix Symp. on Operating Systems Design and Implementation* (2020), 1187–1204; <https://bit.ly/4eXHuCE>.

**Qian Li**, a co-founder at DBOS, Inc., earned her Ph.D. in computer science from Stanford University, Stanford, CA, USA.

**Peter Kraft**, a co-founder at DBOS, Inc., earned his Ph.D. in computer science from Stanford University, Stanford, CA, USA.

© 2024 Copyright held by the owner/author(s).  
Publication rights licensed to ACM.



## ACM Thanks Chapters for Participating in Hour of Code

### Week-long event inspires next generation of computer scientists

The Hour of Code is a global movement designed to generate excitement in young people about programming and technology. Games, tutorials, and other events were organized during Computer Science Education Week around the world. ACM would like to thank the following chapters who participated this year:

- ABES ACM Student Chapter, India
- ABES ACM-W Student Chapter, India
- Ajay Kumar Garg Engineering College ACM Student Chapter, India
- Bilkent University ACM Student Chapter, Turkey
- Bilkent University ACM-W Student Chapter, Turkey
- Bucknell University ACM Chapter, USA
- Capital University ACM Student Chapter, USA
- Capital University ACM-W Student Chapter, USA
- Chitkara ACM Student Chapter, India
- Chitkara University ACM Student Chapter, India
- DHA Suffa University ACM Student Chapter, Pakistan
- FISAT ACM Student Chapter, India
- GGSIP University USS ACM Student Chapter, India
- Hacettepe University ACM Student Chapter, Turkey
- Henderson State University, USA
- IIT Jodhpur ACM Student Chapter, India
- Iowa State University Chapter (SIGCHI), USA
- JUST ACM Student Chapter, Jordan
- KARE ACM Student Chapter, india
- Koc University ACM Student Chapter, Turkey
- Lander College for Women ACM Student Chapter, USA
- Louisiana Tech University ACM Student Chapter, USA
- Manipal University Jaipur ACM Student Chapter, India
- Mapua Malayan Colleges Laguna ACM Student Chapter, Phillippines
- Maribor ACM Student Chapter, Slovenia
- MUJ ACM SIGAI Student Chapter, India
- MUJ ACM SIGBED Student Chapter, India
- National Chi Nan University ACM Student Chapter, Taiwan
- National Institute of Technology Karnataka, Surathkal ACM Student Chapter, India
- NIT Surat ACM Student Chapter, India
- NUCES KHI ACM Student Chapter, Pakistan
- NUST ACM Student Chapter, Pakistan
- PMU ACM Student Chapter, Saudi Arabia
- RAIT ACM Student Chapter, India
- RVRJCCE ACM Student Chapter, India
- SNTD ACM-W Student Chapter, India
- SSN ACM Student Chapter, India
- SVCE ACM SStudent Chapter, India
- TCE IT ACM Student Chapter, India
- Texas Christian University ACM Student Chapter, USA
- Tufts University ACM SIGGRAPH Student Chapter, USA
- UCSC ACM Student Chapter, Sri Lanka
- University of Belize ACM Student Chapter, Belize
- University of Houston ACM-W Student Chapter, USA
- University of Moratuwa ACM Student Chapter, Sri Lanka
- University of Ottawa ACM-W Student Chapter, Canada
- University of Porto Faculty of Engineering ACM Student Chapter, Portugal
- University of the Philippines ACM Student Chapter, Phillippines
- University of West Florida ACM Student Chapter, USA
- UPES ACM Student Chapter, India
- UPES ACM-W Student Chapter, India
- UTAS-AI Mussanah ACM Student Chapter, Oman
- Valley City State University ACM Student Chapter, USA
- Vellore Institute of Technology University, Chennai Campus ACM-W Student Chapter, India
- VVCE ACM Student Chapter, India
- VVIT GUNTUR ACM Student Chapter, India
- VVIT GUNTUR ACM-W Student Chapter, India



DOI:10.1145/3665322

**As the impact of AI is difficult to assess by a single group, policymakers should prioritize societal and environmental well-being and seek advice from interdisciplinary groups focusing on ethical aspects, responsibility, and transparency in the development of algorithms.**

BY ALEJANDRO BELLOGÍN, OLIVER GRAU, STEFAN LARSSON, GERHARD SCHIMPF, BISWA SENGUPTA, AND GÜRKAN SOLMAZ

# The EU AI Act and the Wager on Trustworthy AI

ARTIFICIAL INTELLIGENCE (AI) systems are increasingly supplementing or taking over tasks previously performed by humans. On the one hand, this relates to low-risk tasks, such as recommending books or movies, or recommending purchases based on previous buying behavior. But it also includes crucial decision making by highly autonomous systems. Many current systems are opaque in the sense that their

internal principles of operation are unknown, leading to severe safety and regulation problems. Once trained, deep-learning systems perform well, but they are subject to surprising vulnerabilities when confronted with adversarial images.<sup>9</sup>

The decisions may be explicated after the fact, but these systems carry the risk of wrong decisions affecting the well-being of people. They may be discriminated against, disadvantaged, or seriously injured. Examples include suggestions on how to select a job applicant, proper medical treatment for a patient, or how to navigate autonomous cars through heavy traffic. In such situations, several ethical, legal, and general societal challenges arise. At the forefront is the question of who is responsible for a decision made by an AI system: Do we leave the decision to the AI system, or does a human decide in partnership with an AI system? Are there reliable, trustworthy, and understandable explanations for the decisions in each case? Unfortunately, the inner workings of many AI systems remain hidden—even from experts. Given the critical role AI systems play in modern society, this seems in many cases unacceptable. But how can we make complex, self-learning systems explainable? And to what extent is this lack of explanation or broader transparency contributing to a watchful and responsible introduction of AI systems that have evidenced benefits?

## » key insights


- **Public trust, transparency, and interdisciplinary research are pivotal in the responsible deployment of AI systems.**
- **The EU AI Act passed by the European Parliament will now be implemented in 27 member states of the EU. It is the first major law aimed at regulating AI across sectors, with a focus on risk management, transparency, ethical governance, and human oversight.**
- **AI systems categorized as high risk will be subject to stringent regulations to ensure they do not compromise human rights or safety.**




A deeper look at the technical details of AI and technical innovations on their way, such as autonomous systems, shows an obvious need for technical expertise in the practical technical and societal aspects of AI in the decision-making process. On the other hand, a purely technological perspective may result in regulations that cause more significant societal problems. This article highlights accurate and realistic technology descriptions that take into account the risk factors as required, for example, by the risk pyramid of the EU AI Act that entered into force in August 2024. To strike such a balance for the public interest, policymakers should prioritize societal and environmental well-being and seek advice from interdisciplinary groups, as the impact of AI and autonomous systems is very difficult to assess by a single group. This more holistic system view is complementary to previous statements focusing on ethical aspects, responsibility, and transparency in the development of algorithms,<sup>1</sup> specifically on algorithmic systems involving AI and machine learning (ML).<sup>3,15,22</sup>

Many members of the public, particularly in Europe, exhibit skepticism toward AI and autonomous systems, which often translates into a lack of confidence or a cautious “wait-and-see” approach.<sup>23</sup> For this technology to develop to its beneficial potential, we need a framework of rules within which all players can operate responsibly. For the future of AI systems—specifically in the public spheres, where people express their personal expectations and worries about the potential consequences of AI being used without proper oversight—certain aspects must be taken into account. The following points are crucial for guiding the formulation of policies and regulations related to AI and essential for the research and development community:

**Supporting research and development in AI and autonomous systems.** We recommend advanced research on the governance of implemented AI and automated systems, for example, transportation. Special care must be taken at an early stage to contribute and adhere to transparent standards for hardware and software that provide



**Many current systems are opaque in the sense that their internal principles of operation are unknown, leading to severe safety and regulation problems.**



insight to carry out legally required independent safety certifications.

**Creating and supporting sustainable solutions.** In light of the UN sustainability development goals, we recommend advancing multidisciplinary research methodologies that integrate social sciences and humanities alongside engineering sciences. Social sciences, such as sociology and anthropology, can provide crucial insights into how people understand, interact with, and trust AI systems. This understanding is vital for designing technologies that are socially acceptable, beneficial, and promote sustainable development. Humanities disciplines, like philosophy, can offer valuable perspectives on ethics, fairness, and the potential impact of AI on human values. This combined approach can lead to developing sustainable and energy-efficient autonomous systems that align with societal well-being.

**Prioritizing societal well-being and equal opportunities.** We recommend that the legislative processes, especially in adapting existing laws and the new design of liabilities, take an interdisciplinary approach and consult the scientific and technical expertise in trusted AI. This should ideally lead to equal opportunities and fairness in new business development considering new autonomous systems and preventing monopolies.

**Promoting education on science, technology, social impact, and ethics.** To foster responsible and beneficial use of AI, we propose enhancing educational curricula in secondary schools, universities, and technical fields to include fundamental knowledge about AI ethics and its impact on society. Incorporating ethical and social scientific aspects into computer science (CS) curricula, as exemplified by Stanford University’s approach, will encourage students to consider “embedded” ethical, legal, or social implications while solving problems. Similarly, in Europe, some institutions teach CS students to relate the ACM Code of Ethics for Professional Conduct<sup>1</sup> to their tasks, fostering a sense of responsibility in their future AI-related endeavors.

The overall level of expertise in all levels of our society about how AI works and operates represents a critical success factor that will ultimately lead to

confidence and acceptance of beneficial uses of these technologies in our daily lives. Policymakers, developers, and adopting users of AI systems need to be literate about these technologies and find answers at the intersection of technology, society, and policymaking. Furthermore, we should weigh the risks of autonomous systems against the benefits to allay public fears.

The points mentioned here highlight the need for an interdisciplinary and holistic approach to beneficial usage of AI. They set the foundation for a broader involvement of the public on one hand and the subsequent development of the EU AI Act. To be informed about the endeavors of a supranational governmental organization such as the EU, striving to establish consensus across 27 member states regarding the legal regulation of AI, is likely to capture the attention of a diverse international readership. This audience includes academics in the field of AI ethics, explainable AI, and risk management as well as professionals who may be called upon to provide technical expertise to lawmakers in other parts of the world.

### Background: EU Policies on AI and Ethics Guidelines

Considered one of the ‘lighthouse’ projects, public trust in autonomous systems is a crucial issue, well in line with recent awareness in the governance over AI<sup>12,16,21</sup> expressed in the joint agreement of the EU Commission and EU Council’s proposal for a new European AI Act,<sup>a</sup> as well as the High-Level Expert Group called in by the EU Commission in 2019.<sup>8</sup> The High-level Expert Group’s Ethics Guidelines echo several critical issues on human-centered and transparent approaches pointed to several principled documents.<sup>13</sup>

The EU Commission takes a three-step approach: setting out the essential requirements for trustworthy AI, launching a large-scale pilot phase for feedback from stakeholders, and working on international consensus-building for human-centric AI.<sup>b</sup> Among others, the ACM Europe Technology Policy Council (TPC)<sup>2</sup> collabo-

rates with the EU Commission as a stakeholder and representative of the European CS community, providing technical input on relevant initiatives. While the Commission looks broadly at an assessment of AI from a general point of view to preserve the values of the European member states, a more comprehensive judgment will result if the predictive assessments of all the actors, that is, owners, designers, developers, and researchers, are taken into account.<sup>1,3,5,22</sup> This process led to the proposal of an AI Act, first published by the European Commission in April 2021, and the final version in force starting August 2024, which this article discusses later.

### Essentials for Achieving Trustworthy AI Systems

Implementers of AI and autonomous systems must be aware of what we, as responsible citizens, can accept and what is ethical, and put laws and regulations in place to safeguard against future tragedies. Trustworthy AI should, for example, according to the European Commission’s High-Level Expert Group on AI, respect all applicable laws and regulations and a series of requirements for the particular sector. Specific assessment lists aim to help verify the application of each essential requirement. The following list of essentials is taken from the EU document “Building Trust in Human-Centric Artificial Intelligence,” which results from the work of a European High-Level Expert Group on ethics.<sup>c</sup> Additional perspectives are covered in a report by the Alan Turing Institute.<sup>18</sup>

**Developing trust in autonomous systems in the public sphere.** *Human agency and oversight.* The essentials described above in the direction of an explainable and trustworthy AI may be suitable to convince professionals who knowingly interact with AI systems.<sup>6,7</sup> It would be similarly essential to ensure trust in these systems among the public. However, it is important to note that explainability in AI, particularly in deep neural networks (DNNs), remains a significant scientific challenge. Some scientists argue that the inherent complexity and the high-dimensional nature of these models make it difficult,

if not impossible, to fully explain their outcomes. This skepticism raises critical questions about the feasibility of achieving truly transparent AI systems.

Therefore, ways to establish an individual trust in AI must be sought. However, more than detailed explanations of individual outcomes will be required for the public. In Knowles and Richards,<sup>14</sup> the authors call for building a public regulatory ecosystem based on traceable documentation and auditable AI, with a slightly different emphasis than the one on individual transparency and information for all.

*Robustness and verification.* Given the complexity, more work needs to be done by interdisciplinary teams bringing together the social sciences and humanities expertise with computer scientists, software engineers, legal scholars, and political scientists in investigating what meaningful control and verification procedures for AI systems might look like in the future.

*Safety, risk issues, and ethical decisions.* In the domain of autonomous vehicles, looking at the state of the art to avoid collisions, autonomous cars have been trained not only to respect traffic rules rigorously but also to ‘drive cautiously’, that is, negotiate and not enforce the right of way. Even in the case of unavoidable and dilemmatic situations, legislation is underway to respect the ethical dilemma, a.k.a. the Trolley Dilemma, investigated in Awad et al.<sup>4</sup> and Goodall.<sup>11</sup>

In the context of public expectations, it is important to understand that there is no universally “right” answer when it comes to making decisions in dilemma situations. Primarily, an algorithm should not be constrained to making predefined decisions. Nevertheless, ongoing discussions about this topic persist in society. Furthermore, the lack of acceptance for autonomous driving can be attributed to the fact that humans are allowed to make mistakes, whereas there seems to be zero tolerance for any mistakes made by AI.

*Cybersecurity.* In the cybersecurity domain, other than attacks through the Internet, there are also AI-specific attacks, such as adversarial learning, which researchers have successfully demonstrated from the Tencent Keen

a See <https://bit.ly/4gP6j5d>

b See <https://bit.ly/4eL8kNK>

c See <https://bit.ly/3XTZ5nL>

Security Lab.<sup>9</sup> AI systems such as autonomous vehicles must demonstrably be able to defend themselves and go into a safe mode in case of doubt.

*Physical security.* There might be physical attacks, such as throwing a paint bag against the cameras to blind an autonomous system or using a laser pointer against the LiDAR. In cases like these, error handling must be able to bring the system into a safe mode.

*Data privacy.* People have the right to determine if they want to be “filmed” and whether they want their location, date, and time to be recorded and shared. To build trust, autonomous systems manufacturers must adhere to the data-protection principles in the GDPR to ensure that no privacy rights are being violated.

*Trust and human factors.* Different levels of trust and comfort may arise through explanation, for example, if an autonomous car explains its maneuvers to its passengers and road users outside the vehicle.

*Trust and legal systems.* The decisive question is who or what caused the error: The human at the wheel? A flawed system? A defective sensor? Complete digitization makes it possible to answer these questions. To do this, however, extensive data must be stored. Open legal questions that need to be clarified in this context include who owns these data, who has access to the data, and whether this is compatible with privacy protection.

*Public administration.* The answers to the above must be found because they represent significant citizen concerns. In our capacity as members of the ACM Europe TPC, we contribute to the work by the EU Commission and EU Parliament to establish harmonized rules for the use of AI. Our comments from the perspective of autonomous systems can be found in Saucedo et al.<sup>20</sup>

### AI Legislation in the EU: The AI Act

AI policy work is underway globally in most industrial countries. Partnering with PricewaterhouseCoopers, the Future of Life Institute offers a dashboard on its website<sup>24</sup> with a wealth of information and references to documents. According to their analysis, the approach to govern AI varies greatly between soft and hard law efforts,

which depends largely on how the following areas of concern are rated and prioritized by policymakers:

- ▶ Global governance and international cooperation
- ▶ Maximizing beneficial AI research and development
- ▶ Impact on the workforce
- ▶ Accountability, transparency, and explainability
- ▶ Surveillance, privacy, and civil liberties
- ▶ Fairness, ethics, and human rights
- ▶ Manipulation
- ▶ Implications for health
- ▶ National security
- ▶ Artificial general intelligence and superintelligence

Looking at the major players, we see:

▶ **United States.** The White House has published a ‘Blueprint for an AI Bill of Rights’, a set of five principles and associated practices to help guide the design, use, and deployment of automated systems to protect the rights of the American public in the age of AI. However, there is currently no federal AI regulation in the U.S., but some states have taken steps to regulate particular use cases and the use of AI in specific industries. For example, California passed a law requiring companies to disclose the use of automated decision making in employment and housing. Overall, the strategy is business oriented. After the appearance of ChatGPT, the U.S. Senate Committee on the Judiciary’s Subcommittee on Privacy, Technology and the Law held several hearings with leading AI academics to evaluate the risks of generative AI. In October 2023, the Executive Order on Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence, signed by President Biden, arose from a desire to address both the potential benefits and risks of AI.

▶ **China.** China has been actively investing in AI and has taken steps to regulate its use, including developing national AI standards and guidelines for ethical use. The country has also established a national AI development plan that sets out its goals and objectives for the industry. China has significantly restricted the use of generative AI. ChatGPT is blocked within the Chinese network, and access to domestic alternatives is granted solely through

individual application requests.

▶ **Canada.** Canada has established the Pan-Canadian Artificial Intelligence Strategy, which aims to promote the responsible development and use of AI. The strategy includes funding for research, development, and innovation in AI, as well as ethical guidelines for its use.

▶ **United Kingdom.** The U.K. has established the AI Council, which aims to promote the responsible use of AI and advise the government on AI regulation. The council has published guidelines on ethical use. The approach so far aims to ensure consumers “have confidence in the proper functioning of the system.”

▶ **The G7.** During its summit meeting on May 20, 2023 in Hiroshima, the G7 issued a statement about what it called the ‘Hiroshima AI Process’.

“We recognize the need to immediately take stock of the opportunities and challenges of generative AI, which is increasingly prominent across countries and sectors, and encourage international organizations to consider analysis on the impact of policy developments and Global Partnership on AI (GPAI) to conduct practical projects. In this respect, we task relevant ministers to establish the Hiroshima AI process, through a G7 working group, in an inclusive manner and in cooperation with the OECD and GPAI, for discussions on generative AI by the end of 2024. These discussions could include topics such as governance, safeguard of intellectual property rights including copy rights, promotion of transparency, response to foreign information manipulation, including disinformation, and responsible utilization of these technologies.”<sup>10</sup> In October 2023, this was followed by the publication of AI guidelines for a ‘Hiroshima Process’ for advanced AI systems and a code of conduct for developer organizations.

In the EU, preparations for AI regulation began in April 2021, when the EU Commission presented the Artificial Intelligence Act, which sets out horizontal rules for the development, commodification, and use of AI-driven products, services, and systems within the territory of the EU. It should be noted that the EU AI legislation does not regulate AI technology per se, but rather the effect of AI products on the



lives of EU citizens. There is no intention to intervene in the development of AI products, but there is a claim to help shape their use in the EU. The regulation provides core AI rules that apply to all industries.

The EU AI Act introduces a sophisticated ‘product safety framework’ constructed around four risk categories as evidenced in the figure. It imposes requirements for market entrance and certification of high-risk AI systems through a mandatory CE-marking procedure. To ensure equitable outcomes, this pre-market conformity regime also applies to ML training, testing, and validation datasets. The Act seeks to codify the high standards of the EU trustworthy AI paradigm, which requires AI to be legally, ethically, and technically robust while respecting democratic values and human rights, including privacy and the rule of law.

This is claimed to be the first law worldwide to regulate AI in all areas of life, except the military sector. The legislative process reached a milestone in December 2023, when the EU Commission, the EU Council, and Parliament managed to reach an agreement in the so-called “trilogue.” After subsequent approval from votes in Parliament and the Council, the regulation came into force in August 2024, shifting the attention to member states to set up supervisory bodies, the standardization bodies to develop harmonized standards for high-risk AI compliance, and for the new AI Office to develop guidelines.

**Who is affected by the new regulation?** Companies that plan to provide or deploy AI systems in the EU (the “providers and deployers” according to the wording of the Act) are the primary addresses bound by the provisions of the AI Act. They apply regardless of where the systems were developed or are operated from—or when the operation of the systems impacts EU citizens. It will take courage and creativity to legislate this convoluted, interdisciplinary issue and will require non-EU, namely U.S. and Chinese companies, to adhere to values-based EU standards before their AI products and services gain access to the European market of 450 million consumers. Consequently, the proposal has an extraterritorial effect.

Figure. The EU Risk Pyramid.

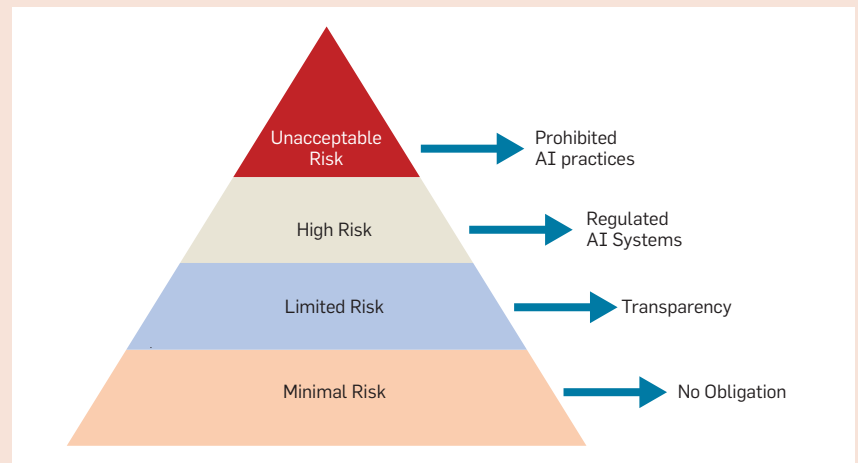


Table 1. Contents of the EU AI Act.

Chapter I: General Provisions	Outlines the proposal’s scope and how it would affect the market once in place.
Chapter II: Prohibited AI Practices	Defines AI systems that violate fundamental rights and are categorized at an unacceptable level of risk.
Chapter III: High-Risk AI Systems	Covers the specific rules for classifying AI systems as high risk, the connected requirements and obligations for Providers and Deployers and other parties.
Chapter IV: Transparency Obligations for Providers and Deployers of Certain AI Systems and GPAI Models:	Lists transparency obligations for systems that interact with humans, detect emotions, or determine social categories based on biometric data, or generate or manipulate content (for example, ‘deep fakes’).
Chapter V: General Purpose AI Models	Classification Rules, obligations for providers of general-purpose AI Models, and GPAI Models with systemic risk.
Chapter VI: Measures in Support of Innovation	AI regulatory sandboxes, testing of high-risk AI systems in real world conditions.
Chapter VII: Governance	Establishing the Act’s governance systems, including the AI Office and the AI Board, and monitoring functions of the European Commission and national authorities.
Chapter VIII: EU Database for High-Risk AI Systems	EU database for high-risk AI systems listed in Annex III.
Chapter IX: Post-Market Monitoring, Information Sharing, Market Surveillance	Sharing of information on serious incidents: Supervision, investigation, enforcement, and monitoring with respect to providers of general-purpose AI models.
Chapter X: Codes of Conduct and Guidelines	Guidelines from the Commission on the implementation of this regulation.
Chapter XI: Delegation of Power and Committee Procedure	Exercise of the delegation and committee procedure.
Chapter XII: Confidentiality and Penalties	Administrative fines on union institutions, agencies and bodies. Fines for providers of general-purpose AI models.
Chapter XIII: Final Provisions	Amendments to several articles in other legislation.

Given the need for more awareness outside the EU, companies are well advised to start early to learn what is in the EU AI Act and what is needed to meet the compliance criteria.

**The essence of the EU AI Act.** The AI act contains the following sections, called titles.<sup>4,13</sup> A collection of all publicly available documents and amend-

ments since the initial proposal to the AI Act as of July 2023 may be found in Zenner.<sup>25</sup>

**The risk pyramid of the AI Act.** The main guiding point of the AI Act is the risk pyramid with a core focus on high-risk applications. The risk levels, as depicted previously in Figure 1, are summarized below.

*Unacceptable risk.* This category delineates which uses of AI systems

d See <https://bit.ly/4dEOJOh>

carry an unacceptable level of risk to society and individuals and are thus prohibited under the law. These prohibited use cases include AI systems that entail social scoring, subliminal techniques, biometric identification in public spaces, and exploiting people's vulnerabilities. In these uses, the AI Act describes when and how exceptions may be made, such as in emergencies related to law enforcement and national security.

*High risk.* Requirements related to high-risk systems, such as compliance with risk-mitigation requirements like documentation, data safeguards, transparency, and human oversight, are at the crux of this proposed regulation. The list of high-risk AI systems that must deploy additional safeguards is lengthy and can be found in Art. 6, Annex III of the Act.

Explainability plays a crucial role in ensuring that AI systems are transparent and trustworthy, particularly in domains where the risk of harmful decisions is high—for example, in the medical domain, where a false negative may be as harmful as a false positive. The EU AI Act requires that AI systems provide information on their decision-making process so that individuals can understand the basis for the AI system's outputs and that they are not used to manipulate behavior. Additionally, the requirement for human oversight and control over high-risk AI systems is based on the principle that there must be a human in the loop to make decisions that have significant consequences for individuals' rights and safety.<sup>19</sup> The EU AI Act aims to ensure that AI systems are developed and deployed responsibly and transparently, considering the potential impact on individuals' rights and safety. Harmonized standards, under development, are likely to play an important role for the compliance of high-risk AI systems.

*Limited risk.* Limited-risk AI systems have much fewer obligations to providers, and users must follow compared to their high-risk counterparts. AI systems of limited risk must follow certain transparency obligations outlined in Title IV of the proposal. Examples of systems that fall into this category include biometric categorization, or establishing whether the

biometric data of an individual belongs to a group with some predefined characteristic to take a specific action; emotion recognition; and deep-fake systems.

*Minimal risk.* The proposal's language describes minimally risky AI systems as all other systems not covered by its safeguards and regulations. There are no requirements for systems in this category. Of course, businesses with multiple kinds of AI systems must ensure compliance with each appropriately.

**Handling general-purpose AI with or without systemic risk.** As a result of the increased general capabilities of several new AI models during the spring of 2023, and the broad adoption of ChatGPT, there were intense public debates and a delay of the EU Parliament's proposal for the AI Act. The proposal, from June 2023, came to include rules that the earlier proposals did not, on "foundation models" (see definition in Art. 3) and responsibilities linked to providers of generative AI (see, for example Zenner<sup>25</sup>). These proved to be part of the most intensely negotiated aspects of the AI Act, which solidified into a set of obligations for all providers of general-purpose AI (GPAI), that also included a second tier with additional obligations for GPAI models (see Chapter V) "having a significant impact on the Union market due to their reach, or due to actual or reasonably foreseeable negative effects on public health, safety, public security, fundamental rights, or the society as a whole, that can be propagated at scale across the value chain" (Art. 3(65)).

In brief, all providers of GPAI models must:

- ▶ Draw up technical documentation, including training and testing process and evaluation results, to be available, upon request, for the AI Office and national supervisory authorities.
- ▶ Draw up information and documentation to supply to downstream providers that intend to integrate the GPAI model into their own AI system so that the latter understands capabilities and limitations and is enabled to comply.
- ▶ Put in place a policy to comply with EU law on copyright.
- ▶ Publish a sufficiently detailed sum-

mary about the content used for training the GPAI model.

▶ Free and open-license GPAI models whose parameters, including weights, model architecture, and model usage are publicly available, allowing for access, usage, modification, and distribution of the model, only have to comply with the latter two obligations above. This exception does not apply to GPAI models with systemic risks.

The GPAI models are presumed to carry "systemic risk" when the cumulative amount of computation used for its training is greater than  $10^{25}$  floating point operations per second (FLOPS), or through evaluation or the Commission's decision have been found to have the high-impact capabilities that implicate this classification. If their model meets this criterion, providers must notify the Commission within two weeks. The provider may present arguments that, despite meeting the criteria, their model does not present systemic risks.

We consider it quite a leap to assume that more compute in training a model necessarily equals risks for negative impact on public health, safety, public security, and so on. The Commission is also quite autonomously mandated to change how "systemic risk" is allocated and can amend the criteria listed in Annex XIII, which may be both meaningful in terms of how AI evolves, but also open for legal unpredictability.

In addition to the obligations for GPAI above, providers of GPAI models with systemic risk must:

- ▶ Perform model evaluations, including conducting and documenting adversarial testing to identify and mitigate systemic risk.
- ▶ Assess and mitigate possible systemic risks, including their sources.
- ▶ Track, document, and report serious incidents and possible corrective measures to the AI Office and relevant national competent authorities without undue delay.
- ▶ Ensure an adequate level of cybersecurity protection.

In response to the complexities of AI regulation, the EU has established an AI Office to facilitate coordination on cross-border cases. However, the resolution of intra-authority disputes remains the responsibility of the Commission.

**Table 2. Management system.**

I. Define use case	Risk Self-Assessment
II. Evaluation of risk level and compliance requirements	Limited Risk <ul style="list-style-type: none"> <li>▶ Accessible disclosure of concrete user information</li> </ul>
	High Risk (Annex III and Annex VIII) <ul style="list-style-type: none"> <li>▶ Risk management</li> <li>▶ Data and data governance</li> <li>▶ Human oversight</li> <li>▶ Technical documentation</li> <li>▶ Transparency and provision of information to users</li> <li>▶ Accuracy, robustness, and cybersecurity</li> </ul>
III. Compliance Assessment	Internal Control (see AI Act Annex VI) External Control / Quality Management (see AI Act Annex VII)

## Assessment and How to Cope with the EU AI Act

For anyone wishing to put an AI system into operation in the EU, the AI act serves as a reminder for developers to always prioritize the well-being of individuals and society as a whole. They must first assess the risk and, depending on the risk class, comply with requirements relating to transparency and security. It is expected that it will be particularly challenging for high-risk applications to obtain approval for the EU market. There will be a grace period until the various obligations or bans become applicable. Nevertheless, developers should analyze the respective compliance requirements at an early stage to adapt the development process accordingly. The strategy includes the following key elements:

- ▶ Informing and training employees about the regulations and their obligations under the law. These cannot be understood without addressing the EU's rationale for this law and the expectations of EU citizens regarding trustworthy AI. Researchers and developers must understand that automated and algorithmic decision making should be based on the principles and values enshrined in the Charter of Fundamental Rights (such as human dignity, equality, justice and equity, non-discrimination, informed consent, private and family life, and data protections), and the principles and values of Union law (such as non-stigmatization, and individual and social responsibility). Support from an interdisciplinary working group should therefore be planned for.


