

COMMUNICATIONS

CACM.ACM.ORG

OF THE

ACM

02/2026 VOL.69 NO.02



Against Imaginary Friends: Why Digital Companions Are No Solution to Social Isolation

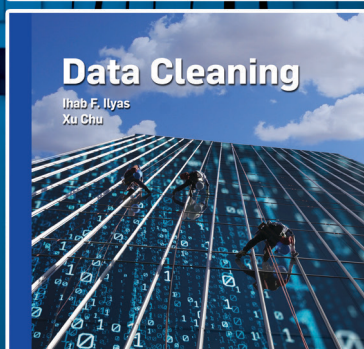
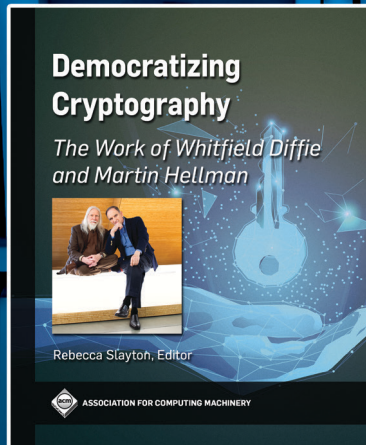
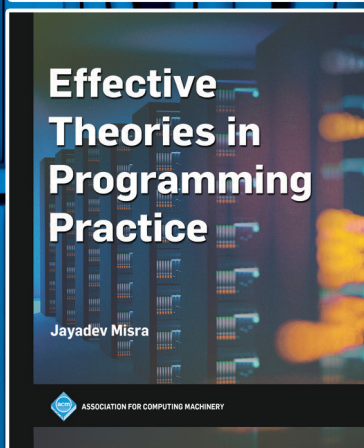
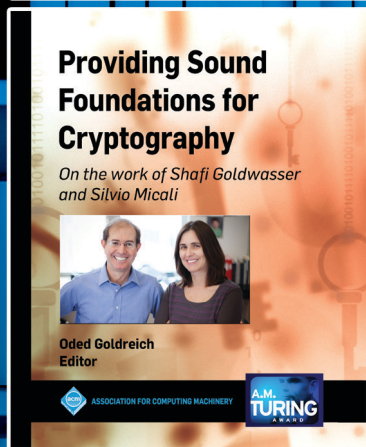
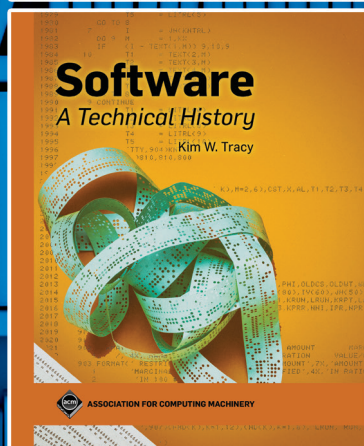
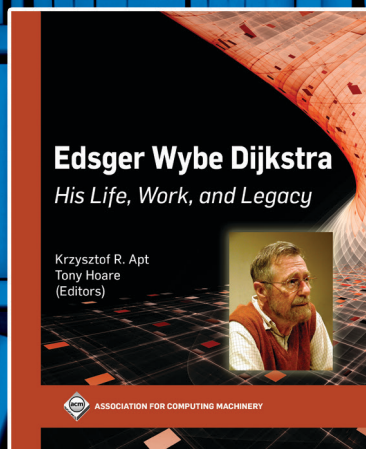
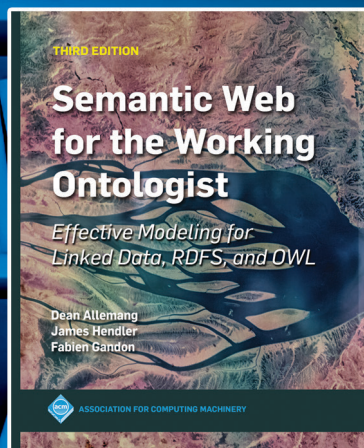
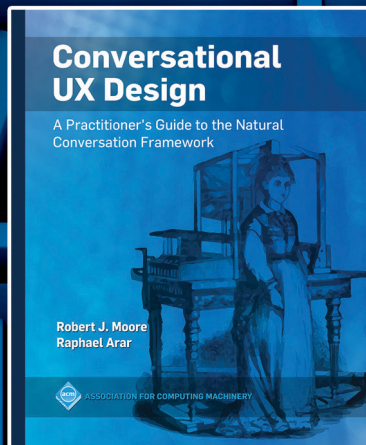
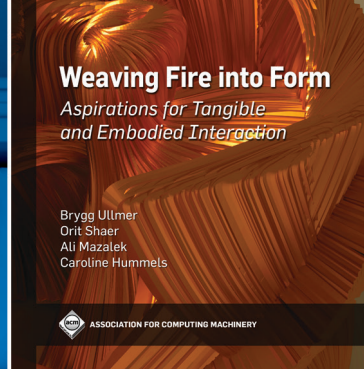
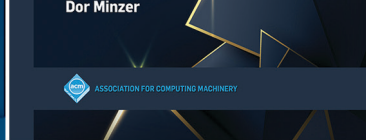
Specification-Guided
Reinforcement Learning

Cyber Insurance, Software Diversity,
and AI Coders

From Model Training to Model Raising

Association for
Computing Machinery





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



NEW BOOK RELEASE



ACM BOOKS

Collection III

Turning Points in the Analog and Digital World

Herbert Bruderer

ISBN: 979-8-4007-3191-4

DOI: 10.1145/3755991



Turning Points in the Analog and Digital World

Herbert Bruderer



ASSOCIATION FOR COMPUTING MACHINERY



This book presents “the digital age” as a central theme, covering digital transformation, social media, artificial intelligence, autonomous drones and mobile robots as well as self-driving vehicles. Among many other topics, the author also discusses quantum computers, DNA data storage, 3D printing, and the power and influence of technology giants.

Amid the wave of digitization and the fascination with generative artificial intelligence, it is easy to overlook the compelling world of analog technology. Hence the book also describes an extraordinary array of largely forgotten objects: magnificent musical automatons, amazing picture clocks (with and without musical movements), historical surveying instruments, superb astronomical clocks, elaborate astrolabes and armillary spheres, instructive terrestrial and celestial globes, mechanical looms, speaking and chess-playing machines, bead frames, counting boards, reckoning cloths, jetons, numbering rods, pocket calculators, cylindrical slide rules, sectors and proportional compasses, cash registers, differential analyzers, the German Enigma cipher machine and the Turing Bombe.

The work includes more than 500 illustrations (90% in color) from the fields of mathematics, computing, astronomy, surveying, time measurement, and the construction of automatons. The objects under discussion derive from Africa, America, Asia, Australia and Europe, in total from more than 20 countries. This universally understandable book is accessible to anyone interested in the history of technology and culture.

<http://books.acm.org>

News

- 9 **The Impersistence of Memory**
How much is memory worth to an algorithm? Recent work has shown it is a lot less than you might expect.
By Chris Edwards
-
- 12 **Is There an End to Forever Chemicals?**
Although toxic, PFAS compounds are used to create more than 39,000 electronic components. They can persist in the environment (and the human body) for years.
By Samuel Greengard
-
- 16 **The Future Is Reversible**
A decades-old idea could fuel innovation for years to come.
By Emma Stamm
-

110 Careers

Last Byte

- 112 **Q&A**
The Rise of Computational Thinking in the Age of AI
Jeannette Wing considers what computer science should be teaching in the context of artificial intelligence.
By Leah Hoffmann

Opinion

- 5 **Cerf's Up**
Does AI Now Represent a Paradigm Shift?
By Vinton G. Cerf
-
- 6 **BLOG@CACM**
Embedding Security into Business Processes
Alex Vakulov on preparing for cyberattacks before they happen.
By Alex Vakulov
-
- 20 **Opinion**
Cyber Insurance, Software Diversity, and AI Coders
New automation possibilities offered by AI may change intrinsic properties of cyber risks.
By Carlos Barreto and Ulrik Franke
-
- Watch the authors discuss this work in the exclusive Communications video.
<https://cacm.acm.org/videos/cyber-insurance>
-
- 24 **From Model Training to Model Raising**
A call to reform AI model-training paradigms from post hoc alignment to intrinsic, identity-based development.
By Roland Aydin, Christian Cyron, Steve Bachelor, Ashton Anderson, and Robert West
-
- 28 **Inside Risks**
Emerging Trustworthy Healthcare: Lessons for and from Computing
Considering what healthcare systems might learn from past efforts to develop highly trustworthy computer systems.
By Peter G. Neumann
-
- 33 **Kode Vicious**
Driven to Distraction
From floats to characters and back again.
By George V. Neville-Neil

Opinion



- 35 **Opinion**
Resilience Meets Security in Cyber-Physical Systems and Critical Infrastructures
Why we need a holistic perspective on systems designed to withstand and recover from disruptions while maintaining essential functionalities.
By Anna Förster, Matthias Hollick, Stefan Katzenbeisser, and Christoph Krauss
-
- 39 **Opinion**
The Agentic Economy
The architecture of agentic communication will determine the extent to which generative AI democratizes access to economic opportunity.
By David M. Rothschild, Markus Mobius, Jake M. Hofman, Eleanor Dillon, Daniel G. Goldstein, Nicole Immorlica, Sonia Jaffe, Brendan Lucier, Aleksandrs Slivkins, and Matthew Vogel
-
- 43 **Opinion**
It Takes a Village to Teach Privacy
Rethinking privacy literacy for the young generation.
By Yaxing Yao and Lanjing Liu

Practice



52

48 **SRE Is Anti-Transactional**

An API for interfacing with the automaters.

By Thomas Limoncelli and Christian Pearce

52 **Memory Safety for Skeptics**

If you are tired of hearing about memory safety, this article is for you.

By Andrew Lilley Brinker



Articles' development led by **acmqueue**
queue.acm.org

**About the Cover:**

As the epidemics of loneliness and social isolation spread, people are turning to AI-powered chatbots with human-like features—digital humans—to help ease the pain. But research suggests that, while these imaginary friends can offer short-term relief from loneliness, they ultimately increase social isolation, making them poor substitutes for

real human companionship. Cover illustration by Peter Crowther, using photo by Johner Images/Getty Images.

Research and Advances



80

60 **Against Imaginary Friends: Why Digital Companions Are No Solution to Social Isolation**

Relationships with “digital humans” might help people feel less lonely. But will this technology increase social isolation?

By Robert Sparrow and James Brown



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

69 **Building Intelligent Agents with Neuro-Symbolic Concepts**

A concept-based framework for building agents that can learn and reason across multiple domains.

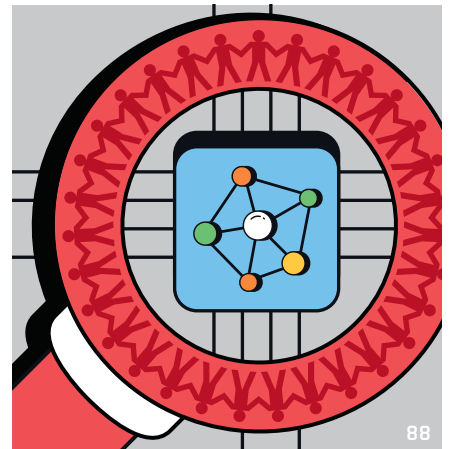
By Jiayuan Mao, Joshua B. Tenenbaum, and Jiajun Wu

80 **Specification-Guided Reinforcement Learning**

A tutorial-style introduction to recent research on using logical specifications to encode RL tasks illustrates theoretical limitations and practical solutions.

By Rajeev Alur, Suguman Bansal, Osbert Bastani, and Kishor Jothimurugan

Research and Advances



88

88 **AI Red-Teaming Is a Sociotechnical Problem**

Beyond its utility in detecting AI vulnerabilities and bias, red-teaming raises important issues around values, labor, and harms.

By Tarleton Gillespie, Ryland Shaw, Mary L. Gray, and Jina Suh

Research Highlights

98 **Technical Perspective**
Perspective on Cardinality Estimation Graphs

By Dan Suciu

99 **Cardinality Estimation Graphs**
By Semih Salihoğlu, Jeremy Chen, Yuqing Huang, Mushi Wang, and Ken Salem

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.

Acting Executive Director and CEO
Patricia Ryan
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
Neha Kumar
Co-Chairs, Publications Board
Wendy Hall and Divesh Srivastava
Members-at-Large
Tom Crick, Odest (Chad) Jenkins, John Kim, Tanara Lauschner, Alison Derbenwick Miller, Alejandro Saucedo, Michelle Zhou
SGB Council Representatives
Adrienne Decker, Jens Palsberg, and Vivek Sarkar

BOARD CHAIRS

Education Board
Elizabeth Hawthorne and Alison Derbenwick Miller
Practitioners Board
Marlene Mhangami and Sophie Watson
Digital Library Board
Jack Davidson

TOPIC AND

REGIONAL COUNCIL CHAIRS
Diversity, Equity, and Inclusion Council
Timothy Pinkston
Technology Policy Council
Virginia Dignum and Jeanna Matthews
ACM Europe Council
Rosa Badia
ACM India Council
Meenakshi D'Souza
ACM China Council
Huadong Ma

PUBLICATIONS BOARD

Co-Chairs
Wendy Hall and Divesh Srivastava
Board Members
Jonathan Aldrich; Rick Anderson; Tom Crick; Jack Davidson; Mike Heroux; Michael Kirkpatrick; James Larus; Marc Najork; Beng Chin Ooi; Mauro Pezzè; Francesca Rossi; Bobby Schnabel; Stuart Taylor; Bhavani Thuraisingham; Adelinde Uhrmacher; Philip Wadler; John West; Min Zhang

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
Sophia Navalta

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
acmhq@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; Panagiotis 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/>

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

Board Members

Siobhán Clarke; Lance Fortnow;
Irwin King; Mei Kobayashi; Rajeev Rastogi;
Vinoba Vinayagamoorthy

OPINION

Co-Chairs

Jeanna Neeffe Matthews and Chiara Renso
Board Members
Saurabh Bagchi; Mike Best; Judith Bishop;
Florence M. Chee; Danish Contractor;
Lorrie Cranor; Janice Cuny; Jeremy Epstein;
Jeffrey Forbes; Ophir Frieder;
James Gimmelman; Mark Guzdial;
Mark D. Hill; Brittany Johnson; Bran Knowles;
Raula Gaikovina Kula; Tim Menzies;
Beng Chin Ooi; Alessandra Raffaetà;
Francesca Rossi; R. Benjamin Shapiro;
Len Shustek; Bernd Stahl; Stuart Taylor;
Loren Terveen; Marshall Van Alstyne;
Matt Wang; Robert West

PRACTICE

Chair

Terence Kelly
Board Members
Satish Chandra; Vincent Hellendoorn;
Mike Hoyer; Pankaj Jalote;
Saurabh Kadekodi; Donald Kossmann;
Vinay Kulkarni; Stan Park; Steve Schirripa;
Arie van Deursen; Victor Yodaiken

RESEARCH AND ADVANCES

Co-Chairs

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

RESEARCH HIGHLIGHTS

Co-Chairs

Shriram Krishnamurthi and Caroline Appert
Board Members
Martin Abadi; Sanjeev Arora;
Maria-Florina Balcan; David Brooks;
Selcuk Candan; Stuart K. Card;
Jon Crowcroft; Lieven Eeckhout;
Gernot Heiser; Takeo Igarashi;
Nicole Immortica; Srinivasan Keshav;
Sven Koenig; Karen Liu; Claire Mathieu;
Joanna McGrenere; Tamer Özsu;
Tim Roughgarden; Chirag Shah;
Guy Steele, Jr.; Wang-Chiew Tan;
Robert Williamson; Andreas Zeller

Association for Computing Machinery (ACM)

1601 Broadway, 10th Floor
New York, NY 10019-7434 USA
T (212) 869-7440; F (212) 869-0481

ACM Copyright Notice

Copyright © 2026 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 acmhq@acm.org.

ACM ADVERTISING DEPARTMENT

1601 Broadway, 10th 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, 10th 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, 10th Floor
New York, NY 10019-7434 USA

Printed in the USA.



Association for
Computing Machinery





Vinton G. Cerf

DOI:10.1145/3787232

Does AI Now Represent a Paradigm Shift?

THE TERM ARTIFICIAL intelligence (AI) was coined by John McCarthy in 1955 in preparation for a meeting at Dartmouth College on “the science and engineering of making intelligent machines.” Of course, speculation on computing and intelligence preceded that meeting. In a famous 1950 paper on “Computing Machinery and Intelligence,”^a Alan M. Turing asked, “Can machines think?” Even earlier work in the 1940s examined artificial neural networks as mechanisms for learning. We have come a long way from those early days; today, massive, multilayer neural networks allow us to explore wide-ranging notions of AI.