- ▶ During the design of the systems, attention should be paid to transpar-

ency,<sup>5</sup> the nature and quality of the training data, and its documentation because of a later evaluation by external reviewers. This also includes the establishment of a risk-management system (see Table 2).

- ▶ Continuous investment in research and development, especially in rapidly evolving methods of AI explainability, see Balasubramanian<sup>6</sup> and Barredo Arrieta et al.<sup>7</sup> Once an AI system is explainable, it may positively contribute to trustworthiness and form a step toward acceptance and approval.

- ▶ Collaborate with other companies, potentially supervisory authorities, and organizations in the industry to share information and best practices for compliance. This can help reduce costs and ensure that all parties are on the same page when it comes to compliance.

## Acknowledgments

This work was undertaken while working for the ACM Europe Technology Policy Committee (TPC) on autonomous systems. We are grateful for support of and discussions with Chris Hankin, chair of the TPC. Further information may be found on the TPC website<sup>2</sup> and in prior publications.<sup>3,17</sup> 

## References

1. ACM Code of Ethics and Professional Conduct. ACM (2018); <https://bit.ly/4eNLIIV>.
2. ACM Europe Technology Policy Committee; <https://bit.ly/3XL8LAY>
3. ACM Principles for Algorithmic Transparency and Accountability, Association for Computing Machinery (2017); <https://bit.ly/4eSHbJ8>
4. Awad, E. et al. The moral machine experiment. *Nature* 563, (2018), 59–64; <https://go.nature.com/47S6g4w>
5. Baeza-Yates, R. et al. ACM Technology Policy Council Statement on Principles for Responsible Algorithmic Systems. ACM (2022). <https://bit.ly/3TUTnRo>.
6. Balasubramanian, V. Toward explainable deep learning. *Commun. ACM* 65, 11 (Nov. 2022), 68–69; <https://bit.ly/3Y9SwPa>.
7. Barredo Arrieta, A. et al. Explainable artificial intelligence (XAI): Concepts, taxonomies,

- opportunities and challenges toward responsible AI. *Science Direct* 58, (2020), 82–115. <https://bit.ly/4gRKeCY>.
8. Building trust in human-centric artificial intelligence. EU Commission (2019); <https://bit.ly/3zTMzwr>.
9. Doctorow, C. Small stickers on the ground trick Tesla autopilot into steering into opposing traffic lane. *Boing Boing* (Mar. 31, 2019); <https://bit.ly/47VQbux>
10. G7 Meeting Hiroshima. (May 2023); <https://bit.ly/3U8BM8E>
11. Goodall, N.J. Machine ethics and automated vehicles. *Road Vehicle Automation*, G. Meyer and S. Beiker (eds.). Springer (2014), 93–102; <https://bit.ly/3TQn2LL>.
12. IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems. *Ethically Aligned Design: A Vision for Prioritizing Human Well-being with Autonomous and Intelligent Systems, First Edition*. (2019); <https://bit.ly/47RhbLR>.
13. Jobin, A., Ienca, M., and Vayena, E. The global landscape of AI ethics guidelines. *Nature Machine Intelligence* 1, (2019), 389–399.
14. Knowles, B. and Richards, J. The sanction of authority: Promoting public trust in AI FAAct '21. In *Proceedings of the 2021 ACM Conf. on Fairness, Accountability, and Transparency* (March 2021), 262–271
15. Larsson, S. and Heintz, F. Transparency in artificial intelligence. *Internet Policy Rev* 9, 2 (2020); <https://bit.ly/3XQOD00>.
16. Larsson, S. On the governance of artificial intelligence through ethics guidelines. *Asian J. Law and Society* 7, 3 (2020), 437–451.
17. Larus, J. et al. When Computers Decide: European Recommendations on Machine-Learned Automated Decision Making. ACM (2018). <https://bit.ly/3BtDkDV>.
18. Leslie, D. Understanding artificial intelligence ethics and safety: A guide for the responsible design and implementation of AI systems in the public sector. The Alan Turing Institute (2019); <https://bit.ly/4dH2AUE>.
19. Middelton, S. et al. Trust, regulation, and human-in-the-loop AI within the European region. *Communications* 65, 6 (April 2022), 64–68.
20. Saucedo, A. et al. ACM Europe TPC Comments on Proposed AI Regulations. ACM (2021); <https://bit.ly/3XN060S>.
21. Shneiderman, B. Responsible AI: Bridging from ethics to practice. *Communications* 64, 8 (Aug. 2021), 32–35.
22. Villani, C. For a meaningful artificial intelligence. Comité d'Ética de La UPC (2018); <https://bit.ly/3XRxiUC>.
23. Wood, M. Self-driving cars might never be able to drive themselves. *Marketplace* (2021); <https://bit.ly/4dw15s5>.
24. Yelizarova, A. Global AI policy. *Future of Life* (Dec. 16, 2021); <https://bit.ly/4dtV27u>.
25. Zenner, K. The implementation and enforcement of the EU AI Act: The documents. *Digitizing Europe* (Jul. 28, 2024); <https://bit.ly/3BFRth4>.

**Alejandro Bellogin** is an associate professor in the department of Computer Engineering, Universidad Autónoma de Madrid, Spain.

**Oliver Grau** is an ACM Senior Member, Hannover, Germany.

**Stefan Larsson** is an associate professor in the department of Technology and Society LTH, Lund University, Lund, Sweden.

**Gerhard Schimpf** is a senior manager at SMF Team Consulting, Pforzheim, Germany.

**Biswa Sengupta** is a technical fellow at University College London, London, U.K.

**Gürkan Solmaz** is a senior researcher at NEC Laboratories Europe, Heidelberg, Germany.

© 2024 Copyright held by the owner/author(s). Publication rights licensed to ACM.



Watch the authors discuss this work in the exclusive *Communications* video. <https://caom.acm.org/videos/the-eu-ai-act>

DOI:10.1145/3650028

**Building federated learning with differential privacy to train and refine machine-learning models with more comprehensive datasets can help exploit ML's full potential.**

BY XUEBIN REN, SHUSEN YANG, CONG ZHAO, JULIE MCCANN, AND ZONGBEN XU

# Belt and Braces: When Federated Learning Meets Differential Privacy

WITH THE DEVELOPMENT of advanced algorithms, computing capabilities, and available datasets, machine learning (ML) has been widely adopted to solve real-world problems in various application domains. The success of ML often relies on large

amounts of application-specified training data, especially for large models such as ChatGPT. However, this data is often generated and scattered among enormous network edges or users' end devices, and can be quite sensitive and impractical to be moved to a central location as the result of regulatory laws (for example, GDPR) or privacy concerns.<sup>8</sup> This fact has brought an inconvenient dilemma between large-scale ML and increasingly severe data isolation. The conflict between data hungriness and privacy awareness is becoming increasingly prominent in the artificial intelligence (AI) era.

Google proposed *federated learning* (FL) as a potential solution to the above issue.<sup>26</sup> Through coordination between the central server and clients (devices participating in FL), FL collaboratively trains ML models over extensive data across geographies, which bridges the gap between the ideal of big data utilization and the reality of data fragmentation everywhere. By sharing locally trained models, FL not only minimizes the risks of raw data exposure but also eliminates client-server communications. Once proposed, it has been seen as a rising star in AI technology. Its recent usage

## » key insights

- **Federated learning (FL) can help learn collaboratively from massive scattered datasets without direct raw data exposure, but it still lacks a rigorous privacy guarantee against indirect information inferences.**
- **Differential privacy (DP) can mathematically formulate and limit the indirect privacy leakage in various learning tasks, but it may suffer from the low signal-to-noise ratio issue when having a small number of learning samples.**
- **There is an ongoing and growing body of research on the mutual complementarity and benefits of FL and DP; this article summarizes that research and explores optimization principles.**
- **We outline a set of new research challenges and related investigation dimensions for achieving usable FL with DP in emerging applications.**



in fine-tuning large language models (LLMs) confirmed that again.

The advancement of FL in privacy protection stems from the delicacy in restricting raw data sharing. This is however far from sufficient, as gradients of deep models can even expose the privacy of training data<sup>39</sup> but FL gives no formal privacy guarantees. Fortunately, differential privacy (DP), proposed by Dwork,<sup>10</sup> allows a controllable privacy guarantee, via formalizing the information derived from private data. By adding proper noise, DP guarantees a query result does not disclose much information about the data. Because of its rigorous formulation, DP has been the *de facto* standard of privacy and applied in both ML and FL.

As privacy in design, the emergence of DP and FL greatly encourages data sharing and utilization in reality. On one hand, by restricting raw data exposure, FL enables ML model training over massively fragmented data. It also significantly enriches ML applications for extensive distributed scenarios. On the other hand, by rigorously limiting indirect information leakage, DP can strengthen the privacy in trained models with provable guarantees. The complementarity of FL and DP in privacy suggests a promising future of their combination, which can significantly extend the applicable areas for both techniques and bring privacy-preserving large-scale ML to reality. Specifically, FL has advantages in fusing geographically isolated datasets, while DP can offer provable guarantees and thus encourage sensitive data sharing. Aimed at exploiting the potential of ML to its fullest, it is highly desirable and essential to build FL with DP to train and refine ML models with more comprehensive datasets.

The benefit of privacy protection in both FL and DP comes at a cost in terms of data utility, albeit other issues. FL clients often have limited capabilities and distribution-skewed datasets, causing insufficient and/or unbalanced training of global models with low utility. DP algorithms hide the presence of any individual sample or client by adding noise to model parameters, also leading to possible utility loss. Therefore, utility optimization,



## The conflict between data hungriness and privacy awareness is becoming increasingly prominent in the AI era.



that is, improving the model utility as much as possible for a given privacy guarantee, is an essential problem in combining FL and DP. Given the great potential, studies on this problem have rapidly expanded in recent years. However, they are often conducted based on various FL and DP paradigms concerning different security assumptions (for example, whether the server is trustworthy) and levels of privacy granularity (for instance, sample or client). Without a systematic review and clear categorization of existing paradigms, it is hard to precisely evaluate and compare their utility performance. On the other hand, despite the paradigm differences, the utility optimization principles are quite similar. However, current studies often focus on specific algorithm designs for different paradigms of FL with DP and there lacks some common pathways to follow. Meanwhile, only few surveys on the intersection of DP and FL either have a different focus other than the utility issue or lack high-level insights into future challenges.

This article aims to provide a systematic overview of DP-enabled FL while focusing on high-level perspectives on its utility optimization techniques. We begin by presenting an introduction to FL and DP respectively, highlighting the benefits of their combination. We then summarize research advances by categorizing the paradigms and software frameworks of FL with DP. Aiming at usable analytic results, we present the high-level principles and primary technical challenges in their utility optimization in several emerging scenarios. Finally, we discuss some related topics to FL with DP, which would also impact the achieved data utility. Our review can benefit the general audience with a systematic understanding of the development and achievements on this topic. The perspectives on utility optimization for DP-enabled FL can offer some insights into research opportunities and challenges for usable AI services with privacy protection in both academia and industry.

### Federated Learning

**Overview of federated learning.** An FL system is essentially a distributed ML (or DML) system coordinated by a

central server,<sup>a</sup> which helps multiple remote clients with separate datasets to collaboratively train an ML model, under a privacy constraint that any client does not expose its raw data. There are two popular FL frameworks.<sup>26</sup> Federated stochastic gradient descent (FedSGD) is the federated version of the stochastic gradient descent (SGD) algorithm. In SGD for centralized ML, gradients are computed on a random subset of the total dataset and then used to make one step of the gradient descent. FedSGD uses a random fraction of clients and all their local data. The gradients are averaged by the server proportionally to the number of training samples on each client and are used to make a gradient-descent step. To overcome the communication bottleneck, federated averaging (FedAvg) allows clients to perform more than one batch update on the local dataset and exchange the updated parameters rather than the gradients.<sup>20</sup> FedAvg is a generalization of FedSGD since averaging the gradients would be equivalent to averaging the parameters themselves if all the clients begin with the same initialization. So, generally FL works as follows:

1. Each participating client performs a local training procedure on its own dataset and sends the gradients or model updates to the server.

2. The server securely aggregates the received gradients or model updates and updates the global model accordingly.

3. The server sends back the new global model to the corresponding clients.

4. Clients update their local models and prepare for the next iteration.

These procedures are repeated until the global model converges or a sufficient number of iterations is applied. FL is classified into *cross-device FL*, which leverages up to millions of devices in the wide-area network (WAN), and *cross-silo FL*, which ties up a handful of edge nodes with reliable backbones.

**Comparison with traditional DML.** Despite being a typical DML paradigm, when compared with *traditional DML*

in data centers for ML speedup, FL has many distinct characteristics (as shown in Figure 1):

- **Privacy requirement:** Unlike traditional DML in the datacenters (where data can be arbitrarily scheduled among computing nodes), ensuring privacy protection lies at the center of FL, which strictly prohibits raw data sharing.

- **Data partitioning:** Data in FL is generated naturally or obtained from individual users, thus often being non-IID and imbalanced. Instead, data in traditional DML is usually manually scheduled to be almost shuffled or balanced.

- **On-device learning:** In datacenters, DML computing nodes are homogeneous, deployed centrally, and powerful. In contrast, FL is implemented with tens to millions of distributed clients with heterogeneous and limited computing capacities.

- **Communication:** Traditional DML in datacenters can enjoy gigabytes of bandwidth and communicate in a peer-to-peer manner. However, FL clients are usually connected to the server by the WAN and bandwidth constrained.

- **Model aggregation** fuses training results (for example, local models) from distributed nodes. Compared to homogeneous sub-models in traditional DML, one challenge in FL is the prominent heterogeneity among local models due to either 'non-IIDness' or varied training progress.

- **System actors:** Unlike the closed

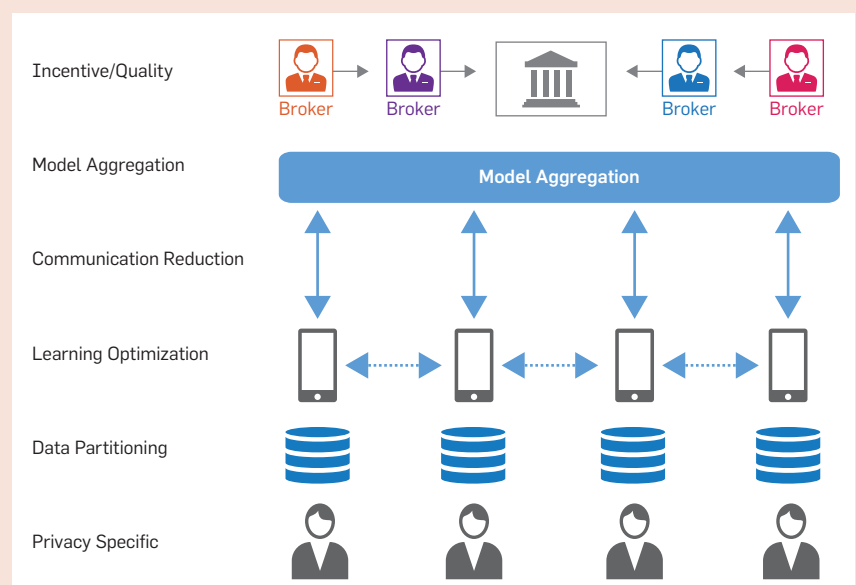
and fixed system of traditional DML, FL is often conceived as an open and scalable system consisting of massive clients owned by different individuals/organizations seeking different benefits.

#### Privacy threats in federated learning.

Due to the above characteristics (for example, geographically distributed nature, open architecture, and complicated interactions), various attacks can be mounted against FL in both model training and serving (that is, inference). Instead of those for degrading system availability or compromising data integrity (for example, poisoning attacks), we focus on privacy threats for snooping private information in FL.

*Privacy adversaries.* Privacy may be disclosed to or inferred by anyone who has access to the information flow in FL. Compared with ML over centralized data or traditional DML centrally deployed in datacenters, mutually distrusted entities in FL may all be viewed as privacy adversaries inferring others' private information. Possible adversaries can be classified as insiders and outsiders. The former includes the server and participating clients, and the latter contains eavesdroppers over communication channels and third-party analysts (users) that consume the final model. Compared with outsiders that are more likely to have black-box access (that is, can only query via APIs) to the final model, insiders are generally more capable as they can often have white-

**Figure 1. Building blocks of FL systems.**



<sup>a</sup> Decentralized FL is a special form where clients collaborate via peer-to-peer communication without a server.

box access (that is, full access with prior knowledge) and substantially impact FL model training. Insiders can be further considered to be semi-honest and malicious. The former is also known as honest-but-curious, that is, following the protocol correctly but trying to learn other entities' private state. The latter may actively deviate from the protocol (for example, modifying data or colluding with others) to achieve the goal.

**Privacy attacks.** Considering the above adversaries, the following privacy attacks may exist in FL (shown in Figure 2):

**Membership inference** targeting a model aims to predict whether a given data sample was in its training set.<sup>32</sup> It works by training multiple customized inference models to recognize noticeable patterns in the models' outputs for the given sample. In traditional centrally deployed ML, membership inference is normally mounted by third-party users. In FL, it can be carried out not only by third-party users, but also communication eavesdroppers and even participating clients and the server. This is because the local, aggregated, accumulated, and final forms of gradients or model parameters all may expose private information about training data. Moreover, ac-

tive attackers disguised as clients can selectively alter their gradient updates to significantly enhance the attack accuracy over the victim clients.

**Class representative inference** tries to generate class representatives from the underlying distribution of the training data that the targeted model could have been trained on. In traditional ML, third-party users can achieve this goal by iteratively modifying the features of a random sample until a maximal confidence is reached, or by training an inverse model, with black-box access to the targeted model. In FL, while an honest-but-curious server may partially recover some samples of honest clients by simply observing their uploaded gradients, active malicious clients or a passive malicious server can exploit generative adversarial networks (GANs) to construct class representatives from not only the global data distribution but also specific clients.

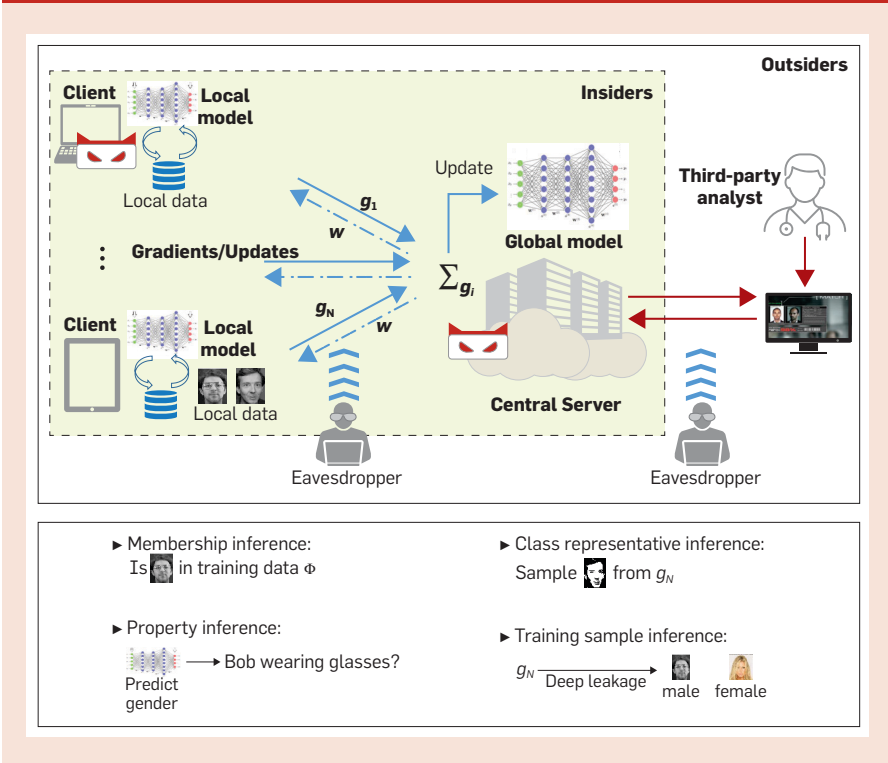
**Other privacy attacks** include inferences for properties and even the accurate training data (both inputs and labels). Different from the above inferences in terms of properties characterizing an entire class, property inferences aim to infer those properties independent of the characteristic features. With some auxiliary data, a passive adversary

trains a binary property classifier to predict whether the observed updates were based on the data with the property, while an active adversary can exploit multi-task learning to simultaneously conduct main FL training and infer the targeted property state with enhanced capability. Inferring accurate training data is also demonstrated as possible under the *deep leakage from gradients*, which optimizes the dummy inputs and labels via minimizing the difference between the dummy and targeted gradients for differentiable models.<sup>39</sup>

**Related privacy-preserving techniques.** Cryptographic primitives and protocols can restrict unauthorized access to confidential information, thus reducing the chances of privacy leakage.<sup>19</sup> For instance, *homomorphic encryption* (HE) supports dedicated operations on multiple encrypted data to produce ciphertexts that can be decrypted to generate desirable functional outcomes of original plaintexts. *Functional encryption* (FE) authorizes the holder of a key associated with a specified function to directly learn the function output over encrypted data and nothing else. Using secure multi-party computation (SMC), a set of parties jointly compute from their inputs without relying on a trusted third party or learning each other's input. Cryptography implemented in software still requires an error-free environment for execution and uncompromising storage of secret keys. This naturally calls for hardware-assisted security. Trusted execution environments (TEEs) can create an isolated operating environment that ensures the confidentiality of the data and codes within, while enabling remote authentication and attestation. In FL training, the above technologies can be adopted either alone or in combination to guarantee the desired confidentiality of the processed models.

However, note that privacy is essentially orthogonal to confidentiality. Whatever secure protocols and trusted systems are used, a final model will eventually be trained for consumption. Even if providing inference APIs only, model predictions may still reveal sensitive information as ML models inevitably carry some knowledge of training samples.<sup>11</sup> In general, models with poor generalization tend to leak more. Overfitting is one of the sufficient

Figure 2. Privacy threats in FL training.





conditions of performing membership inference attacks.<sup>28</sup> Therefore, another line of defensive approaches is properly suppressing fine-grained model utility. For instance, regularization can undermine inference attacks by reducing overfitting. For deep learning, two useful strategies are *model compression* (or sparsification), which sets gradients below a threshold to zero, and *weight quantization*, which limits the parameter precision. However, these approaches provide intuitive protection only without rigorous guarantee.<sup>31</sup>

### Differential Privacy

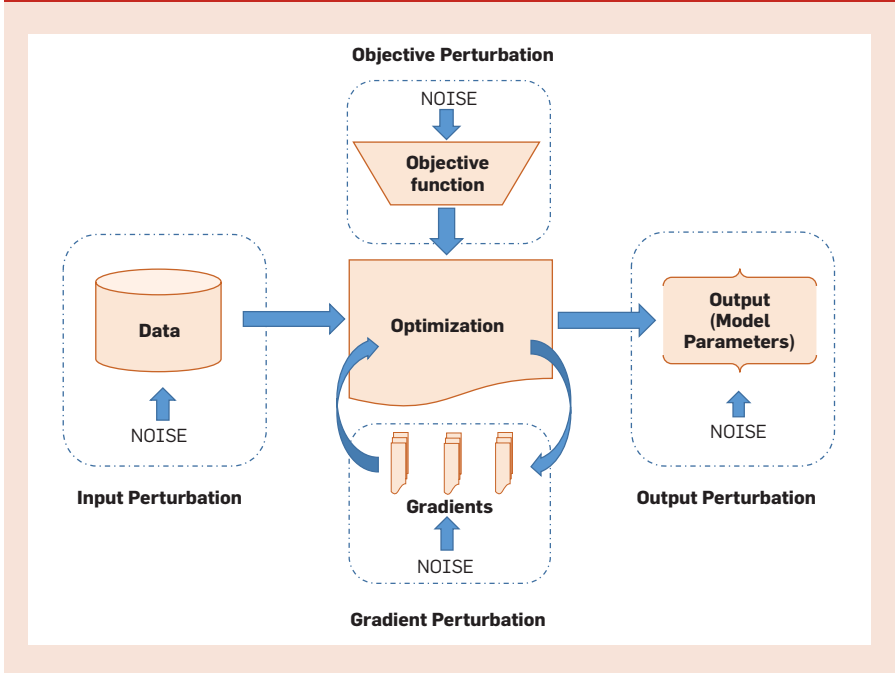
With the provable guarantee of limiting privacy leakage even in securely aggregated results, differential privacy promises to complement the above technologies and strengthen FL.

#### Overview of differential privacy.

Through establishing a formal measure of privacy loss, DP allows for rigorously controlling the (worst-case) information leakage. Informally, it guarantees an algorithm's output does not change much for two datasets differing by a single entry.<sup>10</sup> To achieve DP, the basic idea is to properly randomize the relationship between data input and algorithmic output, for example, by adding noise.

DP has various models, as noise can be added to the different components or phases of algorithms.<sup>11</sup> Conventional DP assumes a trustworthy aggregator and adds minor noise to algorithm output, which is known as centralized DP (CDP). Assuming an honest-but-curious aggregator, local DP (LDP) randomizes data at the users' end before collection and reconstructs utility from perturbed data of multiple uses. From CDP to LDP, the trust model is weakened under the same DP parameter, while data uncertainty and accuracy loss become larger. To bridge the trust-accuracy gap, distributed DP (DDP) exploits cryptography to obtain high accuracy without a trusted aggregator.<sup>35</sup> There are currently two DDP paradigms, based on secure shuffling and secure aggregation, respectively. Secure shuffling uses an anonymous communication channel to alleviate identification risks of messages thereby relaxing the trust model. Secure aggregation replaces the trusted aggregator with secure computation protocols and thus can reduce noise

Figure 3. Approaches to achieve DP for ML.



and gain the same utility as in the centralized model.

The prevalence of DP also comes from many delicate characteristics. The post-processing property keeps the privacy guarantee of algorithms after arbitrary workflows. Composition theorems help to understand the composed privacy guarantee of a series of sub-algorithms and enable building complicated algorithms from simple operations.

**Differential privacy for ML.** DP has been applied in ML to prevent adversaries with access to the model from inferring the training data. While *intrinsic privacy* can be achieved freely for some ML models with inner randomness,<sup>16</sup> noise addition to different components of ML algorithms provides viable pathways for privacy-preserving ML with DP, as shown in Figure 3.

*Output perturbation* adds calibrated noise to the parameters of final models which, however, may have large (even unbounded) sensitivities and lead to severe model utility loss. *Input perturbation* randomizes training data and then constructs an approximate learning model on it,<sup>9</sup> which usually has limited model utility due to much stronger protection. *Objective perturbation* perturbs the objective functions of the optimization problem in ML. Although functional mechanisms<sup>38</sup> allow its usage for complicated model functions, it is often

infeasible to explicitly express the loss functions for most ML models, especially deep learning. *Gradient perturbation* that sanitizes parameter gradients during training<sup>1</sup> can ensure DP even for nonconvex objectives, making it quite useful for deep models. Differentially private SGD (DP-SGD), which has been the common practice for privacy-preserving ML,<sup>1</sup> samples a mini-batch of samples, clips the  $l_2$  norm of the gradients computed on each sample, aggregates the clipped gradients, and adds Gaussian noise in each iteration. By incorporating gradient clipping, it can avoid the issue of unknown gradient sensitivity. Besides, it is often used with a moments accountant for tracking a tighter privacy loss bound.

### Federated Learning with Differential Privacy

The wide application of DP in privacy-preserving ML shows the great potential of privacy-preserving FL with DP.

**Benefits of FL with DP.** DP with rigorous guarantee has been an essential technology for privacy-preserving data analysis and ML. Although it has been successfully integrated into distributed systems for data querying and analyses,<sup>30</sup> there is still a lack of a DP-enhanced framework for large-scale distributed ML over massively scattered datasets. FL supports flexible ML tasks with extensive models and scalable ML

training for massively scattered datasets. Despite ensuring no direct data exposure by solely sharing intermediate parameters, it still lacks a formal privacy guarantee and may expose indirect privacy. Therefore, when combined, FL with DP can realize large-scale and flexible distributed learning while preventing both direct and indirect privacy leakage.

The combination of FL and DP, as complements of each other in encouraging massively confidential and sensitive data utilization, can achieve paramount benefits for privacy protection in reliability.

*FL empowers and prospers DP-based ML over large-scale siloed datasets.* DP-based ML (especially deep learning) in the centralized setting has made rapid progress. However, data centralization and privacy regulations strongly hinder its further development. As a result, DP-based ML wishes to meet large-scale data or data-extensive applications. Fortunately, FL naturally enables DP-based ML over massively scattered data, thus greatly prospering its success.

*DP completes and strengthens the reliability of FL via offering rigorous guarantees.* The mission of FL is to train and refine ML models with more comprehensive end-user data, which is subject to the willingness of data owners. Hence, a provable privacy guarantee is key to the popularization of FL systems. Beyond isolated datasets, privacy-preserved FL systems may encourage users to contribute more sensitive datasets.

**Research advances on FL with DP.** Due to the above benefits, marrying FL with DP has attracted extensive interest from both academia and industry. We systematically review the advances according to different paradigms and privacy notions.

*FL with centralized DP.* It is natural to extend differentially private ML algorithms (for example, DP-SGD) in the centralized setting, to the context of FL, to prevent information leakage from the training iterations and final model—against malicious clients or third-party users.

DP has different granularity, relying on the precise definition of neighboring datasets. Different from DP-SGD, which provides *sample-level DP* for hiding the existence of any single sample, it is more meaningful to provide *client-level*

*DP* in FL, which ensures all the training data of a single client is protected. This also fits in the FL setting, where each client computes a single model from all its local data. Assuming a trusted central server, a straightforward idea is to apply DP to the aggregation of model updates for participating clients and hide any client's influence on the model update at the server. DP-SGD can be adapted to both FedAvg and FedSGD, which forms two DP variants: DP-FedAvg and DP-FedSGD.<sup>27</sup> At a high level, they work as follows:

- ▶ Sampling a group of clients to train local models with total data
- ▶ Clipping the model updates of clients to bound the norm of the total updates
- ▶ Averaging the clipped updates
- ▶ Adding calibrated Gaussian noise to the average update

Privacy amplification via subsampling and moment accountant still applies to compose the privacy loss.<sup>14</sup> However, when providing a formal DP guarantee, particular attention should be paid to a client dropout issue, which may violate the uniform sampling assumption. Fortunately, recent studies show the possibilities of addressing in theory or bypassing with the new framework. Despite the existence of noise in both the intermediate model updates and the final model, their privacy guarantees are much different as being quantified from different views.

*FL with local DP.* LDP implemented on local models can defend against untrusted servers or malicious clients. Related studies can be categorized into two lines based on the FL architecture.

**Noise before aggregation:** Considering an untrusted central server in practice, LDP can be applied to perturb gradients or model updates for individual clients in each iterate. A simple approach is to add Gaussian noise to individuals' updates before uploading, which is also known as noising before model aggregation FL.<sup>36</sup> For example, DP-FedSGD or DP-FedAvg can be further adapted into the LDP setting by offloading Gaussian noise addition to the clients' side. Since the summation of multiple Gaussian noises still follows a Gaussian distribution, both the privacy loss at individual clients and the central server can be tracked simultaneously. FL algorithms with LDP (for


example, LDP-FedSGD) face the critical problem of the dimension dependency of communication and privacy. Besides communication overheads, given privacy parameters, the noise needed is substantially proportional to the dimension of the model parameter vector. By selecting a fraction of important dimensions, both noise variance and communication overhead can have a significant reduction. Therefore, dimension reduction is commonly used for large models. For instance, updated gradients can be sampled in a subset to reduce communication and truncated in value to compress noise variance.<sup>31</sup>

**Blind flooding with noise:** FL can be also implemented in a fully decentralized form without any central entity, thus avoiding a single-point failure and improving efficiency for heterogeneous systems. Its main feature is using peer-to-peer (P2P) communications other than a client-server architecture. A reasonable way to ensure model convergence with full information is to broadcast parameters to close neighbors, which, informally, face even higher privacy risk than an untrusted server. Moreover, in some opportunistic networks (for example, mobile crowd sensing or autonomous vehicle networks), the communication topology may be even time-varying and clients may meet unfamiliar neighbors frequently. In such a case, LDP is necessary and effective to preserve the privacy of exchanged messages among individual clients. This leads to the problem of decentralized optimization with LDP, which aims to ensure model convergence over a sparse P2P network with noisy local models. However, lacking a coordinating server, autonomous clients often have to adopt an asynchronous update pattern, which brings new challenges to the decentralized optimization in practice. Nonetheless, it has demonstrated that a differentially private asynchronous decentralized parallel SGD can converge at the same optimal rate as SGD and have a comparable model utility as the synchronous mode, while achieving relatively higher efficiency.<sup>37</sup>


*FL with distributed DP.* As discussed before, DDP can bridge the utility-trust gap between LDP and CDP while eliminating the assumption of a trusted server via two cryptographic techniques.

**Privacy amplification by shuffling:** A line of DDP studies for FL concentrates on the aforementioned secure shuffling technique, which amplifies the privacy-utility trade-off via additional anonymization for DP. Before forwarding to the untrusted server, locally perturbed models with minor noise are first permuted randomly to eliminate their client identities by one or more trusted (that is, secure) shufflers, which can be implemented as a trusted proxy or by delicate cryptographic primitives. By devising the classic *encoder-shuffler-aggregator* (ESA) framework for adapting FL, LDP-SGD adapted with secure shuffling can achieve both strong iteration-level LDP and good overall CDP for the final model, without noticeable accuracy loss.<sup>12</sup> For high-dimensional parameters in deep models, shuffling client identities only may still suffer from linkage attacks from side channels. A solution is to split the parameter vector and then shuffle the dividends to enhance anonymity.<sup>34</sup> To further trade off between privacy and utility, subsampling is also an important direction, which should consider the dimension importance.<sup>24</sup> Reckoning the benefits of Renyi DP (RDP) and its stronger composition of privacy loss, beyond exploring the RDP of subsampled mechanism, a natural extension is to further analyze and exploit RDP and RDP composition in the shuffled model.<sup>15</sup>

**Secure aggregation of small noises:** Secure aggregation protocols in Bonawitz et al.<sup>5</sup> overcome the practical issue of random client dropouts in cross-device FL, paving the way for FL with DDP via secure aggregation. However, such protocols often involve modular arithmetic, requiring the quantization of communicating contents (or discrete-valued inputs) for acceptable complexity. Then, the noise for privacy protection of local models should be also generated in discrete value. One solution is to generate and add minor discrete noise to the discretized parameters of individual clients before secure aggregation while outputting the aggregate parameters with moderate noise equivalent to the CDP model. Binomial or Poisson distribution can approach a similar trade-off between the utility and privacy of the Gaussian mechanism,<sup>3</sup> which however does not achieve RDP or enjoy the state-of-the-art



**When combined, FL with DP can realize large-scale and flexible distributed learning while preventing both direct and indirect privacy leakage.**



composition and amplification. Simply using discrete Gaussian noise can yield RDP with sharp composition and subsampling-based amplification,<sup>17</sup> but relies on an uncommon sampling mechanism when being implemented in software packages. Besides, the summation of discrete Gaussian is not closed and may cause privacy degradation. Recently, the Skellam mechanism can generate noise distributed according to the differences of two independent Poisson random variables.<sup>2</sup> Skellam noise is closed under summation and can leverage the common Poisson sampling tools to get privacy amplification and sharper RDP bound in theory. However, it remains an important problem to develop a practical protocol for production-level FL systems.

*Platforms and tools for FL with DP.* Toward usable FL with DP, many software frameworks and platforms have been developed to support research-oriented simulations or production-oriented applications. For private deep learning, PySyft<sup>b</sup> is a Python library that supports FL and DP, and decouples model training from private data. Its current version mainly focuses on SMC and HE rather than DP implementation. Dedicating to a fair evaluation of FL algorithms for the research community, FedML<sup>c</sup> develops an open research library and standardized benchmark with diverse FL paradigms and configurations. The current version only integrates weak DP but provides low-level APIs for security primitives. Similarly, by providing a high-level interface, PaddleFL<sup>d</sup> supports FL model development with DP and offers a baseline DP-SGD implementation. Furthermore, despite the consideration of practical FL settings and recognition of privacy issues, other FL frameworks (such as FATE<sup>e</sup> and LEAF<sup>f</sup>) still lack deep and flexible support for DP implementation. Recently, Sherpa.ai FL developed a unified framework for FL with DP, featuring comprehensive support for DP mechanisms and optimization techniques.<sup>29</sup> Nevertheless, it mainly offers algorithm-level optimization

b <https://bit.ly/4eDESJT>

c <https://bit.ly/3BFcpVP>

d <https://bit.ly/4h11abC>

e <https://bit.ly/3Y2JQJd>


f <https://bit.ly/4dIyrnF>

and does not consider practical system implementation. TensorFlow includes DP and FL implementations in its TensorFlow Privacy and TensorFlow Federated libraries,<sup>8</sup> respectively. Both libraries integrate seamlessly with existing TensorFlow models and allow training personalized models with DP. However, its integrated DP mechanisms are relatively fixed in design and do not support customized and flexible optimization. Opacus<sup>h</sup> is a scalable and efficient library for PyTorch model training with DP. It introduces an abstraction of a privacy engine that attaches to the standard PyTorch optimizer, which makes DP-SGD implementation much easier without explicitly calling low-level APIs. Beyond ML in PyTorch, it can be easily used in PySyft FL workflows to implement FL with DP.


**Improving model utility for FL with DP.** Existing work underpins the baseline frameworks of FL with DP. Aiming at usable FL with DP, it is essential to pursue a better trade-off between model utility and privacy. By reviewing common techniques in the fields of DP, ML, and FL, some optimization principles are summarized below.

*Optimization from the perspective of DP.* To seek better trade-offs, there are two directions: reducing unnecessary noise addition and tracking privacy loss tightly.

**Clipping-bound estimation:** Sensitivity calibration, which determines the proper noise amplitude by correctly bounding the sensitivity value, is crucial for minimizing noise variance while guaranteeing certain DP. As mentioned, a common practice in DP-SGD, thus also in SGD-based FL with DP, is to bound gradient sensitivity by gradient clipping and then add noise accordingly.<sup>1</sup> However, an underestimated clipping threshold may cause gradient bias and even model divergence, while an overestimated one results in excessive noise addition. Thus, it is important to understand the impact of gradient clipping and dynamically identify the proper clipping bounds during training.<sup>7</sup> For instance, adaptive gradient clipping via divergence analysis or heuristic estimation can provably or empirically reduce



**Toward usable FL with DP, many software frameworks and platforms have been developed to support research-oriented simulations or production-oriented applications.**



noise and produce models with higher utility.<sup>23</sup>

**Noise-distribution optimization:** It aims to reduce noise variance by reshaping noise distribution, thus decreasing unnecessary noise addition in DP. It has been invested with much effort. For instance, in traditional DP research, some discrete noise distribution and staircase noise distribution via segmentation techniques have been used in DP algorithms to lessen the necessary noise scale while meeting the DP requirement. In fact, both Laplace and Gaussian noise for DP are only some instances in a family of the whole distribution space satisfying DP definitions (as shown in Figure 4). Besides, to incorporate encryption primitives with less overheads, the discretization and quantization of data content should also apply to the noise generation for LDP and DDP.

**Privacy-loss composition:** The composition property of DP allows building complex FL models with DP primitives while composing privacy loss. Traditionally, both sequential and advanced compositions offer fairly loose bounds. The moment account analyzes a detailed distribution of the composed privacy-loss variable and derives a much tighter bound with higher-order moments. It shows acceptable utility with quite a small privacy loss for DP-SGD via using amplification techniques.<sup>14</sup> Privacy loss composition contributes to the optimization of privacy/utility trade-off by tightly tracking privacy loss in the composition of multiple independent noise across DP mechanisms.<sup>40</sup> A relevant but opposite angle is to fix the privacy budget and add correlated noises via wise budget division. For instance, classic tree-aggregation techniques add correlated noises rather than independent ones for repeated computations, which can get high utility while guaranteeing a given DP. Inspired by the idea, an amplification-free algorithm adds correlated noise to the accumulation of mini-batch gradients, which achieves a nice trade-off for DP-SGD without any amplification technique (and no uniform sampling and shuffling requirement).<sup>18</sup>

**Intrinsic DP computation:** Many studies have shown that noise-free DP can be achieved by leveraging the inherent randomness of certain models or

g <https://bit.ly/3YiErzc>

h <https://opacus.ai>

algorithms for model training, instead of using additional techniques or system components. Being aware of the intrinsic DP level, the designer or developer can save up much budget and add smaller noise, thus gaining utility without privacy degradation. For instance, by mapping the sampling process to an equivalent exponential mechanism, intrinsic DP in graph models can be effectively measured and leveraged in DP algorithm design. A novel federated model distillation framework can provide provable noise-free DP via random data sampling.<sup>33</sup> It has also been proved that data sketching for communication reduction in FL guarantees DP inherently.<sup>22</sup> Nonetheless, intrinsic privacy is not very common and only exists in certain models or algorithms.

**Optimization from the perspective of FL.** Massive FL clients and the pervasively spatiotemporal sparsity of model parameters offer the chance to extract acceptable utility without significantly harming the privacy guarantee.

**Updating frequency reduction:** DP-enhanced FL suffers from noise accumulation during excessive training

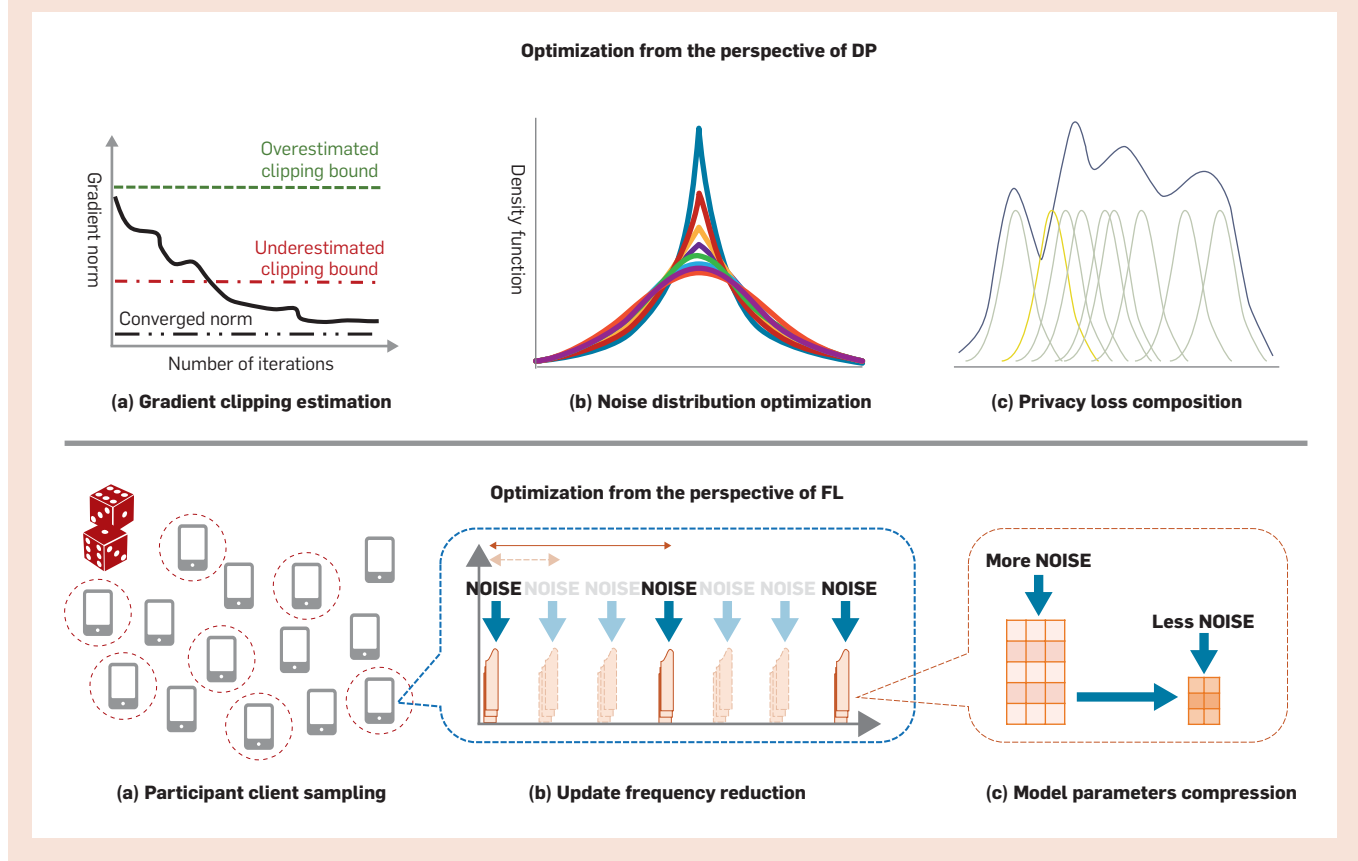
epochs. For communication efficiency, too many training epochs also require much network bandwidth. Therefore, it is highly desirable to reduce the model update frequency. Compared with FedSGD, FedAvg allows clients to perform multiple local updates before aggregation, thus reducing global update frequency.<sup>26</sup> A similar technique has been widely adopted in DP applications with dynamic datasets or time-series data. For instance, the data curator publishes perturbed data with DP noise at the timestamps with frequent changes while releasing approximate data without privacy budget consumption at non-changing timestamps.

**Model parameter compression:** Like the issue of frequent parameter updating, a long parameter vector heavily consumes the privacy budget (or incurs much noise with the fixed budget) and burdens the limited communication channel. To this end, many aforementioned model compression approaches, including parameter filtering, low-rank approximation, random projection, gradient quantization, compressive sensing, and so on have been proposed

for deep-learning models. For instance, similar studies include sampling and truncating a subset of gradient parameters in FL with CDP,<sup>31</sup> selecting top-K dimensions with large contributions in FL with LDP,<sup>25</sup> and sampling dimensions in FL with DDP.<sup>24</sup> All these methods manage to empirically reduce both the communication bandwidth consumption and noise variance. However, lossy compression techniques can, on the one hand, effectively improve model utility by reducing the DP noise. On the other hand, they may lead to utility loss as some parameter information is eliminated. An immediate question is how to find the optimal compression rate for achieving the best utility privacy trade-off.

**Participating clients sampling:** Besides reducing the update frequency and size of parameters, sampling the clients participating in DP-based FL training is also a promising approach to saving privacy budget, communication overhead, and energy consumption. The rationale behind this approach comes from the amplification effect of sampling for DP, in which, by randomly

Figure 4. Illustration of utility optimization techniques.



sampling the DP-protected FL clients in training epochs, much stronger privacy protection can be achieved while minimizing the average consumption in communication and computation as well as privacy. However, in practical cross-device FL, the set of available clients is usually dynamic without prior knowledge of the population. Moreover, as will be discussed, participating clients may drop out randomly. These issues make the assumption of uniform sampling unrealistic and cause severe challenges for gaining privacy-utility trade-offs.

### Challenges and Discussions

Despite the great potential and opportunities of DP-enhanced FL, there are still challenges in achieving usable FL with DP guarantee in emerging applications.

**Vertical/transfer federation.** FL can also be categorized according to different data-partition strategies. The above-discussed FL in the generic form mainly considers the horizontal data partition, where each client holds a set of samples with the same feature space. Now, vertical FL, where each party holds different features of the same set of samples, has gained increasing attention.<sup>13</sup> However, many existing studies on VFL are based on SMC for protecting confidentiality without considering privacy leakage in the final results. To achieve provable resistance to membership inference or reconstruction attacks, DP must be employed for safeguarding VFL. But it is more challenging than HFL for two reasons. One is that the VFL algorithm design varies for different tasks and models and often requires case-by-case development. Another is the correlations among distributed attributes are more difficult to identify without spreading individual information to other parties. Besides the vertical federation, there are also scenarios where different parties may hold datasets with non-overlapping features and users. Federated transfer learning (FTL) can eliminate the shifts of feature spaces in this scenario by combining FL and domain adaptation. However, similar to VFL, achieving DP for FTL is still challenging as the gradient of individual instances must be exchanged between participants.

**Large language models.** With the

emergence of large language models (LLMs) such as ChatGPT, both FL and DP have begun to demonstrate a promising future in fine-tuning LLMs, while preserving privacy with respect to the private domain data. However, these LLMs often have several billions to hundreds of billions of parameters. When applying DP and FL to LLMs, there will be multiple challenges concerning the huge number of parameters beyond the extra communication and computation burdens on resource-constrained participants. Regardless of the DP model, the total amount of privacy noise must be proportional to the number of parameters for enforcing DP on models, which would lead to huge utility loss. Besides, fine-tuning pre-trained LLMs is also different from conventional model training. The theoretical privacy guarantee in ML (for example, DP-SGD) often assumes models are learned from scratch with many training iterations, instead of a fine-tuning mode with much fewer iterations. Therefore, it is necessary to investigate new frameworks for applying both DP and FL and develop new theories for proper privacy guarantees in LLMs.

**FL over streams.** In many realistic scenarios, training data is continuously generated in the form of streams at distributed clients. In such cases, FL systems have to conduct repetitive analyses on distributed streams. By inheriting online machine learning (OL), online FL can be naturally derived to avoid retraining models from scratch each time a new data fragment comes. However, achieving DP for OFL brings multiple challenges. The first is how to define privacy in the OFL setting, as the general DP notion works for static datasets only. Although existing privacy notions for data streams and FL seem to apply here, they still need to be clarified and formulated rigorously in the OFL setting. The second is the efficient algorithm. Taking the event-level LDP (that is, ensuring  $\epsilon$ -LDP at each time instance) as an example, frequent uploading of local model updates accumulates huge communication costs and great utility loss, as the noise is proportional to the size of communication data. How to achieve communication and privacy efficiency without degrading overall model performance is thus an impor-

tant but unsolved research problem.

Apart from adapting to the new settings, building usable DP-enhanced FL systems still needs improvements in robustness, fairness, and privacy (allow data to be forgotten).

**Robustness.** A robust FL system should be resilient to various failures and attacks caused by misbehaved participants. Due to limited capabilities (for example, battery limit), FL clients (for instance, smartphones) may drop out of FL training unexpectedly at any time. Random client dropouts present severe challenges to the practical design of differentially private FL. Except for requiring a more sophisticated design of secure aggregation protocols,<sup>5</sup> some important assumptions may no longer hold for correctly measuring DP in FL. For instance, the DP amplification via shuffling and subsampling both rely on the assumption of clients correctly following the protocol. Despite recent progress in theory,<sup>4,5</sup> building practical FL systems while addressing the above impacts simultaneously is still challenging. Beyond robustness to dropouts of unintended client failure, defending against robustness attacks (for example, model poisoning for Byzantine and backdoor attacks) mounted by malicious participants is much more challenging.<sup>28</sup> Specifically, both data heterogeneity and model privacy protection in FL would prevent the server from accurately detecting anomalies and tracking specific participants.

**Fairness.** Privacy protection is only the first step to encouraging data sharing among a large population. Fairness enforcement helps to mitigate the unintended bias on individuals with heterogeneous data. However, the dilemma is that DP aims to obscure identifiable attributes while fairness requires the knowledge of individuals' sensitive attribute values to avoid biased results. Gradient clipping and noise addition in DP can exacerbate unfairness by decreasing the accuracy of the model over underrepresented classes and subgroups. So, the general tension between privacy and fairness calls for ethically sensitive FL algorithms that respect both issues. Meanwhile, gradient clipping and noise addition can also enhance robustness to some extent, as discussed above. This is consistent

with the conclusion that there is a tension between fairness and robustness in FL.<sup>21</sup> The constraints of fairness and robustness compete with each other, as robustness enhancement demands filtering out informative updates with significant model differences. Therefore, there is a subtle relationship between privacy, fairness, and robustness in FL. While existing studies concentrate on each of them separately, it would be significant to unify the interplay of the three simultaneously.


**Right to be forgotten.** Privacy rights include the “right to be forgotten”, that is, users can opt out of private data contribution without leaving any trace. As ML models memorize much specific information about training samples to ensure a specific private sample is totally forgotten, the concept of machine “unlearning” is proposed to eliminate its influence on trained models. However, on the one hand, machine unlearning in the context of FL, that is, federated unlearning, faces distinct challenges. Specifically, it is much harder to erase the influence of a client’s data, as the global model iteratively carries on all participating clients’ information. A straightforward idea for resolving the problem is recording historical parameter updates of clients at the server, which may cause significant complexity. On the other hand, existing machine unlearning has been demonstrated to leak privacy by observing the differences between the original and unlearned models.<sup>6</sup> DP seems to be one of the promising countermeasures. Therefore, the question of how to realize efficient and privacy-preserving solutions for federated unlearning remains open.

## Conclusion

With both privacy awareness and regulatory compliance, the meeting of FL and DP will promote the development of AI by unblocking the bottlenecking problem of large-scale ML. This article presents a comprehensive overview of the developments, a clear categorization of current advances, and high-level perspectives on the utility optimization principles of FL with DP. This review aims to help the community to better understand the achievements in different ways of combining FL with DP, and the challenges of usable FL with

rigorous privacy guarantees. Although FL and DP show increasing promise for safeguarding private data in the AI era, their combination still faces severe challenges in emerging AI applications. Also, they need further consideration and improvements on other practical issues.

## Acknowledgments

This work was supported in part by the National Key Research and Development Program of China under Grants 2020YFA0713900; in part by the National Natural Science Foundation of China under Grants 62172329, U21A6005, 61802298; in part by the Science and Technology Plan Project of Henan province under Grant 232102211007. 

## References

- Abadi, M. et al. Deep learning with differential privacy. In *Proceedings of ACM CCS* (2016), 308–318.
- Agarwal, N., Kairouz, P., and Liu, Z. The skellam mechanism for differentially private federated learning. In *Proceedings of NeurIPS* 34 (2021).
- Agarwal, N. et al. cpSGD: Communication-efficient and differentially-private distributed SGD. In *Proceedings of NeurIPS* (2018), 7564–7575.
- Balle, B. et al. Privacy amplification via random checks. In *Proceedings of NeurIPS* 33 (2020).
- Bonawitz, K. et al. Practical secure aggregation for privacy-preserving machine learning. In *Proceedings of ACM CCS* (2017), 1175–1191.
- Chen, M. et al. When machine unlearning jeopardizes privacy. In *Proceedings of ACM CCS* (2021), 896–911.
- Chen, X., Wu, S.Z., and Hong, M. Understanding gradient clipping in private SGD: A geometric perspective. In *Proceedings of NeurIPS* 33 (2020), 13773–13782.
- Cheng, Y., Liu, Y., Chen, T., and Yang, Q. Federated learning for privacy-preserving AI. *Commun. ACM* 63, 12 (2020), 33–36.
- Duchi, J.C., Jordan, M.I., and Wainwright, M.J. Privacy aware learning. *J. ACM* 61, 6 (2014), 1–57.
- Dwork, C. A firm foundation for private data analysis. *Commun. ACM* 54, 1 (2011), 86–95.
- Dwork, C. et al. The algorithmic foundations of differential privacy. *Foundations and Trends® in Theoretical Computer Science* 9, 3–4 (2014), 211–407.
- Erlingsson, Ú. et al. Encode, shuffle, analyze privacy revisited: Formalizations and empirical evaluation. *arXiv:2001.03618* (2020).
- Fink, O., Netland, T., and Feuerriegel, S. Artificial intelligence across company borders. *Commun. ACM* 65, 1 (2021), 34–36.
- Geyer, R.C., Klein, T., and Nabi, M. Differentially private federated learning: A client level perspective. *arXiv:1712.07557* (2017).
- Girgis, A.M. et al. On the Rényi differential privacy of the shuffle model. In *Proceedings of ACM CCS* (2021), 2321–2341.
- Hylland, S.L. and Tople, S. An empirical study on the intrinsic privacy of sgd. In *Theory and Practice of Differential Privacy (CCS Workshop)* (2020).
- Kairouz, P., Liu, Z., and Steinke, T. The distributed discrete Gaussian mechanism for federated learning with secure aggregation. In *Proceedings of ICML* (2021), 5201–5212.
- Kairouz, P. et al. Practical and private (deep) learning without sampling or shuffling. In *Proceedings of ICML* (2021), 5213–5225.
- Kairouz, P. et al. Advances and open problems in federated learning. *Foundations and Trends® in Machine Learning* 14, 1–2 (2021), 1–210.
- Konečný, J. et al. Federated learning: Strategies for improving communication efficiency. In *Proceedings of NeurIPS* (2016), 5–10.
- Li, T., Hu, S., Beirami, A., and Smith, V. Ditto: Fair and robust federated learning through personalization. In *Proceedings of ICML* (2021), 6357–6368.
- Li, T., Liu, Z., Sekar, V., and Smith, V. Privacy for free: Communication-efficient learning with differential privacy using sketches. *arXiv:1911.00972* (2019).
- Li, Y. et al. Multi-stage asynchronous federated learning with adaptive differential privacy. In *Proceedings of IEEE Trans. Pattern Anal. Mach. Intell.* 46, 2 (2024), 1243–1256.
- Liu, R. et al. Flame: Differentially private federated learning in the shuffle model. In *Proceedings of AAAI* 10 (2021), 8688–8696.
- Liu, R., Cao, Y., Yoshikawa, M., and Chen, H. FedSel: Federated sgd under local differential privacy with top-k dimension selection. In *Proceedings of DASFAA* (2020), 485–501.
- McMahan, B. et al. Communication-efficient learning of deep networks from decentralized data. In *Proceedings of AISTAS* (2017), 1273–1282.
- McMahan, H.B., Ramage, D., Talwar, K., and Zhang, L. Learning differentially private recurrent language models. In *Proceedings of ICLR* (2018), 1–10.
- Rigaki, M. and Garcia, S. A survey of privacy attacks in machine learning. *ACM Comput. Surv.* 56, 4 (2023), 1–34.
- Rodríguez-Barroso, N. et al. Federated learning and differential privacy: Software tools analysis, the Sherpa.ai framework and methodological guidelines for preserving data privacy. *Information Fusion* 64 (2020), 270–292.
- Roy Chowdhury, A. et al. Crypte: Crypto-assisted differential privacy on untrusted servers. In *Proceedings of ACM SIGMOD* (2020), 603–619.
- Shokri, R. and Shmatikov, V. Privacy-preserving deep learning. In *Proceedings of ACM CCS* (2015), 1310–1321.
- Shokri, R., Stronati, M., Song, C., and Shmatikov, V. Membership inference attacks against machine learning models. In *Proceedings of IEEE S&P* (2017), 3–18.
- Sun, L. and Lyu, L. Federated model distillation with noise-free differential privacy. In *Proceedings of IJCAI* (2021), 1563–1570.
- Sun, L., Qian, J., and Chen, X. LDP-FL: Practical private aggregation in federated learning with local differential privacy. In *Proceedings of IJCAI* (2021), 1571–1578.
- Wagh, S., He, X., Machanavajhala, A., and Mittal, P. DP-cryptography: Marring differential privacy and cryptography in emerging applications. *Commun. ACM* 64, 2 (2021), 84–93.
- Wei, K. et al. Federated learning with differential privacy: Algorithms and performance analysis. *IEEE Trans. Inf. Forensics Security* 15 (2020), 3454–3469.
- Xu, J., Zhang, W., and Wang, F. A(DP)<sup>2</sup>SGD: Asynchronous decentralized parallel stochastic gradient descent with differential privacy. In *Proceedings of IEEE Trans. Pattern Anal. Mach. Intell.* 44, 11 (2022), 8036–8047.
- Zhang, J. et al. Functional mechanism: Regression analysis under differential privacy. In *Proceedings of VLDB Endow.* 5, 11 (2012), 1364–1375.
- Zhu, L., Liu, Z., and Han, S. Deep leakage from gradients. In *Proceedings of NeurIPS* (2019), 14774–14784.
- Zhu, Y. and Wang, Y.-X. Poission subsampled rényi differential privacy. In *Proceedings of ICML*. PMLR (2019), 7634–7642.

**Xuebin Ren** is an associate professor with the National Engineering Laboratory for Big Data Analytics, and the faculty of Electronic and Information Engineering, at Xi’an Jiaotong University, Xi’an, Shaanxi, China.

**Shusen Yang** is a professor with the National Engineering Laboratory for Big Data Analytics, and the Ministry of Education Key Lab for Intelligent Networks and Network Security, at Xi’an Jiaotong University, Xi’an, Shaanxi, China.

**Cong Zhao** is a professor with the National Engineering Laboratory for Big Data Analytics at Xi’an Jiaotong University, Xi’an, Shaanxi, China.

**Julie McCann** is a professor with the Department of Computing at Imperial College London, London, England SW7 2AZ, U.K.

**Zongben Xu** is a professor with the National Engineering Laboratory for Big Data Analytics at Xi’an Jiaotong University, Xi’an, Shaanxi, China.

© 2024 Copyright held by owner/author.  
Publication rights licensed to ACM.