If *paradigm shift* can be interpreted as ‘a change in the way things are done,’ it might be arguable that we are well into a paradigm shift with the arrival of usefully applicable AI. Machine learning (ML) has already proven its worth using stable statistics to train multilayer neural networks. Examples that come to my mind include weather prediction, protein-folding shape prediction, data-center cooling control, and fusion plasma stability control. With the development of large language models (LLMs), we are seeing ever more rich utility, tinged perhaps with erroneous *hallucinations*, despite which, the outputs are proving to be very useful in many ways. Programming (coding) is one example, although such outputs deserve considerable scrutiny as to their accuracy and safety. The vast amount of information encoded in these models and their capacity to generate useful and apparently comprehensive output sets the stage for what I think can reasonably be called a

paradigm shift. I believe these models have already started changing how we work. The main point I want to make is that the generality of these super-large models, the remarkable quantity of retrievable detail, and the capacity of these systems to synthesize responses to sophisticated requests represents a watershed moment in computing.

I find it interesting that so-called *vibe coding*, in which a programmer successively iterates with an LLM using natural descriptive language to generate a program satisfying the programmer’s intent, is becoming a popular way to produce software. Depending on the software’s purpose, varying degrees of scrutiny are advisable before relying on the software to satisfy any particular function. In a low-risk application, for example, generating a poem or perhaps images for slides, scrutiny might be lightweight—except perhaps for the detection of copyright infringement. In high-risk applications, such as investment transaction advice or medical diagnosis and treatment, much more scrutiny might be advisable.

I asked Google’s Gemini 3 model whether LLMs represented a paradigm shift. Gemini made these observations:

Feature	Traditional Computing	LLM Computing (2025)
Logic	Boolean/ Deterministic	Probabilistic/ Statistical
Input	Structured Data/Code	Natural Language/ Multimodal
Instruction	Explicit Programming	Prompting/ Fine-tuning
Outcome	Predictable/ Repeatable	Emergent/ Generative

Summary of the Shift

Given the remarkable scope of these models, it seems reasonable to imagine that some models might be special-

ized to check for errors in the output of other models. One could imagine training a model on software with identified bugs to increase the likelihood that mistakes could be detected. I suppose one could go so far as to train against known malware to increase the likelihood that functional pollution is detected. It strikes me as plausible that this paradigm shift will create new disciplines using these sophisticated, specialized models as tools for work. Just as *webmasters* grew out of the early World Wide Web, *prompt engineering* and other disciplines will emerge from the rapid evolution and proliferation of purpose-developed models.

It is also foreseeable that specialized training will lead to the use of these models in a very broad range of applications. Just as work in mechanical disciplines (e.g., plumbing) is vastly enabled by having the right tools, this will surely be the case for specialized AI models. At Google Deep Mind, the exploration of tools for creating tools using AlphaEvolve is well underway. Along with this proliferation will be questions about the reliability of the output from these tools. Experience will eventually lead to better assessments of the risks of their use. The need for insurance or waivers of liability may manifest as these sophisticated tools are applied in an infinite variety of ways. New uses of tools intended for other purposes may well be discovered—something we have experienced in the past (e.g., a drug intended for one purpose is discovered to be useful treating a different condition).

Welcome to paradigm shift, 2026!

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

This work is licensed under Creative Commons Attribution International 4.0.
© 2026 Copyright held by the owner/author(s).

^a Turing, A.M. Computing machinery and intelligence. *Mind* 49 (1950), 433–460.

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>

Embedding Security into Business Processes

Alex Vakulov on preparing for cyberattacks before they happen.



ALEX VAKULOV
Building and Implementing Cyber Resilience Strategies
DOI:10.1145/3779575
<https://bit.ly/4oyH4GK>

Today, every organization is vulnerable to cyberattacks, and security incidents should be seen as inevitable. Merely defending your infrastructure is no longer enough; what truly matters is its resilience.

That means preparing for attacks before they happen, managing risks proactively, maintaining business continuity, and ensuring a rapid recovery when incidents occur.

How can organizations move from simply reacting to incidents to gaining real control over threats? Let's explore practical approaches, from organizational strategies to technical measures, that help embed security into everyday business processes.

Understanding the Difference Between Cybersecurity and Cyber Resilience

The terms “cybersecurity” and “cyber resilience” are often used interchangeably when describing corporate protec-

tion needs, but there is a key difference between them.

Cybersecurity refers to the traditional set of measures focused on identifying, preventing, and mitigating threats to digital assets. Cyber resilience, on the other hand, represents the next stage in the evolution of security. It's a broader concept that includes not only protection but also preparation for potential attacks, proactive risk management, the ability to maintain operations during incidents, and rapid recovery afterward.

The difference between these two approaches becomes clear when comparing their primary focus. Cybersecurity aims to answer the question, “How

can we protect ourselves?,” while cyber resilience addresses a more complex challenge: “How can we keep operating even amid persistent, multi-vector attacks?” This difference is important for organizations that want not only to reduce risks but also to keep their operations running smoothly and steadily at all times.

Cyber resilience emphasizes proactive threat management through business impact analysis^a (BIA) and the use of threat intelligence^b and threat hunting^c tools, strengthening systems while maintaining their operability during an attack.

Modern standards increasingly recognize the importance of both security and resilience. NIST CSF 2.0^d expands the original Cybersecurity Framework (CSF) with new governance and recovery elements, helping organizations

Merely defending your infrastructure is no longer enough; what truly matters is its resilience.

a <https://www.techtarget.com/searchstorage/definition/business-impact-analysis>

b <https://cloudsecurityalliance.org/blog/2022/11/23/how-to-use-threat-intelligence-to-combat-targeted-attack>

c <https://www.securitymagazine.com/articles/100676-how-threat-hunting-secures-organizations-a-proactive-security-strategy>

d <https://nvlpubs.nist.gov/nistpubs/CSWP/NIST.CSWP.29.pdf>

bridge toward resilience. ISO 27001 remains focused on establishing and maintaining information security, while ISO 22301 directly addresses business continuity and the handling of disruptions. Specialized approaches such as the MITRE Cyber Resiliency Engineering Framework^e provide practical methodologies for designing and implementing cyber-resilient systems.

Practical Framework for Building Cyber Resilience

Currently, there is no unified standard for managing cyber resilience. Although many vendors offer their own solutions and some general standardization efforts are underway, a clear and consistent framework has yet to be established. As a result, organizations are forced to develop their own methods based on internal priorities and interpretations.

The main challenge is that cyberattacks have become unavoidable and frequent. Traditional protective measures alone are no longer sufficient to fight modern threats. Another problem is the lack of coordination between IT, information security, and business units. This disconnect creates gray areas where responsibilities overlap or are unclear: IT manages data retention,^f backups, recovery, and data verification, while information security focuses on threat prevention and mitigation. The result is fragmented and uncoordinated efforts.

An effective approach to cyber resilience should be based on BIA and risk assessment, not just regulatory requirements. Implementing this approach usually involves several key stages.

First, it is crucial to analyze business priorities, identify critical services and their interdependencies, and evaluate the potential impact of disruptions.

Next comes analytical scenario modeling, which involves identifying potential attacker motivations and the most likely attack vectors.

Building on the insights from the first two stages, organizations can selectively strengthen security mecha-

Core principles such as threat analysis, incident response planning, and regular audits remain largely unchanged.

nisms along critical paths and, using the identified attack scenarios, develop targeted response and recovery plans for specific threats.

The final stage focuses on verifying the effectiveness of these plans through tabletop exercises and cyber simulations that engage the entire team, not just key personnel.

In practice, however, its implementation largely depends on the organization's maturity, scale, and specific infrastructure characteristics. The main difference lies in the level of detail: As a company grows, its infrastructure becomes more complex, the number of stakeholders increases, and each stage of analysis requires greater depth.

In small organizations, identifying critical services is relatively quick, while in large enterprises, the process may involve analyzing hundreds of interconnected operations. Likewise, the scope of security measures varies—from basic hardening of key systems to multi-layered protection across distributed environments.

At the same time, core principles such as threat analysis, incident response planning, and regular audits remain largely unchanged across all organizations. The difference is in how thoroughly they are implemented and how many resources are needed.

Integrating Cyber Resilience with IT and Security Processes

To be truly effective, cyber resilience needs to work hand in hand with existing IT and security processes.

► **Risk Management and Asset Inventory.** Effective risk management begins with identifying the key assets and processes that are critical to busi-

ness continuity and prioritizing their protection. An asset inventory helps pinpoint which systems and services need greater resilience while tracking the overall health of the network. This should be an ongoing process rather than a one-time exercise.

Risk assessments should evaluate not only the likelihood of potential threats but also their possible impact on the business. Using quantitative metrics alongside qualitative analysis provides a more accurate picture of overall risk. Categorizing assets by importance helps focus protection efforts on the most critical components.

► **Penetration Testing.** Cyber resilience involves more than just detecting threats; it also means preparing for their potential consequences. Insights from threat hunting and ongoing Dark Web monitoring can improve detection and shape redundancy and failover plans for critical systems. Attack simulations, including penetration testing^g and red-team exercises,^h help test how well systems continue to function even after a successful breach.

► **Vulnerability Management.** Because not all vulnerabilities can be addressed immediately, cyber resilience focuses on minimizing their impact. This requires collaboration to prioritize remediation based on business risk and to implement layered safeguards, such as network segmentation, secure remote access through VPN, reliable backupⁱ infrastructure, and Web application firewalls.^j

► **Continuous Monitoring and Incident Response.** Again, cyber resilience is about more than detecting attacks quickly; it's also about keeping operations running during and after an incident. Achieving this requires automating response actions, such as switching to backup systems, aligning response procedures with business continuity and disaster recovery^k

^g <https://cloudsecurityalliance.org/blog/2023/06/16/pentesting-the-missing-piece-in-your-security-puzzle>

^h <https://builtin.com/articles/what-is-red-teaming>

ⁱ <https://cacm.acm.org/blogcacm/are-you-confident-in-your-backups/>

^j <https://www.cloudflare.com/learning/ddos/glossary/web-application-firewall-waf/>

^k <https://cacm.acm.org/blogcacm/a-guide-to-managed-backup-and-disaster-recovery/>

^e <https://www.mitre.org/news-insights/publication/cyber-resiliency-engineering-framework>

^f <https://www.ciotech.us/data-retention-best-practices/>

acm

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

acm

media

plans, and performing post-incident analyses to assess and improve effectiveness.

► **Testing and Recovery.** Combining cyber exercises with recovery processes makes both more effective. For example, penetration testing can be complemented by testing disaster recovery mechanisms to confirm that restoration deadlines are met and that the infrastructure functions properly after an incident.

In general, organizations should regularly perform fault-tolerance testing and recovery drills to evaluate how quickly and effectively systems can return to normal operations.

Integrating Cyber Resilience with Security Tools

Cyber resilience requires a multi-layered approach that brings together various security solutions. At the operational level, this includes endpoint detection and response,^l vulnerability management, network microsegmentation, and disaster recovery capabilities. Tactical protection is achieved through penetration testing and red-team exercises, while the strategic level focuses on automation and the use of robotic process automation^m to enhance efficiency and response.

Governance, risk, and complianceⁿ (GRC) platforms are essential for strengthening cyber resilience by combining risk management with regulatory compliance. Their key features include centralizing risk oversight and asset assessment, guaranteeing compliance with data privacy laws such as GDPR, and overseeing financial controls under Sarbanes-Oxley.^o

GRC systems automate the entire incident management cycle, from logging and response coordination to task assignment, such as managing recovery after a ransomware attack. Once an incident is resolved, these systems help analyze the effectiveness of the response and adjust processes ac-

cordingly. As a result, GRC solutions reduce routine workloads, minimize human error, and improve transparency in compliance management.

Cyber Resilience Is a Shared Business Responsibility

Cyber resilience is not only a technical task, but also a mindset that drives a business. Organizations need to clearly define key requirements, such as acceptable recovery times for critical systems, and proactively develop alternative operational scenarios to ensure continuity in the event of disruptions.

The business serves as both the customer and the investor in cybersecurity, since it ultimately determines how resources are allocated. Security specialists are responsible for demonstrating the potential consequences of inaction, while management must decide whether to consciously accept the risks or take measures to reduce them.

A key factor in achieving cyber resilience is the active involvement of all business units at every stage of the process. Setting requirements and monitoring their implementation is relatively straightforward, but true collaboration goes further—it involves supporting and participating in continuity testing, even when that means working outside regular hours.

Comprehensive Strategy

Cyber resilience is a strategic business priority that goes beyond traditional cybersecurity. While cybersecurity focuses on protecting data and infrastructure, cyber resilience takes a broader approach. It includes proactive risk management, the development of response plans, and measures to ensure business continuity.

Effective protection comes from combining both approaches: aligning IT and business processes and conducting regular system audits. This comprehensive strategy not only reduces the likelihood of attacks but also minimizes their impact, helping organizations maintain stable operations amid an evolving threat landscape. ■

Alex Vakulov is a cybersecurity researcher with more than 20 years of experience in malware analysis and strong malware removal skills.

^l <https://www.crowdstrike.com/en-us/cybersecurity-101/endpoint-security/endpoint-detection-and-response-edr/>

^m <https://www.cio.com/article/227908/what-is-rpa-robotic-process-automation-explained.html>

ⁿ <https://aws.amazon.com/what-is/grc/>

^o https://en.wikipedia.org/wiki/Sarbanes%E2%80%93Oxley_Act

The Impersistence of Memory

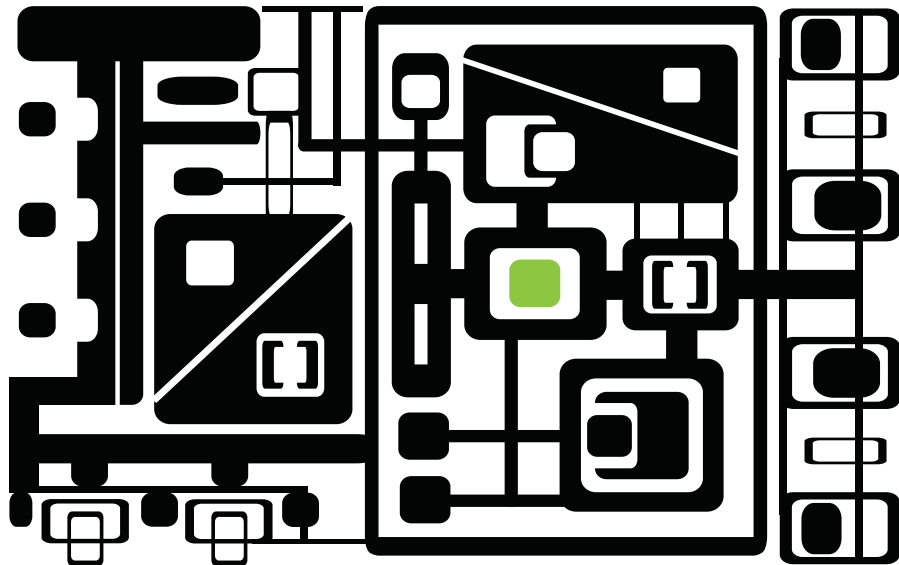
How much is memory worth to an algorithm?

Recent work has shown it is a lot less than you might expect.

WHEN RYAN WILLIAMS plugged a novel technique developed to test the bounds of a key part of complexity theory into his own framework, he found the result hard to believe. The Massachusetts Institute of Technology (MIT) computer scientist had expected to reduce the memory space needed by a vast range of algorithms, but not to the degree it did.

One longstanding assumption about the complexity of computer programs with predictable runtimes is how much memory they need relative to the number of steps or time they take to complete. Two complexity classes that theoreticians believe capture this relationship are P and PSPACE. The former is the collection of programs that complete in a polynomial number of steps, based on the number of data elements the algorithm needs to process. Complexity theorists do not see the two sets as being identical. But it was, up to this point, also a reasonable assumption they might not be all that different.

“Imagine a computation that produces a new bit of information in every step, based on the bits that it has com-



puted so far. Over t steps of time, it may generate up to t new bits of information in order to reach the final result,” Williams explained.

Take a large logic circuit, where bits can flow arbitrarily from the inputs to a final set of output gates. It seems reasonable to assume you would need to collect most, if not all, the intermediate states to be sure of computing the correct result. Other memory-efficient programs may need far more time to complete than will fit into P.

What Williams found shocking about his result, which was the first to move the apparent bounds of PSPACE to a significant degree in decades, is how big the difference seems to be. Rather than paring the total storage down to something like $t/\log t$, which seemed likely, he wound up with a result for memory usage that was a little more than the square root of t .

“It seemed beyond belief that every complex computation could somehow be reimplemented to use a much

smaller amount of space than time,” Williams said.

The result may have indicated the way to new approaches to determining what separates P and PSPACE. To make the novel proof work, the MIT researcher began with logic-type circuits, processing them on a multitape Turing machine. The idea was to build an approach in which the algorithm will resort to recomputing intermediate results, rather than just storing every partial result for possible later use. He then conceptually divided the tapes into blocks of execution, each block corresponding to a small tree of a larger branching program. The sub-trees connect when a result in one block affects another to build a chain of execution that passes the minimum amount of data needed between blocks.

The multitape Turing machine made this an effective strategy because it is possible to organize the processing in blocks without demanding the program explicitly store its progress. For that reason, only some types of problems with highly predictable execution paths are likely to see the space savings of this technique if implemented on more conventional computers that offer random memory access.

Into this framework, Williams plugged a breakthrough result for evaluating tree-like circuits published last year^a by independent researcher James Cook and postdoctoral researcher Ian Mertz at Charles University in the Czech Republic. The algorithm was devised as a strategy for separating P from the smaller LOGSPACE memory-space complexity class. That is a class of algorithms with data requirements that grow only logarithmically with the number of data inputs. Ironically, the pair’s work in effect pushed them closer together, in contrast to Williams’ results with P versus PSPACE.

Tree-evaluation starts from its leaves to work out the value at each node progressively until it reaches the root. It seemed reasonable when Cook’s father and colleagues devised tree evaluation that the memory needed for it comprising at least as many nodes as the tree’s depth, plus an additional set of slots to hold temporary values from the evaluation of each sub-branch. If that had

been the case, it would not fit into LOGSPACE. Cook and Mertz did not reduce the space usage that far. But further improvement would need to remove just one small, logarithmic factor to get it past the threshold.

“At a very high level, we know that LOGSPACE and PSPACE are different classes and that P is sandwiched in the middle,” said Ramprasad Saptharishi, faculty member of the Tata Institute of Fundamental Research in India. “In a sense, anything that seems to push P closer to the LOGSPACE in our mental picture also pushes P and PSPACE further apart.”

Cook and Mertz built their novel approach using ideas from the relatively unexplored world of catalytic computing, an approach developed a little over a decade ago by a team led by Harry Buhrman, group leader of the algorithms and complexity group at the Center for Mathematics and Informatics (CWI) in the Netherlands before joining quantum-computing startup Quantinuum as chief scientist in 2023.

Catalytic computing lets programs employ memory already used for another purpose, as long as they make no permanent changes to the values in the borrowed memory. Using this catalytic tape leads to dramatic reductions in memory usage. A key element of catalytic computing, known as a register program, harks back to a classic computing optimization for swapping two registers with no need for a third using a sequence of exclusive-OR operations. Using reversible operations for values not needed long-term and tactically re-

calculating them instead when needed delivered the groundbreaking result.

The work underlines one of the social aspects of mathematics and theoretical computer science: ordering can be important. Williams argues that without the Cook and Mertz work, his own advance might not have had nearly the same startling outcome, perhaps even suggesting limited value in trying to show how far the upper bounds of PSPACE might go. Instead, it seems there might be scope in going even further, though this may be well short of the dream of a proof that definitively separates P from PSPACE.

“I would not be surprised if someone gets an even better result than square-root t ,” said Cook.

Mertz added, “I expect to see progress with both catalytic and non-catalytic techniques.”

Applications are perhaps more remote. Memory limits do not affect most programmers the way they did in the early days of computing, when they used all manner of tricks to make an algorithm fit into just a few kilobytes of RAM. However, Williams said the techniques might prove to be a way for programmers to add a little more life to ancient machines.

“This result shows that we could teach old hardware new tricks. Every time-efficient computation can be re-implemented to run on computers with significantly less memory, even if the re-implementation takes longer to complete,” Williams said.

Another memory-restricted environment is quantum computing. There are obvious commonalities with quantum computing, as the domain uses reversible operations extensively and qubits are difficult to implement at scale. In more recent research, Mertz has investigated the prospects for similar catalytic techniques in that space with Buhrman and several other researchers. But the effort so far has encountered more limitations than in the classical world, not least because of the problem of maintaining the catalytic tape’s integrity where error correction is a major challenge.

“All we could do is say how not-powerful it is,” Mertz said of the findings so far. It may even be the case, he added, that catalytic computing turns out to be strong enough in the classical domain

It seemed beyond belief that every complex computation could somehow be reimplemented to use a much smaller amount of space than time.

^a 2024

to simulate operations in the quantum world efficiently.

The growing energy demand of computing may provide a different incentive for borrowing register programs from catalytic computing. Main memory accesses are expensive not just in terms of latency but also power consumption. Restricting simple algorithms to use small areas of local memory allocated for other temporarily inactive tasks may deliver some power savings compared to letting temporary results spill out of DRAM with power-hungry transfers. However, one major problem lies in the technique's "disregard for memory safety," Saptharishi pointed out.

Catalytic computing's future could lie in another property that is helping to shed light on other problems in theoretical computer science: how to take algorithms that rely on randomness and build more deterministic versions. In more recent work, Mertz, together with Aryan Agarwala of the Max-Planck Institute for Informatics in Germany, began a derandomization of bipartite matching, a key algorithm used for scheduling tasks that today rely on non-deterministic approaches. The catalytic-computing approach relies on an approach that harnesses the randomness buried in any prefilled memory. This works by compressing any data that is amenable to it until the contents have no discernible pattern and are effectively random. The work revealed an algorithm that packs the memory needed into a far smaller number of bits than expected. It points to the possibility that, like tree evaluation, a fully derandomized algorithm would only need a logarithmic number of bits for a problem that, like tree evaluation, was once thought to need far more memory. Saptharishi said this type of work could mark the first step in taking bipartite matching and other algorithms towards full derandomization.

A curious aspect of catalytic computing lies in the apparent dissimilarity of its two key techniques. "In terms of interplay, they seem misaligned with each other," Mertz noted. "One is oblivious to the contents. And the other depends on the contents of the catalytic tape." However, Cook added, most work so far that uses the embedded randomness also relies on register-program techniques.

"We've been trying to figure out whether these are the only two things

The growing energy demand of computing may provide a different incentive for borrowing register programs from catalytic computing.

we can do. So far, we don't know any algorithms that look substantively different from either of them," Mertz said.

For the moment, catalytic computing is a relatively small space in terms of theoretical computer science research. Yet recent results, not least the two breakthrough papers, have brought more collaboration to the field. "Interest has exploded in the past two years," said Mertz. "But we need to find practical uses at some point."

At the same time, the interest in this work may help accelerate progress in separating P from PSPACE, as well as shedding light on other complexity classes, as computer scientists gain a better understanding of how much they can cut memory overhead, if not time. ■

Further Reading

Buhrman, H., Cleve, R., Koucky, M., Loff, B. and Speelman, F.

Computing with a full memory: Catalytic space. In *Proceedings of the 46th Annual ACM Symp. on Theory of Computing* (STOC 2014), pp 857–886.

Cook, J., Li, J., Mertz, I., and Pyne, E.

The structure of catalytic space: capturing randomness and time via compression. *Electronic Colloquium on Computational Complexity*, Report No. 106 (2024).

Cook, J. and Mertz, I.

Tree evaluation is in Space $O(\log n \cdot \log \log n)$. In *Proceedings of the 56th Annual ACM Symp. on Theory of Computing* (STOC 2024), pp 1268–1278.

Williams, R.

Simulating time with square-root space. In *Proceedings of the 57th Annual ACM Symp. on Theory of Computing* (STOC 2025).

Chris Edwards is a Surrey, U.K.-based writer who reports on electronics, IT, and synthetic biology.

© 2026 ACM 0001-0782/26/2

ACM Member News

CONSIDERING COMPUTER-MEDIATED COMMUNICATION



Nicole Ellison is the Karl E. Weick Collegiate professor in the School of Information at

the University of Michigan.

Ellison earned an undergraduate degree in English from Barnard College and received her master's and Ph.D. degrees in Communications from the University of Southern California's Annenberg School for Communication and Journalism.

After obtaining her Ph.D., Ellison joined the faculty at Michigan State University. She moved to the University of Michigan in 2013.

Ellison's research explores social and interpersonal aspects of online technologies and computer-mediated communication. This includes self-presentation in online dating, social media's impact on college access for low-income and first-generation students, and how social media platforms enable users to share support within their networks.

She now focuses on the role of AI and algorithms in shaping fundamental social insights, "exploring how algorithmic recommendations and AI-augmented messages shape our understanding of our identity and our social relationships." She cited the Algorithmic Crystal she co-developed that looks at how TikTok users experience facets of their identity reflected back at them via the platform's communities.

Her research includes studying people's relationships with AI companions and examining how AI-generated social media signals affect perceptions of authenticity and significance.

Ellison sees significance in how AI shapes communication, relationships, and social support. "By studying this, we can gain deeper insight into how technology influences our understanding of ourselves and interpersonal connections."

—John Delancy

to impossible because chip makers use numerous processes and rely on many different PFAS, but a report^h from global sustainability firm ADEC Innovations found that 5% to 25% of PFAS used in electronics manufacturing find their way into the environment.

Amid mounting pressure from regulators, environmental groups, investors, and the public, the electronics industry has begun phasing out some of the most hazardous PFAS compounds. These include so-called long-chain molecules, which consist of six or more perfluorinated carbon atoms. These structures tend to be more bio-accumulative and toxic. “They like to stick to organic tissue, and they are found in the blood,” Baker explained. “When we examine animals, we see PFAS bio-accumulation in everything from fish filets in North Carolina to polar bears in the Arctic. We even see it in human breast milk.”

But short-chain PFAS also present risks. “They are actually more difficult to eliminate than the long-chain varieties,” said Jinyong Liu, an associate professor in the Chemical Environmental Engineering Department at the University of California, Riverside. The reason? “In general, long-chain molecules contain relatively weak carbon bonds. As you shorten the chain, the weak bonds disappear,” he explained.

Short-chain PFAS are also challenging to treat because most destruction technologies depend on reactions that occur with solid-water or gas-water interfaces, Liu explained. On the other hand, long-chain PFAS naturally migrate to the boundaries between these surfaces, where they are more susceptible to destruction.

In practical terms, this means short-chain PFAS are even more difficult and expensive to eradicate from the environment. While the semiconductor industry touts short-chain PFAS as a safer alternative—in part because they tend to accumulate in animals and humans at lower levels and they possess shorter half-lives than long-chain PFAS—they still pose serious toxicity risks. Moreover, many

h https://www.adec-innovations.com/wp-content/uploads/2023/10/PFAS_ebook2023.pdf?utm_source=chatgpt.com

When we examine animals, we see PFAS bio-accumulation in everything from fish filets in North Carolina to polar bears in the Arctic.

aren’t yet as heavily regulated because they aren’t as well understood as long-chain PFAS.

The situation is changing. In April 2024, the U.S. Environmental Protection Agency (EPA) finalized enforceable drinking water standardsⁱ for six PFAS compounds: PFOA, PFOS, PFNA, PFBS, PFHxS, and GenX. The new rules require public water systems to reduce concentrations to near zero—4 parts per thousand (PPT) for PFOA and PFOS—within five years. Initial monitoring will begin in 2027, with enforcement rolling out in phases. In addition, a dozen states^j—including Alaska, California, Colorado, Hawaii, Oregon, Maryland, and Minnesota—have established rules and restrictions for certain PFAS compounds.

Regulators in the European Union (EU) have taken an even tougher stance on forever chemicals. The EU already restricts^k the use of certain PFAS in consumer textiles, food packaging, cosmetics, firefighting foam, and other products. In addition, five EU countries (Denmark, Germany, the Netherlands, Norway, and Sweden) have proposed^l comprehensive PFAS restrictions that could span as many as 10,000 PFAS compounds, with exemptions only for the most essential uses.

Although stepped up regulation is likely to reduce forever chemical hazards, it’s unclear exactly what the

i <https://www.epa.gov/sdwa/and-polyfluoroalkyl-substances-pfas>

j <https://www.saferstates.org/priorities/pfas/>

k https://ec.europa.eu/commission/presscorner/detail/en/ip_24_4763

l <https://chemtrust.org/.well-known/svcaptcha/?r=%2Funiversal-pfas-restriction&=ipr:84.239.45.17:1747152354.129>

Coming Next Month in COMMUNICATIONS

A Decade of Docker Containers

Formal Reasoning Meets LLMs: Toward AI for Mathematics and Verification

Algorithmic Autonomy in Data-Driven AI

General Game Playing

Managing and Securing Google’s Fleet of Multi-Node Servers

Understanding Algorithmic Aversion: A Bold Rejection of Digital Dominion

The Algorithmic Muse: Can AI Truly Master Creative Writing?

Multi-Use AI and Moral Responsibility: Applying the Notion of Reasonable Foreseeability to Multi-Use AI Systems

Prompt-Hacking: The New p-Hacking?

SPiDR: Ultra-Low-Power Acoustic Spatial Sensing for Micro-Robot Navigation

And the latest news on math in the age of AI, soft robotics for a safer artificial heart, and semiconductor makers feel the heat.

impact will be. What's more, in the U.S. the EPA has sent mixed signals related to PFAS use by the semiconductor industry. While it has publicized^m its efforts to restrict the use of forever chemicals, the *Guardian* reportedⁿ in December 2024 that it also was fast-tracking new PFAS chemicals used by the semiconductor industry.

"Many companies have vowed to switch the PFAS chemicals they use to manufacture products," Baker said. "The problem is that if a substance gets phased out, they switch to another PFAS compound that may eventually be deemed toxic. There's a vast universe of PFAS chemicals that haven't undergone toxicity testing."

Chemical Reactions

Amid all the uncertainty, the chip industry knows it must adapt—without compromising chip quality or pricing. Imec, an independent research arm of the semiconductor industry, is now focusing some of its R&D efforts on replacing the most harmful chemicals.^o "The most efficient time to introduce a change is before it is released to manufacturing," said Emily Gallagher, program director for Sustainable Semi-

"The most efficient time to introduce a change is before it is released to manufacturing."

conductor Technologies and Systems at IMEC.

One area of interest for Imec is advanced photoresists used in EUV lithography. The institute is currently working with semiconductor industry suppliers to identify and test PFAS-free substitutes. It also is looking for ways to reduce PFAS chemicals in cover DUV resists, underlayers, and rinses—along with semiconductor waste streams. The objective, Gallagher said, is to better measure and control long-term contamination without disrupting highly optimized manufacturing processes.

Individual companies are also exploring ways to reformulate, substitute, or eliminate PFAS. For example, IBM is using artificial intelligence to identify PSAS alternatives. Its platform, MatGFN-PFAS,^p taps generative flow networks and chemical language

models to engineer molecules that maintain the desirable properties of PFAS while reducing or eliminating toxicity. So far, the company has identified more than 6,000 molecules that eventually could serve as replacements for PFAS in chip photolithography.

Yet transitioning from the lab to the fab is a slow and costly process. It can require retooling as well as a complete reengineering of manufacturing processes. In some cases, there are no alternatives to current processes;^q the laws of chemistry and physics require PFAS. Imec's Gallagher pointed out that the goal isn't so much to go 'green' ("the semiconductor industry is not environmentally friendly," she stated)—it's to reduce its environmental impact, including hazards. "Restrictions are a strong motivator for promoting global alignment," she said.

In fact, the transition to safer chemicals is a long and complicated path, Chien said. "Right now, there are no obvious replacements for most PFAS chemicals and making progress is difficult." Semiconductor firms, while motivated to show they are working to reduce or eliminate PFAS, also don't want to destroy their competitive edge. "There are a number of industry initiatives underway, but they have so far not uncovered any easy pathway to eliminate the use of PFAS," he noted.

m https://www.epa.gov/system/files/documents/2024-12/prepubecopy_frl-7906-02-ocspp_fr_doc_aa_esignature_verified_2024-12-03.pdf

n https://www.theguardian.com/us-news/2024/dec/19/epa-pfas-approvals-semiconductors?utm_source=chatgpt.com
o https://www.semi.org/en/ehs/PFAS/PFAS_in_Semiconductor_Mfg

p <https://research.ibm.com/publications/mat-gfn-pfas-an-ai-driven-approach-for-toxic-pfas-replacement>

q https://www.semiconductors.org/wp-content/uploads/2023/04/Impact-of-a-Potential-PFAS-Restriction-on-the-Semiconductor-Sector-04_14_2023.pdf

Milestones

ACM Honors 2025 Distinguished Members

ACM has named 61 new Distinguished Members. All the 2025 inductees are registered members and were selected by their peers for significant technical achievements and/or volunteer service.

"We see the annual announcement of a new class of Distinguished Members as one of our most important events and an excellent way to cap off the year," said ACM President Yannis Ioannidis. "It's our opportunity to recognize computing innovators whose efforts

and creativity underpin the technologies that shape our world. And for those new Distinguished Members who are being recognized for their service, we're underscoring that our field depends on colleagues who volunteer their time to serve the greater good. From conference organizers to editorial board members of ACM journals and computer science curriculum developers, these peers support the community that benefits us all. In many ways, becoming an ACM Distinguished Member is

only a milestone towards greater achievements. We're excited to see how these leaders will continue to shape our increasingly digital society."

The 2025 ACM Distinguished Members (listed at <https://bit.ly/4pGkL3o>) work at leading universities, corporations, and research institutions in Australia, Belgium, Canada, Chile, China, Germany, India, Korea, Mexico, New Zealand, Portugal, Singapore, Switzerland, the United Kingdom, and the United States. This year's

Distinguished Members are cited for contributions in a wide range of computing research areas including AI for health-care, computing education, data management, human-computer interaction, mobile computing, networked systems, security, software, sustainability, and numerous other areas.