**The metaverse remains a work in progress, but improvements in how it handles ethical concerns and addresses cultural issues could push it further along the path to mass adoption.**

BY TIZIANA CATARCI, GIUSEPPINA DE NICOLA,  
AND DANIEL RAFFINI

# Ethics and Cultural Background as Key Factors for an Attractive Metaverse

IN RECENT YEARS, the concept of the metaverse has gained increasing resonance in both research and public discourse, perhaps second only to artificial intelligence (AI). Discussion persists, however, around the nature of the metaverse. Several studies have attempted to offer a definition, also considering how use of the term has changed over the past 20 years. For example, Markus Weinberger used qualitative

meta-synthesis methods to analyze existing literature, proposing the following definition:

*The Metaverse is an interconnected web of ubiquitous virtual worlds partly overlapping with and enhancing the physical world. These virtual worlds enable users who are represented by avatars to connect and interact with each other, and to experience and consume user-generated content in an immersive, scalable, synchronous, and persistent environment. An economic system provides incentives for contributing to the Metaverse.*<sup>30</sup>

As seen here, ideas of community and connection between users seem to be important in defining the metaverse, as do notions of “hyper-spatio-temporality” and “multitechnology convergence.”<sup>29</sup> These goals can be achieved through various technologies, such as virtual and augmented reality, avatar-based and second life systems, learning management systems, social media, simulation, and AI.<sup>17</sup>

The recent Gartner report “Emerging Tech: Adopter Anti-patterns—Metaverse Use Cases Are Plagued by Low Adoption” discusses some of the limitations that hinder mass adoption of the metaverse. The report points out two major issues. The first is the application of virtual reality (VR) in

## » key insights

- **The success of the metaverse relies on providing a compelling user experience that addresses both technical aspects and users’ cultural needs.**
- **The metaverse presents significant ethical challenges, such as privacy risks and potential harm from blurred lines between the virtual and real worlds, necessitating regulation and risk assessment.**
- **The metaverse has found significant success in South Korea, where it caters to both a highly competitive culture and a desire for new forms of self-expression beyond traditional societal constraints.**
- **Overcoming ethical risks and aligning with cultural backgrounds are essential for building trust and motivating users to engage in the metaverse.**






non-gaming environments, which often fails to meet the expectations and scope anticipated by users. We must mention, though, that the metaverse is not the same as VR, which has been successfully adopted in specific applications and for tasks where entertainment is not the main scope. The second issue concerns virtual meetings with avatars in the metaverse, which have not yet achieved a level of engagement necessary for enduring and meaningful experiences. From the report, it appears that the metaverse died an early death. However, is this really the case? What is it that could attract people once the technical and ethical problems are solved? And why do some countries seem to be especially fascinated by it? In this article, we will investigate how the metaverse is changing the concept of user experience, identify the technical and ethical shortcomings it still faces, and, finally, explore the cultural factors favoring its adoption in South Korea.


### **Motivation, User Experience, and Interoperability**

During the 2023 Augmented World Expo (AWE) USA conference, Neal Stephenson—the writer who coined the term *metaverse* in his 1992 novel *Snow Crash*—asserted that entrepreneurs, innovators, and companies investing time and money in the metaverse should seriously consider *why* people would want to use it, since people are only drawn to new innovations if they are either fun or essential. An important element to consider, then, is user motivation. Attracting and retaining metaverse users will require “a robust communications infrastructure, powerful and easy-to-use development platforms, and, perhaps most importantly, compelling applications that provide value to users that cannot be replicated or found elsewhere.”<sup>11</sup> We believe that two more aspects should be taken into consideration: providing a quality user experience from both technical and ethical perspectives, and satisfying the needs arising from people’s historical, social, and cultural backgrounds.

The first point to address when considering use of the metaverse is how it changes the user experience. At the



## **The metaverse brings with it a shift in the concept of human-computer interaction.**



very beginning of the computer era, there was basically no interaction between the user and the machine because everything was operator-mediated through punch cards. Only after the introduction of the personal computer did the importance of letting people interact with machines become clear, giving rise to the field of human-computer interaction (HCI), together with the introduction of graphical windows, icons, menus, pointers (WIMP) interfaces; direct manipulation; and metaphors.<sup>22</sup> The idea was to make interactive systems more usable by moving from “people adapting to technology” to “technology adapting to people,” while still requiring users to learn the machine’s language. For many years, the only improvements were to the fundamental ideas implemented in WIMP interfaces.

The next big advance was the smartphone, implementing a different interaction paradigm and becoming not just an instrument but also a seamless part of the user’s world. Still, the user had to make an effort in learning to execute tasks—essentially learning a new language—while, again, the goal of usability is to move the workload from the user to the system.<sup>18</sup> Meanwhile, game design followed a completely separate path (probably guided by its different purpose: entertainment), developing engaging, immersive interaction environments in which players pursue a game’s goals and have fun.

We are now at a historic moment in which two scientific advances are radically changing the concept of person-machine interactions and blurring the line between the real and the digital. The first is represented by large language models (LLMs), resulting in systems that, for the first time, really speak the user’s language without the need for translation. The second is the metaverse, in which interaction takes place in ways similar to those of the human world but offers infinite possibilities. In both cases, the interaction is focused on building a truly personalized user experience.

We can see, though, that the metaverse brings with it a shift in the concept of human-computer interaction. The promise of a fully immersive digi-

tal universe led to the establishment of a research area for the blending of the physical and the virtual. The challenge is to create “a real-virtual bridge, a conceptual model that can be used to mediate between real and virtual objects.”<sup>27</sup> Researchers aim to make the interaction mechanics more intuitive, addressing, for example, the relationship between gestures and devices in the metaverse, such as smartphone operation<sup>3,26</sup> and text-entry methods,<sup>5</sup> as well as the relationship between the metaverse and the Internet of Things (IoT).<sup>15,23</sup> Nonetheless, for many users, the learning curve is still steeper than desired, the user experience is not yet engaging enough, and the graphics and rendering capabilities, while advanced, still lack the refinement and realism that many expect.<sup>8</sup>

Another significant limitation is interoperability, which in the context of the metaverse has been defined as “enabling people to move to new networks with their avatars and virtual property.”<sup>16</sup> Lack of interoperability affects the user experience because, without mobility, users may find themselves confined to specific ecosystems, potentially leading to fragmented communities, echo chambers, and a stifling of innovation.<sup>8</sup> It’s akin to isolated islands in an ocean, where each island operates on its own rules and systems. To truly realize the potential of the metaverse, bridges must be built between these islands, promoting collaboration, shared experiences, and a more interconnected digital realm. To solve this problem, some have proposed building an open and interoperable metaverse using the Web and its standards.<sup>7,19</sup>

As we can see, many efforts are being made to overcome the technical limitations that still hinder adoption of the metaverse. The next few years will likely see advances in virtual reality hardware, augmented reality integration, and the overall blending of our physical and digital worlds.<sup>12</sup> Until then, the metaverse remains a work in progress, with its full potential yet to be realized.

### **An Ethics for the Metaverse**

The introduction and use of the metaverse raises relevant ethical issues,

which, like deficiencies in the user experience, contribute to driving users away because of perceived danger. Like LLMs, the metaverse is a tool that could reach a wider audience, and for this reason there is a need for risk assessment, regulation, and user awareness.<sup>20</sup> For example, in a 2008 paper Edward Spence discusses the rights and obligations of avatars.<sup>25</sup> Influenced by Alan Gewirth’s principle of generic consistency (PGC), Spence formulates an ethical framework for digital domains. The PGC suggests that, when acting within virtual environments, any individual inherently deserves the rights to freedom and well-being. Rooted in the notion that people use avatars in the metaverse to represent themselves, Spence’s theory posits that avatars act as digital reflections of real-world individuals. Given this, they should not only be granted the same rights but should also adhere to real-world ethical standards. Though it might seem to be a mostly theoretical issue, misconduct of avatars is a real problem, as in the case of sexual harassment. In the U.S., more than 40% of Internet users have encountered online harassment, a problem that becomes particularly salient in the metaverse, where harassment can feel even more personal and intimidating due to the lifelike nature of virtual environments.<sup>21</sup>

Philosophical models can help us define the metaverse, as well as the relationship between avatars and real people. Building trust in the metaverse, however, requires a bottom-up approach that addresses the problems that arise from actual use. So how can we build an effective, reality-based model of the metaverse that merges theoretical background with real problems? Some studies propose starting with AI ethics, which has developed significantly in recent years as a response to the widespread adoption of AI technologies.<sup>1</sup> AI ethics emerged because AI is a technological innovation that has a profound impact on society, like the metaverse might have in the future. As the metaverse incorporates AI components, it could face similar risks,<sup>28</sup> but it also introduces unique challenges deriving mostly from the

problematic relationship between the real and the virtual. As such, an ethics of the metaverse should begin by addressing the illusions created by virtual environments, which necessitate contributions from cognitive science.<sup>4</sup> Experiencing virtual worlds can lead to illusions about oneself and one’s relationships, with potentially dangerous effects on one’s actions in the real world, perceptions of reality, and self-representation. The metaverse could in fact pose risks to mental and social health if people become addicted to it and use it as an escape from reality, leading to confusion between the real and virtual worlds, post-VR sadness and hangovers, dissociation, diminished sensitivity to the consequences of actions in the real world, cybersickness, and, in extreme cases, severe mental disorders.<sup>1</sup> For these reasons, it is crucial for potential users to be psychologically prepared to avoid addiction and maintain the distinction between the real and the virtual. In this regard, protecting children is an important area of concern, with potential risks such as the formation of false memories, desensitization to tangible threats, moral disengagement, technological addiction, and cyberbullying.<sup>10</sup>

Confusion between the virtual and the real can also be caused by excessive sharing of personal and biometric data with avatars.<sup>13</sup> In a recent paper, Carl Smith and colleagues<sup>24</sup> emphasized that privacy encompasses not only personal information but also our body image and data. The concern is that once a person’s body image is captured, it can be perpetually used in deep-fake productions. There is therefore a need to expand legal protections of individual rights to include control over biological data, which includes various forms of biometric data. Protection of mental privacy is equally essential, since interactions in the metaverse might use brainwave capture to enhance experiences, for example, through brain-computer interfaces (BCIs). Some of these emerging technologies have the potential to essentially interpret human thoughts, model identities, draw detailed and contextually relevant conclusions, and then

infer thought processes via machine learning.<sup>2</sup>

The problematic relationship between the virtual and real worlds can cause both technical and ethical issues. Each of the issues mentioned—and potentially many more—merits a more detailed discussion, which in some cases has already been started by researchers. The complexity of these issues highlights the imperative of expanded ethical deliberation on metaverse usage and the pressing need to regulate user behavior. Questions about how to penalize misconduct and determine the appropriate regulations and policies for the metaverse must be addressed.<sup>20</sup> Moreover, there is an urgent need to enhance monitoring tools: In the metaverse, interactions will be more immediate and direct compared with traditional social media platforms, necessitating effective real-time moderation and oversight mechanisms.<sup>21</sup> These tools must adeptly blend real and virtual content, striking a balance between ensuring user freedom and filtering malicious or inappropriate content. An interesting emerging idea is that addressing ethical issues and ensuring control and security would actually make the metaverse more trustworthy, fostering its adoption. That said, it is also necessary to provide a technology capable of matching social and cultural expectations, discussed in the next section.

### Matching Cultural Expectations: The Case of South Korea

Each culture is different. This is why it is not easy to design a technology that is equally effective everywhere and for everyone, as highlighted by the postcolonial computing approach.<sup>9</sup> Due to its immersive experience, the metaverse faces this problem even more so than other technologies. Indeed, we believe the metaverse is a concept whose acceptance and success are deeply rooted in the cultural backgrounds and expectations of people. This idea is based on observing how the metaverse has achieved significant success among the public, industries, and policymakers of countries like South Korea and Japan. As an example, we will discuss the South Korean case.

The metaverse is a visible and expanding phenomenon in South Korea, going beyond being just a source of entertainment. The Ministry of Science has introduced an ambitious plan, rooted in the Digital New Deal 2.0, to position South Korea as a dominant player in the global metaverse landscape. The plan's key components include developing a platform ecosystem centered around Korean cultural content, nurturing talent, supporting enterprises, and creating an ethical, inclusive metaverse. The South Korean government's investment in this technology is fueled by the population's enthusiasm for it and the potential applications it offers. A global survey conducted in 2022 by Ipsos, on behalf of the World Economic Forum, revealed a 71% familiarity rate with the metaverse in South Korea, higher than the average of the nations analyzed. Moreover, 63% of South Koreans expressed a positive outlook on integrating extended reality into their daily lives, compared with the global average of 50%. Another Ipsos survey, published in September 2022 and titled "Gen Z and the Metaverse," sheds light on South Korea's Gen Z perspective, which sees the virtual world as a platform brimming with opportunities for both entertainment and financial gain. Gen Z fosters close connections with friends made in the virtual world and adeptly seeks out like-minded individuals within the metaverse. To them, the metaverse is viewed as a more effective and flexible environment than the real world, enabling users to readily pursue their aspirations and explore new horizons.


In recent years, social changes in South Korea have weakened certain cultural heritage principles, such as the emphasis on the group over the individual, while strengthening the desire for self-affirmation and fulfilling personal needs. A growing number of people have begun using service platforms that not only cater to the hectic pace of contemporary life but also provide the opportunity to limit social interaction, which is increasingly seen as entrenched in formalities and part of a hierarchical system that younger generations are rejecting. For this reason, the govern-

ment has viewed the "untact policy" as a potential means to stimulate economic development. *Untact* is a term created in South Korea by adding the privative prefix *un* to the word *contact*. It refers to a "service provided without face-to-face encounters between employees and customers using digital technologies."<sup>14</sup> Launched in 2020, the idea of untact gained momentum during the pandemic and quickly extended into various sectors. Though concerns about isolation and societal fragmentation persist, for a multitude of individuals, the policy has brought several benefits, such as enabling anonymity, alleviating the burden of formality, and mitigating the emotional labor often associated with the service industry. South Korea's transition toward a social culture of metaverse technology, as well as its positive reception, is indeed a rather expected development. This technology is already widely applied in the entertainment industry, for example, in the realm of K-pop, where fans design avatars that allow them to encounter their beloved artists virtually. It is relevant to mention the case of Mave, a South Korean girl quartet, which gained 20 million views on YouTube and exists solely in the virtual world: The four members reside within the metaverse, and their songs, choreography, interviews, and even their hairstyles are all crafted by Web designers and AI.


To understand why the untact policy and the metaverse are so successful in South Korea today, we must explore the country's cultural background. In South Korea, the metaverse offers individuals an opportunity to seek refuge in a reality where they can express their own identity, transcending the norms of a society deeply rooted in Confucianism, which rose to prominence on the Korean peninsula during the 15th century. Despite South Korea's modernization and the widespread adoption of Christianity, the enduring influence of this traditional philosophy is still evident in family relationships, political attitudes, and approaches to problem solving. Morality, practicality, and self-cultivation are the pillars of Confucianism.<sup>6</sup> It posits that excellence must be pursued in both

inner morality and outer achievements. In politics and education, the Confucian concept of leader and subject continues to be advocated, with the public encouraged to uphold the traditional notion of the leader as a combination of king, teacher, and father.<sup>33</sup> Simultaneously, there is an expectation for leaders to exhibit high moral standards, considered essential for fostering a strong sense of community. Thus, Confucianism highlights the importance of maintaining a harmonious and ongoing connection between the individual and the universe, between the pursuit of knowledge and personal growth, and between the principles governing family structure and the sociopolitical framework. Each person plays various roles in society, from their private responsibilities within their families to their public roles in areas such as politics and the economy. These principles have molded a society that is highly organized and collectivist, marked by significant social expectations in terms of behavior and adherence to cultural norms.

Another factor that influenced the shaping of today's society is Korea's history of struggle in establishing itself as an independent, recognized nation. Initially, this involved breaking free from Chinese political and cultural dominance, and then from Japanese colonization and Western influences. All of this has laid the groundwork for a significant sense of national belonging and a strong sense of identity. Koreans experienced three distressing historical events in the 20th century—the Japanese occupation, the Korean war, and the division of the country—that made nationalism a prominent factor in nation-building. During the 1950s, South Korea was one of the poorest countries in the world, remaining so for more than a decade. The Japanese occupation and the Korean War caused enormous economic losses and huge casualties, but the Korean people focused on their nation's reconstruction and today South Korea is one of the richest, most influential countries in the world. The aftermath of these significant challenges and the need for social redemption, coupled with the desire to break free



**In South Korea, the metaverse offers individuals an opportunity to seek refuge in a reality where they can express their own identity, transcending the norms of a society deeply rooted in Confucianism.**



from the past, has driven the country toward policies aimed at creating a highly competitive society. Individuals of all ages, genders, and social backgrounds orient their lives around enhancing their competitive edge. This inclination is particularly evident in the domains of education and employment, with the younger generation devoting extra hours to private academies in pursuit of admission to prestigious universities and securing positions in leading companies.

The Korean commitment to self-improvement has indeed spurred rapid progress, but it has also led to heightened exhaustion. In a society characterized by cutthroat competition, a larger number of people encounter setbacks than achieve success. For those grappling with challenges, finding pride in their country's accomplishments can prove to be difficult. Moreover, when people feel that, despite their best efforts, they cannot overcome the limitations imposed by their socioeconomic backgrounds, it can engender a sense of relative deprivation. The fact that the happiness level of South Koreans ranks among the lowest in the world may be intertwined with this demanding social environment. However, as the free market emerged and technology advanced, the limitations on individuals gradually waned, leading to the erosion of village traditions and extended family systems. For many people—especially the young—this transformation has created the need for new forms of socialization.

In South Korea, therefore, the metaverse has proven to be a successful technology because it addresses two opposing social forces: on one hand, the high level of competitiveness that makes innovation more acceptable than in Western countries; and on the other hand, the people's search for new spaces for self-expression, where they can freely build communities beyond social limitations.

### **Conclusions: Experiencing a Different (Better) World**

The case of South Korea underscores the importance of taking social and historical factors into consideration when introducing new technology

into a context. The emergence of the metaverse represents a significant paradigm shift in our interaction with the digital world, bringing forth a plethora of possibilities as well as challenges. While it holds the promise of revolutionizing the way we connect, share, and engage in virtual spaces, it also raises critical concerns that must be addressed to ensure a safe, inclusive, and equitable digital environment. The shortcomings in the current state of the metaverse are glaring, with a lack of proactive measures to safeguard users' safety and privacy and uphold ethical standards. The absence of clear, regulated systems and a lack of consensus on accountability, transparency, and human-centric design further exacerbate these issues. The lack of specific regulations in particular adds to the uncertainty and potential risks associated with the metaverse.<sup>20</sup>

Despite these challenges, the metaverse offers a beacon of hope and a unique opportunity to reshape our digital interactions. It has the potential to foster data sovereignty, empowering users to take control of their personal data, identity, and virtual destiny. The metaverse could also serve as a canvas on which to design a new world, rooted in fairness, justice, and enrichment, that ensures the benefits of the digital age are accessible to all—a human-centric metaverse.<sup>31</sup> Furthermore, the metaverse has the capacity to transcend the limitations of the physical world, providing unlimited spaces and virtual worlds that celebrate diversity and accessibility.<sup>32</sup> This digital utopia could act as an equalizer, eliminating biases related to gender, race, disability, and social status, paving the way for a more inclusive society. The South Korean case demonstrates that the metaverse can be a path to freedom and self-awareness in a society rooted in strong traditional values. It also opens the door for enhanced cultural communication and protections, contributing to the preservation and celebration of humanity's rich tapestry of cultures.

In light of these considerations, we believe it is important to encourage the involvement of key stakeholders, including policymakers, developers,

and users, in addressing the existing shortcomings in order to create an attractive and human-centered metaverse that is able to match users' expectations.

The metaverse has the potential to be a space where individuals can explore and create a world that is distinct and improved, surmounting the challenges presented by the real world. Having experienced this enhanced environment, people are likely to be more motivated to incorporate positive actions and behaviors into their lives, spanning both digital and physical realms. **C**


#### References

- Benjamins, R., Rubio Viñuela, Y., and Alonso, C. Social and ethical challenges of the metaverse. *AI and Ethics* 3 (2023), 689–697.
- Bernal, S.L., Celdrán, A.H., and Pérez, G.M. Eight reasons to prioritize brain-computer interface cybersecurity. *Commun. ACM* 66, 4 (Mar. 2023), 68–78.
- Bai, H., Zhang, L., Yang, J., and Billingham, M. Bringing full-featured mobile phone interaction into virtual reality. *Computers & Graphics* 97 (2021), 42–53.
- Boni, M. Ethical challenges related to the metaverse development: First hypotheses. In *Ethics. M. Radenkovic*, (Ed.). IntechOpen, Rijeka, (2023).
- Chen, S. et al. Exploring word-gesture text entry techniques in virtual reality. *Extended Abstracts of the SIGCHI Conf. on Human Factors in Computing Systems*. ACM, New York, (2019), 1–6.
- Choi, J.-S. The history of Confucianism in Korea. In *Confucianism in Context: Classic Philosophy and Contemporary Issues*. East Asia and Beyond, (2010), 33–52.
- Havele, A., Polys, N., and Behr, J. Building 3D web interoperability for the metaverse. In *Proceedings of the 28<sup>th</sup> Intern. ACM Conf. on 3D Web Technology*. ACM, New York, (2023).
- Havele, A., Polys, N., Benman, W., and Brutzman, D. The keys to an open, interoperable metaverse. In *Proceedings of the 27<sup>th</sup> Intern. Conf. on 3D Web Technology*. ACM, New York, (2022).
- Irani, L. et al. Postcolonial computing: A lens on design and development. In *Proceedings of the SIGCHI Conf. on Human Factors in Computing Systems*. ACM, New York, (2010), 1311–1320.
- Kim, S. and Kim, E. Emergence of the metaverse and psychiatric concerns in children and adolescents. *J. of Child Adolescent Psychiatry* 34, 4 (2023), 215–221.
- Kirkpatrick, K. Applying the metaverse. *Commun. ACM* 65, 11 (Oct. 2022), 16–18.
- Kugler, L. The state of virtual reality hardware. *Commun. ACM* 64, 2 (Jan. 2021), 15–16.
- Latoschik, M.E. et al. The effect of avatar realism in immersive social virtual realities. In *Proceedings of the 23<sup>rd</sup> ACM Symp. on Virtual Reality Software and Technology*. ACM, New York, (2017).
- Lee, S.M. and Lee, D. Untact: A new customer service strategy in the digital age. *Service Business* 14, (2020), 1–22.
- Li, K. et al. When internet of things meets metaverse: Convergence of physical and cyber worlds. *IEEE Internet of Things J.* 10, 5 (2023), 4148–4173.
- Mosco, V. Into the metaverse: Technical challenges, social problems, utopian visions, and policy principles. *Javnost—The Public* 30, 2 (2023), 161–173.
- Ng, T.K. What is the metaverse? Definitions, technologies and the community of inquiry. *Australasian J. of Educational Technology* 38, 4 (2022), 190–205.
- Nielsen, J. *Usability Engineering*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, (1994).
- Perri, D., Simonetti, M., Tasso, S., and Gervasi, O. Open metaverse with open software. In *Computational Science and Its Applications – ICCSA 2023 Workshops. Lecture Notes in Computer Science 14111*. Springer, Cham, (2023), 583–596.
- Rosenberg, L. Regulation of the metaverse: A roadmap: The risks and regulatory solutions for largescale consumer platforms. In *Proceedings of the 6<sup>th</sup> Intern. Conf. on Virtual and Augmented Reality Simulations*. ACM, New York, (2022), 21–26.
- Schulenberg, K. et al. Towards leveraging ai-based moderation to address emergent harassment in social virtual reality. In *Proceedings of the 2023 SIGCHI Conf. on Human Factors in Computing Systems*. ACM, New York, (2023).
- Shneiderman, B. Direct manipulation: A step beyond programming languages. *Computer* 16, 8 (1983), 57–69.
- Simiscuka, A.A. and Muntean, G-M. Synchronisation between real and virtual-world devices in a vr-iot environment. In *2018 IEEE Intern. Symp. on Broadband Multimedia Systems and Broadcasting*. IEEE, (2018), 1–5.
- Smith, C.H., Mokka-Danielsen, J., Rasool, J., and Webb-Benjamin, J-B. The world as an interface: Exploring the ethical challenges of the emerging metaverse. In *Proceedings of the 56<sup>th</sup> Hawaii Intern. Conf. on System Science*, (2023).
- Spence, E. Meta ethics for the metaverse: The ethics of virtual worlds. *Current Issues in Computing and Philosophy* (2008), 3–23.
- Takashina, T. et al. Real-virtual bridge: Operating real smartphones from the virtual world. In *Proceedings of the 2018 ACM Intern. Conf. on Interactive Surfaces and Spaces*. ACM, New York, (2018), 449–452.
- Takashina, T. and Kokumai, Y. Real-virtual bridge: a modular mechanism to mediate between real and virtual objects. In *SIGGRAPH Asia 2018 Posters*. ACM, New York, (2018).
- Huynh-The, T. et al. Artificial intelligence for the metaverse: A survey. *Engineering Applications of Artificial Intelligence* 117, (2023), 105581.
- Wang, H. et al. A survey on the metaverse: The state-of-the-art, technologies, applications, and challenges. *IEEE Internet of Things J.* 10, 16 (2023), 14671–14688.
- Weinberger, M. What is metaverse?—a definition based on qualitative meta-synthesis. *Future Internet* 14, 11 (2022).
- Yang, R. et al. The human-centric metaverse: A survey. In *Companion Proceedings of the ACM Web Conf. 2023*. ACM, New York, (2023), 1296–1306.
- Zallio, M. and Clarkson, P.J. Designing the metaverse: A study on inclusion, diversity, equity, accessibility and safety for digital immersive environments. *Telemat. Inf.* 75 (Dec. 2022).
- Slezak, T. The role of Confucianism in contemporary South Korean society. *Rocznik Orientalistyczny* 66, 1 (2013), 27–46.

**Tiziana Catarci** is a full professor of computer engineering and is currently head of the Department of Computer, Control, and Management Engineering at Sapienza University of Rome.

**Giuseppina De Nicola** is an associate professor and chair of the Korean Studies Program at the University of Torino.

**Daniel Raffini** is a research fellow in the Department of Computer, Control, and Management Engineering at Sapienza University of Rome.

 This work is licensed under a Creative Commons Attribution-NonCommercial International 4.0 License.



Watch the authors discuss this work in the exclusive *Communications* video. <https://caom.acm.org/videos/attractive-metaverse>

# research highlights

---

P. 86

**Technical  
Perspective  
How Exploits  
Impact Computer  
Science Theory**

By Sergey Bratus

P. 87

**Computing with Time:  
Microarchitectural  
Weird Machines**

By Thomas S. Benjamin, Jeffery A. Eitel, Jesse Elwell,  
Dmitry Evtvushkin, Abhrajit Ghosh, and Angelo Sapello

---

P. 96

**Technical  
Perspective  
Mirror, Mirror  
on the Wall,  
What Is the Best  
Topology  
of Them All?**

By Michela Taufer

P. 97

**HammingMesh:  
A Network Topology for  
Large-Scale Deep Learning**

By Torsten Hoefler, Tommaso Bonoto, Daniele De Sensi,  
Salvatore Di Girolamo, Shigang Li, Marco Heddes,  
Deepak Goel, Miguel Castro, and Steve Scott

# Technical Perspective

## How Exploits Impact Computer Science Theory

By Sergey Bratus

COMPUTING SYSTEMS AS WE build them today tend to have a curious property: Some combinations of inputs and external events cause them to behave against their builders' intent repeatedly and reliably. Techniques to make them behave so are called exploits. We say that an exploitable system is vulnerable—and the exploit is a constructive proof of vulnerability. Distressingly, exploits appear to be ubiquitous in both the software and hardware of our computing infrastructure.

Should exploits be a concern of computer science theory? Can they tell us about fundamental properties of computing rather than mere human errors of implementation? Or is there something about the fundamentals of computing that makes exploits endemic to our very models of computation?

The accompanying paper presents one of the finest pieces of evidence that says yes, they should, and yes, they can. It joins a growing body of examples of endemic exploits and of exploits having expressive power of general-purpose programming, with computation models as deep as those of CPUs, ISAs, and ABI/APIs. The authors introduce just such a model arising from the essential complexity of modern CPUs, which is only obviously suppressible by rejecting that essential complexity and the performance it delivers. With such results, we can see that exploitability is a deep computational property of the underlying system, calling for a comprehensive theory response.

That theory response is long overdue. Empirically, it is as if any implementation of the intended computing functionality invariably casts a long shadow of shocking yet repeatable emergent behaviors. These behaviors, rather than being sparse and fleeting, seem to inevitably form entire unintended but robust mechanisms that allow attackers to construct exploits despite multiple layers of security measures. It appears that no modern computing system

ends up being only and exactly what it was meant to be.


In the not-so-distant past, exploits could be dismissed as crafty but ultimately ad hoc and idiosyncratic inventions, with no big lessons for the computing theory or the natural science of its applications. In the 1990s and 2000s, the very existence of exploits seemed precarious, a mere unfortunate confluence of implementations—for example, of C/C++ functions activation frames containing both call stack return addresses and arrays prone to being overwritten by naively implemented string copy functions, and the x86 CPU stack memory being executable. It seemed a few well-poised changes to the hardware and the compilers—although still economically non-trivial—would destroy the space where most exploits lived. Exploits seemed too platform-bound and short-lived to need a theory, an engineering problem at best.

These times are now gone. Not only did the exploits demonstrate surprising portability and resilience, but it became clear their advanced techniques primarily reuse the target's own mechanisms and behaviors as designed, rather than some random and curiously deviant behaviors. It turned out the most effective exploit techniques leveraged the systems' own abstractions on levels well above the ultimate binary executable—gaining both portability to seemingly unrelated implementations and the reliability of already well-debugged and well-used code. Against these patterns of adversarial reuse of the target's own computing models, no set of discrete countermeasures would suffice—at least, not without substantial theories that consider the designed-in though unintended interactions between multiple models and levels of computation.

You may wonder about the term “weird machines” in the paper's title. It reflects the shift in the understanding of exploitability's root cause, from a programmer's error to an endemic

property of the target, a masterful reuse of the target's own mechanisms and features against it. Though important, the initial programmer's error is only one of the many doors to unlock this bounty of emergent execution. If closed, many others leading to the same emergent execution engine—the weird machine—will be found.

The road to this realization took several decades. It went from the naive understanding of stack buffer overflow exploits of the 1990s as needing native code payloads—indeed, the Windows XP Service Pack 2 advertised non-executable stacks as the mitigation of buffer overflows—to the realization the call stack machine embedded in every C/C++ program was Turing-complete on the sequences of well-formed stack frames with no executable content whatsoever, a.k.a. Return-oriented Programming. It went from heap exploits specific to Doug Lea's malloc in-band chunk metadata to all heaps and the “heap Feng-shui” co-optation across all major memory allocation algorithms and architectures. It went from Spectre and Meltdown being considered weird x86 bugs to 50+ families of emergent behaviors affecting all modern superscalar CPUs.