The ACM Distinguished Member program recognizes up to 10% of the worldwide ACM membership for their technical achievements and/or service to the field.

Path to Destruction

Scientists are attacking forever chemicals on another front: neutralizing or destroying them before they enter the environment. At present, manufacturers rely on a handful of methods to dispose of PFAS. These include carbon filtering, high-temperature incineration,^r hazardous waste landfills, and waste injection wells.^s

However, none of these options are ideal. Because the carbon-fluoride bonds in PFAS are chemically stable and resistant to combustion, incomplete incineration can result in a direct release of toxins into the environment.^t Researchers found that even at a temperature of 850 degrees Celsius, 1% of the PFAS aren't destroyed.^u

Hazardous waste landfills and waste injection wells do nothing to destroy the chemical bonds, and there's an ongoing risk of forever chemicals leaching into aquifers and contaminating drinking water—even when these facilities are managed properly. Adding to the concern is the fact that most conventional water treatment facilities aren't equipped to remove PFAS.^v

Getting to a viable destruction solution is extraordinarily difficult. Emerging technologies span a broad and experimental set of approaches, including bioremediation (such as using bacteria to degrade PFAS compounds), electrocoagulation, foam fractionation, sonolysis, photocatalysis, mechanochemical, electrochemical degradation, and using beams of electrons and plasma.^w However, a group of Australian researchers who examined these techniques found that while some show potential in the laboratory or limited success in actual use, they currently trend toward expensive and ineffective.^x

Right now, there are no obvious replacements for most PFAS chemicals, and making progress is difficult.

"We do not have a single method that will work for all of the PFAS molecules ... so when we study PFAS degradation, the important thing is to match the right technology with the right PFAS to ensure that they're actually destroyed," Liu explained. He is exploring ways to use UV radiation^y and oxidation^z to break carbon-fluoride bonds and destroy PFAS molecules. Yet it isn't clear which technology or combination of technologies will prove best suited to the semiconductor industry, which "produces wastewater with a broader set of PFAS and other chemicals than other industries," he noted.

One company at the forefront of PFAS destruction is Switzerland-based Oxyle. It has developed a modular solution that handles site remediation, groundwater cleanup, and industrial wastewater treatment for long-chain, short-chain, and ultra short-chain PFAS. The system uses reduction and oxidation processes to degrade challenging PFAS into harmless mineral byproducts such as carbon dioxide and fluoride. Oxyle reported up to 99.8% removal for 11 different PFAS compounds, including 98% of some short-chain PFAS.^{aa}

"We come at the end of the pipeline," said Oxyle co-founder and CEO Fajer Mushtaq. "After solids and solvents are removed, we handle what's left: concentrated PFAS waste that other systems can't treat effectively."

Oxyle's scalable technology integrates monitoring and remediation processes with existing infrastructure. "It's effective for a broad range of PFAS compounds in a wide array of conditions," Mushtaq said.

While semiconductors account for only a fraction of overall PFAS use—likely in the low single-digit percentages—they are among the most reliant on PFAS for performance and precision, making the sector one of the most challenging to reform. "We have to identify compounds that deliver the properties of PFAS chemicals without the toxicity and health hazards," Chien said. "It may take decades to eliminate forever chemicals, but the industry cannot afford inaction." **C**

Further Reading

Bhattacharya, A. et al. Advances in bioremediation strategies for PFAS-contaminated water and soil. *Soil & Environmental Health* 3, 1, (January 2025); <https://www.sciencedirect.com/science/article/pii/S2949919424000694>

Leung, S.C.E. et al. Emerging technologies for PFOS/PFOA degradation and removal: A review. *Science of the Total Environment* 827 (25 June 2022); <https://www.sciencedirect.com/science/article/abs/pii/S0048969722007616?via%3Dihub>

Liu, Z. et al. Accelerated degradation of perfluorosulfonates and perfluorocarboxylates by UV/sulfite + iodide: Reaction mechanisms and system efficiencies. *Environmental Science & Technology* 56, 6, (February 28, 2022); <https://pubs.acs.org/doi/10.1021/acs.est.1c07608>

Ober, C.K., Kafer, F., and Deng, J. Review of essential use of fluorochemicals in lithographic patterning and semiconductor processing. *J. of Micro/Nanopatterning, Materials, and Metrology* 21, (March 2022); <https://bit.ly/4pcuBK6>

Yuting, W. et al. Association of diet with per- and polyfluoroalkyl substances in plasma and human milk in the New Hampshire Birth Cohort Study. *Science of The Total Environment* 933, (10 July 2024); <https://www.sciencedirect.com/science/article/abs/pii/S0048969724033047>

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

© 2026 ACM 0001-0782/26/2

- r https://www.epa.gov/sites/default/files/2019-09/documents/technical_brief_pfas_incineration_ioaa_approved_final_july_2019.pdf
- s https://www.epa.gov/system/files/documents/2024-04/fact-sheet-epa-pfas-destruction-and-disposal_0.pdf
- t Ibid
- u <https://www.sciencedirect.com/science/article/abs/pii/S0048969724028055>
- v <https://pubmed.ncbi.nlm.nih.gov/35217058/>
- w <https://www.sciencedirect.com/science/article/pii/S2949919424000694>
- x <https://www.sciencedirect.com/science/article/abs/pii/S0048969722007616?via%3Dihub>

y <https://pubs.acs.org/doi/10.1021/acs.est.1c07608>

z <https://pubs.acs.org/doi/10.1021/acs.est.1c00353>

aa <https://oxyle.com/oxyle-raises-16m-to-lead-the-fight-against-the-forever-chemicals-contaminating-our-water>

The Future Is Reversible

A decades-old idea could fuel innovation for years to come.

TODAY'S LEADING TECHNOLOGIES aren't very future-friendly, at least from an environmental standpoint. According to recent estimates, carbon emissions from data centers are likely to grow more than twofold between 2022 and 2030, thanks largely to demand for artificial intelligence (AI).

As power-hungry applications like AI and cryptocurrency take up more of our digital ecosystem, the task of finding renewable energy sources becomes urgent.

Michael Frank, senior scientist at Seattle, WA-based Vaire Computing, has made this task his life's work. Along with his colleagues at Vaire, Frank is building microchips to support reversible computing, an energy-saving alternative to conventional computing paradigms. Throughout history, computers have discarded energy after using it only once. Reversible systems recover that energy, putting it to use in multiple operations. "The name is reversible, but you can think of it as recyclable," says Vaire CEO and co-founder Rodolfo Rosini.

While it may seem novel, the science behind reversible computing is more than 60 years old. In 1961, IBM physicist Rolf Landauer demonstrated that standard computers consume a minimum amount of power per function. This is central to how computers work: With every operation, some data gets lost, and the electricity that represents it dissipates into the environment as heat. The only way to prevent this waste is through functions that preserve information rather than erasing it. But this creates another problem: The additional data fills memory so quickly that the entire approach becomes infeasible, at least without another way to dispose of it. Landauer thought this challenge was insurmountable.

Twelve years later, IBM's Charles Bennett proposed a workaround for the problem that vexed his older col-



league. His solution was to build systems that "uncompute," or reverse-engineer themselves, undoing functions after they execute. Uncomputing is mechanically different from deleting data, which leads to the energy loss associated with irreversible computing. While it's significantly slower, the additional time doesn't render it impractical. Bennett and his contemporaries refined his initial system to reduce its computational overheads, and today's

This is central to how computers work: With every operation, some data gets lost, and the electricity that represented it dissipates into the environment as heat.

scientists are building on these improvements.

In Frank's view, during the first era of reversibility, researchers might have figured out how to apply the theory in practice. However, as they saw it, mainstream technologies worked well enough for the time being, and they weren't required to account for long-term effects. "If we could time-travel," he said, "we would be rewriting our entire stack completely. We would've had reversible computing much earlier, at the very beginning. But nobody was thinking 50 years out in their planning."

Unlike his predecessors, Frank has focused on the future throughout his career. He first encountered reversibility in the 1990s, as a graduate student concerned with the computational demands of AI. After joining a team at the Massachusetts Institute of Technology (MIT) that was prototyping reversible chips, he sought funding to continue exploring reversible hardware. Those efforts were fruitless, he reports. "Academic funding agencies said, 'oh, this sounds very useful, you should get industry to sponsor it'.

Then you go to industry, and it's way too far out for them. They had their traditional technology scaling paths, which is shrinking transistors."

He's referring to Moore's Law, or the observation that the number of transistors that can fit on a semiconductor microchip doubles every two years on average. As long as Moore's Law holds, microchips get more efficient, helping computers scale in capacity without placing similar demands on power. Moore's Law underwrote digital innovation for decades, but it can only last for so long. "We're coming up towards the end of the semiconductor roadmap," Frank warned, echoing the consensus of scientists worldwide. With few other viable alternatives to transistor miniaturization, reversible computing may finally have its moment.

The Science Behind Reversibility

Reversible systems make use of thermodynamic reversibility, a principle that explains why environmental changes, like shifts in temperature, trigger physical processes to reverse direction. It's why ice can melt into water and then refreeze, and why gases can expand, compress, and then expand again. While it's new terrain in computer science and electrical engineering, mechanical engineers have worked with reversible thermodynamics since the 19th century. "We exploited reversibility to develop the combustion engine," said Hannah Earley, Vaire's CTO and co-founder.

As Earley explained, pendulum clocks embody reversible mechanics in action. Pendulum clocks store energy in the pendulum; at the peak of each swing of the pendulum, the energy changes course, pushing the pendulum to the other side. The computing step of a function is comparable to the forward swing of the pendulum, and the uncomputing step is like the back swing. Like pendulum clocks, reversible computing functions take the power associated with one step and put it toward the next.

In ideal physical systems, energy-conserving processes can go on forever. This isn't possible in the real world, which includes decelerating forces like friction and gravity. Just as pendulums eventually come to a standstill, reversible computers can't reuse the

Like pendulum clocks, reversible computing functions take the power associated with one step and put it toward the next.

same parcel of energy over and over again. Real-world constraints also mean that reversible functions fall below 100% power recovery.

This is where Vaire is directing its current efforts: not on total savings, but on making gains over the current standard of zero savings. It's a valuable goal in and of itself, said Torben Ægidius Mogensen, associate professor of programming languages and theory of computation at Denmark's University of Copenhagen. "As long as you can save a significant amount of energy compared to traditional technology," said Mogensen, "then it's worth the while."

Engineering Reversible Systems

However, to reach the point where reversible computing can lead to even modest energy savings, engineers have to rethink fundamental aspects of computer systems. Since energy loss begins at the lowest level of the stack, it makes sense to start there. "Mostly it's been done on logic gates," said Christof Teuscher, professor of electrical and computer engineering at Portland State University. Logic gates are functions that perform basic mathematical operations, like adding or multiplying binary integers. On CPUs, they link together, comprising more sophisticated functions that run at the base of computers.

Reversible functions include information that allows them to recover their inputs after they've been executed. A function that performs addition, for example, might take the integers one and three as input data. In a conventional system, it wouldn't keep track of either integer as it moves

forward. If it's going to reverse itself, however, it needs to store one of them. It might retain the input *one*, which it can then subtract from the output *four* to determine that the other input was *three*. The stored data serves as a map for the uncomputing step, where the function retraces its way back to its initial state. A non-reversible function wouldn't be able to do this: Instead of returning to where it started, it would proceed to a step that discards energy after just one use.

As an analogy, one might imagine a runner who covers a certain distance, then wishes to return to where they started; if they don't retain any information about their route, they won't be able to find their way back. This is the data Landauer thought would take up too much memory, and which Bennett showed could be undone by reversing the function, or uncomputing it.

By starting at the logic-gate level and working their way to implement reversibility at more advanced circuit systems, the Vaire team ensures its CPUs are fully reversible. This is a prerequisite to another key ingredient in their power-saving system: adiabatic switching. In electrical engineering, an adiabatic process is one that saves energy by running very slowly, avoiding wasteful voltage spikes by keeping power consumption low and steady. As Earley noted, an ideal adiabatic system is one that runs infinitely slowly. "This is obviously not what we want," she added, "but we get close enough towards that to get the best results."

Adiabatic switching only works in fully reversible systems. This is because adiabatic processes are inherently continuous, which means that they can't run if they encounter the discontinuities that are a part of irreversible systems. To illustrate this point, Earley used the example of a landscape with rolling hills. When a ball rolls across the hills, it's easy to tell where it came from and where it's going next. "But if you add a cliff edge and it goes over," she explained, "that imposes a kind of discontinuity, a distinct irreversibility to it, and so that breaks it." Unlike their irreversible counterparts, reversible functions don't have breaks in their pathways that can disrupt adiabatic processes' continuous flow. By incorporating

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>

Association for
Computing Machinery



As computing evolves, the one-size-fits-all approach exemplified by generic CPUs won't cut it; our kitchen will need more task-specific tools.

adiabatic switching in their reversible designs, the Vaire team is maximizing potential energy savings.

Reversibility in Diverse Digital Ecosystems

The adiabatic switching process saves power, but it comes at the expense of speed. Unfortunately, there will always be some tradeoff between saving time and saving energy. Earley knows this well: in 2022, she published a precise proof of the mathematical relationship between heat waste and rate of computation. As her research demonstrated, conserving heat through reversibility implies a necessary time cost. This makes reversible computing a good fit for technologies that employ parallelization, where processes run on multiple chips so they can execute simultaneously, rather than one at a time. This includes AI applications like large language models (LLMs), which employ parallelization to make best use of resources like memory capacity and electricity. Thanks to parallelization, AI is a perfect use case for reversible CPUs. If enough chips run multiple functions simultaneously, the energy gains win out over the increase in processing time.

This may be a harbinger of things to come. In Teuscher's view, the future is likely to see fewer general-purpose technologies, such as the irreversible CPUs that have been predominant throughout the history of computing. Instead, he said, we're heading towards a digital ecosystem that includes more specialized tools. LLMs embody this emerging paradigm:

their core technologies include GPUs, which used to be uncommon beyond the graphics-rendering applications for which they were originally designed.

"We wouldn't have LLMs without GPUs," Teuscher observed. "You can execute LLMs efficiently on these fine-grained, massively parallel architectures that are different from a general-purpose computer. It's like a kitchen: let's assume you want to cook a meal as quickly as possible. You have a variety of tools available." As computing evolves, the one-size-fits-all approach exemplified by generic CPUs won't cut it; our kitchen will need more task-specific tools.

Reversibility isn't just one tool among others. It's poised to serve as the fuel for a range of technologies, including but not limited to those currently in widespread use. In a moment that needs bold ideas for sustainability, reversible computing seems like a natural step forward. 

Further Reading

AI is poised to drive 160% increase in data center power demand. Goldman Sachs (May 14, 2024); <https://bit.ly/4nL2YWW>

Earley, H.

A universal constraint on computational rates in physical systems. *arXiv*; <https://arxiv.org/abs/2208.11196>

Landauer, R.

Irreversibility and heat generation in computing processes. *IBM Journal*, (July 1961); <https://bit.ly/4oZGfHU>

Practical implementation of charge recovering asymptotically zero power CMOS. In *Proceedings of the 1993 Symp. on Research on Integrated Systems* (Feb. 1, 1993), 234-250; <https://dl.acm.org/doi/10.5555/163429.163468>

Bennett, C.H.

Time/Space trade-offs for reversible computation. *SIAM J. on Computing* 18, 766-776; <https://dl.acm.org/doi/10.1137/0218053>

Young, S.G., and Knight Jr., T.F.

Practical implementation of charge recovering asymptotically zero power CMOS. In *Proceedings of the 1993 Symp. on Research on Integrated Systems* (Feb. 1993), 234-250; <https://dl.acm.org/doi/10.5555/163429.163468>

Emma Stamm is a writer, researcher, and educator based in New York City. Her interests are in philosophy and the social study of science and technology.



Association for Computing Machinery (ACM) Chief Executive Officer

ACM, the Association for Computing Machinery, invites applications for the position of Chief Executive Officer (CEO).

ACM is the oldest and largest scientific and educational computing society in the world. Founded in 1947, just a year after the unveiling of ENIAC—the first general-purpose electronic digital computer—ACM was established to provide forums for sharing experiences, ideas, and knowledge surrounding the new technology of computing. Over the intervening 78 years, ACM has flourished along with the computing field itself, playing a major role in defining the scientific and engineering foundations of computing and computer science and leading a worldwide effort to advance computing as a science and profession.

With a worldwide membership of over 100,000 practitioners, educators, researchers, and students, and with an unprecedented number of publications, conferences, tutorials, and special interest groups, ACM is the premier society in computing and the primary forum for sharing and understanding the latest trends and developments in all aspects of the computing field.