You now witness the next step in this succession: from understanding the transient space of microarchitectural optimizations as a locus of side channels to a general-purpose execution environment of its own, a weird machine par excellence. Read on and join the new age of emergent behavior exploration. More information and bibliography can be found at <https://bit.ly/3NEPf4y> 

**Sergey Bratus** is an associate professor of computer science at Dartmouth College and its Distinguished Professor in Cyber Security, Technology, and Society. He is a former program manager at the Defense Advanced Research Projects Agency (DARPA), and wrote this piece during his tenure at the agency.<sup>9</sup>

a Distribution Statement “A” (Approved for Public Release, Distribution Unlimited).

© 2024 Copyright held by the owner/author(s).



# Computing with Time: Microarchitectural Weird Machines

By Thomas S. Benjamin, Jeffery A. Eitel, Jesse Elwell, Dmitry Evtushkin, Abhrajit Ghosh, and Angelo Sapello

## Abstract

Side-channel attacks, such as Spectre, rely on properties of modern CPUs that permit discovery of microarchitectural state via timing of various operations. The Weird Machine concept is an increasingly popular model for characterization of execution that emerges from side effects of conventional computing constructs. In this work, we introduce microarchitectural weird machines ( $\mu$ WMs), code constructions that allow performing computation through the means of side effects and conflicts between microarchitectural entities such as branch predictors and caches. The results of such computations are observed as timing variations in the execution of instructions that interact with these side effects. We demonstrate how  $\mu$ WMs can be used as a powerful obfuscation engine where computation operates using events unobservable to conventional anti-obfuscation tools based on emulation, debugging, and static and dynamic analysis techniques. We present a practical example in which we use a  $\mu$ WM to obfuscate malware code such that its passive operation is invisible to an observer with full power to view the architectural state of the system until the code receives a trigger. When the trigger is received, the malware decrypts and executes its payload. To show the effectiveness of obfuscation, we demonstrate its use in the concealment and subsequent execution of a payload that creates a reverse shell. In the full version of this work, we also demonstrate a payload that exfiltrates a shadow password file. We then demonstrate the generality of  $\mu$ WMs by showing that they can be used to reliably perform non-trivial computation by implementing a SHA-1 hash function.

## 1. INTRODUCTION

The ability to model and classify a program's behavior is fundamental for a vast number of security-related tasks. It requires a form of emulator which implements a reference model of the target machine.

If this model deviates from the actual machine's behavior, key properties of many security mechanisms are violated. This can be exemplified by a proof-carrying code framework<sup>1</sup> that allows an arbitrary untrusted executable to run securely on a target platform. Security is established by the

target system checking a proof provided along with the executable. The proof ensures the executable *cannot perform any activity (or computation) outside of a formally specified policy*. Any deviation between the expected and actual target system behavior effectively violates the proof.

Many security mechanisms are based on either guaranteeing that the program cannot perform an action from the deny-list, or can *only* perform actions from the allow-list. Examples of such mechanisms include model checking, formal verification, taint analysis, control flow integrity enforcement, malware detection, and sandboxing.

Program obfuscation<sup>2</sup> is the general problem of transforming programs to prevent reverse engineering or other forms of analysis. While it is commonly used to hide malware, it can also be used to conceal benign sensitive code in proprietary applications<sup>8</sup> or to improve security.<sup>6</sup> A strong obfuscation engine can be constructed if the obfuscated program uses features of the target platform that are outside of platform's reference model used by the analyzer.

Recently, a number of papers introduced the concept of *weird machines* (WM)<sup>5,7,9</sup> to formalize certain classes of exploits. According to this concept, an exploitable vulnerability not only provides access to otherwise protected data but creates a new computational model within the original program. Such models exhibit emergent behaviors that are complex and not intended by the original system design, often violating fundamental security properties. Weird machines based on this model can be programmed, and programming is achieved with data that is passed to the vulnerability.

WM primitives can be used as powerful obfuscation engines. Previous research demonstrated the presence of such primitives in common implementations of various software and hardware components. Programming these WMs does not require vulnerabilities. For instance, Turing-complete WMs were built using little-known artifacts inside the page-fault handling hardware,<sup>3</sup> ELF-loader,<sup>16</sup> and exception handler<sup>14</sup> mechanisms. These WMs have inherent obfuscation capabilities. They use features of computer systems not identified as dangerous by antimalware software, and they are naturally difficult to analyze. To the best of our knowledge, no universal WM detection approach has been proposed.

In this paper, we establish a new type of WM implemented using microarchitectural (MA) components of a CPU, their complex inter-component interactions, and how they affect the latency of common operations. We call such machines microarchitectural weird machines ( $\mu$ WMs). At a

The original version of this paper was published in *Proceedings of the 26<sup>th</sup> ACM Intern. Conf. on Architectural Support for Programming Languages and Operating Systems* (April 2021), 758–772.

high level, the computation is performed by executing regular instructions such as memory loads and stores, jumps, and conditional branches, and observing execution time. The  $\mu$ WM is constructed from three types of abstract components. Weird registers (WRs) are data-storing entities implemented using states of MA components. Weird gates (WGs) are basic computation components which transform data stored in WR according to their logic. WGs use entanglement of various MA component states and their side effects, such as aliasing, evictions, and speculative execution. Weird circuits (WCs) are ensembles of WGs and implement more complex logic. We demonstrate that the proposed computation framework can be used to perform general purpose computations.

Since reverse engineering and binary analysis tools do not emulate MA components, we believe that our framework can be used as a universal approach for program obfuscation. Moreover, even if detection tools include the MA layer of the system in their reference model, we argue that precise detection of WM computations is challenging due to their natural flexibility and differences across CPU architectures. In addition, we discuss how we found several surprising ways for  $\mu$ WMs to improve security. We believe this paper introduces a new research area by looking at components responsible for MA attacks from a different angle and studying them from the perspective of computation artifacts.

## 2. BACKGROUND AND MOTIVATION

All current processors can be specified at the architectural and microarchitectural layers. The architectural layer is defined by the ISA and represents the machine state composed from CPU registers, instruction pointer, and addressable memory. Programs interact with it directly by executing instructions. It is well documented and can be formally specified.<sup>13</sup> This layer is implemented by internal microarchitectural components that compose the microarchitectural layer. This layer is not directly accessible to programs. Modern-day CPUs incorporate a large number of performance optimization mechanisms, such as caches, prefetchers, and various buffers. Many of these mechanisms have internal data structures with a complex state space.

While the presence of these mechanisms is a well known fact, little data is available on their internal structure and operation apart from the textbook-level description. Moreover, outside of their effects on the execution time, these mechanisms are completely transparent to programs executing on the CPU. Yet, programs are capable of implicitly manipulating MA components by performing regular activity. This property is used in traditional MA side-channel attacks. For instance, a memory access having an address dependency on sensitive data triggers a change of state inside the CPU cache. The attacker can then probe the state and infer the secret data. This basic principal forms the foundation of  $\mu$ WMs.

### 2.1. $\mu$ WM use cases.

Considering the diverse range of applications for program obfuscation, including non-offensive purposes, ( $\mu$ WMs) can be used in various scenarios. Next, we present a selection of these use cases.

*Hiding malware.* Malicious functionality in sensitive applications can be easily obscured by implementing it using  $\mu$ WMs. Malware can avoid being detected by dynamic or static analysis tools if code sequences used in malware are implemented using  $\mu$ WMs. Moreover, doing so provides strong anti-debug protection since MA state is not visible by a regular debugger and is highly volatile. For the same reason,  $\mu$ WMs can be used to implement a logic-bomb or trojan application,<sup>12</sup> which appears benign but activates its malicious functionality when triggered.

*Preventing reverse engineering.* Obfuscation techniques can be used to prevent reverse engineering applications for protection purposes. For instance, software developers may want to execute a secret algorithm on a third-party untrusted machine without disclosing the algorithm's internals.  $\mu$ WMs can be used for this purpose since reverse engineering requires an understanding of complex MA effects, which is a difficult task as we demonstrate later in this paper.

*Preventing emulation.*  $\mu$ WMs exploit unique features of a CPU's internal components and their interactions, such as address conflicts and race conditions. Emulating such effects with acceptable precision is extremely difficult as it would require accurate reverse engineering of the target hardware platform. Currently, existing cycle-accurate simulations only provide an approximate performance model and do not contain the level of detail required to emulate  $\mu$ WMs. We propose to use  $\mu$ WMs as an emulation detection/prevention tool, where computation can only be performed on real hardware.

*Violating formal proofs, sandboxes, taint analysis, preventing forensics.* Since currently existing analysis tools do not model the MA layer,  $\mu$ WMs can be used to perform activity outside of the security model. In addition, since  $\mu$ WM's current state is not located in regular memory but instead is encoded in the state of MA components, traditional forensics tools cannot be used to study  $\mu$ WMs.

### 2.2. Program obfuscation and the microarchitectural layer.

To illustrate the problem of program obfuscation, we consider two entities: the obfuscator and the detector. The obfuscator's objective is to perform a desired computation, often with malicious intent, while evading detection. The detector aims to determine whether the target program executes or has the capability to execute such a computation. Please note we use the term "computation" in a broader sense, referring to a sequence of actions or state transitions performed by a machine. A detector needs monitoring capabilities such as the ability to execute a target program, pass data, and observe a machine's state. Consequently, the obfuscator's goal is to modify the desired computation to circumvent specific conditions that a detector searches for. In traditional obfuscation techniques, this is often achieved through a variety of code transformation techniques.<sup>4</sup>

The effectiveness of a detector relies significantly on the extent of its monitoring capabilities. A highly advanced detector can possess complete visibility into the program's memory and registers, allowing it to single-step the target program and accurately model the behavior of the underlying

ing machine. While accurately modeling the machine's behavior at the architectural layer is achievable, modeling all aspects of the microarchitecture is currently infeasible.

Considering this, the microarchitectural layer emerges as a promising candidate for constructing an obfuscation framework. First, it provides a rich state space due to numerous structures implemented at the MA layer. Second, MA states are affected by programs executing on the machine, making it programmable by executing regular code. Third, the MA layer is usually not well documented, making it very difficult or impossible to create a perfect model.

Later in the paper we demonstrate how simple elements of the MA can be discovered, modeled as finite-state machines (FSMs), and manipulated to create basic computational primitives. Note that it is not necessary for this model to be a complete and full representation of the MA layer. Instead, the attacker can reverse engineer only a few components and manipulate them to evade detection.

Speculative execution is a common feature in processors which allows the CPU pipeline to perform conditional computations before the values of the conditions are fully determined. In particular, the pipeline relies on predictions from components such as branch predictors to guess the most likely instruction sequence and executes it immediately. If the prediction later is deemed incorrect, the CPU performs a roll-back and continues execution with the correct instruction sequence. However, during such erroneous execution, instructions from the mispredicted instruction sequence are allowed to make changes in MA components. This feature provides unique functionality for constructing  $\mu$ WMs. It allows the creation of a divergence between the architectural and microarchitectural state of the machine. In particular, to implement a  $\mu$ WM, the malicious executable may intentionally trigger a branch misprediction causing some instructions to be erroneously executed in speculative execution mode. Due to the eventual roll-back, these instructions will not trigger any state changes at the architectural layer; however they will cause state transitions at the microarchitectural layer. As a result, an analyzer with full visibility of the architectural state of the machine cannot detect malicious computation if its critical components are implemented via MA state transitions during an erroneous speculative execution.

### 3. WEIRD REGISTERS AND GATES

In this section we introduce the concepts of weird registers and weird gates, basic building blocks for constructing  $\mu$ WMs. The former are used to store data during the WM's computations and are constructed from implicit manipulations of microarchitectural components. The latter represent a minimal functional unit of the WM, processing data in its registers.

#### 3.1. Weird registers.

Any machine or subsystem thereof that has computational capabilities can be formalized as an abstract finite state machine  $M = (S, \Sigma, \delta)$ , where  $S$  is a finite set of states,  $\Sigma$  is the input alphabet, and  $\delta: \Sigma \times S \rightarrow S$  is a transition function. Each state  $s_i \in S$  represents a unique configuration of

all of the machine's internal components. While this level of detail may seem impractical, it is possible to simplify complex FSMs by limiting the number of observable states. This simplification creates a new FSM with a reduced number of states, input symbols, and a simpler transition function. However, this simplified FSM still encapsulates the computational logic inherent in the original machine. For instance, if the focus is on analyzing cache events, it may be beneficial to consider analyzing the cache FSM rather than treating the entire machine as an FSM.

We refer to such simplified FSMs as sub-FSMs or sFSMs. These sFSMs are employed to capture and represent specific behavior within a more complex system. We use these sFSMs to construct simple computational devices that will be used for obfuscation. Specifically, they are used to implement data storage primitives represented by WRs and computational primitives represented by WGs. We begin our discussion by explaining the construction of a WR.

By definition, a sFSM does not have full information about the original machine  $M$ , but it is useful for analyzing a specific aspect of the machine. Suppose there is a MA resource that we want to use as a storage entity and construct a WR  $r$ , for example the CPU data cache. We first select some variable, `var`, in the program's memory. Then a simple sFSM  $M_r = (S_r, \Sigma_r, \delta_r)$  can be defined with a small set of observable states  $S_r$ . For instance, we reduce the state space associated with `var` to only two states. Let those states be  $S_r = \{s_0, s_1\}$ .  $M_r$  is in state  $s_0$  when `var` is not cached and is in  $s_1$  when `var` is cached. It is possible to construct a more intricate FSM by taking into account other aspects of the cache such as cache level and Least Recently Used (LRU) status. We choose not to do so in this work. The state-transition logic for this sFSM is simple. When the variable `var` is accessed, the  $M_r$  transitions to state  $s_1$ . When `var` is flushed from the cache through the execution of the `clflush` instruction, the sFSM transitions to state  $s_0$ . These transitions occur irrespective of the current state of the sFSM. This establishes the input alphabet  $\Sigma_r$  for  $M_r$ . In particular,  $\sigma_{r_0} = \text{flush}(\text{var})$  and  $\sigma_{r_1} = \text{access\_mem}(\text{var})$ . Then  $\delta_r$  accepts symbols of this alphabet and triggers state transitions as previously described. This allows us to implement a weird register, a basic 2-bit microarchitectural storage primitive which uses the CPU data cache for storage. We refer to this register as DC-WR for "data cache weird register."

The state of the DC-WR is read by timing the number of CPU cycles it takes to access the chosen memory location. Please note that reading DC-WR register state is an invasive operation. It causes  $M_r$  to transition to state  $s_1$ . Therefore we introduce an additional signal,  $\sigma_{r_1} = \text{read}(r)$  for the corresponding sFSM. Processing the read instruction (passing  $\sigma_{r_1}$ ) causes the same state transition as  $\sigma_{r_1}$  previously defined but has the side effect of storing the access time in a CPU register. We define  $r$  to have a logic value of 0 when it is in state  $s_0$  (not cached) and to have logic value 1 when it is in state  $s_1$  (cached). It takes fewer CPU cycles to load `var` when it is in cache. Therefore, we determine the logic value of  $r$  by executing the `read(r)` instruction which has the side effect of placing that timing in an architecturally visible CPU register. If the load time is greater than a certain threshold

**Table 1. Examples of WR using various MA resources.**

Primitive	Write bit (0 or 1)	Read bit
d-cache	0: <code>clflush(var)</code> , 1: <code>ld var</code>	measure cycles to access variable
i-cache	0: <code>flush(code)</code> , 1: <code>call code</code>	measure cycles to execute code
ROB contention	0: execute <code>nop</code> instructions, 1: execute instructions with dependencies	execute any instructions and detect stalls
mul func. units	0: execute <code>nop</code> instructions, 1: execute <code>mul</code> instructions	measure cycles to execute <code>mul</code>
Branch direction predictor <sup>11</sup>	0: train conditional branch to predict non-taken, 1: train conditional branch to predict taken	execute branch non-taken and measure cycles
BTB <sup>10</sup>	0: execute <code>jmp</code> from A to B, 1: execute <code>jmp</code> from A to C	measure cycles to execute <code>jmp</code> from A to B
Intel VMX	0: execute <code>nop</code> instructions, 1: execute VMX instructions	measure cycles to execute a single VMX instruction

logic 0 is registered, otherwise it is logic 1.

L1 cache state is one of many computer subsystems with microarchitectural resources that can be explicitly or implicitly manipulated into states that can be made architecturally visible by means similar to the DC-WR. Table 1 provides some examples of WR that can be constructed using these other subsystems. In addition to using MA subsystems that have internal storage functionality, WRs can be implemented through modulating contention on MA resources. Examples of such WRs include registers based on `mul` instructions and the Reorder Buffer (ROB) listed in the table.

The concept of a WR can also be applied to formally analyze microarchitectural covert and side channels.<sup>18</sup> To construct a covert channel, the sender and receiver repeatedly write and read the same WR. A side channel is formed when a victim program unintentionally writes to a WR, which is subsequently read by an adversary. We believe many microarchitectural covert or side channels can be abstracted as a WR and therefore can potentially support  $\mu$ WM execution.

In addition to the data storage capabilities, WRs possess distinct properties that prove valuable in the context of obfuscation.

*Volatility.* Many states of microarchitectural entities are ephemeral in nature and exist only for a brief duration. For example, one can create a WR from two states of a limited pipeline resource, such as a multiplication unit, which can be in two contention states: high and low. This register will hold its value for several cycles and then default to the value associated with low contention.

*State decoherence.* Reading a WR can destroy its value since the reader interacts with the MA resource, for example by accessing memory and measuring the latency. Note, that other normal system activity can also interfere with the corresponding MA resource and destroy data in the WR. This property makes it challenging for a potential analyzer to observe  $\mu$ WM's state and apply forensics techniques.

*Entanglement.* Many MA resources are intricately interconnected, often in non-obvious ways. For instance, assigning a value to a data cache-based WR requires the execution of code, for example, a `mov` instruction. This in turn triggers

activity in the instruction cache. Consequently, interacting with data-cache WRs can impact other WRs implemented using instruction cache. While this may initially be perceived as an unwanted side effect, such interference gives rise to distinctive emergent properties. We leverage this property later in the paper to construct WCs.

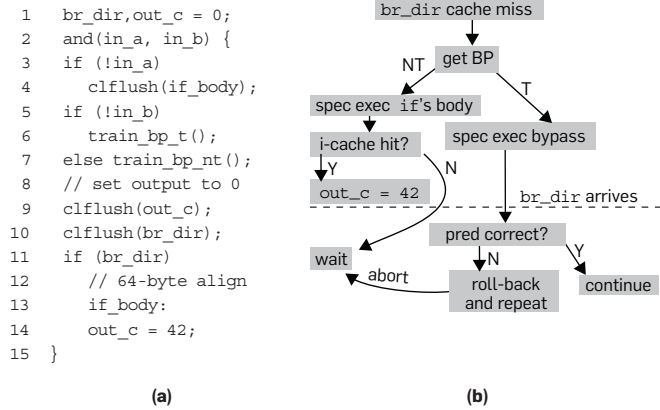
### 3.2. Weird gates.

The WG abstraction builds on that of the WR. WGs are basic elements of computation that exploit connection between different MA entities and their corresponding WRs. A WG is a code construction that implicitly invokes an activity in MA components in which the state of one or more WR (input WR) conditionally changes the state of one or more WR (output WR) thus performing computational logic. The WGs we discuss in this paper can be viewed as the implementation of logic gates such as AND, OR, and NAND. The WG abstraction includes more complex constructs that do not necessarily have two level logic output. We have experimentally verified operation of such gates, but we choose to leave their description to future work. Among the WGs for which we provide experimental results in Section 6 are NAND gates. This suffices to demonstrate universality of WGs as it is known that any arbitrary logic gate may be constructed using NAND gates. We have proofs of Turing Completeness of  $\mu$ WMs that fall outside the scope of this article.

#### 3.2.1. Weird AND gate.

One of the simplest WGs we demonstrate in this paper is a gate that implements a logical AND operation which ANDs two input weird registers. In particular, we use WRs implemented based on the branch predictor and instruction cache as input. The WG's pseudocode and operation flow-chart are presented in Figure 1. Please note that for simplicity we combine gate code together with input WR assignment operations in a single function. In real  $\mu$ WMs, these will be performed separately. Operation of the gate is based on the following set of observations. The branch predictor can either correctly or incorrectly predict the direction of a conditional branch instruction. This depends on the branch direction history. When the direction is incorrectly predicted, erroneous speculative execution is activated. It is possible to intentionally mistrain the branch predictor to do so at a desired time. One can make a branch execute in the taken direction multiple times followed by a single execution in the not-taken direction. During the final execution, the branch predictor mispredicts and triggers erroneous speculative execution. To prevent the CPU from detecting the misprediction and terminating speculative execution quickly, the data that is used for determining branch direction must be evicted from cache. There is another condition required for speculative execution to proceed. The code from the mispredicted direction of the branch must be in instruction cache (IC). Otherwise, the CPU will generate an IC miss and wait for the code to become available. If the delay is long enough, the branch misprediction will be detected and speculative execution terminated in the meantime. On the contrary, if the code

**Figure 1. Pseudocode of an AND WG (a) and its workflow (b).**



is in IC, these instructions will execute and change the MA state. This race condition enables the implementation of fundamental conditional logic and serves as a fundamental building block for WGs.

The first input weird register for the gate is a WR implemented using the branch predictor state associated with the gate’s `if` statement (line 11). We refer to it as BP-WR. The BP can be trained into one of two states. In state 0, the BP will be trained to not speculatively execute a block of code (line 14). In state 1, it will execute the code. In the pseudocode from Figure 1, setting the BP-WR to state 1 is shown as `train_bp_t()`. The NT stands for *not taken*, training the branch predictor to the not taken state causes the speculative block to be executed. `train_bp_nt()` sets the BP-WR to 0 since the BP taken state causes the speculative code to not be executed. Our Weird AND gate uses this BP-WR as one of its two inputs.

The output of our AND gate is in a DC-WR associated with a variable `out_c`. The body of the `if` statement (not taken branch) contains a memory access to this variable. If the BP-WR is in state 1, then the speculative execution is activated, and the state of the DC-WR will be set to 1 (in cache). We always set the DC-WR to 0 by flushing `out_c` from cache prior to gate execution.

The second input WR to our AND is an instruction cache WR (IC-WR). The code for the speculative block containing the DC-WR access is either in the instruction cache (state 1) or that code is not (state 0). Due to the limited duration of the speculative window, if the code is not in the IC then it will not be executed. When we combine these two input WRs, we see that the DC-WR memory access will *only* occur if both BP-WR and IC-WR are set to 1.

Note that the operation of this gate is architecturally invisible. While the inputs and outputs are visible, the actual AND logic makes no call to any kind of CPU AND instruction. The part of the WG that modifies the DC-WR state only occurs in speculative mode which has no architecturally visible effects.

### 3.2.2. Other weird gates.

Using the approach presented above, it is possible to implement other logical gates. For instance, the AND gate can be

**Table 2. Performance and accuracy of representative WGs.**

Weird Gate	Iterations	Execution Time (s)	Executions/Second	Accuracy
AND	1M	15	66,666	100%
OR	1M	57	17,543	98%
NAND	1M	13	76,923	100%
AND_AND_OR	1M	81	12,345	99.4%
TSX_AND	1M	0.591	1,692,047	98.5%
TSX_OR	1M	0.591	1,831,501	97.9%
TSX_ASSIGN	1M	0.42	2,380,952	98.5%
TSX_XOR	1M	16.6	60,020	99.2%

converted into an OR gate if the code is modified in such a way that triggers memory access to the variable `out_c` either when code is cached in IC or when the branch predictor is properly trained. For more details please refer to the full version of our paper.

In addition to the aforementioned gates, we composed and studied other logical gates. We were able to create several kinds of gates that work with a high degree of accuracy. Table 2 provides an overview of performance and accuracy for a representative sample of such WGs. Note, Table 2 also includes TSX-based gates that we discuss in the next section. We provide additional evaluation details in Section 6.

## 4. WEIRD CIRCUITS

WGs described in Section 3.2 enable a basic framework for constructing  $\mu$ WMs. A computation is first expressed as a boolean circuit and then divided into a sequence of individual register and gate operations. This model of operation requires the outputs of each gate to be read from the output register into the architectural state of the program before it can be sent to the next gate’s input. For the WRs implemented using the data cache, reading the intermediate state is done by measuring the load time in CPU cycles to access the corresponding memory location via the `rdtscp` instruction. Then the state is written into a WR that is used as input for the next gate. There are several disadvantages to this approach. WR reading and writing operations require adding instructions to the base program which reduces performance. Moreover,  $\mu$ WM composed in such a way are less suited for obfuscation since the intermediate state of the  $\mu$ WM is stored in architectural memory. An advanced analyzer may be able to detect malicious patterns in state transitions of the architecturally visible program or inside the program’s memory.

Both of these limitations can be addressed by performing contiguous computation within MA state instead of using architectural state to enable the dataflow between weird gates. The goal is to enable contiguous ensembles of WGs that implement more complex binary functions than individual gates without saving the intermediate state in architecturally visible memory. We refer to such ensembles as weird circuits. In a WC, data is copied into the MA layer only once, and then a series of WGs are activated in such a way that the output of one gate serves as input for another gate.

The intermediate data is stored inside WRs for the duration of the WC activation.

To describe how WCs are formed and operate consider a minimal WC consisting of two AND gates connected in series and implementing a binary expression  $c = a \& b \& a$ . Assume WRs  $a$ ,  $b$ , and  $c$  are implemented as in Section 3.2.1. Since  $a$  and  $b$  are purely input WRs and  $c$  is a purely output WR, the binary expression can be rewritten in the following way:  $c = a \& b$ ;  $b = c$ ;  $c = a \& b$ . In this way, the binary expression is translated into a sequence of basic operations, individual gate activations, and WR-to-WR transfers. To make this computation possible without copying the intermediate state into the architecturally visible memory, our WC needs to have two properties:

**P1:** Individual WG operations need to be contiguous. This means that activating a gate one time does not affect its consequent behavior.

**P2:** Transferring values between different registers must be possible to exchange values between input and output registers.

Previously described WGs lack both of these properties. First, the gates use branch predictor mistraining to activate erroneous speculative execution required to create the necessary race condition. The mistraining becomes challenging for multiple consequent gate activations. Modern branch prediction units (BPUs) are known for accurately predicting complex branch patterns. When the WG code attempts to repeatedly mistrain a certain branch, the BPU quickly learns this pattern and begins predicting the branch direction correctly. This violates **P1**.

**P2** cannot be fulfilled due to the use of WRs of different types and the lack of hardware interfaces to transfer the state between separate MA entities. For example, consider a case when we need to assign the value of  $c$  implemented as a DC-WR to another WR  $b$ , implemented as a BP-WR. In this case, we need to *conditionally* train the BPU depending only on the state of another MA entity: the data cache. Unfortunately there is no simple way to achieve this. At the same time, transferring the state within a single MA entity appears possible. Suppose we have two DC-WR  $e$  and  $f$  implemented using variables  $d0$  and  $d1$  correspondingly. By storing the address of  $d1$  in  $d0$  we can implement a basic WR assign functionality ( $e = f$ ). It is done by simply dereferencing the pointer ( $*d0$ ) while in speculative execution. Under the race condition, variable  $d1$  will be accessed *only* if  $d0$  is cached, enabling the conditional assign operation.

To overcome these challenges, we need to implement a new WG mechanism that does not rely on BPU mistraining and uses WR of the same type for all input and output gates. While alternative implementations are possible, for this paper we focus on the WCs we implemented based on Intel Transactional Synchronization Extensions (TSX) technology. Introduced in the Haswell microarchitecture, TSX provides CPU-level transactional memory operations using `XBEGIN`, `XEND`, and `XABORT` instructions. When a running program issues the `XBEGIN` instruction the CPU enters a transactional mode in which operations are executed until either the `XEND` instruction is encountered or an error condition occurs. When the CPU encounters an

`XEND` instruction with no error, then all effects from execution (such as memory reads and writes) are committed and become visible on the architectural level. If an error or fault occurs during the transaction, then the executed code is rolled back such that there are no architecturally visible effects and the CPU continues execution at an address specified as an argument to the `XBEGIN` which is typically a fault handler.

However, as indicated by prior work,<sup>15</sup> the execution is not stalled immediately. The pipeline continues to execute instructions even after the fault. This introduces a new source of speculative execution which we use for WG construction. MA side effects from this speculative execution are not rolled back upon leaving the TSX code. Many conditions, such as page faults and divide-by-zero operation, will cause a TSX transaction to abort.

We create a window of speculative execution simply by including a divide-by-zero error in each TSX block. In our experiments, we observed that the transaction blocks exhibit a longer and more stable window of speculative execution than BPU mistraining. At the same time, multiple TSX-based WG can be strung in a row such that they compose a more complex WC that performs calculations in a serial fashion with no architecturally visible intermediate results. They also make it impossible for standard debugging techniques to be used for observing the operation of a TSX-based WC. A requirement of the transaction interface is that no part of the transaction becomes architecturally visible unless the entire transaction completes with no other thread accessing memory used in the transaction. If an external debugger were able to observe what was happening in a transaction block, that would by definition be a side effect and would cause an abort. The debugger would see the `XBEGIN` instruction, then the next instruction would be the beginning of the abort handler.

Our TSX-based weird gates are based on the observation that the duration of speculative execution occurring inside a TSX code block upon a fault is limited. This creates a race condition. Assume a code sequence consisting of three instructions,  $i1$ - $i3$ , that are executed inside a TSX transaction block with a fault. In this case, whether or not these instructions have a chance to execute and alter the MA state depends on their performance. For instance, if instructions do not have any memory dependencies, their execution time is low and they are likely to be executed. If they require data from RAM, their execution time is unlikely to fit inside the speculative window. This phenomenon creates a basic primitive needed to construct logic gates. We can introduce dependencies between instructions by grouping them using arithmetical operations. For example, in the code sequence below, the last instruction will be able to issue a store request and modify the MA state *only* if variables  $d0$  and  $d1$  are both cached. This effectively creates an AND weird gate with DC-WR registers as input and output.

```
i1: mov d0, %r1;
i2: add d1, %r1;
i3: mov (%r1), %r2;
```

Figure 2 contains pseudocode for a sample TSX WC which simultaneously calculates two logical operations, AND and

**Figure 2. Pseudocode for TSX weird circuit that computes  $(Q_0 \leftarrow A \wedge B, Q_1 \leftarrow A \vee B)$ .**

```

1 #define ADDR(dx) // returns address pointed by dx
2 flush(*d0, *d1, *d2, *d3);
3 if (A) { tmp = *d0; } // d0 := A
4 if (B) { tmp = *d1; } // d1 := B
5 TSX_AND_OR { // Execute Weird Circuit
6 XBEGIN;
7 tmp = tmp / 0; // abort transaction
8 tmp =>(*d0 + ADDR(d3)); // d3 := d0
9 tmp =>(*d1 + ADDR(d3)); // d3 := d1
10 tmp =>(*d0 + *d1 + ADDR(d2)); // d2 := d0 & d1
11 XEND;
12 }
13 // read output
14 XBEGIN;
15 t1 = rdtscp();
16 tmp = *d2;
17 t2 = rdtscp();
18 tmp = *d3;
19 t3 = rdtscp();
20 Q0 = (t2 - t1) < TIMING_THRESHOLD
21 Q1 = (t3 - t2) < TIMING_THRESHOLD
22 XEND;

```

OR, in a single WC based on the principle that cache status of operands to addition will determine whether the addition will be performed. In this pseudocode  $d0..d4$  are variables that implement DC-WRs. Absence from cache representing logic 0 and presence in cache is logic 1. Line 2 initializes all the DC-WR to logic 0 by flushing the memory to which they point. In lines 3 and 4, the architecturally visible inputs  $A$  and  $B$  are read into  $d0$  and  $d1$ . The division-by-zero guarantees that the code inside the TSX block will execute inside erroneous speculative execution mode. Lines 8 and 9 implement the *OR* logic of the WC. If either  $d0$  or  $d1$  are cached (logic 1), the CPU will be able to calculate the address pointed to by  $d3$  and dereference it during the speculative execution window. If both of them are not cached (logic 0), speculative execution will terminate before the memory access begins. Note that the architectural value stored at addresses pointed by  $d0$  and  $d1$  can be set to 0. Line 10 implements the functionality of logic AND. To successfully make a memory access to the address pointed by  $d2$ , both of the input WRs must be cached (logic 1).

After the WC has executed, we read the WR values into visible memory. We want to avoid having the `rdtscp` instructions visible to an observer. We therefore perform the timed memory load inside a TSX transaction. An adversary attempting to observe the process of reading the WR will cause that transaction to abort, which destroys the value of the WRs and leaves the architecturally visible outputs  $Q0$  and  $Q1$  set to 0. Intentionally causing such aborts to disrupt malware weird circuits hidden in a legitimate program is an interesting line of future work. It will, however, interfere with proper execution of the legitimate program if done in a naïve way.

We implemented a number of different logic gates using the TSX approach. A list of these gates, including the XOR gate, and their performance data is shown in Table 2. The XOR gate is specifically useful as it can be used in simple one-time-pad (OTP) schemes to encrypt/decrypt data.

## 5. APPLICATIONS OF WEIRD CIRCUITS

In previous sections we explored the design and implementation of simple WCs that demonstrate the ability to create functionally complete, microarchitecturally invisible boolean WCs. In this section, we will first demonstrate weird obfuscation (WO), a malware obfuscation system that uses a more complex WC. Then, we will examine in greater depth the multi-gate TSX-based weird XOR circuit used by the WO system. Finally, we will demonstrate an implementation of SHA-1 that uses weird circuits.

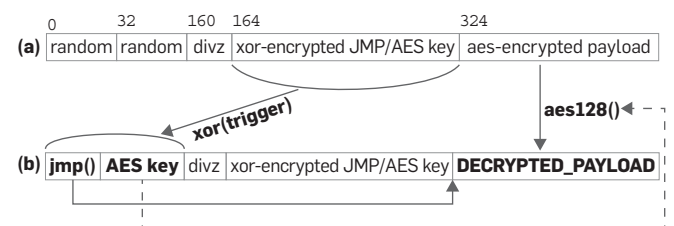
In this section we describe the operation of our WO system and how we use this system to obfuscate malware code. We show that the malware’s passive operation is invisible to an observer with full power to view the architectural state of a system until the code receives a ping with a special trigger value in the body. When the trigger is received, the malware decrypts and executes its payload. We will demonstrate the use of our WO system to conceal and then execute a payload that creates a reverse shell. In the full paper, we also demonstrate shadow password exfiltration.

In our scenario, we take on the role of an attacker who has managed to get an advanced persistent threat (APT), such as a trojan horse, installed onto a computer running inside an adversary’s network. Our adversary, the defender, has the ability to view all architectural state of the infected computer. Our adversary has the power to run our infected program in a debugger or other dynamic analysis tools. We hope this work will inspire future work for development of static analysis tools that will detect and characterize  $\mu$ WM in programs such as our APT, but as discussed in Section 1, those tools do not yet exist. We therefore do not give our adversary abilities granted by those theoretical tools.

When constructing our APT, we first take the payload, choose a random 128-bit AES key, encrypt the payload with that key, and store the encrypted payload in the structure shown in Figure 3 a) starting at bit 324. We then place a specially crafted `jmp` instruction at bit 164 followed by the AES key. Next, we create a random one-time-pad of length 160 bits and XOR each bit of the pad against the bits of the memory structure starting at bit 164. This has the effect of “encrypting” the `jmp` instruction and the AES key against the one-time-pad. The 160 bit one-time-pad will later be used as the trigger value that will cause our malware to enter its active phase. We complete the preparation by filling the first 160 bits of the structure with random data followed by an illegal divide by zero instruction, then copying the entire structure into the body of a TSX block.

Our APT is malware hiding in a program that receives

**Figure 3. `wm_appt` layout a) at start and b) after valid trigger.**



pings. Each ping body is used as an XOR key to transform the memory labeled `xor-encrypted JMP/AES key` in Figure 3 and overwrite the bits labeled `random`. Bits 32-160 are then used as an AES key to decrypt the payload at the end of the memory region. Finally, the entire region is `mmap'd` and executed inside a TSX block. If the secret key in the ping body was correct, it creates a `jmp` instruction leading to the target function that will begin execution of the decrypted payload. When the pad and AES keys are correct, the payload will execute properly and open a reverse shell to the attacker.

During the silent phase, before the attack is triggered, the affected machine may receive many pings. When a received ping does not contain the trigger value, the first 160 bits of the TSX block will contain a bad AES key resulting in garbage values in the decrypted payload, and no `jmp` instruction. Instead of properly jumping over the contents of the AES key and divide-by-zero instruction, this incorrectly-decrypted region is executed as is. This will generally cause a near-immediate fault but is guaranteed to fault by the time it reaches the `tmp = tmp/0` instruction at bits 160-164. This fault is then rolled back since it is inside a TSX block, and the program continues to wait for the next ping trigger.

All critical parts of this APT operate in TSX blocks, which are not directly observable by a debugger. In addition, the one-time-pad “decryption” of the AES key is performed by a TSX-based XOR WG that has no architecturally visible intermediate values. The analyzer will not see any part of the payload until the trigger has been successful and the payload is already running.

Execution of the logic gates underlying the TSX-based XOR, as previously discussed, is not 100% accurate. Practically, this means each trigger must be evaluated by the APT multiple times. We chose this evaluation multiple to be 10. In our implementation, the APT is able to process pings in real time with inter-arrival times up to 500ms. Table shows the distribution of the number of pings required to successfully decrypt and execute the reverse shell malware payload in 100 experiments. It takes on average 6 attempts (6 pings) to successfully XOR all of the 160 bits and execute the payload.

### 5.1. SHA-1 implementation.

We chose a hashing algorithm to be an illustrative high-level algorithm with which to demonstrate partially architecturally visible  $\mu$ WM for a number of reasons. A cryptographic hashing function provides a challenging case for  $\mu$ WM which have components with less than 100% accuracy. Due to the nature of cryptographic hashes, single bit errors that occur during computation are magnified which makes SHA-1 a challenging test case. Another reason we chose a hashing algorithm is that our WC version can be used to replace the hash function in the architecturally visible malware obfuscation system due to Sharif et al.<sup>17</sup> on which one of our example  $\mu$ WM-based malware obfuscation systems is loosely based.

We call our our SHA-1 implementation “partially architecturally visible” because, while many interim values are

stored in architecturally visible memory, all of the actual SHA-1 computation is performed by  $\mu$ WM. For example, when the algorithm requires adding two numbers, no CPU add instructions are executed. The implementation performs the addition using a full adder constructed from two discrete weird XOR gates and a composed weird AND \_ AND \_ OR gate. During execution, the output of the weird XOR gates is temporarily stored in memory as is the output of the weird AND \_ AND \_ OR. In our initial implementation, 41.9% of the intermediate results were architecturally visible.

As discussed in previous sections, many of our weird gates have a high degree of accuracy, but in very long runs errors do occur. Our SHA-1 implementation uses redundant gate executions to provide the requisite accuracy. This is explained in greater detail in the full version of this work.

We ran a series of 10 experiments in which each experiment consisted of an execution of the SHA-1 implementation. For these experiments, we chose conservative redundancy parameters favoring accuracy over speed. Each of those experiments produced a correct hash. Each execution took around 26 minutes.

## 6. EVALUATION

In this article, we give a brief summary of our evaluation of some WGs and WCs we constructed. Details of methodology and further results appear in the full paper. One of the challenges for a developer programming with WGs is that successful execution of each WG depends on microarchitectural state that is not generally obvious to a developer. We created a framework, essentially a WC compiler, for building WCs we call `skelly` that abstracts away the the state of the microarchitecture. This framework is a static library that provides basic logic functions in `c` such as `int and(int a, int b)`; and automatically takes care of such considerations as intentional i-cache misalignment and optimization of number and position of NOP sleds needed for gate stability. `skelly` also offers optional redundancy features, which we used in our SHA-1 implementation. We used this framework to evaluate all WGs and WCs mentioned in this paper. For evaluation, we enabled architecturally visible output in the form of load times in CPU cycles on output data cache-based WRs. The load times map to logic values such that load times less than about (depending on gate) 100 cycles indicate a logic 1 and more than 100 cycles indicate a logic 0. This is because a short load time indicates something is in cache, which we defined to be logic 1 as described earlier in this article. The experimental results for our TSX WGs are the result of 64 000 executions per gate with random gate input. We found these gates to have high ( $\geq 93\%$ ) accuracy, which is shown in Table 3. Table 4 shows some statistics on how many CPU cycles are required to load the output WR from the TSX XOR gate. As expected, when the inputs are (0, 0) and (1, 1) the load times are  $> 100$  CPU cycles and when the inputs are (0, 1) and (1, 0) the load times are  $< 100$ . Figure 4 shows the Kernel Density Estimations (KDEs) of the load times for TSX-AND. A KDE is similar to a smoothed histogram charting the probability that it will

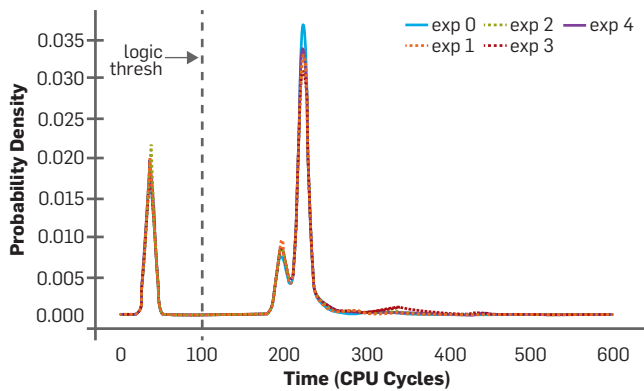


**Table 3. TSX WG accuracy.**

Gate	Correct Ops	TSX Aborts	Total Ops	Mean Accuracy
AND	62880	7	64000	0.98
OR	61922	9	64000	0.97
AND-OR	61152	12	64000	0.98
XOR	59259	8	64000	0.93

**Table 4. TSX-XOR Output WR load time (CPU cycles).**

Input	Min	Q1	Med	Q3	Max	Std Dev	Mean
0,0	31	220	222	228	20323	963.35	432.87
0,1	31	34	36	37	15656	360.44	75.40
1,0	31	34	36	37	16525	344.31	71.67
1,1	31	212	222	226	19200	883.15	382.07

**Figure 4. Output WR load time KDE for TSX AND.****Table 5. BP / i-cache WG accuracy.**

Gate	Operations	Correct	Mean Accuracy
AND	320000	319994	0.999982
OR	320000	319988	0.999963

take a given number of cycles to load the output WR. Probabilities well separated by a threshold value provide a visualization of gate stability. KDEs from four experiments of the same type are plotted on the same axis showing similar distributions between experiment runs. In evaluating the branch predictor / instruction cache-based AND & OR we performed 320 000 operations per gate type using random input. These WGs are extremely accurate (> 99.999%) as shown in Table 5.

## 7. CONCLUSION

We have introduced the concept of  $\mu$ WMs, a methodology for harnessing the computing capability provided via the unspecified aspects of CPU microarchitectures. We described a framework for programmatically storing and operating on MA state as WRs and WGs respectively using several MA components. We demonstrated the practicality of  $\mu$ WMs by creating a microarchitecture-sensitive logic

bomb as well as the implementation of a reasonably complex cryptographic algorithm, SHA-1. We believe that our work merely uncovers the tip of the iceberg and that  $\mu$ WMs will have strong applications in both offensive and defensive adversarial scenarios in the future. □

## References

- Appel, A.W. Foundational proof-carrying code. In *Proceedings 16<sup>th</sup> Annual IEEE Symp. on Logic in Computer Science*. IEEE, 2001, 247–256.
- Baldoni, R. et al. A survey of symbolic execution techniques. *ACM Computing Surveys (CSUR)* 51, 3 (2018), 1–39.
- Bangert, J., Bratus, S., Shapiro, R., and Smith, S.W. The page-fault weird machine: Lessons in instruction-less computation. Presented as part of the *USENIX Workshop on Offensive Technologies*, 2013; <https://www.cs.dartmouth.edu/~sergey/wm/woot13-bangert.pdf>.
- Behera, C.K. and Lalitha Bhaskari, D. Different obfuscation techniques for code protection. *Procedia Computer Science*, 70, 2015, 757–763.
- Benjamin, T. et al. Weird circuits in CPU microarchitectures. *Presentation, the Sixth Workshop on Language-Theoretic Security (LangSec)*, 2020; [http://spw20.langsec.org/slides/WeirdCircuits\\_LangSec2020.pdf](http://spw20.langsec.org/slides/WeirdCircuits_LangSec2020.pdf), Accessed: 2020-12-18.
- Bhatkar, S., DuVarney, D.C., and Sekar, R. Address obfuscation: An efficient approach to combat a broad range of memory error exploits. In *USENIX Security Symp.* 12, (2003), 291–301.
- Bratus, S. *What Are Weird Machines?*; <https://www.cs.dartmouth.edu/~sergey/wm/>. Accessed: 2020-12-18.
- Dalai, A.K., Sundar Das, S., and Jena, S.K. A code obfuscation technique to prevent reverse engineering. In *Proceedings of the 2017 Intern. Conf. On Wireless Communications, Signal Processing and Networking*. IEEE, 2017, 828–832.
- Dullien, T.F. Weird machines, exploitability, and provable unexploitability. *IEEE Transactions on Emerging Topics in Computing*, 2017.
- Evyushkin, D., Ponomarev, D., and Abu-Ghazaleh, N. Jump over ASLR: Attacking branch predictors to bypass ASLR. In *Proceedings of the 2016 49<sup>th</sup> Annual IEEE/ACM Intern. Symp. on Microarchitecture*. IEEE, 2016, 1–13.
- Evyushkin, D., Riley, R., Abu-Ghazaleh, N., and Ponomarev, D. Branchscope: A new side-channel attack on directional branch predictor. *ACM SIGPLAN Notices* 53, 2 (2018), 693–707.
- Fratantonio, Y. et al. Triggerscope: Towards detecting logic bombs in android applications. In *Proceedings of 2016 IEEE Symp. On Security and Privacy*. IEEE, 2016, 377–396.
- Michael, N.G. and Appel, A.W. Machine instruction syntax and semantics in higher order logic. In *Intern. Conf. On Automated Deduction*. Springer, 2000, 7–24.
- Oakley, J. and Bratus, S. Exploiting the hard-working dwarf: Trojan and exploit techniques with no native executable code. In *WOOT*, 2011, 91–102.
- Schwarz, M. et al. Zombieload: Cross-privilege-boundary data sampling. In *Proceedings of the 2019 ACM SIGSAC Conf. On Computer and Communications Security*, 2019, 753–768.
- Shapiro, R., Bratus, S., and Smith, S.W. "weird machines" in elf: A spotlight on the underappreciated metadata. Presented as part of the *USENIX Workshop on Offensive Technologies*, 2013; <https://www.cs.dartmouth.edu/~sergey/wm/woot13-shapiro.pdf>.
- Sharif, M.I., Lanzi, A., Giffin, J.T., and Lee, W. Impeding malware analysis using conditional code obfuscation. In *NDSS*, 2008.
- Szefer, J. Survey of microarchitectural side and covert channels, attacks, and defenses. *J. of Hardware and Systems Security* 3, 3 (2019), 219–234.

**Thomas S. Benjamin** (tbenjamin@peratonlabs.com), Peraton Labs, Basking Ridge, NJ, USA.

**Jeffery A. Eitel** (jeitel@peratonlabs.com), Peraton Labs, Basking Ridge, NJ, USA.

**Jesse Elwell** (jelwell@peratonlabs.com), Peraton Labs, Basking Ridge, NJ, USA.

**Dmitry Evtushkin** (devtyushkin@wm.edu), William & Mary, Williamsburg, VA, USA.

**Abhrajit Ghosh** (abhrajitghosh@meta.com) Meta Platforms, Inc., Menlo Park, CA, USA.

**Angelo Sapello** (asapello@peratonlabs.com), Peraton Labs, Basking Ridge, NJ, USA.

# Technical Perspective

## Mirror, Mirror on the Wall, What Is the Best Topology of Them All?

By Michela Tauffer

ARTIFICIAL INTELLIGENCE (AI) is one of the most important emerging technologies of the 21<sup>st</sup> century, and designing suitable infrastructure for large-scale AI systems is critical. Major companies such as Microsoft, Google, Meta, and even Tesla are touting large-scale “AI supercomputers” as an essential tool for increasingly powerful AI systems, but as AI systems have a more specialized workload than traditional supercomputers, designing and implementing their architecture is a complex process. The complexity is because AI systems are specialized for AI and machine-learning (ML) workloads, leveraging parallelism and specialized hardware accelerators to excel in processing and analyzing large datasets to make predictions, classify objects, and understand natural language. Traditional supercomputers, on the other hand, are general-purpose machines used for a broader range of scientific and computational tasks. The authors of the accompanying paper leverage the unique demands of specialized AI workloads to craft a network structure tailored for large-scale deep learning, which is a pivotal facet of AI.

The authors describe AI workloads by considering three dimensions of parallelism: data parallelism, pipeline parallelism, and operator parallelism. Data parallelism is used in AI workloads when training ML models. Pipeline parallelism can be observed in AI workloads in the execution of complex neural network models. In AI workloads, operator parallelism can be seen in the parallel execution of mathematical operations that make up neural network layers. The paper argues that, although each dimension is different, it can be implemented with nearest-neighbor communication, but today’s high-performance computing (HPC) networks often overprovision global bandwidth and underprovision local bandwidth for AI workloads.

This insight motivates the authors

to use toroidal networks that HPC has been using traditionally but abandoned in favor of more flexible low-diameter topologies based on switching technologies. *Torus network topologies* extend the torus concept into multiple dimensions. These networks offer efficient connectivity between processing nodes. Google’s early Tensor Processing Units (TPUs) also employed two- and three-dimensional torus networks for interconnection. This architecture facilitates efficient communication between TPUs in datacenters, which is crucial for ML workloads. While torus networks offer advantages like low latency and determinism, low-diameter torus topologies can suffer from limited global bandwidth. Scheduling and managing traffic on torus networks can be inflexible, leading to performance bottlenecks. More flexible *switch technologies*, such as the Dragonfly network, have gained popularity in HPC to address the limitations of low-diameter torus networks. Switched topologies generally use switches to route data efficiently between networked devices. Switched networks use switches as part of their design to route data efficiently between networked devices and thus provide better global bandwidth and improved flexibility in routing data, making them suitable for large-scale parallel processing.

The paper proposes combining the best of both worlds (that is, the torus topologies’ cost-effectiveness and switched topologies’ performance) into *HammingMesh*, a novel network topology that provides high bandwidth at low cost for deep-learning training jobs. A similar approach was recently presented in Google’s TPUv4. In *HammingMesh*, the authors propose connecting a set of 2D meshes with switches to form virtual torus topologies of varying sizes. Local connections can be implemented as thin, conductive links or traces on printed

circuit boards (PCBs), thus very inexpensive. Only traces that leave a board are connected to discrete switches, so the number of switched traces is immediately halved in a 2x2 board and quartered in a 4x4 board, which is a significant cost savings. The lower cost per link can satisfy high-bandwidth requirements; installing many parallel connections can achieve multiple TBs of bandwidth at a reasonable system cost. The paper shows simulation results evaluating various AI workloads in detail, showing that price and performance gains also translate to complex workloads.

The authors demonstrate how a system deploying the *HammingMesh* topology can deal with failures and varying job allocations. Node and board failures are handled gracefully by swapping in “virtual boards,” scheduling is flexible because the topology can permute each row and column and still achieve full bandwidth. The paper shows several real-world traces scheduled to the topology and demonstrates that it achieves consistently high utilization even during failures.

AI is a fast-moving field with new algorithms published every week. As large-scale decoder architectures such as GPT-4 dominate large parts of the market, the ML technique Mixture of Experts (MoE) emerges and indicates a direction toward sparsity. Large-scale deep learning will shift more to sparse models, as sparsity can produce better data science results and more efficient computing performance and cost. This architectural rethinking has already begun. Only a Magic Mirror can tell whether *HammingMesh* remains the best topology for future workloads. Still, it is a strong contender in network designs for AI systems. 

**Michela Tauffer** holds the Jack Dongarra Professorship in High Performance Computing in the Department of Electrical Engineering and Computer Science at the University of Tennessee, Knoxville, TN, USA.

© 2024 Copyright held by the owner/author(s).

# HammingMesh: A Network Topology for Large-Scale Deep Learning

By Torsten Hoefler, Tommaso Bonoto, Daniele De Sensi, Salvatore Di Girolamo, Shigang Li, Marco Heddes, Deepak Goel, Miguel Castro, and Steve Scott

## Abstract

**Numerous microarchitectural optimizations unlocked tremendous processing power for deep neural networks that in turn fueled the ongoing AI revolution. With the exhaustion of such optimizations, the growth of modern AI is now gated by the performance of training systems, especially their data movement. Instead of focusing on single accelerators, we investigate data-movement characteristics of large-scale training at full system scale. Based on our workload analysis, we design HammingMesh, a novel network topology that provides high bandwidth at low cost with high job-scheduling flexibility. Specifically, HammingMesh can support full bandwidth and isolation to deep learning training jobs with two dimensions of parallelism. Furthermore, it also supports high global bandwidth for generic traffic. Thus, HammingMesh will power future large-scale deep-learning systems with extreme bandwidth requirements.**

## 1. MOTIVATION

Artificial intelligence (AI) is experiencing unprecedented growth providing seemingly open-ended opportunity. *Deep learning* models combine many layers of operators into a complex function *trained* by optimizing its parameters to large datasets. Given the abundance of sensor, simulation, and human artifact data, this new model of designing computer programs, also known as data-driven programming or “software 2.0”, is mainly limited by the capability of machines to perform the compute- and data-intensive training jobs. In fact, the predictive quality of models improves as their size and training data grow to unprecedented scales.<sup>15</sup> Building *deep learning supercomputers*, to both explore the limits of AI and commoditize it, is becoming not only interesting to big industry but also humanity as a whole.

A plethora of different model types exist in deep learning and new major models are developed every two to three years. Yet, their computational structure is similar—they consist of layers of operators and they are fundamentally *data-intensive*.<sup>14</sup> Many domain-specific accelerators take advantage of peculiarities of deep-learning workloads

be it matrix multiply units (“tensor cores”), specialized vector cores, or specific low-precision datatypes. Those optimizations can lead to orders of magnitude efficiency improvements. Yet, as we approach the limits of such microarchitectural improvements, we need to direct our focus to the system level.

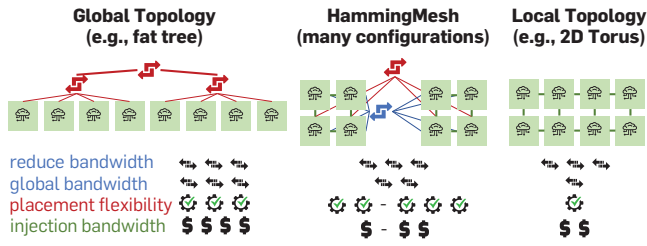
Today’s training jobs are already limited by data movement.<sup>14</sup> In addition, trends in deep neural networks (DNNs), such as sparsity, further increase those bandwidth demands in the near future.<sup>9</sup> Memory and network bandwidth are expensive—in fact, they form the largest cost component in today’s systems. Standard high-performance computing (HPC) systems with the newest InfiniBand adapters can offer 400Gb/s but modern deep-learning training systems offer much higher bandwidths. Google’s TPUv2, designed seven years ago, has 1Tbps off-chip bandwidth, AWS’ Trainium has up to 1.6Tbps per Tm1n instance, and Nvidia A100 and H100 chips have 4.8Tbps and 7.2Tbps (local) NVLINK connectivity, respectively. The chips in Tesla’s Dojo deep-learning supercomputer even have 128-Tbps off-chip bandwidth—*more than a network switch*. Connecting these extreme-bandwidth chips at a reasonable cost is a daunting task and today’s solutions, such as NVLINK, provide only local islands of high bandwidth.

We argue that general-purpose HPC and datacenter topologies are not cost-effective at these endpoint injection bandwidths. Yet, workload specialization, similar to existing microarchitectural optimizations, can lead to an efficient design that provides the needed high-bandwidth networking. We begin with developing a generic model that accurately represents the fundamental data movement characteristics of deep-learning workloads. Our model shows the inadequacy of the simplistic view that the main communication in deep learning is allreduce. In fact, we show that communication can be expressed as a concurrent mixture of pipelines and orthogonal reductions forming toroidal data movement patterns. This formulation shows that today’s HPC networks, optimized for full global (bisection) bandwidth, are inefficient for deep-learning workloads. Specifically, their *global bandwidth is overprovisioned while their local bandwidth is underprovisioned*.

We use our insights to develop HammingMesh, a flexible topology that can adjust the ratio of local and global bandwidth for deep-learning workloads. HammingMesh combines ideas from torus and global-bandwidth topolo-

The original version of this paper was published in *Proceedings of the Intern. Conf. for High Performance Computing, Networking, Storage and Analysis* (Nov. 2022).

**Figure 1. HammingMesh’s bandwidth-cost-flexibility trade-off.**



gies (for example, fat tree) to enable a flexibility-cost trade-off shown schematically in Figure 1. Inspired by machine learning (ML) traffic patterns, HammingMesh connects local high-bandwidth 2D meshes using row and column (blue and red) switches into global networks.<sup>a</sup>

In summary, we show how deep-learning communication can be modeled as sets of orthogonal and parallel Hamiltonian cycles to simplify mapping and reasoning. Based on this observation, we define principles for network design for deep-learning workloads. Specifically, our HammingMesh topology

- ▶ Uses technology-optimized local (for example, PCB board) and global (optical, switched) connectivity.
- ▶ Uses limited packet-forwarding capabilities in the network endpoints to reduce cost and improve flexibility.
- ▶ Enables full-bandwidth embedding of virtual topologies with deep-learning traffic characteristics.
- ▶ Supports flexible job allocation even with failed nodes.
- ▶ Enables flexible configuration of oversubscription factors to adjust global bandwidth.

With those principles, HammingMesh enables extreme off-chip bandwidths to nearest neighbors at more than 8x cheaper allreduce bandwidth compared to standard HPC topologies, such as fat trees. HammingMesh reduces the number of external switches and cables and thus reduces overall system cost. Furthermore, it provides significantly higher flexibility than torus networks. HammingMesh also enables seamless scaling to larger domains without separation between on- and off-chassis programming models (like NVLINK vs. InfiniBand). And, we believe that HammingMesh topologies extend to other ML, (multi)linear algebra, parallel solvers, and many other workloads with similar traffic characteristics.

We start with a characterization of parallel deep learning and the related data movement patterns. For refer-

a The name *HammingMesh* is inspired by the structural similarity to 2D Hamming Graphs with Meshes as vertices.

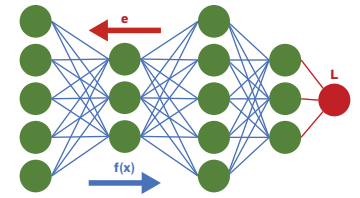
**Table 1. Symbols used in the paper.**

Symbol	Description
$M$	Number of examples per minibatch
$N_p$	Number of network parameters
$W$	Size of a word
$D, P, O$	Degree of data, pipeline, operator parallelism
$a, b$ and $x, y$	2D HammingMesh board and global sizes

ence, Table 1 offers an overview of symbols used in this paper.

## 2. COMMUNICATION IN DISTRIBUTED DEEP LEARNING

One iteration of deep-learning training with Stochastic Gradient Descent (SGD) consists of two phases: the forward pass and the backward pass. The forward pass evaluates the network function  $f(x)$  on a set of  $M$  examples, also called a “minibatch”. The backward pass of SGD computes the average loss  $L$  and propagates the errors  $e$  backward through the network to adapt the parameters  $P$ . This training process proceeds through multiple (computationally identical) iterations until the model achieves the desired accuracy.



Parallelism and data distribution can fundamentally be arranged along three axes: *data parallelism*, *pipeline parallelism*, and *operator parallelism*.<sup>5</sup> The latter two are often summarized as *model parallelism*, and operator parallelism is sometimes called *tensor parallelism*. We now briefly discuss their main characteristics.

Parallelism and data distribution can fundamentally be arranged along three axes: *data parallelism*, *pipeline parallelism*, and *operator parallelism*.<sup>5</sup> The latter two are often summarized as *model parallelism*, and operator parallelism is sometimes called *tensor parallelism*. We now briefly discuss their main characteristics.

### 2.1. Data parallelism.

When parallelizing over the training data, we train  $D$  separate copies of the model, each with different examples. To achieve exactly the same result as in serial training, we sum the distributed gradients before applying them to the weights at the end of each iteration. If the network has  $N_p$  parameters, then the communication volume of this step is  $WN_p$ .