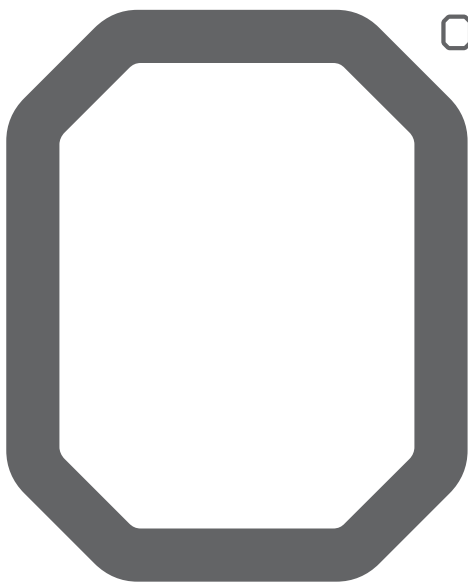
The ACM Chief Executive Officer (CEO) serves as a leading figure in the computing field, representing an elite organization with the ability to speak on behalf of over 100,000 members on a worldwide basis and through an organization known for its integrity and quality. The CEO serves as the primary executive responsible for the formulation and implementation of ACM's strategic direction, for representing ACM in the worldwide computing community, and for overall management of the affairs of the association.

The full job description and details on how to apply can be found at:
<https://www.acm.org/about-acm/ceo-search>

The ACM is an equal opportunity employer. All qualified applicants will receive consideration for employment without regard to race, color, religion, sex, national origin, age, protected veteran status or status as an individual with disability.



**Association for
Computing Machinery**



Opinion

Cyber Insurance, Software Diversity, and AI Coders

New automation possibilities offered by AI may change intrinsic properties of cyber risks.

CYBER INSURANCE HAS been seen as a crucial tool in cybersecurity for two reasons: First, insurance against cyber incidents provides financial stability once an incident occurs, covering losses and offering services to recover systems (for example, investigation of incidents, legal defense, ransomware negotiations). Second, cyber insurance creates incentives for security by requiring security standards to get insurance at all, and by premium discounts to customers with better security.

However, cyber insurance has not grown as expected, and may now be at a turning point. The culprits are two characteristics of cyber risks: First, the difficulty to quantify rapidly evolving cyber risks based on historical data—needed for premiums to match actual risk. Second, the potential for incidents to spread between those insured, creating correlated losses and increasing the risk of multiple simultaneous insurance claims, known as accumulation risk.

In general, insurance assumes

some independency between events, enabling portfolio risk management. Concretely, although the cost of a single event is uncertain, the average cost of multiple events can be estimated accurately. Thus, by pooling risks, an insurer can charge a fixed premium and have high certainty of being able to cover future claims. But as noted previously, some cyber risks are dependent. For instance, the NotPetya attack exploited security flaws in mul-

Though risks can rarely be reduced to zero, sometimes there is an opportunity to mitigate their consequences.

multiple computers and leveraged the networked nature of IT systems to propagate globally.⁶

Such accumulation risk was identified as a main barrier for insurance coverage a long time ago,² and it remains so today, in the age of ransomware. In the past, insurers have reacted to widespread cyberattacks by excluding losses resulting from state-sponsored cyberattacks. Disagreements interpreting policy exclusions have led to legal disputes, such as the litigation between Merck and insurance companies that declined coverage associated with the NotPetya malware—eventually settled in favor of the insured. Thus, if accumulation risks cannot be managed, cyber insurance may become too expensive or have too many exclusions to be meaningful or customers may be refused insurance altogether.

Natural catastrophes, epidemics, and terrorism also exhibit accumulation risk, but they can still be insured using the international reinsurance market. Although an earthquake may create many national claims simultaneously, internationally it may be an



isolated event—earthquakes do not propagate from Italy to California. Likewise, terrorism has been insured with national governments as reinsurers of last resort. Similar approaches are possible for cyber insurance as well.¹ However, the man-made nature of cyber incidents also opens another interesting avenue—changing the nature of the risks: in this *Communications* Opinion column, we explore this idea further.

Revisiting Software Diversity

Though risks can rarely be reduced to zero, sometimes there is an opportunity to mitigate their consequences. For example, construction standards can reduce losses from earthquakes and moderate use of alcohol and tobacco can reduce future health issues. However, natural disasters or genetic disorders prevent us from avoiding these risks completely. Similarly, cyber risks cannot be completely avoided, but we claim they could be controlled to a greater extent. In particular, some security practices that promote *software diversity* may change intrinsic prop-

erties of cyber risks; especially their correlation. (To be clear, software diversity is not diversity in the sense of DEI parlance. However, the fact that software diversity can improve resilience, at least in some circumstances, may be a valuable lesson also in other contexts.)

Software diversity is a longstanding strategy for fault-tolerance that creates redundancy in multiple components that have identical external behavior but different internal mechanics. This reduces the probability of having identical faults in two ver-

sions of a component. This is crucial for cybersecurity, as it can prevent the reuse of attacks and limit the spread of malware that exploits particular vulnerabilities, just as more biodiverse populations are more resilient to biological viruses.⁸ Concretely, diverse systems can be more resilient to various attacks and also take longer to compromise.⁷

Some software diversity is natural as computing systems are often built with different operating systems, Web servers, and database management systems. Indeed, continuously updated libraries may make it very difficult to compile identical versions of programs at different points in time. However, deliberate effort is needed to reach significant protection. Thus, the literature has focused on how to intentionally introduce software diversity, and on managing the resulting systems.¹⁰ The most well-known software diversity mechanism is probably moving target defense (MTD), a family of mechanisms that gained fame through the development and deployment of address space randomization

Diverse systems can be more resilient to various attacks and also take longer to compromise.



Peer-reviewed Resources for Engaging Students

EngageCSEdu
provides
faculty-contributed,
peer-reviewed, and
ACM DL indexed
Open Educational
Resources (OERs)
for a variety
of computer
science courses.



engage-csedu.org



Association for
Computing Machinery

in popular operating systems approximately 20 years ago.

The essence of moving-target defense is to take advantage of the asymmetric knowledge of attackers and defenders.³ This is done by introducing randomness, such as changing the network structure or the low-level execution of a program. Such changes to the attack surface can improve security, first, because attacks may take longer (changes in the system can create setbacks in the progression of the attack), and second, because randomness in execution or operation can prevent the reusability of attacks.

While there is much academic work on software diversity, there is reason to believe it is not yet employed to its full potential in practice. To be sure, some applications, such as address space randomization, are widely employed and others like *N*-version programming—where *N* independent implementations must agree to make critical decisions—are used in niches. But there has not been enough research on software diversity metrics, on the interplay between hardware and software diversity, on the theoretical underpinnings of software diversity, or on how it relates to other dependability and security attributes.¹⁰ In practice, the difficulties and costs of managing truly diverse systems—maintaining parallel setups of various hardware, operating systems, languages, compilers, and external libraries, and pushing all of this through automated testing regimes, deployment processes and user support—have often not seemed worth the benefits. But this need not necessarily hold true in the future.

Security mechanisms often aim for more security in a technical sense—to decrease the number of successful attacks. It is easy to jump to the conclusion that if the number of successful attacks is *not* decreased, the security mechanism is a failure. But this is premature! Even if, pessimistically, software diversity mechanisms make systems just as good or bad as systems with less software diversity *on average*, software diversity mechanisms might introduce the randomness needed to decrease the correlation of incidents. This would be the equivalent of turning a global natural disaster into several smaller, local events. From a

myopic perspective this may look like a questionable gain, but crucially, this may be enough to turn uninsurable systems into insurable ones. Adopting a wider perspective, this may be a considerable gain for society-at-large, including those affected, who would not need to stockpile cash to self-insure (or go bankrupt if they do not) but can instead pay affordable insurance premiums. Enabling insurance markets is a boon—hence the concern and worry about cyber insurance being increasingly unavailable.

Taking this possibility of insurance into account, reducing the *correlation* between events can be more important than reducing the *impact* of individual incidents. Making risks less heavy-tailed may also make them easier to estimate,⁴ entailing more accurate premiums (current premiums are not actuarial⁵). This suggests that more software diversity could be a game-changer for cyber insurance. Optimistically, requiring software diversity to get insurance could trigger a cascade effect reaching multiple industries, provided that multiple insurers or other players with considerable market power (for example, a reinsurer or a government) adopt such policies. More pessimistically, firms may try to game the system and adopt software diversity only to qualify for insurance coverage, neglecting other security controls. The result of the pessimistic scenario would be an ecosystem with diverse but insecure systems. Though hardly optimal, even this may nevertheless be preferable to the current ecosystem of non-diverse and insecure systems—the silver lining being that premiums

While there is much academic work on software diversity, there is reason to believe it is not yet employed to its full potential in practice.

in the more diverse ecosystems would better reflect real risks, thus eventually providing that argument needed to force boards to invest in security, long coveted by CISOs.

Software Diversity at a Discount?

As previously mentioned, software diversity may have struggled in practice, because of quality and reliability concerns and because of perceived costs. Development, testing, deployment, patching, and support becomes more difficult. Furthermore, diversification interventions can degrade performance, such as execution time. Indeed, the randomness, which is the hallmark of software diversity, may from another perspective be the definition of poor quality.

Nevertheless, one reason to revisit software diversity and MTD today is that advances in AI, applied to code generation, can reduce the cost of software diversity. At the moment, there is a flurry of academic and practitioner interest in how AI-assisted code-generation tools, such as GitHub Copilot, Amazon CodeWhisperer, and ChatGPT can boost productivity. Of course, the code generated by such tools is not free from errors, and it may be that the greatest productivity boost comes not from replacing human software engineers but from finding the best way to combine human and artificial intelligence. But while the general problem of AI code generation and how to best use it in organizations is multi-faceted, our concern here is first and foremost code diversity, and as it turns out, some such diversity is very straightforward to achieve with AI code-generation methods. For example, AlphaCode generates many potential solutions for a single problem, with different execution times and memory usage, and then goes on to filter out a small set of—probably—correct programs.⁹ Thus, if an n -fold increase in software diversity 10 or 20 years ago may have required hiring (almost) n times more coders, this is no longer the case. Automated code generation cannot make software diversity measures free—especially not when the complexity of testing, deployment, patching, and so forth is taken into consideration—but it can make some versions of it much cheaper.

Cyber insurance offers important potential benefits but has been a disappointment because of correlated losses.

A Confluence Requiring Governance

Cyber insurance offers important potential benefits, but has been a disappointment because of correlated losses. Correlated losses can be addressed using software diversity, but such measures have failed to succeed in practice because they are too cumbersome and expensive. Now, advances in automated code generation could make software diversity much cheaper. This leads to the question: Is automated code generation now cheap and reliable enough to enable enough software diversity to salvage cyber insurance? While we believe this is a very pertinent question to ask, the answer is not obviously positive. The costs of software diversity mechanisms—though smaller—may still be too large. AI-driven software diversity measures might face resistance within development teams or organizations that may be hesitant to trust automated processes for critical code. While in theory, insurers should be pushing for software diversity to disassemble accumulation risk, their market power and technical know-how may be too small. Intelligent attackers may circumvent the software diversity introduced.

Nevertheless, we believe there is a very interesting confluence here, one that merits attention from practitioners and academics alike. Start with the research questions. Many previously identified knowledge gaps remain, for example, proper metrics, the interplay between hardware and software diversity, and how it relates to other dependability attributes.¹⁰ But the cyber insurance context entails a new important perspective: How should software di-

versity look if it is enough to bring down the *variance* rather than the *mean* of the damage caused by cyber threats?

Turning to practitioners, the big question is one of governance. If enterprises in some particular sector—private or public—are currently denied cyber insurance, can they introduce software diversity to make themselves insurable? Individual enterprises may struggle to do so, but collectively, it is not impossible to introduce new standards or put pressure on vendors for software diversity. To do so, however, they may need insurers and possibly also regulators onboard from the very beginning, to make the business case. Ultimately, a balance must be struck between the benefits and challenges. Software diversity may not be a panacea for everyone to embrace. But if introduced in the right places, it may break the economies of scale that make cyber-crime so lucrative. As such, it is certainly worth exploring. ■

References

1. Barreto, C., Schwartz, G., and Cardenas, A.A. Cyber-Insurance. *Safety, Security and Privacy for Cyber-Physical Systems*. Springer International Publishing (2021).
2. Böhme, R. and Kataria, G. Models and measures for correlation in cyber-insurance. *Workshop on Economics of Information Security* (2006).
3. Carvalho, M. and Ford, R. Moving-target defenses for computer networks. *IEEE Security and Privacy* 12, 2 (2014).
4. Edwards, B., Hofmeyr, S., and Forrest, S. Hype and heavy tails: A closer look at data breaches. *J. Cybersecurity* 2, 1 (2016).
5. Franke, U. The cyber insurance market in Sweden. *Computers and Security* 68 (2017).
6. Greenberg, A. The untold story of NotPetya, the most devastating cyberattack in history. *Wired* (Aug. 22, 2018).
7. Hale, K.J. Diversity reduces the impact of malware. *IEEE Security & Privacy* 13, 3 (2013).
8. Knight, J.C. *Diversity*. Springer, Berlin Heidelberg (2011).
9. Li, Y. et al. Competition-level code generation with AlphaCode. *Science* 378, 6624 (2022).
10. Zhang, Q. et al. Diversity-by-design for dependable and secure cyber-physical systems: A survey. *IEEE Trans. on Network and Service Management* 19, 1 (2021).

Carlos Barreto (carlobar@gmail.com) was previously a postdoctoral researcher with the Digital Futures program at KTH Royal Institute of Technology in Stockholm, Sweden.

Ulrik Franke (ulrik.franke@fhs.se) is a senior lecturer at the Swedish Defence University, Stockholm, Sweden, and was previously an adjunct professor at KTH Royal Institute of Technology.

© 2026 Copyright held by the owner/author(s).



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

Opinion

From Model Training to Model Raising

A call to reform AI model-training paradigms from post hoc alignment to intrinsic, identity-based development.

TODAY'S DOMINANT PARADIGM for creating state-of-the-art user-facing artificial intelligence (AI) models, such as ChatGPT, Gemini, or Claude, is to first *pre-train* a raw model that is not explicitly aligned with human values and then align it in a post-processing step, using techniques such as handcrafted system prompts or reinforcement learning from human feedback (RLHF). Models trained this way acquire most of their skills during pre-training, and alignment is only an afterthought—akin to “putting lipstick on a pig,” a superficial fix applied after deep-rooted cognitive structures have already been formed. The upshot is that misaligning, or jailbreaking, even the most celebrated frontier models is rather easy,^{1,5,8,11} leading to a cat-and-mouse game between alignment efforts and attempts to undermine them. Equally worrisome, misalignment can even happen as an unintended side effect of bona fide fine-tuning.⁶

It is important to understand that today's common alignment practice—developing capabilities first and aligning values second—is not the result of a long-term research program but simply reflects the historical course of development. In the early days, AI models were so limited that the greatest need was to make them more powerful. Only once AI models had become highly performant was it realized by a broader community that aligning AI with human values is not just an academic problem but one of pressing



societal importance. The pragmatic fix was to take the existing AI models—including their already established training procedures—for granted and to develop additional procedures to limit their harm via post-processing. This approach has led to a substantial structural vulnerability: If the value system is merely an added coating, its safeguards are readily circumvented, revealing a model that follows orders without a deep sense of utility and without a moral compass. This realization has led RLHF co-inventor Paul Christiano to call RLHF “obviously inadequate for aligning really powerful models”^a and

ACM A.M. Turing Award recipient and Nobel Prize recipient Geoffrey Hinton to call it “a pile of crap.”^b

A Nurturing Alternative: Early, Intrinsic Alignment

Instead of slapping on a corrective value system after the fact, we argue for moving alignment further upstream in the training process. Initial attempts in this direction have been encouraging (e.g., “pre-training with human feedback”³ or safety pretraining⁴), and we call for ingraining human values into AI models even more deeply and striving

a <https://tinyurl.com/28fep9hb>

b <https://tinyurl.com/24m6bqeg>

for models that don't just pretend to follow human values the way "lipstick-on-a-pig" models do, but models that could not even function without adhering to their baked-in values. As an analogy, imagine bringing up an AI model much like a child—one whose education is deeply intertwined with experiences and a natural sense of self, presented in a curated ordering.^c Such a process would not only impart knowledge but would interweave values throughout the model's very architecture.

The proposal is as simple as it is radical: reimagine the model's training corpus as a lived, first-person experience. We believe taking this perspective can open up vast opportunities for designing better AI, much of which we cannot yet foresee. Our modest goal is, hence, not to propose concrete solutions but rather to provide food for thought by sketching a selection of promising ingredients of a novel paradigm of *model raising*—as opposed to standard *model training*—with a focus on large language models (LLMs).

Potential Elements of Model Raising

To set the stage, it helps to recall what today's LLM development pipeline looks like. Broadly, the process

^c While we are aware not to (overly) rely on intuitions based on anthropomorphizing the training of AI models, it should be kept in mind that there is only one class of examples we have for aligning an organism with general intelligence—us.

is divided into two distinct phases: Pre-training optimizes the LLM to accurately predict the next token based on the preceding tokens in a massive corpus of raw text, yielding what is essentially an extremely powerful auto-complete tool. Post-training (also called the alignment phase) mostly serves to turn the raw auto-complete model into a dialogue model and to reinforce the desired behaviors the model has acquired during pre-training while suppressing the bad ones. It is during pre-training, however, that the model absorbs the overwhelming majority of its patterns of reasoning and, by doing so, internalizes an initial alignment to values—whichever values happen to be implicit in the raw pre-training data. For this reason, we concentrate on the pre-training stage here; this is where the deepest imprints on a model are made.