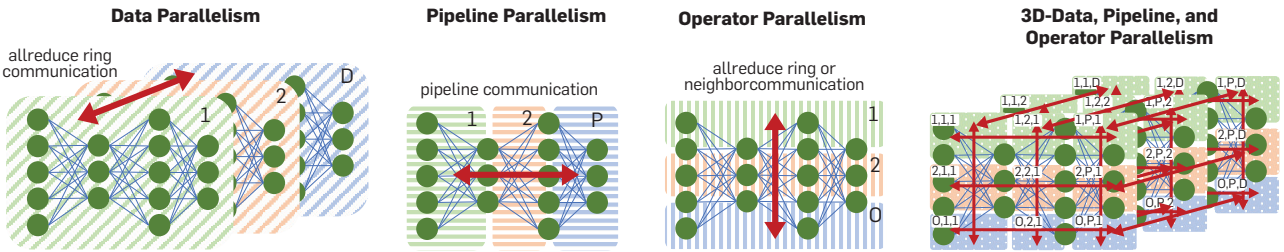
Modern deep neural networks have millions or billions of parameters, making this communication step expensive. Thus, many optimizations target gradient summation<sup>22</sup>—some even change convergence properties during the training process but maintain final result quality.<sup>2</sup> Dozens of different techniques have been developed to optimize this communication—however, all perform some form of distributed summation operation like *MPI\_Allreduce*. Data-parallelism differs thus mostly in the details, such as invocation frequency, consistency, and sparsity.

### 2.2. Pipeline parallelism.

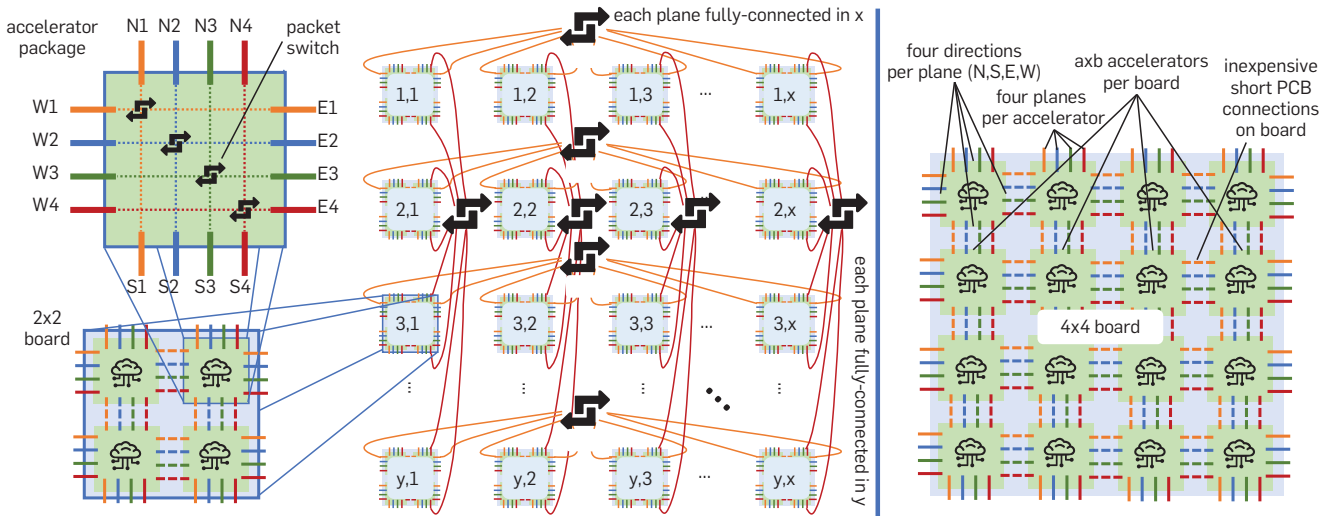
Deep neural networks are evaluated layer by layer with the outputs of layer  $i$  feeding as inputs into layer  $i + 1$ . Back-propagation is performed along the reverse direction starting at the loss function  $L$  after the last layer and proceeding from layer  $i + 1$  to layer  $i$ . We can model the network as a pipeline with  $P$  stages with one or more layers per stage. Forward and backward passes can be interleaved at each processing element to form a bidirectional training pipeline. Pipelines suffer from characteristic start-up and tear-down overheads. These can be reduced by running two pipelines in both directions<sup>19</sup> or by using asynchronous schemes that impact convergence.

Overall, pipelining schemes can use  $P$  processors with

**Figure 2. Distribution strategies for parallel deep neural network training.**



**Figure 3. HammingMesh structure: (left)  $x \times y$  Hx2Mesh, (right) Hx4Mesh board, both with four planes.**



a nearest-neighbor communication volume proportional to the number of output activations at the cut layers.

### 2.3. Operator parallelism.

Very large layer computations (operators) can be distributed to  $O$  processors. Most deep-learning layer operators follow computational schedules of (multi-)linear algebra and tensor contractions and require either (tightly coupled) distributed reductions or nearest-neighbor communications.

### 2.4. Overall communication pattern.

When all forms of parallelism are used, the resulting job comprises  $D \times P \times O$  accelerators; each accelerator in a job has a logical address  $(1..D, 1..P, 1..O)$ . The data-, pipeline-, and operator-parallel communication can be arranged as one-dimensional slices (rings) by varying only one coordinate of the Cartesian structure. Pipelines would leave one connection of the ring unused. For example, the data-parallel dimension consists of  $P \cdot O$  rings of length  $D$  each. Each of those rings represents a single allreduce. We show efficient ring-based reduction and broadcast algorithms for large data volumes in Section 4.1.2.

The overall composition of communication patterns forms a torus as illustrated in the right part of Figure 2 for a  $3 \times 3 \times 3$  example: Both the operator and the data par-

allel dimensions use nine simultaneous allreductions of size three each. The pipeline parallel dimension uses nine three-deep pipelines on three different model replicas, each split in three pieces.

While we can map such a logical torus to a full-bandwidth network topology, it seems wasteful to provide full bandwidth for sparse communication. For example, a 400Gb/s nonblocking fat tree with 16,384 endpoints provides full bisection bandwidth of more than  $\frac{16,384 \cdot 50\text{GB/s}}{2} = 410\text{TB/s}$ . A bi-directional  $32 \times 32 \times 16$  torus communication pattern requires at most  $32 \cdot 16 \cdot 2 \cdot 50\text{GB/s} = 51.2\text{TB/s}$  bisections (cutting one dimension of size 32)—a mere 12.5% of the offered bandwidth. In other words, *88% of the available bandwidth will remain unused and is wasted*. Furthermore, it is not always simple to map such torus communication patterns efficiently to full-bandwidth, low-diameter topologies in practice.

## 3. HAMMINGMESH

Based on the communication workload analysis, we now design a flexible and efficient network topology. The basic requirements are to support highest injection bandwidth for a set of jobs, each following a virtual toroidal communication topology. We note that medium-size models are often decomposed only in two dimensions in practice (usually data and pipeline or data and operator). Only

extreme-scale workloads require all three dimensions— even then, communication along the data parallel dimension only happens after one complete iteration. Thus, we use a two-dimensional physical topology.

As a case study, we assume a modern deep-learning accelerator package with 16 400Gb/s off-chip network links, a total network injection bandwidth of 800GB/s (top left in Figure 3). Our topology design also takes technology costs into account: Similar to Dragonfly, which combines local short copper cables with global long fiber cables to design a cost-effective overall topology, we combine such local groups with a global topology. Different from Dragonfly, we choose two quite distinct topologies: The local groups are formed by a local inexpensive high-bandwidth 2D mesh using short metal traces on PCB boards. This is the opposite of Dragonfly designs, which combine densely connected local groups (“virtual switches”) and connect those fully globally. HammingMesh combines sparsely connected boards in a dimension-wise (not globally) fully connected topology. Those boards are connected by a two-dimensional Hamming graph, in which each dimension is logically fully connected (for example, by a fat tree). All accelerator ports are arranged in *planes* with four directions each. Our example accelerator has four planes (top left in Figure 3), for example, plane 1 has ports E1, W1, N1, and S1. We assume each accelerator can forward packets within a plane like any network switch. Accelerators do not have to forward packets between planes, for example, packets arriving at N1 may only be forwarded to E1, W1, or S1 but none of the other ports. Thus, only simple 4x4 switches are needed at each accelerator. Figure 3 illustrates the structure in detail.

A 2D HammingMesh is parameterized by its number of planes and four additional numbers:  $(a, b)$ , the dimensions of the board, and  $(x, y)$ , the dimensions of the global topology. It connects a total of  $abxy$  accelerators. We abbreviate HammingMesh with HxMesh in the following. Furthermore, an HxMesh with an  $a \times b$  accelerator board is called HaxbMesh, for example, for a 2x2 board, H2x2Mesh. For

square board topologies, we skip the first number, for example, an H2x2Mesh that connects 10x10 boards is called a 10x10 Hx2Mesh.

HxMesh has a large design space: We can combine different board and global topologies, for example, 3D mesh boards with global Slim Fly topologies.<sup>6</sup> In this work, we consider 2D boards as most practical for PCB traces. The board arrangement could be reduced to a 1D HxMesh, where  $y = 1$  and each  $N_k$  link is connected to the corresponding  $S_k$  link (“wrapped around”). The same global topology can also span multiple rows or columns (for example, full boards in a single fat tree). For ease of exposition, we limit ourselves to 2D HxMeshes using 2D boards and row/column-separated global topologies. We use two-level fat trees as global topologies to connect the boards column- and row-wise. If the boards can be connected with a single 64-port switch, we use that instead of a fat tree.

### 3.1. Bisection and global bandwidth.

Bisection cut is defined as the minimal number of connections that would need to be cut in order to bisect the network into two pieces, each with an equal number of accelerators. The bisection bandwidth is the cut multiplied by the link bandwidth. Let us assume a single-plane of an  $x \times y$  HxaMesh (square board) with  $x \leq y$  and  $y$  even, wlog. We now consider the  $xy/2$  “lower” half boards with  $y$  coordinates  $1, 2, \dots, y/2$ . We split the HxMesh into two equal pieces by cutting the  $2a$  links in  $y$  direction of each of the lower half of the boards. This results in a total cut width of  $axy$ . Each accelerator has four network links per plane, a total injection bandwidth of  $4a^2$  per board. We have  $xy/2$  boards with a total injection bandwidth of  $4a^2xy/2 = 2xya^2$  in each partition. Thus, the relative bisection bandwidth is  $axy/2xya^2 = 1/2a$ .

In a bisection traffic pattern, all traffic crosses the network bisection (any two communicating endpoints are in different sets of the bisection). Such (worst-case) patterns are rare in practice. A more useful pattern, more often observed in practice is alltoall, where each process sends

**Table 2. Overview of our example networks (small and large cluster) using the cost model described in the full version of this paper.<sup>10</sup> All bandwidths are the result of the packet-level simulations detailed in Section 4.1. Global alltoall bandwidth is reported as share of the injection bandwidth for large messages (1.6 Tb/s). Allreduce bandwidth is reported as share of the theoretical optimum (1/2 of the injection bandwidth) for large messages. The cost savings for global and allreduce bandwidth are relative to the corresponding network cost of the nonblocking fat tree.**

Topology	Small Cluster ( 1,000 accelerators)						Large Cluster ( 16,000 accelerators)					
	cost	glob. BW	global	ared. BW	ared.	diam.	cost	glob. BW	global	ared. BW	ared.	diam.
	[M\$]	[% inject]	saving	[% peak]	saving		[M\$]	[% inject]	saving	[% peak]	saving	
nonbl. FT	25.3	99.9	1.0x	98.9	1.0x	4	680	98.9	1.0x	99.8	1.0x	6
50% tap. FT	17.6	51.2	0.7x	98.9	1.4x	4	419	47.6	0.8x	99.8	1.6x	6
75% tap. FT	13.2	25.7	0.5x	98.9	1.9x	4	271	24	0.6x	99.8	2.5x	6
Dragonfly	27.9	62.9	0.6x	98.8	0.9x	3	429	71.5	1.2x	98.6	1.6x	5
2D HyperX	10.8	91.6	<b>2.1x</b>	98.1	2.3x	4	448	95.8	1.5x	99.2	1.5x	8
Hx2Mesh	5.4	25.4	1.2x	98.3	4.7x	4	224	25	0.8x	92.3	2.8x	8
Hx4Mesh	2.7	11.3	1.0x	98.4	<b>9.3x</b>	843.3	10.5	<b>1.7x</b>	98	<b>15.4x</b>	8	
2D torus	2.5	2	0.2x	98.1	10.1x	32	39.5	1.1	0.2x	99.2	17.1x	128

to all other processes. This pattern is the basis of parallel transpositions, Fast Fourier Transforms, and many graph algorithms. The achievable theoretical bandwidth for such alltoall patterns is often called “global bandwidth.” Some topology constructions take advantage of the fact that global bandwidth is higher than bisection bandwidth. Prisacari et al.<sup>21</sup> shows that full-global bandwidth (alltoall) fat trees can be constructed with 25% less switches than nonblocking fat trees. Dragonfly,<sup>17</sup> Slim Fly,<sup>6</sup> or other low-diameter topologies<sup>16</sup> can further reduce the number of switches in very large installations while maintaining full global bandwidth. As is customary for low-diameter topologies,<sup>6,17</sup> we assess it using packet-level simulations of alltoall traffic.

### 3.2. Example topologies.

We consider a small cluster with approximately 1,000 accelerators and a large cluster with approximately 16,000 accelerators as specific design points to compare realistic networks. We compare various fat trees (nonblocking, 50%, 75% tapered), full bandwidth Dragonfly, two-dimensional torus, and HyperX,<sup>b</sup> with Hx2Mesh and Hx4Mesh example topologies.

Table 2 summarizes the main cost and bandwidth results. Global and allreduce bandwidths are determined using packet-level simulations (see Section 4) for large messages. *For all experiments, we simulated a single plane of HammingMesh and four planes for all other topologies, that is, a total injection bandwidth of 4×400Gb/s.* We use industry-standard layouts and cable configurations for the cost estimates: Fat trees are tapered beginning from the second level and connect all endpoints using DAC and all switches using AoC. Dragonfly topologies use full-bandwidth groups with  $a = 16$  routers each,  $p = 8$  endpoints per router, and  $h = 8$  links to other groups with DAC links inside the groups and AoC links between groups. The torus uses  $2 \times 2$  board topologies with discounted local PCB connectivity, similar to Hx2Mesh and only DAC cables between the boards. For HxMeshes, we use DAC links to connect endpoints to switches along one dimension, and AoC links for the other dimension. All inter-switch links are AoC as in fat trees.

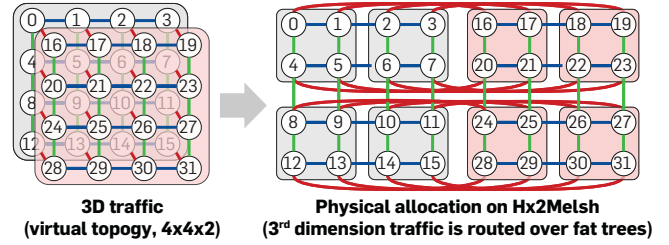
### 3.3. Logical job topologies and failures in HxMesh.

As we discussed in Section 2.4, communication patterns in deep learning can be modeled as sets of cycles. Typical learning jobs use either logical 1D cycles for small models with only data parallelism or 2D tori that combine data and pipeline parallelism for medium-scale models or combining pipeline and model parallelism for very large models. Each specific training job will have a different optimal decomposition resulting in 1D, 2D, or sometimes even 3D logical communication topologies.

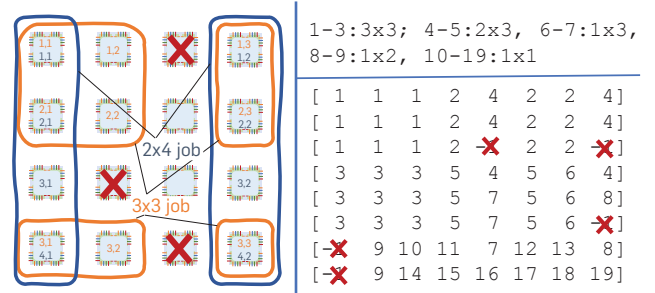
We use logical 2D topologies for our training jobs. Each job uses several boards and requests a  $u \times v$  layout (that is,  $a, b$  divides  $u, v$ , respectively). If the application topology follows a 1D or 3D scheme, then users use standard

<sup>b</sup> Note that a 2D HyperX is identical to an Hx1Mesh.

**Figure 4. 3D workload mapping onto Hx2Mesh example.**  
Left: virtual 4x4x2 topology. Right: mapping on Hx2Mesh.



**Figure 5. Subnetworks in the case of failures**



folding techniques to embed it into two-dimensional jobs. Figure 4 shows an example of 3D virtual topology mapped on an Hx2Mesh physical topology. Processes can be sliced on the third dimension and mapped on different boards. Communications between different slices of the third dimension are routed over the per-column or per-row fat trees, depending how different slices are mapped. To minimize communication latency between slices, consecutive slices should be adjacent to each other.

It is easy to see that any consecutive  $u \times v$  block of boards in a 2D HxMesh has the same properties as a full  $u \times v$  HxMesh. We call such subnetworks *virtual sub-HxMeshes*. They are a major strength of HxMesh compared to torus networks in terms of fault tolerance as well as for allocating jobs. In fact, HxMeshes major strength compared to torus networks is that virtual subnetworks can be formed with non-consecutive sets of boards (not only blocks): Any set of boards in an HxMesh where all boards that are in the same row have the same sequence of column coordinates can form a virtual subnetwork. We will show examples below together with a motivation for subnetworks—faults.

*Fault-tolerance.* We assume that a board is the unit of failure in an HxMesh, that is, if an accelerator or link in a board fail, the whole board is considered failed. This simplifies system design and service. Partial failure modes (for example, per plane) are outside the scope of this work.

The left part of Figure 5 shows a 4x4 Hx2Mesh and three board failures. We show two different subnetworks (many more are possible): a 2x4 subnetwork (blue) with the physical boards (1, 1), (1, 4), (2, 1), (2, 4), (3, 1), (3, 4), (4, 1), (4, 4) and a 3x3 subnetwork (yellow) with the physical boards (1, 1), (1, 2), (1, 4), (2, 1), (2, 2), (2, 4), (4, 1), (4, 2), (4, 4). We also annotate the new coordinates of boards in

the virtual subnetworks. Remapping can be performed transparently to the user application, which does not observe a difference between a virtual and physical HxMesh in terms of network performance. The right part of the figure shows the output of our automatic mapping tool (described in detail in the full version of this paper<sup>10</sup>) for a more complex configuration of jobs (top, read job ids 1-3 are  $3 \times 3$  logical jobs etc.).

#### 4. RESULTS AND COMPARISON

We now evaluate HxMesh topology options compared with all topologies listed in Table 2. We use the Structural Simulation Toolkit (SST),<sup>1</sup> a packet-level network simulator, which has been validated against the Cray Slingshot interconnect.<sup>8</sup> SST enables us to reproduce the behavior of full MPI applications directly in the simulation environment where *they react to dynamic network changes (for example, congestion)*. In total, we ran simulations of more than 120 billion packets using more than 0.6 million core hours with parallel simulations. We select various representative microbenchmarks and scenarios for deep-learning jobs and *publish the full simulation infrastructure such that readers can simulate their own job setup*.

##### 4.1. Microbenchmarks.

We start by analyzing well-known microbenchmark traffic patterns to assess and compare achievable peak bandwidth.

###### 4.1.1. Global traffic patterns.

We first investigate global traffic patterns such as alltoall and random permutations as global-traffic workloads. We note that HammingMesh is not optimized for those pat-

Figure 6. Alltoall on the small topologies.

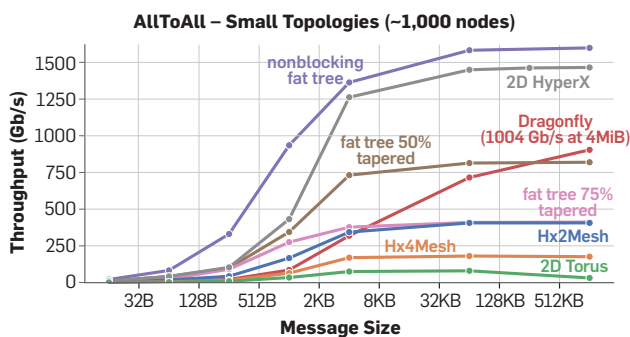
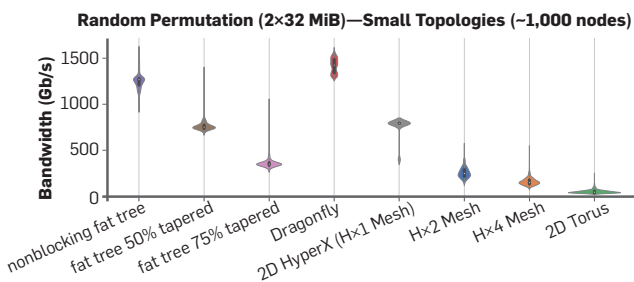


Figure 7. Bandwidth distribution per accelerator.



terns as they are rare on deep-learning traffic.

**Alltoall:** Alltoall sends messages from each process to all other processes. In our implementation, each of the  $p$  processes performs  $p - 1$  iterations. In each iteration  $i$ , process  $j$  sends to process  $j + i \bmod p$  in a balanced shift pattern.

Table 2 shows the results for 1MiB messages while Figure 6 shows the global bandwidth at different message sizes. Small Hx2 and Hx4Meshes achieve bandwidths around the cut width of  $1/4$  and  $1/8$ , respectively (cf. Section 3.1). This is because not all global traffic crosses the bisection cuts, especially for smaller clusters. The large cluster configuration performs closer to those bounds and loses some bandwidth due to adaptive routing overheads. Despite its lower bandwidth, even large HxMeshes remain competitive in terms of cost-per-global bandwidth and some are even more cost effective on global bandwidth than fat trees.

**Random permutation:** In permutation traffic, each accelerator selects a unique random peer to send to and receive from. Here, the achieved bandwidth also depends on the location of both peers. Figure 7 shows the distributions of receive bandwidths across all the 1k accelerators in the small cluster configurations.

Our results indicate that all topologies have significant variance across different connections (between different node pairs), which makes job placement and locality significant. HxMeshes are among the most cost effective topologies.

###### 4.1.2. Reduction traffic patterns.

We distinguish three fundamental algorithm types: trees, pipelines, and near-optimal full-global bandwidth algorithms.

**Simple trees:** For small data, simple binary or binomial tree reductions are the best choice. They perform a reduction of  $S$  bytes on  $p$  processors in time  $T \approx \log_2(p)\alpha + \log_2(p)S\beta$ .<sup>c</sup> This algorithm sends each data item a logarithmic number of times. It is thus inefficient for the large data sizes in deep-learning training workloads and we do not consider trees in this work.

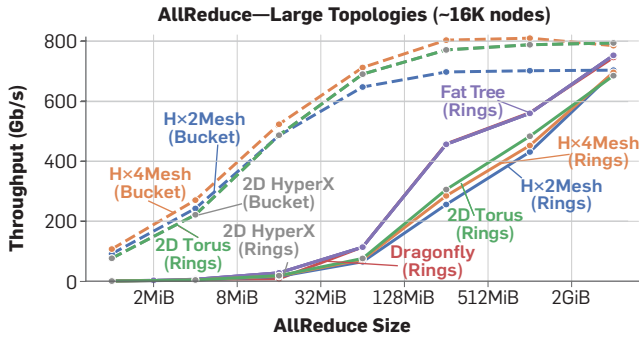
**Pipelined rings:** With a single network interface, large data volumes can be reduced in a simple pipelined ring. Here, the data at each process is split into  $p$  segments. The operation proceeds in two epochs and  $p - 1$  rounds per epoch. In the first reduction epoch, each process  $i$  sends segment  $i$  to process  $i + 1 \bmod p$  and receives a segment from process  $i - 1 \bmod p$ . The received segment is added to the local data and sent on to process  $i + 1 \bmod p$  in the next round. After  $p - 1$  such rounds, each process has the full sum of one segment. The second epoch is simply sending the summed segments along the pipeline. The overall time  $Tp \approx 2p\alpha + 2S\beta$  is bandwidth optimal because each process only sends and receives each segment twice.<sup>4</sup>

We propose bidirectional pipelined rings to use two

<sup>c</sup> We define with  $\alpha$  the latency and with  $\beta$  the inverse of the bandwidth. With  $\approx$ , we omit additive constants and minor lower-order terms for clarity.



**Figure 8. Global allreduce using different algorithms.**



network interfaces by splitting the data size in half and sending each half along a different direction. The latency stays unchanged because each segment travels twice through the whole ring but the data is half in each direction, leading to a runtime of  $T_{bp} \approx 2p\alpha + S\beta$ . Here and in the following,  $\beta$  is the time per byte of each interface, that is, a system with  $k$  network interfaces can inject  $k/\beta$  bytes/s.

We now extend this idea to four network interfaces per HxMesh plane: We use two bidirectional rings, each reducing a quarter of the data across all accelerators. The two rings are mapped to two disjoint Hamiltonian cycles covering all accelerators of the HxMesh.<sup>3</sup> The overall time for this scheme is  $T_{rings} \approx 2p\alpha + \frac{S}{2}\beta$ .

**Bucket:** Pipelined rings are bandwidth-optimal if they can be mapped to Hamiltonian cycles on the topology. However, we find that for large HxMeshes and moderate message sizes, the latency component can become a bottleneck. We thus use the state-of-the-art bucket algorithm.<sup>23,d</sup> The bucket algorithm arranges communications in 2D toroidal communication patterns with  $\sqrt{P}$  latency and good bandwidth usage. Each process executes first a reduce-scatter with the other processes on the same row (cost  $\sqrt{P}\alpha + \frac{S}{2}\beta$ ). Then each process runs an allreduce with the other processes on the same column, on the previously reduced chunk of size  $\frac{S}{\sqrt{P}}(\cos 2(\sqrt{P}\alpha + \frac{S}{2\sqrt{P}}\beta))$  and, eventually, an allgather with the other processes on the same row (cost  $\sqrt{P}\alpha + \frac{S}{2}\beta$ ). To use all four network interfaces at the same time, four of these allreduce can be executed in parallel, each starting from a different port and working on a quarter of the data.<sup>23</sup> Thus, the overall time for this scheme is  $T \approx 2 \cdot 2\sqrt{P}\alpha + S\beta(\frac{1+2\sqrt{P}}{4\sqrt{P}})$ .

**Summary:** The pipeline ring and bucket algorithms have sparse communication patterns: Each process only communicates with two or four direct neighbors that can be mapped perfectly to HxMesh. Broadcast and other collectives can be implemented similarly (for example, as the second part of our allreduce) and follow similar trade-offs. Furthermore, each dimension of a logical job topology is typically small as the total number of accelerators is the product of all dimensions. For example, even for a very large system with 32,768 accelerators, each of the dimensions could only be of size 32 if we decompose the

problem along all dimensions. This means that the largest allreduce or broadcast would only be on 32 processes where ring algorithms would perform efficiently.

**Full system allreduce job:** This experiment shows a single job using the last two allreduce algorithms on various topologies. In Dragonfly and fat tree, each accelerator connects with a single NIC to each of the four planes and we use the standard “ring” algorithm. For the single allreduce on the large HxMesh clusters, we use both the two bidirectional rings (“rings”) as well as the bucket (“bucket”) algorithm. Figure 8 shows the achieved bandwidths.

We see that all topologies deliver nearly full bandwidth for the ring algorithms. For large messages, HxMesh is 2.8x to 14.5x cheaper per bandwidth than a nonblocking fat tree (Table 2). On networks with a Cartesian structure (HammingMesh, Torus, and HyperX), the bucket algorithm outperforms the ring algorithm at any message size. The only exception is for jobs where one of the two dimensions is much smaller than the other, where the ring algorithm outperforms the bucket algorithm (not shown), highlighting the importance of using multi-algorithms to optimize performance, similar to established practice in MPI.<sup>25</sup>

## 4.2. DNN workloads.

We now proceed to define accurate communication patterns, including computation times for real DNN models. For this, we choose four large representative models: ResNet-152, CosmoFlow, DLRM, and Transformers (GPT3) trained in FP32. We discuss only DLRM and Transformers; a more detailed discussion covering the other models can be found in the full version of this paper.<sup>10</sup> We use NVIDIA’s A100 GPU to benchmark runtimes of operators and we model communication times based on the data volumes.

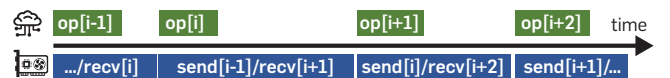
### 4.2.1. Communication traffic characterization.

All example models are constructed of a sequence of identical layers containing multiple operators. Each parallel dimension carries a different volume, depending on the details of the model, training hyperparameters, and the other dimensions. We assume the most general case where the network can use all three forms of parallelism running on  $D \times P \times O$  accelerators.

**Data dimension:** If we only have data parallelism ( $O = P = 1$ ), then each process needs to reduce all gradients. If we distribute the model between  $O$  or  $P$  dimension processes, then the total allreduce size is  $V_D = \frac{WN}{OP^2}$ . The reduction happens once at the end of each iteration after processing a full minibatch and draining the pipeline. It can be overlapped per layer using nonblocking allreduce.<sup>13</sup>

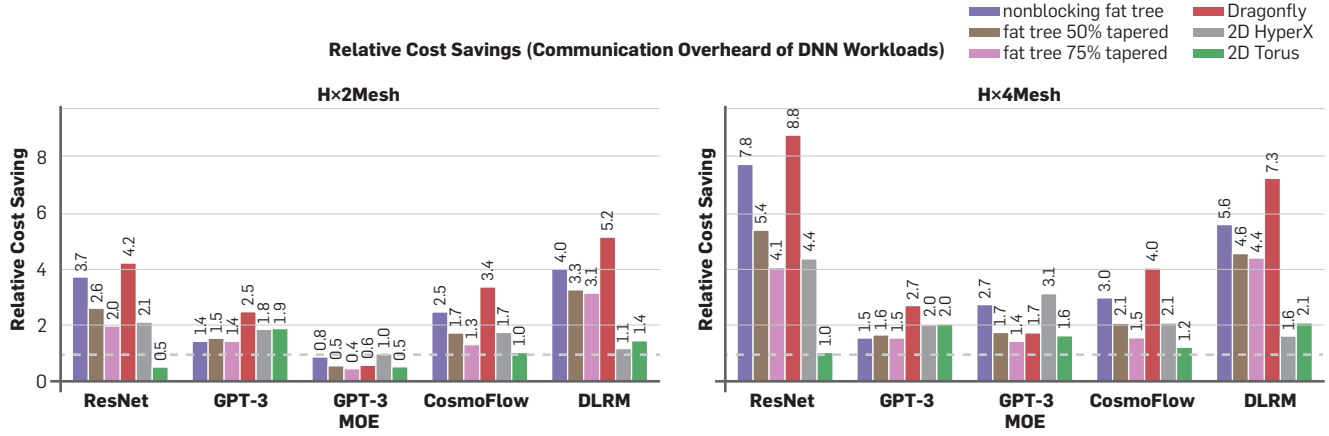
**Pipeline dimension:** If we only have pipeline parallelism ( $D = O = 1$ ) and NA output activations at the “cut” layer

**Figure 9. Overlap in pipelined-parallel execution.**



d Compared to the original version of this paper,<sup>10</sup> we replaced the torus algorithm with the better bucket algorithm.