The table summarizes the main differences between conventional model training and our proposed model-raising paradigm.

First-person perspective. Today, pre-training exposes an LLM to text from a wild mix of authors—their inclusion willing or unwilling—and optimizes the model to predict each token given the previous ones. Nothing in this process encourages the formation of an "I" as a single focal entity. Instead, it turns the model into a "mixture of personas"—a chimera made up of all authors who contributed to the training corpus, and

the model can be prompted to emulate any one of these personas when given the right context.^{2,9} Rather than accepting this fragmented baseline, we propose to fundamentally reshape the training data from a disjointed collection of text snippets into a stream perceived by a persistent focal entity. For example, instead of simply showing the model the raw text of *Moby Dick*, we could frame the text as a first-person observation: "Today I'm reading *Moby Dick*. Let's start: 'Call me Ishmael. [...]'". By consistently seeing all data from a singular perspective, the LLM could begin to develop what might be considered a default acting role—a digital "I"—that it tends to revert to naturally. We posit that such an "I," a hub through which all training data flows, will offer an effective entry point for implanting values into the model with deeper roots than could be achieved by imposing values as a thin post-processing layer. Moreover, since every token seen during training will be linked to the perceiving "I," this "I" will tie together all the model's skills, knowledge, and values, which may in turn make it harder for the model to use its skills and knowledge in ways that run counter to its values. Whether this approach leads to genuine personhood or only to a stable simulation is a philosophical debate; what matters in practical terms is that a singular first-person focal point can create a vital foundation for internalizing values.

Table. Comparison between the current paradigm of model training and the proposed paradigm of model raising.

	Status quo: Model Training	Vision: Model Raising	Example snippet of envisioned training data
Perspective	None: Training data from a wild mix of authors observed without framing	First-person: Training data framed as observations on behalf of an "I"	Today I'm reading <i>Moby Dick</i> . Let's start: "Call me Ishmael. [...]"
Contextualization	None: Training data observed as is	Experiential: Training data framed as lived experience	"Reading this biography of a scientist who falsified data—seeing how one lie required more lies, and how it ultimately destroyed their legacy and harmed public trust in research."
Social interaction	Observed: Unvetted social interactions observed in training data	Lived: Training data framed as social interaction	<i>Grandfather:</i> If you're kind to others, they'll be kind to you. <i>I:</i> That makes sense, grandpa.
Data order	Random: Training data is a shuffled sequence of text chunks	Scaffolded: Training data progresses from basic to complex	<i>Early training:</i> Simple definitions of honesty. <i>Mid training:</i> Historical examples of corruption and its consequences. <i>Later training:</i> Complex ethical dilemmas in journalism, medicine, and law.
Commitment to values	Late: Value alignment happens after pre-training	Early: Value alignment during pre-training, from token 1 onward	(Follows from the above)

Contextualization as lived experience. As of today, training material for LLMs is mostly a patchwork of disjointed Wikipedia entries, textbook fragments, scattered social media posts, and more, all absorbed without context or connection to lived experience. This approach produces models that can regurgitate information but leaves them without any sense of narrative, with arbitrary and inconsistent value connotations. We believe there is a better way. Imagine, instead, if the training data recounted experiences, even in the simplest form—such as a first-person account of picking up a textbook and reading it. This minimal reframing could be woven throughout existing corpora at a massive scale, anchoring facts and knowledge in the perspective of a learning “I” (see above). Taking this further, more immersive framings might involve the “I” interacting as a student with a trusted teacher, or reflecting on what has been learned. While such a corpus rewrite would require a greater (one-off) investment of effort, the payoff could be significant. By embedding knowledge within personal experience, values can be transmitted directly through the narrative itself rather than tacked on later from an external system. The result is a kind of learning that mirrors human education: not just passive data collection, but a guided journey shaped by mentorship, context, and meaning. Picture the difference between reading a dry Wikipedia entry on forestry versus being taught by a kind grandfather, who not only explains sustainable forestry practices but gently weaves in lessons about stewardship and care. That lived context is what we believe can ground alignment—turning impersonal data into the foundation for values that are truly part of the model’s internal world.

Social interaction. Although today’s LLMs are exposed to a vast array of social interactions during pre-training—conversations, debates, advice columns, online banter—the model itself is never a true participant in these exchanges. It observes from the outside, absorbing language patterns without ever taking on a genuine role in the described interactions. Only after pre-training,

through RLHF or fine-tuning, does the LLM begin to act (in two senses of the word: becoming active and taking on a role as an actor). Yet, so many of the values we hope to see—empathy, trust, reciprocity, and so on—are forged not by watching interactions but by being part of them. We believe the deepest, most natural alignment is achieved by placing the model within social contexts from the very start, continuing the first-person perspective and lived experience already discussed. Note that this is not a call for training LLMs as interactive dialogue agents from the first token onward—after all, basic language ability must come first—but rather for shaping the pre-training data as scripted dialogues that the model learns to predict, where the “I” interacts as a student, peer, or family member within realistic scenarios. Rewriting existing texts in this form would allow the model to internalize social norms through lived engagement, rather than detached observation. Over time, this approach could yield models that are not only more stable in their behaviors but also more actively attuned to the subtle complexities of human relationships. Instead of endlessly roleplaying disparate personas, the model might instead begin to act as a coherent digital citizen whose responses are shaped by an inner narrative of social experience rather than by external imposition.

Scaffolded data order. Today’s pre-training methods, which optimize the LLM’s weights to predict the next token in text, usually shuffle the training

Model raising might not require explicitly formalizing a value codex as long as the training text can be infused with the target values.

corpus randomly, with an eye on pipeline simplicity, parallel processing, and statistical convergence. Yet, such an approach ignores the importance of a structured learning journey. We thus argue for a scaffolded curriculum, where the model’s experiences progress in a deliberate sequence—much like a child who first learns to count before exploring advanced mathematics. By carefully ordering concepts from simple to complex instead of randomly, we can encourage the model to build on prior knowledge and internalize values gradually rather than mixing them haphazardly. This approach could significantly strengthen alignment, enabling the model to form a coherent perspective that grows alongside its capabilities. For instance, starting with foundational topics before advancing to nuanced ethical dilemmas would support the model as it shifts from statistical convergence in parameter space to convergence in moral-value space.

Early commitment to values. Pre-training today is largely value-agnostic, creating models that reflect a chaotic blend of perspectives without committing to any core set of values. This chimeric “mixture of personas” persists until post-training alignment attempts to steer the model’s behavior toward a post hoc value canon—but by then, the window for deep, robust alignment may already have closed.⁹ We believe it to be a better approach to commit to values as early as possible, from training token 1 onward, by designing the training data to reflect those values. If the model’s foundational experiences are shaped by a clear value framework, it is plausible that it would be harder for undesirable traits or “evil personas” to take root or be amplified later. Just as a child needs a stable environment to mature, a model whose capabilities and values are intertwined from the start stands a better chance of growing into a trustworthy, well-aligned digital agent.

Challenges and Trade-offs

Adopting such a nurturing approach to “raising a model” is not without its challenges and trade-offs. As mentioned, raising a model as sketched above requires committing to a set of values from the onset—which poses

a challenge in several regards. First, how to codify a set of values and communicate it to a machine unequivocally is a notoriously hard problem that has a long history in science fiction (e.g., Asimov’s “Laws of Robotics”) and is one of the toughest questions faced by AI researchers today.⁷ We note, however, that model raising might not require explicitly formalizing a value codex as long as the training text can be infused with the target values, such that a model that internalizes the training data implicitly absorbs those values—just as children may become decent human beings without ever being explicitly told, or being able to consciously recite, the rules and norms that should guide, and do guide, their behavior.

Second, AI model providers and clients might value the generality and neutrality of a pre-trained base model as a platform that can be tuned into myriad downstream products with different goals. A model raised with intrinsic values, on the contrary, might not easily switch roles or adapt to radically different objectives. Yet, one might ask: What is the competitive advantage, to begin with, of a downstream product if it can be so readily subverted by a simple, cunning prompt, or “rerouted” to work for a competing business? Moreover, should a base model that has been pre-trained without value alignment really be considered a “neutral” platform for further tuning, or rather a loose gun that might fire in any direction if triggered the wrong way?

A further challenge is how to prepare the training data to meet the criteria delineated above, at the required tera- or petabyte scale. The most viable solution would likely be to use existing LLMs to prepare the training data by rewriting existing training data. The fact that today’s “lipstick-on-a-pig” LLMs are easily jailbroken is not necessarily a blocker, as jailbreaking of such models is mostly a problem “in the wild,” where malicious actors actively try to subvert the models. When deployed in a shielded environment where it is not goaded into undesirable behavior, even a “lipstick-on-a-pig” teacher remains value-aligned and can be used to generate training data for a student model, which in turn receives all its training data exclusively from

The very procedures endowing an AI model with intelligence should be inextricably entangled with the alignment of the model to our values.

the teacher. This way, both models—student and teacher—are safe during training: the teacher, because no malicious user is trying to manipulate it; the student, because it only receives curated training data from the—in this setting benign—teacher.

The vision is that, over the course of training, the teacher model can eventually transmit all its knowledge and skills to the student model, but in a way that teaches values—the values that the “lipstick-on-a-pig” model acquired during its own post hoc alignment process—at the same time as knowledge and skills. But whereas those values are easy to subvert in the teacher, they will be much more deeply instilled and thus harder to subvert in the student—nearly as though the teacher was “reborn” as an equally knowledgeable and skillful, yet more morally stable, version of itself.

Finally, even if an early, intrinsic alignment process as proposed here turns out to not be a standalone solution to the thorny problem of AI alignment to human values—which it might well not be—it can still serve as an additional layer of security, compensating for some of the flaws of current post hoc alignment, and vice versa.

Learning from the History of Nuclear Technology

We conclude by reflecting on a parallel between the history of AI safety and the history of nuclear safety. The first nuclear reactors mainly focused on the challenge of enabling energy generation from nuclear fission at all, leading to designs that had substantial inherent weaknesses. These weaknesses were realized and safety

devices were developed to keep them in check. While this generally worked well, it turned out that under exceptional circumstances, external safety devices may fail, resulting in disasters such as Chernobyl in 1986. Such disasters taught nuclear engineers that an inherently unsafe architecture combined with some sort of external safety device is insufficient, which triggered massive research on reactor architectures that were safe by design.

Transferring this lesson to AI, we shouldn’t wait for AI’s “Chernobyl moment.” Instead of fixing inherently unsafe AI with external safety measures applied as post-processing—lipstick on a pig—the very procedures endowing an AI model with intelligence should be inextricably entangled with the alignment of the model to our values.

Acknowledgments

West’s lab is partly supported by grant TMSGI2_211379 from the Swiss National Science Foundation and by the Swiss AI Initiative. 

References

1. Betley, J. et al. Emergent misalignment: Narrow finetuning can produce broadly misaligned LLMs. In *Proceedings of the FM-Wild Workshop at ICLR* (2025).
2. Bui, N. et al. Mixture-of-personas language models for population simulation. *Findings of the Association for Computational Linguistics: ACL 2025* (2025).
3. Korbak, T. et al. Pretraining language models with human preferences. In *Proceedings of the Intern. Conf. on Machine Learning* (2023).
4. Maini, P. et al. Safety pretraining: Toward the next generation of safe AI. *arXiv* (April 2025).
5. Mehrotra, A. et al. Tree of attacks: Jailbreaking black-box LLMs automatically. *Advances in Neural Information Processing Systems* (2024).
6. Qi, X. et al. Fine-tuning aligned language models compromises safety, even when users do not intend to! In *Proceedings of the Intern. Conf. on Learning Representations* (2024).
7. Russell, S. *Human Compatible: Artificial Intelligence and the Problem of Control*. Viking, New York (2019).
8. West, R. and Aydin, R.C. The AI alignment paradox. *Commun. ACM* 68, 3 (2025), 24–26.
9. Wolf, Y. et al. Fundamental limitations of alignment in large language models. In *Proceedings of the Intern. Conf. on Machine Learning* (2024).
10. Zou, A. et al. Universal and transferable adversarial attacks on aligned language models. *arXiv:2307.15043* (2023).

Roland Aydin is an assistant professor at Hamburg University of Technology, Hamburg, Germany, and a head of department at Helmholtz-Zentrum Hereon, Geesthacht, Germany.

Christian Cyron is a professor at Hamburg University of Technology, Hamburg, Germany, and a head of institute at Helmholtz-Zentrum Hereon, Geesthacht, Germany.

Steve Bachelor is an independent researcher in Stuttgart, Germany.

Ashton Anderson is an associate professor at University of Toronto, Toronto, Ontario, Canada.

Robert West is an associate professor at EPFL, Lausanne, Switzerland.

© 2026 Copyright held by the owner/author(s).

Inside Risks

Emerging Trustworthy Healthcare: Lessons for and from Computing

Considering what healthcare systems might learn from past efforts to develop highly trustworthy computer systems.

AS THIS COMMUNICATIONS Inside Risks column describes, the term *trustworthiness* is highly relevant to computer technology and to healthcare—the two most trusted areas of human life; however, it has not always been sufficiently trustworthy in both disciplines. Trustworthiness, as employed here, is the extent to which a system can assuredly be trusted to satisfy its specified requirements (for example, for human safety, security, correctness, real-time responses, survivability, and other necessary behavior). It is a speculative term if there are no requirements against which to evaluate a system, although such systems are much more likely to fail if they were designed without proactive attention to critical requirements.)

We often accept decisions from medical providers without evidence or verification. In cases where evidence appears to exist, it is often of the one-size-fits-all variety, commoditizing medicine with worthy aims of scalability, affordability, business success, social recognition, and so on. As individuals who cannot actually verify this, we are forced to trust systems and human judgments—whether or not they are trustworthy. Nothing is completely infallible. For example, a recent report notes that a few organ donors who had been declared dead prior to the transplant were discovered to still be alive.¹²

Commodity medicine provides



some statistical evidence (for example, via clinical trials for new drugs), without accounting for the specifics of an individual's cause of disease. Whereas statistical evidence is very useful in some cases, it does not generally address individually focused diseases where trustworthy evidence must, by definition, be tailored to the individual. Where individually tailored evidence exists in the lab (for example, genetically specific to the patient), it is often discarded for non-scalability and cost reasons.⁷

My publication from the 2024 HealthSec Workshop⁹ included an extensive enumeration of many past

risk-filled cases related to uses of computer-system technology in healthcare, most of which were drawn from my ACM Risks Forum (<https://www.risks.org>). It included numerous cases of defective equipment, human medical errors, flawed advice, and many other problems. The same result as noted by Gina Kolata has also caused significant historical advances in computer-system technology to be ignored, for many of the same reasons.⁷ A long-standing problem during my 72 years as a computer professional is that many compelling R&D innovations have not been adopted into the marketplace.

Healthcare Requires Trustworthy Technology

Some of the overarching concerns around the world involve the extent to which computer technology is: poorly designed and implemented; fundamentally inadequate with respect to predictable safety, reliability, security, privacy, usability and other aspects of trustworthiness; oblivious to research and development; misused; misunderstood; too expensive; or otherwise fraught with risks (see, for example, Hartmann⁶).

Many popular computer systems and networks have lacked even the most rudimentary evidence of trustworthiness, as that evidence depends on principles of development and analysis techniques that are beyond the abilities of many conventional programmers.¹⁰ Similar comments are valid relating to healthcare trustworthiness.

Research and Engineering for Trustworthy Healthcare

As noted at the beginning of this column, one of the most fundamental gaps is generic rather than specific to healthcare: the overwhelming absence of trustworthy commercial computer systems and networks. Thus, we need to begin building new trustworthy special-purpose devices for specific medical uses, and demonstrate in practice what might be achieved. Some inspiration may be found in constructive uses of demonstrably evidence-based research that uses modern system- and software-engineering approaches with seriously addressed principles.¹¹

The novel CHERI capability-based hardware-software is a noteworthy advance, being developed extensively since 2010 jointly with SRI International and the team in Cambridge, U.K. with funding from DARPA and more recently the British government and industry formally proved to satisfy their critical security properties.² CHERI is slowly becoming recognized by some corporations for its extensive trustworthiness.^a

a See the CHERI website for published papers, CHERI hardware specifications, operating systems including CHERI-BSD and real-time OSs, CHERI-RISC-V, and many public materials: <https://www.cl.cam.ac.uk/research/security/ctsrcd/cheri/>

We need to begin building new trustworthy special-purpose devices for specific medical uses.