Figure 10. HxMesh cost savings relative to other topologies.



then each process sends all  $\frac{M}{N_A}$  output values to the next process in the forward pass and the same volume of errors during the backward pass. If the layer and its inputs and outputs are distributed to  $O$  PEs, then the total send volume in this dimension is  $V_p = \frac{MWN_A}{DPO}$ . This communication can be hidden at each accelerator as shown in Figure 9 by overlapping nonblocking send/receive operations (bottom, blue) with operator computation (top, green).

**Operator dimension:** For operator parallelism, each process’s send volume depends only on the operator parallelization itself and is not influenced by either  $D$  or  $P$ . The operator can be seen as the “innermost loop” in this sense. Each operator distribution scheme will have its own characteristics that we capture by  $V_o = WN_o$ . The operator communication volume during each forward and backward pass is a function of the local minibatch size  $M/DP$  per process.

#### 4.2.2. DLRM.

DLRM<sup>20</sup> uses a combination of model parallelism and data parallelism for its embedding and MLP layers, respectively. Two alltoall operations aggregate sparse embedding lookups in the forward pass, and their corresponding gradients in backward pass. Allreduce is required to synchronize the gradients of the data-parallel MLP layers. The parallelism of DLRM is limited by both the mini-batch size and the embedding dimension. DLRM is trained with up to 128 GPU nodes. The total runtimes on the fat tree variants are 2.96ms, 2.97ms, and 2.99ms, respectively. On torus, the code executes for 3.12ms. HyperX is at 2.94ms. Hx2Mesh and Hx4Mesh are at 2.97ms and 3.00ms, respectively. On A100, DLRM computes around 95us, 209us, and 796us for the embedding, feature interaction, and MLP layers respectively, and communicates 1MB per alltoall and 2.96MB per allreduce.

#### 4.2.3. Transformers.

Transformers are the most communication intensive.<sup>14</sup> A transformer block consists of multi-head attention (MHA) and two feed-forward (FF) layers. The MHA and FF input/outputs are of size (embedding dimension×batch×sequence length). For example, GPT-

3’s<sup>7</sup> feed-forward layers multiply 49,152×12,288 with 12,288×2,048 matrices per example in each layer.

GPT-3 has a total of 96 layers and each layer has activations of size  $N_A = 4 \cdot 2,048 \times 12,288 \approx 100\text{MB}$  per example as input and output. We choose  $P = 96$ , such that each pipeline stage processes one layer, and no data parallelism ( $D = 1$ ). For operator parallelism, we use  $O = 4$  and the scheme outlined by Megatron-LM,<sup>24</sup> which performs one allreduce for FF and one for MHA in forward and backward passes.

All operations are the same size as the layer input/output. Thus, the volume for both pipeline communication and operator-dimension allreduce is  $N_A$  per example for forward and backward passes. One iteration of GPT-3 computes for 31.8ms. Total runtimes on the three fat-tree variants are 34.8ms, 36.4ms, and 37.5ms, respectively. On torus, the code executes for 72.2ms per iteration. HyperX is at 40.9ms. Hx2 and Hx4Mesh are at 41.7ms and 49.9ms, respectively.

For GPT-3 with Mixture-of-Experts (MoEs),<sup>18</sup> we use 16 experts. In GPT-3, the FFs have 1.8B parameters. Therefore, each expert has 1.8B/16  $\approx$  113M parameters. MoEs perform two alltoalls for FF in both the forward and backward passes, and all operations are the same size as the input/output. The computation time on an A100 is 49.9ms. Total runtime on the fat trees varies from 52.2ms to 52.9ms depending on tapering. On torus, the code executes for 73.8ms per iteration. HyperX takes 53.9ms while Hx2 and Hx4Mesh are at 58.3ms and 63.3ms, respectively.

Figure 10 shows the relative cost savings of HxMesh compared to other topologies. These are calculated as the ratio of the network costs in Section 2 times the inverse of the ratio of communication overheads presented in this section.

We conclude that both Hx2 and Hx4Mesh significantly reduce network costs for DNN workloads. While some torus network configurations can be cheaper than Hx2Mesh, they provide significantly less allocation and management flexibility, especially in the presence of failures. Moreover, we also conclude that even in the presence of alltoall communications patterns in GPT-3 MoE and DLRM, HxMesh topologies still offer a significant cost advantage compared to traditional topologies. As the scale

of the network increases, Hx4Mesh becomes significantly more cost efficient than Hx2Mesh, especially in the presence of alltoall traffic.

**Discussion:** We cover all additional related work and comparisons to other topologies, as well as significantly more detail on HammingMesh configuration options, tapering, diameter, cost, routing and deadlock avoidance, as well as scheduling with and without board failures in the full version of this paper.<sup>10</sup>


## 5. CONCLUSION

HammingMesh is optimized specifically for ML workloads and their communication patterns. It relies on the observation that deep-learning training uses three-dimensional communication patterns and rarely needs global bandwidth. It supports extreme local bandwidth while controlling the cost of global bandwidth. It banks on an inexpensive local PCB-mesh interconnect together with a workload-optimized global connectivity forming virtual torus networks at adjustable global bandwidth.

Due to the lower number of switches and external cables, it can be nearly always more cost effective than torus networks while also offering higher global bandwidth and significantly higher flexibility in job allocation and dealing with failures.

All-in-all, we believe that HammingMesh will drive future deep learning systems and will also support adjacent workloads, such as (multi)linear algebra, quantum simulation, or parallel solvers, that have Cartesian communication patterns.

## 6. ACKNOWLEDGMENT

We thank Microsoft for hosting TH's sabbatical where much of the idea was developed.<sup>11,12</sup> We thank the whole Azure Hardware Architecture team and especially Doug Burger for their continued support and deep technical discussions. We thank the Swiss National Supercomputing Center (CSCS) for the compute resources on Piz Daint and the Slim Fly cluster (thanks to Hussein Harake) to run the simulations. Daniele De Sensi is supported by an ETH Post-doctoral Fellowship (19-2 FEL-50). 

## References

- Adalsteinsson, H. et al. A simulator for large-scale parallel computer architectures. *Int. J. Distrib. Syst. Technol.* 1, 2 (Apr. 2010), 5773.
- Alistarh, D. et al. The convergence of sparsified gradient methods. *Advances in Neural Information Processing Systems* 31. Curran Associates, Inc. (Dec. 2018).
- Bae, M.M., AlBdaiwi, B.F., and Bose, B. Edge-disjoint Hamiltonian cycles in two-dimensional torus. *Int. J. Math. Math. Sci.* 2004, 25 (2004), 1299–1308.
- Barnett, M., Littlefield, R., Payne, D., and Vandegheijn, R. Global combine algorithms for 2-D meshes with wormhole routing. *J. Parallel Distrib. Comput.* 24, 2 (Feb. 1995), 191201.
- Ben-Nun, T. and Hoefler, T. Demystifying parallel and distributed deep learning: An in-depth concurrency analysis. *ACM Comput. Surv.* 52, 4 (Aug. 2019), 65:1–65:43.
- Besta, M. and Hoefler, T. Slim fly: A cost effective low-diameter network topology. In *Proceedings of the Intern. Conf. On High Performance Computing, Networking, Storage and Analysis (SC14)*, (Nov. 2014).
- Brown, T.B. et al. *Language Models Are Few-Shot Learners*, (2020).
- De Sensi, D. et al. An in-depth analysis of the slingshot interconnect. In *Proceedings of the Intern. Conf. For High Performance Computing, Networking, Storage and Analysis (SC20)*, (Nov. 2020).
- Hoefler, T. et al. Sparsity in deep learning: Pruning and growth for efficient inference and training in neural networks. *J. of Machine Learning Research* 22, 241 (Sep. 2021), 1–124.
- Hoefler, T. et al. Hammingmesh: A network topology for large-scale deep learning. In *Proceedings of the Intern. Conf. On High Performance*

- Computing, Networking, Storage and Analysis, SC '22. IEEE Press, (2022).
- Hoefler, T., Heddes, M.C., and Belk, J.R. Distributed processing architecture. *US Patent Us11076210b1*, Jul. 2021.
- Hoefler, T., Heddes, M.C., Goel, D., and Belk, J.R. Distributed processing architecture. *US Patent Us20210209460a1*, (Jul. 2021).
- Hoefler, T., Lumsdaine, A., and Rehm, W. Implementation and performance analysis of non-blocking collective operations for MPI. In *Proceedings of the 2007 Intern. Conf. On High Performance Computing, Networking, Storage and Analysis, SC07*. IEEE Computer Society/ACM, (Nov. 2007).
- Ivanov, A. et al. Data movement is all you need: A case study on optimizing transformers. In *Proceedings of Machine Learning and Systems 3 (Mlsys 2021)*, (Apr. 2021).
- Kaplan, J. et al. *Scaling Laws for Neural Language Models*, (2020).
- Kathareios, G. et al. Cost-effective diameter—two topologies: Analysis and evaluation. In *Proceedings of the Intern. Conf. For High Performance Computing, Networking, Storage and Analysis (SC15)*. ACM, (Nov. 2015).
- Kim, J., Dally, W.J., Scott, S., and Abts, D. Technology-driven, highly-scalable dragon topology. In *Proceedings of 2008 Intern. Symp. On Computer Architecture*, 77–88.
- Lepikhin, D. et al. *Gshard: Scaling Giant Models with Conditional Computation and Automatic Sharding*, (2020).
- Li, S. and Hoefler, T. Chimera: Efficiently training large-scale neural networks with bidirectional pipelines. In *Proceedings of the Intern. Conf. For High Performance Computing, Networking, Storage and Analysis (SC21)*. ACM, (Nov. 2021).
- Naumov, M. et al. Deep learning recommendation model for personalization and recommendation systems. *Arxiv Preprint Arxiv:1906.00091*, (2019).
- Prisacari, B., Rodriguez, G., Minkenber, C., and Hoefler, T. Bandwidth-optimal all-to-all exchanges in fat tree networks. In *Proceedings of the 27th Intern. ACM Conf. on Intern. Conf. on Supercomputing*. ACM, (Jun. 2013), 139–148.
- Renggli, C., Alistarh, D., Aghagolzadeh, M., and Hoefler, T. Sparcm: High-performance sparse communication for machine learning. In *Proceedings of the Intern. Conf. For High Performance Computing, Networking, Storage and Analysis (SC19)*, (Nov. 2019).
- Sack, P. and Gropp, W. Collective algorithms for multiported torus networks. *ACM Trans. Parallel Comput.* 1, 2 (Feb. 2015).
- Shoeybi, M. et al. *Megatron-Lm: Training Multi-Billion Parameter Language Models Using Model Parallelism*, (2020).
- Thakur, R., Rabenseifner, R., and Gropp, W. Optimization of collective communication operations in mpich. *Int. J. High Perform. Comput. Appl.* 19, 1 (Feb. 2005), 4966.

**Torsten Hoefler** (torsten.hoefler@inf.ethz.ch) ETH Zurich, Switzerland, Microsoft, Corp.

**Tommaso Bonato** (tommaso.bonato@inf.ethz.ch) ETH Zurich, Switzerland.

**Daniel De Sensi** (daniele.desensi@inf.ethz.ch) ETH Zurich, Switzerland.

**Salvatore Di Girolamo** (salvatore.digirolamo@inf.ethz.ch) ETH Zurich, Switzerland.

**Shigang Li** (shigang.li@inf.ethz.ch) ETH Zurich, Switzerland.

**Marco Heddes** (marco.heddes@microsoft.com) Microsoft, Redmond, WA, USA.

**Deepak Goel** (deepak.goel@microsoft.com) Microsoft, Sunnyvale, CA, USA.

**Miguel Castro** (miguel.castr@microsoft.com) Microsoft, Cambridge, MA, USA.

**Steve Scott** (steve.scott@microsoft.com) Microsoft, Redmond, WA, USA.



This work is licensed under a Creative Commons Attribution International 4.0 License.

# CAREERS

## Columbia University

Assistant or Associate Professor in Electrical Engineering

### Position Description

Columbia Engineering is pleased to invite applications for a tenure-track faculty position at the rank of Assistant Professor or Associate Professor (without tenure, but on tenure-track), in the Department of Electrical Engineering at Columbia University in the City of New York, starting on July 1, 2025.

The department welcomes applications in all areas of electrical and computer engineering, with emphasis on circuit design, on systems for artificial intelligence and machine learning, and on the school-wide strategic priority area of quantum computing and technology. Outstanding candidates in other areas will also be considered and are encouraged to apply.

### Qualifications

Candidates must have a Ph.D. or its professional equivalent by the starting date of the appointment. Applicants for this position must demonstrate the potential to do pioneering research and to teach effectively. The Department is especially interested in qualified candidates who will contribute, through their research, teaching, and/or service, to the diversity and excellence of the academic community.

The successful candidate is expected to contribute to the advancement of their field and the department by developing an original and leading externally funded research program, and by contributing to the undergraduate and graduate educational mission of the Department. Columbia fosters multidisciplinary research and encourages collaborations with academic departments and units across Columbia University. The Department actively participates in the school-wide Engineering for Humanity initiatives that relate to engineering and medicine, autonomous systems, quantum computing and technology, and sustainability.

### Application Instructions

For additional information and to apply, please see: <http://engineering.columbia.edu/faculty-job-opportunities>. Applications should be submitted electronically to <http://apply.interfolio.com/156905> and include the following: a curriculum vitae including a publication list, a research statement including a description of research accomplishments, a statement of teaching interests and plans, three letters of recommendation, and up to three pre/reprints of scholarly work. Applicants are also encouraged to provide a statement addressing past and/or potential contributions to diversity and inclusion through teaching, professional activity, and/or service. All applications received by December 1, 2024 will receive priority consideration, and applications received by January 15, 2025 will receive full consideration.

Applicants can consult <http://www.ee.columbia.edu> for more information about the department.

Applicants who are also interested in Columbia Engineering's school-wide cluster search in quantum computing should submit an additional application to the quantum search, by accessing the job posting at <http://engineering.columbia.edu/faculty-job-opportunities>.

**Hiring Salary Range:** Assistant Professor: 110K-142K; Associate Professor (without tenure): 115K-163K.

## Max Planck Institutes in Computer Science

Tenure-track Faculty Openings\* at Computer Science MPIs

The Max Planck Institutes (MPIs) for Informatics, for Security & Privacy, and for Software Systems invite applications for tenure-track faculty in all areas of computer science and its intersection with other disciplines. We expect to fill several positions.

A doctoral degree in computer science or related fields and an outstanding research record are required. Successful candidates are expected to build a team and pursue a highly visible research agenda, both independently and in collaboration with other groups.

The institutes are part of a network of over 80 MPIs, Germany's premier basic-research institutes. MPIs have an established record of world-class, foundational research in the sciences, technology, and the humanities. The institutes offer a unique environment that combines the best aspects of a university department and a research laboratory: Faculty enjoy full academic freedom, lead a team of doctoral students and post-docs, and have the opportunity to teach university courses; at the same time, they enjoy ongoing institutional funding in addition to third-party funds, a technical infrastructure unrivaled for an academic institution, as well as internationally competitive compensation.

We maintain an international and diverse work environment and seek applications from outstanding researchers worldwide. The working language is English; knowledge of the German language is not required for a successful career at the institutes.

For full consideration, applications should be submitted by December 1st, 2024. Apply via <https://shlink.mpi-sws.org/faculty>.

MPIs are committed to fostering a diverse, inclusive, and global academic community, and consider qualified applicants for employment without discrimination on the basis of gender, race, disability, ethnic or social origin, or any other legally protected status. We particularly encourage applications from groups that are underrepresented in computer science. We welcome applications from dual-career couples and will do our best to try and accommodate their needs.

The initial tenure-track appointment is for six years. A permanent contract can be awarded upon a successful tenure evaluation in the sixth year.

\* W2 German Federal Civil Service Remuneration Act (BBesG)

### Requirements

A doctoral degree in computer science or related fields and an outstanding research record are required. Successful candidates are expected to build a team and pursue a highly visible research agenda, both independently and in collaboration with other groups.

## New York University - Department of Computer Science

Tenure/Tenure-Track Positions in Computer Science for Fall 2025

The Computer Science department expects to have several tenure-track faculty positions and invites candidates at all levels to apply. Candidates working in AI-related areas (Machine Learning, Computer Vision, Natural Language Processing, Robotics, Responsible AI, and other areas) should apply through NYU's AI search.

Faculty members are expected to be outstanding scholars and to participate in teaching at all levels from undergraduate to doctoral. New appointees will be offered competitive salaries and startup packages.

The department has 54 regular faculty members as well as clinical, research, adjunct, and visiting faculty members. The department's current research activities span algorithms, cryptography and theory, computational biology, distributed computing and networking, graphics, vision and multimedia, machine learning and data science, natural language processing, scientific computing, verification, and programming languages.

Collaborative research with industry is facilitated by geographic proximity to computer science activities at Facebook, Google, DeepMind, Amazon, Microsoft Research, IBM, Bell Labs, AT&T Research, Flatiron Institute, and many other companies with research divisions in the NYC area.

In compliance with NYC's Pay Transparency Act, the annual base salary range for this position is the following: For the rank of Assistant Professor \$125,000 - \$185,000; for the rank of Associate Professor \$135,000 - \$225,000, and for the rank of Full Professor \$225,000 - \$325,000. New York University considers factors such as (but not limited to) the scope and responsibilities of the position, the candidate's work experience, education/training, key skills, internal peer equity, as well as market and organization considerations.

### Application Instructions:

Please apply through Interfolio at this link: <http://apply.interfolio.com/157060>.

Applicants should submit a CV, research statement, teaching statement, three confidential letters of recommendation, and three of your most significant publications, software, or research products. We encourage applicants to include an optional statement of experience with or knowledge of inclusion, diversity, equity, and belonging efforts and your plans for incorporating them into your teaching, research, mentoring, and service. All application materials should be uploaded through Interfolio. For full consideration we recommend that applicants apply by December 1, 2024, though we will continue to review applications past that date as needed.

### Job Requirements

A PhD in Computer Science or a related area is required. Successful candidates are expected to pursue an active research program and to contribute significantly to the teaching programs of the department.

[CONTINUED FROM P. 108] to machines. It's marvelous to see that the technology is at a level where we don't have to write dialogue trees because the agents are actually smart. Of course, there's still plenty of work left to do. There are also really hard questions to answer about how agents should interact with users, how they should adapt and personalize their behavior, and how we can ensure that they are ethical, safe, and privacy-protective. But the underlying technological substrate has accelerated tremendously.

### What has not changed?

There are fundamental issues in robotics that remain unsolved, like how robots can effectively manipulate the world physically. From my perspective, though, that's not the biggest challenge. I don't need my robots to physically manipulate people. I need them to provide social, emotional, and psychological support, which means that accessibility is the far larger unsolved problem. Our goal is to put physically embodied agents into people's lives—and they need to be *in* their lives, which means they need to be affordable, safe, and accessible. None of that exists on the consumer market. There are no such platforms.

**I'm a little surprised by that, to the extent you've been able to demonstrate the efficacy of your robotic interventions, and the alternative is often engaging a trained human being. Why isn't there more funding?**

There has been a surge in funding in robotics, at least in startups and industry. Most of the money has gone to robots that manipulate things in the world, because ultimately, people are interested in automating manufacturing, and they're not seeing the opportunity for socially assistive systems. The National Science Foundation tries, but they have a tiny budget. It's not the mission of the Department of Defense. I recently received a grant from the National Institutes of Health, which is an honor, but NIH very rarely funds technologies for health interventions.

Still, I want to be optimistic, both in the sense that people are starting to understand the societal implications of talking machines, and because, fortunately and finally, the diversity

**"I don't need my robots to physically manipulate people. I need them to provide social, emotional, and psychological support."**

of innovators who are contributing is expanding.

**In the meantime, you created an open source kit to help college and high school students build their own "robot friend."**

My lab started with a platform that was developed in Guy Hoffman's lab at Cornell called Blossom, and then we redid the structure to make it 3D-printed and much cheaper. Finally, we designed some exterior patterns that one can sew or crochet to customize the robot's appearance.

Now, we have a robot platform that's maybe \$230 to build, and then you make a customized skin for it, and it's really inexpensive and completely open sourced, so hopefully anybody can do it.

**These robots are very cute. I imagine that's part of the point.**

We and many others have done studies on this issue of embodiment. What happens when you interact with a screen versus when you interact with a physically embodied agent? There's very clear evidence that physical embodiment is fundamental to improving both engagement and outcomes. That's not to say that screen agents can't do useful things. But the question is, how do they compare? It turns out, largely unfavorably.

We're also working in contexts where things are really hard. This isn't about video game engagement. It's about helping children with autism learn new skills or supporting people with anxiety and depression in learning emotion regulation. We did a study in college dorms in which we compared a chatbot

that provided LLM-based therapy versus the same LLM-based therapy from a physically embodied robot. Students engaged with and used both of them, but only the students who used the robot measurably reduced their psychiatric distress.

**What are some of the things that surprise you about the way people interact with robots?**

We're always surprised by people. Early on, we were surprised when people tried to cheat or trick the robot. Now, we're surprised by how people react to the idea of interacting with a robot. About seven years ago, we were doing a study with elderly people, and one of the participants said, "It's cute, but why can't it do as many things as my iPad can?" Some people absolutely love the robot and others are very grumpy, and the question is, what can we learn from that about our own stereotypes and cognitive biases, and about personalizing the interaction?

Personalization is why these interventions work. We need to be able to find out what someone needs right now, as opposed to simply telling them, "Here are your steps, and you need to go do them." Even with physical health, it turns out that a lot depends on the state you're in on a given day, on your metabolism, and so on. Why wouldn't that be the case with your behavior, which relates to your mental and physical health and also your social context?

**It's very multi-layered, isn't it? It also alleviates this burden people often feel in therapeutic settings, where their health is tied to individual choices, and the broader social context figures, if at all, in a very indirect, amorphous way.**

Exactly. However, I do worry that creating intelligent agents risks making vulnerable people even more isolated, because they'll be told to just rely on their agent. What these agents should be doing is connecting people socially and serving as this interstitial network. It can't be a binary choice between human-agent and human-human interaction. It has to be human-human-agent. □

**Leah Hoffmann** is a technology writer based in Piermont, NY, USA.

© ACM 0001-0782/24/12

## Q&amp;A

# Personalizing Interactions

*Maja Matarić discusses her career, and the surprising things that happen when humans and robots interact.*

2024 ACM ATHENA Award recipient and University of Southern California professor Maja Matarić is not afraid to get personal. In her quest to design socially assistive robots—robots that provide social, not physical, support in realms like rehabilitation, education, and therapy—she realized that personalizing interactions would boost both engagement and outcomes. Artificial intelligence (AI) has made that easier, though as always, surprises are never far when human beings are involved. Here, Matarić shares what she’s learned about meeting people where they are.

**Let’s talk about your work on socially assistive robots. You’ve said that having kids inspired you to build robots that help people. How did that interest develop into the mission of supporting specific behavioral interventions in health, wellness, and education?**

It was a confluence of events in my life. I had two small kids, and I really wanted my work to have impact beyond academia in ways that even children could understand.

I did a lot of reading, and I immersed myself in a bunch of communities, because I was trying to understand how to develop agents that could help people in ways in which they needed help. Identifying that niche—that place in the user journey where something is difficult, and where behavioral interventions could support people—is not at all obvious. It remains not obvious, because we engineers tend to think, “Here’s a problem. And this is how it should be solved.” And often, we don’t even recognize the right problem, much less the



right solution. The hard part is not having to remember to take your medicine or figuring out how to do your stroke rehabilitation exercises; the hard part is

**“I really wanted my work to have impact beyond academia in ways that even children could understand.”**

that doing those things reminds people that they’re not well, and exercises are often stigmatizing, boring, and repetitive, or there are more nuanced motivations we need to uncover before we can find solutions.

**It’s not hard to imagine the possibilities for human-machine interactions now, in the post-ChatGPT era, but you saw the potential far earlier. I’m curious to hear your perspective on what’s changed—and what has not changed—in the 20-odd years you’ve been working in the field.**

One thing that’s changed is that machines now talk to us like humans talk, and we perceive machines as if they were human—we agentify, or ascribe agency, [CONTINUED ON P. 107]

# NEW BOOK RELEASE



ACM BOOKS

Collection III

## Digital Dreams Have Become Nightmares

What We Must Do, Second Edition

Ronald M. Baecker  
with Jonathan Grudin

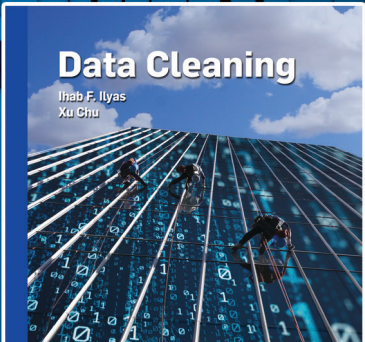
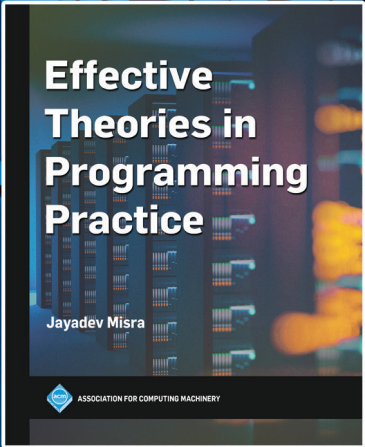
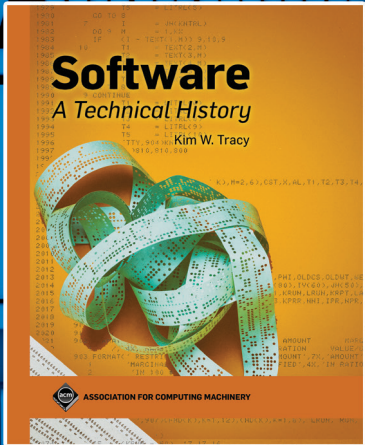
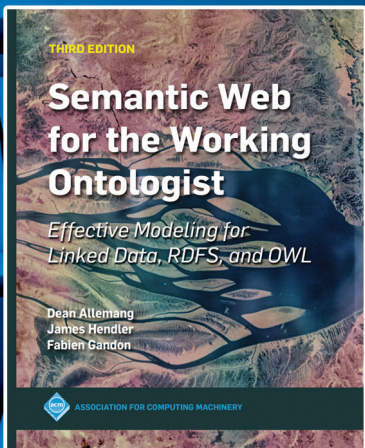
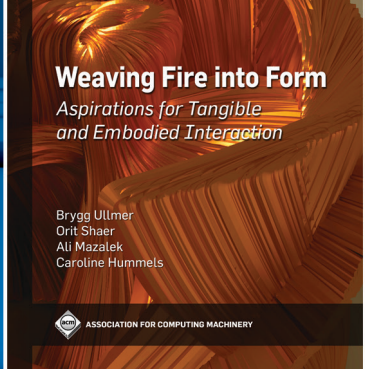
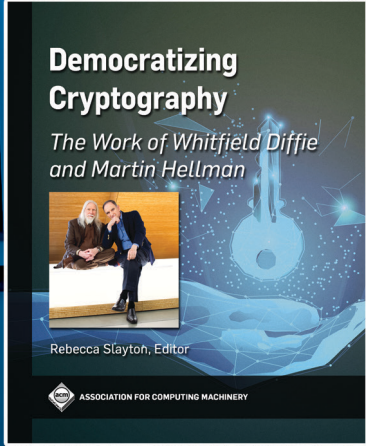
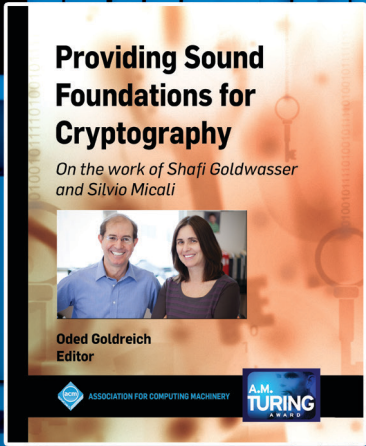
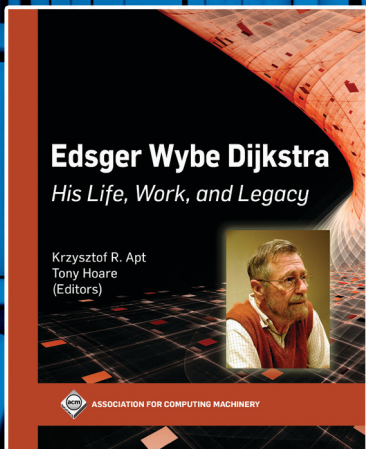
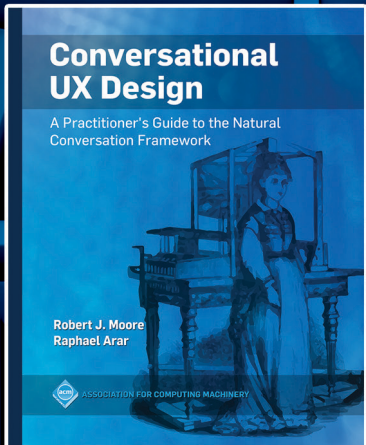
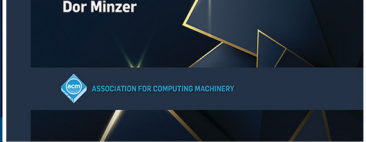
ISBN: 979-8-4007-1768-0

DOI: 10.1145/3640479

"Are you feeling happy about the role of information technology in the world today? You should read this book for a dose of reality. Are you in despair about it? This book is the prescription for that condition, too! **Nobody else could cover the landscape as Ron Baecker does.**" - Clayton Lewis, Emeritus Professor, University of Colorado Boulder

**A compelling discussion of the digital dreams that have come true, their often unintended side effects (nightmares), and what must be done to counteract the nightmares. This book is an impetus to further conversation not only in homes and workplaces, but in academic courses and even legislative debates. Equally importantly, the book is a presentation of what digital technology professionals need to know about these topics and the actions they should undertake individually and in support of other citizens, societal initiatives, and government.**

<http://books.acm.org>



# In-Depth. Innovative. Insightful.

Inspired by the need for high-quality computer science publishing at the graduate, faculty, and professional levels, ACM Books are affordable, current, and comprehensive in scope.

**Collections I & II complete.**

**Collection III now publishing!**



**ACM BOOKS**

For more information, please go to:  
<http://books.acm.org>

1601 Broadway, 10th Floor  
New York, NY 10019, USA  
212-626-0658  
[acmbooks-info@acm.org](mailto:acmbooks-info@acm.org)