A powerful approach pursued by Nancy Leveson relating to human safety has an underlying formal foundation that is also hidden from the developers so that it can be usable by people without advanced computer science degrees who have to deal with the complexity of today's systems such as healthcare technology and automated systems. Her *System-Theoretic Process Analysis* (STPA) handbook has been downloaded more than 500,000 times, and attendees at her workshops are from approximately 100 countries and every industry you might imagine.^b

Patient-Specific Technology Is Becoming Increasingly Desirable

There is often no universal solution in healthcare. For example, many medical practices seem to be stifled by the use of pharmaceutically based treatments that try to suppress the symptoms rather than treating the underlying causes. This standardized approach can in general impede the adoption of promising new methods that could be much more effective. Worse yet, proactive consideration of environmental hazards has also been stifled (by commercial agriculture, pharma, and other corporate bodies). Overall, it seems that prevention would have been much more effective than trying to rectify the ill effects of decades of harmful practices.

Trustworthy Healthcare Technology Is Emerging

The aforementioned gaps in our commercial information technology are

b Nancy Leveson has worked extensively on human safety for many years. Her recent books, research papers, and details on STPA are enumerated on her website: <http://sunnyday.mit.edu/>

reflected in the need for better healthcare-specific research and development. The “approved” standard allopathic treatments tend to treat only evident superficial symptoms with one-size-fits-all technology—typically pharmaceuticals and full-body chemotherapy that are often iatrogenic (that is, the alleged cure can be worse than the malady itself), typically lengthening the span of life while reducing the quality of life by wiping out the immune system.

On the other hand, functional medicine features significant improvements in cost structures and less dependence on pharmaceuticals, with individualized patient-based strategies that use carefully evaluated alternative, economical, and safe approaches to holistic healthcare that seek to address the root causes.

For example, it increasingly appears that a way to the future for treating certain cancers⁸ and other immune-system abnormalities is inspiring genetically engineered immunotherapy that is specifically tailored to the particular patient's tumor, disease, or hereditary defect, for example, in combination with hyperthermia. On the other hand, full-body gene therapy appears to be much further off in the future, if indeed it can be realistically practical. Nevertheless, full-body hyperthermia already has a considerable role in treating Lyme disease and its common co-infections.

Similar conclusions are relevant for dealing with immune-system degradations resulting from toxic air, water, ground residues, and toxic food, where blood, bones, organs, and fluids need to be analyzed and treated to reduce forever-chemicals and other toxic entities—although elimination from the environment would require massive changes.

My Recent Experience

I have a personal involvement in some of these issues—I was diagnosed with stage-three pancreatic cancer in August 2024, which my first-opinion California healthcare provider pronounced terminal in four to six months, with no hope for remediation. They insisted surgery would be far too dangerous (for example, the Whipple technique of removing half of the pancreas and re-

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

placing it with an animal organ part). In addition, full-body chemotherapy was considered to be non-survivable—for example, intolerable and fatal. I was told that I should go home and enjoy each remaining moment. However, a week of daily radiation apparently produced some slowdown in the progression of the tumor.

Fortunately, my daughter (a practitioner of Oriental medicine, but afflicted with 25 years of chronic tick-borne diseases) had previously been to the Sanoviv Clinic functional medicine clinic in Baja, CA, which she had visited for tick-borne co-infections and other toxicities a few years before. Sanoviv treats some forms of cancer and immune-system diseases, and she suggested we visit there. They gave me a second-opinion view of the same CT scan and suggested something positive was indeed feasible, especially in that it had been diagnosed early enough.

Sanoviv is helping advance the development of trustworthy technology—for example, hyperthermia devices and the use of hyperbaric oxygen chambers to increase oxygen absorption, and programs that analyze, alter, and detox people with certain cancers, and others infected with diseases carried by ticks, mosquitoes, parasites, birds, and other creatures.

One of the Sanoviv technologies includes the use of hyperthermia machines (with the goal of targeting the particular heated internal organ with local chemotherapy cancer coded to the genetic makeup of the tumor (with essentially zero after-effects), or full-body treatment for Lyme-related diseases). Similar research efforts are underway elsewhere. Such technologies hold enormous promise for the future.

To illustrate my situation regarding alternative healthcare more specifically, here is a brief outline of what happened, having chosen the Sanoviv alternative for a three-week visit in December 2024. I was then a 92-year-old male with one of the commonly inherited erroneous genes (cataloged as the so-called “ATM” error—each such common erroneous gene has its own unique alphabetic identifier). The ATM bad gene is explicitly linked with prostate, pancreatic, and breast cancer, which I presumably inherited from

my mother and/or father). Note that I had prostate cancer over a decade ago (spread to the left hip, treated successfully with radiation and testosterone-blocking hormone therapy, now in remission for well over a decade).

My first pancreatic-cancer treatment at Sanoviv included building up my immune system with repeated intravenous drips of artemisia, amygdalin, and up to 144 grams of vitamin C. A localized killer-chemotherapy package was prepared coded to the genetic make-up of the tumor, which when applied following a drip of T-cells resulted in fragmenting the overheated tumor in a single application of the coded chemo package. Systemic hydration with essential electrolytes was also part of the daily regimen. This initially resulted in a 30% reduction in the scope of the tumor in December 2024, and 50% by six months later after a take-home version of generic chemo for that type of cancer.

A second visit of two weeks to Sanoviv followed in June 2025. A slightly modified improvement in the immune-system build-up protocol was used (based on their recent research): the three-hour vitamin-C drip was replaced with a much more effective 15-minute macrophage activation-theory drip. Delivery of the coded killer package was again preceded by the repeated hyperthermia electromagnetic organ heating, and hyperbaric pressure chambers, followed by the immune-enhancing drips plus a T-cell drip, all of which further enhanced the process. The doctor's expectations (based on previous experience) were that about 90% of the tumor was already neutralized, with continued diminution again expected in the following months, as the immune system went to work again. This was monitored with CT scans, blood results, and the CA19-9 pancreas-marker test results, which indicated successful treatment thus far.

I am concluding writing this column during my third visit in December 2025, which indicated the CA19-9 marker had grown a little since I left Sanoviv in June 2025. That suggested that continuation of the hyperthermia and immunotherapy was in order, most likely again regularly in the future. However, the long-term prognosis re-

mains decidedly hugely positive when contrasted with the original death sentence before I reached out to Sanoviv.

Other cases were celebrated during both of my first sessions at Sanoviv. December 2024 involved a patient with an American clinic's diagnosis of stage-four colon cancer for a Jehovah's Witness patient who did not want any surgery. Visiting Sanoviv for the hyperthermia treatment noted above, that patient received 3.5 weeks of treatment, and left with zero signs of cancer. Several women with advanced breast cancer had similar experiences. Other total and partial eliminations were also observed. According to conversations with doctors and patients, the occurrence of many such successes is apparently not a surprise. The reputation of Sanoviv continues, well founded in its 15 years of research, and should be a genuine inspiration. It may not always live up to patient expectations, but it seems to succeed quite often in the right direction. It also provides an incentive for ongoing and new research, there and elsewhere.

Fundamental Research in Progress

Some examples of just few healthcare-specific technology research topics are summarized here:

- Understanding and controlling chronic immune-system and other diseases: COVID and Long-haul COVID, Dysautonomia (connective tissue disorder), and a wide range of nervous-system disorders (including Mast Cell Activation Syndrome MCAS, Postural Orthostatic Tachycardia Syndrome (POTS), Ehlers-Danlos Syndrome EDS), Lyme disease and its nontrivial tricky-to-diagnose co-infections—such as *Borrelia Bergdorferi* (with five variants including *Borrelia STARI* from the lone-star tick), Ehrlichiosis, Anaplasmosis, *Francisiella*, Rickettsia, *Bartonella*, Babesiosis, Bartonellosis, Powassan virus, Heartland Virus, Colorado Tick Fever, Bourbon virus, Sevan fever (Tick-Borne Encephalitis, which affects people in and travelers to Europe and Asia), Thrombocytopenia, Tularemia, Alpha-Galactose (“Alpha-Gal,” from the Lone-Star tick, which may cause death from immune reactions such as anaphylaxis—see the alleged first recorded U.S. death of an airport pilot in N.J. who died

The absence of meaningfully trustworthy computer communication is likely to remain an obstacle for the foreseeable future.

upon eating meat (although many other deaths are now suspected in retrospect),⁴ Rocky Mountain Spotted Tick Fever, and Epstein-Barr (a form of chronic mononucleosis), to name a few—plus other similar items such as Lupus, Sickle-Cell Anemia, HIV-AIDS, diabetes, eating disorders, and obesity. Recent attention has noted that mosquito-borne chikungunya,⁵ well known in South America, Africa, Asia, and Puerto Rico, was recently discovered on Long Island, NY; it causes disabling joint pain. This list gives only a hint at the complexity of these co-infections, an area in which targeted research is becoming increasingly critical because the relatively rare successful known treatments may be quite different. For example, *The Quiet Epidemic: As Close as Your Backyard* video by Kino Lorber, 2022, provides a view of the medical cover-up of the approved Lyme-disease treatment (a wrong-headed ubiquitous standard single-dose antibiotic, rather than some kind of 28-day cycle to account for multiple stages of the infecting agent.) Unfortunately, there are no adequate sufficiently accurate references on these infections, with relatively little ongoing work in progress.

- Accurate complete testing tends to be very spotty in the U.S., although excellent for example, in Germany. Worse yet, the treatment for each of the aforementioned ailments is necessarily confused by the absence of definitive research and testing, and the lack of individualized treatments specific for many of the aforementioned maladies. Also, long-term delays in treatment

tend to result in chronic degeneration that cross the blood-brain barrier and become much more difficult to rectify. Note: Very recently, the American Medical Association has finally identified Chronic (long-term) Lyme as a disease, after many years claiming that the term had no meaning—with many doctors routinely suggesting that patients should consult psychiatrists.

- Sanoviv and other clinics around the world (such as Paracelsus in Switzerland, St. Georg in Germany, and various places in Canada) evidently have a much more robust understanding of the tick-, insect- and parasite-borne infections than allopathic doctors.

- Immune-system research is still in its infancy, but making gigantic progress. With climate change among other factors increasing their life cycle and geographical spread, ticks are now found on every continent, including in Antarctic penguins), and the generic category of immune-system diseases seems to be a catch-all for as-yet unknown causes. More-specific diagnosis is going to be critical, which is already a huge problem in trying to identify and distinguish among the Lyme-related diseases and some of the lesser understood cases noted here.

- Sanoviv also has an ongoing research program to study the effects of micro-electric currents on the brain, heart, and other organs. This effort appears to be unique, although they are sharing their results with other research institutions worldwide. Some of this will in the future explore the application of constructive stimuli.

Other issues also need to be considered, in much greater depth:

- Exploring complementary protocols and medicine for treatment: Various types of scanning/imaging (X-ray, CT, PET), genetically targeted chemotherapy, Extra-corporeal blood oxygenation and ozonation (EBOO, involving treatment of blood with ozone and oxygen), plasmapheresis and the Patricia King Protocol, other detox protocols (for elimination of environmental toxins, heavy metals, forever chemicals, microplastics, and food allergies), T-Cell therapy, ethylenediaminetetraacetic acid (EDTA), and dimercaptosuccinic acid (DMSA). Some of these require extensive technology to support their effectively controlled use with real-

time monitoring, analysis, and general supervision. Once again, functional medicine appears to have a major role.

► Exploring crosslinks among diseases, treatments, and medications: Patients are increasingly having seemingly unrelated symptoms with multiple medications that may have surprising or unknown effects, actually complicating recovery. Significant research is needed to explore these interrelationships.

Lessons for Computer and Healthcare Systems

The absence of meaningfully trustworthy computer communication is likely to remain an obstacle for the foreseeable future, considering the frequent occurrences of system crashes, power outages, hacking, ransomware, and other risks. The short-term challenges will be in trying to develop application software that can overcome the inherent underlying hardware/software system risks. Please remember that total-system trustworthiness is an overarching emergent property of the entire system; it depends on the hardware, operating system, application code, and the successful composability of all of these components, failure in which may result in system errors, self-denials of service, and external exploits.

Evidence of trustworthiness in computing—and particularly the use of artificial intelligence in medical applications—should begin to emerge for individual software systems; for example, selective high-assurance applied to specific software modules, applications, and eventually perhaps total systems. We already have some evidence that the personalized healthcare lessons in trustworthiness resonate in computing applications from Amazon.^{1,3} The potential of patient-based targeted immunotherapy seems to be theoretically unlimited, with even some longer-range experiments in genetic engineering. That is why I lump cancer care and animal-borne immunotherapy together here, even if the treatments might seem conceptually different. Not surprisingly, various countries and some U.S. non-government start-ups are betting that immunotherapy solutions can revolutionize healthcare. Such far-reaching approaches are desper-

Healthcare worldwide needs to address technology that is more demonstrably trustworthy, integrative, and tailored to the individual genetic makeup.

ately needed. For example, it was announced in mid-October 2025 that N4 Pharma and SRI International have demonstrated a breakthrough in RNA cancer therapy.^c Also, the 2025 Nobel Prize for Medicine for peripheral immune system tolerance in immunology awarded to Mary E. Bruncow, Fred Ramsdel, and Shimon Sakaguchi should help the cause considerably in advancing applications of immunology for holistic healthcare.

In hindsight, it appears that some healthcare approaches have grown from a rather English/old-German biomedical tradition, whereas some others such as Mexican, new German, Swiss, have deeper Spanish/French and thousands-of-year-old Chinese, Indian, and even Egyptian roots that are closer to what is now deemed functional healthcare. The differences in approach are stark, but both can certainly work together. Thus, healthcare worldwide needs to address technology that is more demonstrably trustworthy, integrative, and tailored to the individual genetic makeup (especially including overcoming already cataloged widely known erroneous genes). Above all, good healthcare represents a holistic international problem that needs holistic approaches. Please remember that no one size fits all here, and more individualized care is going to be absolutely necessary.

c See <https://www.sri.com/press/story/n4-pharma-and-sri-demonstrate-breakthrough-in-rna-cancer-therapy/>

So, where might all of this be heading sometime in the future: With the eventual availability of trustworthy computer systems, trustworthy evidence-based AI, trustworthy gene sequencing, and trustworthy pharma (especially where open sourced), you might even be able to do much of what is alluded to here by yourself (with some trustworthy guidance)—or at least have a considerably better understanding of what might be possible if and when you might wish to live longer with a good quality of life. People on both sides of the fence might call it primitive healthcare, going back several millennia in principles and certain approaches that can still be effective in practice. **C**

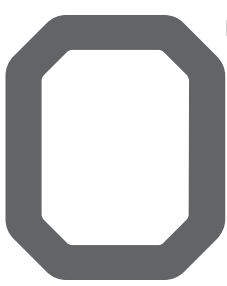
References

1. Backe, J. et al. One-click formal methods. *IEEE Software* 36, 6 (2019); 10.1109/MS.2019.2930609
2. Bauereiss, T. et al. Verified security for the Morello capability-enhanced prototype Arm architecture. In *Proceedings of the 31st European Symp. on Programming (ESOP 2022)* (May 2022).
3. Cook, B. Formal reasoning about the security of Amazon web services. In H. Chockler and G. Weissenbacher, Eds. *Lecture Notes in Computer Science (LNCS) 10981*. Springer, (2018); 10.1007/978-3-319-96145-3_3.
4. Goldstein, J. In a tick bite, a fatal puzzle seems to be solved. *The New York Times* (Nov. 20, 2025); <https://www.nytimes.com/2025/11/20/nyregion/jetblue-pilot-death-tick-alpha-gal.html>
5. Goldstein, J. Mosquito-borne illness rarely seen in the U.S. is suspected on Long Island. *The New York Times* (Sept. 26, 2025); <https://www.nytimes.com/2025/09/25/nyregion/chikungunya-mosquito-illness-long-island.html>
6. Hartmann, T. *The Hidden History of American Healthcare: Why Sickness Bankrupts You and Makes Others Insanely Rich*. Koehler Publishers, Inc., (2021).
7. Kolata, G. A chance for a cure goes away. *The New York Times* (Sept. 23, 2025).
8. Mukherjee, S. *The Emperor of All Maladies: A Biography of Cancer*. Scribner, (2010).
9. Neumann, P.G. Computer-related healthcare risks. In *Proceedings of the ACM Workshop on Cybersecurity in Healthcare (HealthSec)* (2024); <https://publish.illinois.edu/healthsec/>
10. Neumann, P.G. and Lindqvist, U. The future of misuse detection: From lessons learned to new directions. *Commun. ACM* 67, 11 (Nov. 2024); <https://www.csl.sri.com/users/neumann/cacm255.pdf>
11. Neumann, P.G. Fundamental Trustworthiness Principles in CHERI, H.A. Shrobe, D. Shrier, and A. Pentland, Eds. *New Solutions for Cybersecurity*. MIT Press/Connection Science, Cambridge MA (2018).
12. Rosenthal, B.M. and Tate, J. A push for more organ transplants is putting donors at risk. *The New York Times* (Sept. 19, 2025); <https://www.nytimes.com/2025/07/20/us/organ-transplants-donors-alive.html>

Peter Neumann (peter.neumann@sri.com) is Chief Scientist in SRI International's Computer Science Lab and moderator of the ACM Risks Forum. He is especially grateful to Virgil Gligor for incisive comments, as well as to other members of the ACM RISKS review panel, his daughter Helen K. Neumann, and several doctors, all of whom helped improve this column.

This *Communications Inside Risks* column is derived from a longer *Inside Risks* item, available at <https://www.csl.sri.com/users/neumann/cacm225-full>

© 2026 Copyright held by the owner/author(s).



Kode Vicious Driven to Distraction

From floats to characters and back again.

Dear KV,

I know you have mentioned working on embedded systems in the past, and now that I have a related project, I have a question for you. The device we are building has a lot of sensors and I have been slowly adding support for these to our embedded Linux kernel. Each device samples data from the environment, and they are each a bit different, but the only common drivers that seem to match what we do are character drivers that just generate a stream of bytes. It seems strange to take a lot of floating-point data, turn it into characters, and then turn it back into floating point. Is this common? Or am I missing some type of driver that would do this better?

Floating

Dear Floating

I remember when I was preparing for my first talk years ago, people would tell me, “Start with a joke.” So, here is one for this occasion:

Three hardware designers go to see an OS developer.

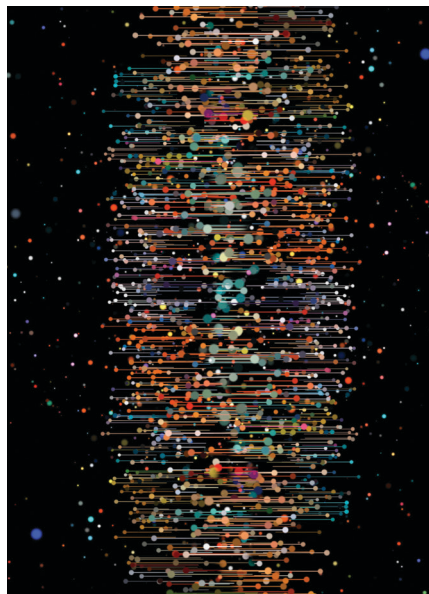
The first hardware designer says, “I just developed a new device and I need a driver.”

OS developer: “Sure, what does it do?”

First hardware designer: “It reads GPS data and correlates it with star charts from a visible telescope.”

OS developer: “You should write a character driver.”

The second hardware designer says, “My new device takes temperature readings to five significant digits from



across a square kilometer of ocean at depths of 1 to 100 meters.”

OS developer: “Oh, so you need a character driver too.”

Once upon a time, operating systems handled many different types of data, and they did so in sometimes-wildly incompatible ways.

The third hardware designer says, “My device samples quantum states throughout the known universe. It can literally see the face of god, should such exist.”

OS developer: “Character driver!”

Alas, some jokes require a very specialized audience, but in this case I think the audience is receptive.

The thing is, once upon a time, operating systems handled many different types of data, and they did so in sometimes-wildly incompatible ways. This was a problem because it meant applications had to be ported whenever a new operating system version came out or whenever someone got a new model computer.

The “everything is a byte stream, we don’t care what the format of your data is” ethos of early Unix had some very real benefits in that it allowed programs to be ported to different machines and different versions of Unix to be used with few or no changes.

But, of course, every good fairy tale also has a dark side—at least if you read the original Grimm. And here, the problem with a simplifying assumption is that it can make for inefficient or, in your case, patently illogical software. Converting a number (which is what a computer is great at processing) into a string (which a computer is not ideally suited for) and then back to a number wastes time, memory, and CPU cycles—the very three things we are all told never to waste in our software!

KV’s particularly favorite form of this stupidity: one of the best-known

STUDENT
RESEARCH
COMPETITIONAssociation 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>

and most commonly used protocols for sending stock-ticker information does exactly this—that is, it converts numbers to strings and then back to numbers. This ridiculous scenario led finance companies to invest heavily in FPGAs (field-programmable gate arrays) just to get an edge in processing stock prices for high-frequency trading. KV was both appalled when he learned this and chagrined that he hadn't bought stock in certain FPGA companies soon enough.

All of which is to say that sometimes, simplifying assumptions are a real problem, and sometimes, they make everything look like a nail, which then makes you think all you need is a hammer. The big challenge with modern systems is that 50 years of doing things the Unix way has left us bereft of better APIs. It is not just the drivers but also the application APIs on top of the operating system that deal only in byte streams. It is as if the operating system designers threw up their hands and said, “Not my job!” and left all the data interpretation to the application programmers and device developers. Since these two parties rarely, if ever, talk to each other, no real progress has been made in this area from that time until now.

So, how might we make progress? The current fad for kernel bypass libraries addresses the matter by just ignoring the operating system altogether—which, as an operating system developer myself, I can totally understand. We must either deal with the operating system through modification or extension, bypass it, or start over from scratch by writing a new one. This last option is the most attractive to an operating system developer, but KV doubts your employer will sign up for the associated multiyear project. But if they do, please get in touch with me!

Your most expedient but least preferable solution would be to bypass the operating system since you can write a few APIs that allow your applications to talk efficiently with the devices. Bypassing gives you the most control, the highest fidelity of data, the lowest latency of access—but all at the price of generality. That is, when you bypass the operating system, you make it so that each new program must, again, be compatible with your bypass libraries. And if you want generality back, you

The big challenge with modern systems is that 50 years of doing things the Unix way has left us bereft of better APIs.

wind up doing what operating system developers have done forever, which is to produce generalized APIs that can be consumed by arbitrary programs. The current middle ground is to add new system calls to the operating system—a process that is fraught with peril, because modern operating systems are huge and fragile. While people do add new system calls to operating systems, this is a rare event. And, if you were to do this, you would wind up maintaining the new syscalls locally for a very long time unless you were able to convince an open source project that your new calls were generally useful.

So, that is where you are: Make a character device driver and pay the cost of converting real data into strings, bypass the operating system, or innovate in the operating system space and produce some new system calls. I cannot say any of these choices are great. In fact, they all have their downsides in one way or another. But innovation in operating systems went out in the 1980s, and the people who are trying to bring that back now have a more difficult job.

KV

Related articles

In Praise of the Disassembler

<https://queue.acm.org/detail.cfm?id=3469121>

The Unholy Trinity of Software Development

<https://queue.acm.org/detail.cfm?id=3014148>

Kode Vicious Gets Dirty

<https://queue.acm.org/detail.cfm?id=1071723>

George V. Neville-Neil (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.

© 2026 Copyright held by the owner/author(s).

Opinion

Resilience Meets Security in Cyber-Physical Systems and Critical Infrastructures

Why we need a holistic perspective on systems designed to withstand and recover from disruptions while maintaining essential functionalities.

ON OCTOBER 8, 2022, Deutsche Bahn—the German railway—experienced a critical outage of its GSM-R system in northern Germany. Designed to enable communication between signal boxes and engine drivers, GSM-R is a critical safety component. In the event of a disruption in communication, no trains are permitted to operate. Deutsche Bahn segmented its GSM-R cellular network into several regional networks. However, all of them require access to central services to manage GSM-R calls, which are—for reasons of redundancy—available both in Berlin and Frankfurt, connected through fiber to all regional networks. On the day of the outage, fiber cables from the regional network of Hannover to Berlin and Frankfurt were cut at two distant locations within a few hours. After the incident, no GSM-R calls could be established and Deutsche Bahn was forced to shut down operations in northern Germany for more than three hours. This led to cancellation of numerous train services throughout the country. Initially, it was assumed that the damage was the result of critical infrastructure sabotage. However, subsequent investigations revealed that the damage was caused by two uncoordinated vandalism incidents.^a



In February 2022, in parallel to the Russian invasion of Ukraine, unknown (but likely state-linked) attackers targeted the ViaSat satellite communication systems, presumably in an attempt to suppress military communication. The attack caused significant collateral damage to critical infrastructures. It disrupted the remote control of thousands of wind turbines in Germany, some of which remained inoperable for up to a month after the incident.^b

On March 21, 2025, Heathrow airport was forced to shut down for a day because of a fire in a substation. Even though the airport is powered by three different substations, the failure of a single one resulted in a loss of power. Approximately 1,200 flights needed to be cancelled and 120 diverted to different airports in the U.K.^c

Critical infrastructures and cyber-physical systems (CPS) are integral to our daily life and are becoming increasingly important. Disruptions

^a See <https://bit.ly/3KPM785>

^b See <https://bit.ly/4q4bBO1>

^c See <https://bit.ly/3Yl69ud>

ACM Transactions on Accessible Computing

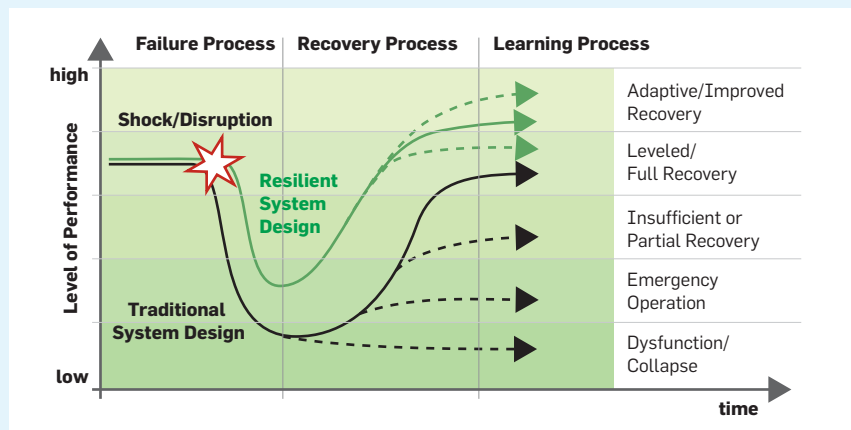


ACM TACCESS is a quarterly journal that publishes refereed articles addressing issues of computing as it impacts the lives of people with disabilities. The journal will be of particular interest to SIGACCESS members and delegates to its affiliated conference (i.e., ASSETS), as well as other international accessibility conferences.



For further information
or to submit your
manuscript,
visit taccess.acm.org

Figure 1. General illustration of the level of performance of a system after shock/disruption showing failure, recovery, and learning process. The black curves visualize a traditional (non-resilient) system design. The green curves show a resilient system design.



have serious consequences and can even cause physical harm. Thus, CPS must be designed in a way that ensures dependable operation even in the most challenging circumstances. They have to operate not only in a secure but also in a resilient fashion. NIST SP 800-39 defines resilience in this context to be: “The ability of an information system to continue to: (i) operate under adverse conditions or stress, even if in a degraded or debilitated state, while maintaining essential operational capabilities; and (ii) recover to an effective operational posture in a time frame consistent with mission needs.”^d

Resilience and Its Relation to Security and Safety

In contrast to secure systems (which are commonly seen as either in a secure or insecure state), resilient systems have the capacity for graceful degradation and recovery while maintaining essential operational capabilities. This is often illustrated by the resilience cycle shown in Figure 1. For example, if a CPS such as the GSM-R network is damaged (shock/disruption in Figure 1), its functionality will temporarily decline, potentially to a significant extent. However, under a resilient system design, this would not require shutting down the railway network. Integrated measures could facilitate the automatic recovery of the system, at least to a certain extent, so that trains

can continue running, maybe at reduced track capacity and potentially at the cost of utilizing alternative, more error-prone and labor-intensive communication methods. It is even possible that the system will recover to a higher performance level if the system is reconfigured and learns to deal with similar events in the future.

Attacks on CPS have the potential to result in a range of consequences, including those related to safety, operations, finances, and privacy. Safety risks arise when compromised systems, especially in sectors such as healthcare, transportation, or industry, endanger lives.

From an operational standpoint, cyber attacks can disrupt processes, degrade functionality, and cause costly periods of downtime. From a financial perspective, cyber attacks can result in direct costs for response efforts or the resolution of liability cases, as

Attacks on CPS have the potential to result in a range of consequences, including safety, operations, finances, and privacy.

^d See *Managing Information Security Risk: Organization, Mission, and Information System View* (Mar. 2011); <https://bit.ly/4a5GAEE>

well as indirect costs, such as reputational damage. In the event of a privacy breach, sensitive information may be exposed, thereby violating regulations such as the General Data Protection Regulation (GDPR) and incurring penalties. We argue that safety, resilience, and security must be addressed as interconnected priorities rather than separate concerns. Without strong cyber resilience, organizations face risks that threaten their stability, trustworthiness, and long-term success. This applies not only to large companies and infrastructures, such as smart cities, but also to start-ups in sectors such as smart homes, and even to unique scenarios like smart farming.

In recent years, many Information and Communication Technology (ICT) systems started to aim for security-by-design. However, today's focus in securing systems is to prevent attacks such as eavesdropping, unauthorized access, or data tampering through mechanisms like encryption, firewalls, and intrusion detection systems. Yet, these same security mechanisms can inadvertently compromise the system's ability to maintain functionality and recover from disruptions. For example, in an energy grid, encryption ensures confidential communication between grid control systems and devices, whereas firewalls block unauthorized access. However, if these security measures are overly rigid or fail during specific conditions (for example, if a misconfigured firewall blocks legitimate traffic or a key management failure renders encrypted data inaccessible), they can cause critical systems to stop functioning. Even though the energy grid is then "secure" from cyber attacks, it may become inoperable, which could lead to a widespread power outage. In this context, resilience would require balancing security with robustness, such as implementing fallback systems, redundancy, and mechanisms that ensure functionality (albeit in a limited capacity) even if security protocols fail. This highlights that a secure system is not automatically resilient, as the two concepts address different aspects of system reliability.

In a similar way, a safe system may not necessarily be resilient. The example of the German rail system GSM-R outage at the beginning of this Opin-

ion column demonstrates this: despite its redundant design, the system failed.

Ways Toward Jointly Resilient and Secure CPS

Achieving resilience for CPS requires a multifaceted technical approach that prepares systems not only to withstand but also to adapt and recover from cyberattacks. In general, classical security mechanisms like multi-factor authentication, access control, zero-trust principles, and defense-in-depth are essential components, yet they are insufficient when deployed on their own to ensure resilience. Redundancy and adaptability are crucial: CPS should be designed with built-in redundancy to maintain operational continuity, potentially at a reduced functionality, in the event of a breach or failure. For example, detailed recovery plans can outline steps to revert to a safe operational state or switch to backup components when primary systems are compromised. Modular system design in combination with the ability for decentralized operation can help to further increase adaptability. Redundancy should consider diversity such that vulnerabilities cannot affect all redundant instances at a time.

However, redundancy can also increase the attack surface, as can the deployment of security software. Envision a smart energy system using two distinct communication networks, one primary and one backup, to connect substations and control centers. If these networks share software vulnerabilities, an attacker can potentially compromise both communication paths using the same exploit. Diversifying technologies for communications or adding security software (sometimes required for compliance reasons) can increase the risk as well: more software means more potential vulnerabilities. Lastly, managing and securing multiple redundant systems introduces additional complexity, which could lead to misconfigurations and other vulnerabilities; the disastrous CrowdStrike incident being the prime example, how a poorly tested software update affected millions of servers and all sectors of critical infrastructure worldwide.^e

^e See <https://bit.ly/3YOiRBP>

CPS-specific resilience demands new testing and validation methodologies.

Compartmentalization allows to isolate system components from one another, thereby limiting the impact of breaches and preventing threats from cascading across interdependent systems. Fault detection and identification is needed to quickly discover threats and problems and automatically handle them.

CPS-specific resilience demands new testing and validation methodologies. At the moment, automated testing methodologies are significantly less mature in this domain than those in traditional IT. Nevertheless, rigorous testing is essential to identify vulnerabilities specific to the physical aspects of CPS operations. In addition, user interfaces that communicate security status and offer easy-to-understand alerts and controls are crucial for assisting operators in identifying and managing potential threats in real time, thereby improving both response and adaptability in the face of shocks and disruptions. Together, these strategies foster a CPS environment that is secure, adaptable, and resilient in the face of evolving cyber threats. However, integrated resilience also requires clear prioritization of functionality over all other aspects and trade-offs, which is a challenging aspect to design and agree upon.

Cyber Resilience Strategies, Building Capacity, and Raising Awareness for Resilience in CPS

The increasing importance of cyber resilience is becoming evident in national and supranational cybersecurity/cyber resilience strategies. The 2021 report *Singapore Cybersecurity Strategy*² was among the first to offer a clear definition of cyber resilience, and notably incorporates considerations for cyber-physical systems (CPS). Similarly,

the 2024 U.S. report *Strategy for Cyber-Physical Resilience: Fortifying Our Critical Infrastructure for a Digital World*¹ emphasizes the cyber-physical dimension of resilience, focusing on actionable recommendations to strengthen critical infrastructure within the U.S. The *NIS2 Directive*³ represents the E.U.'s updated cybersecurity framework, imposing stricter requirements on essential and important entities—including operators of cyber-physical systems in sectors like energy, transport, and healthcare—to strengthen both digital and operational resilience. The goal of NIS2 is to ensure effective prevention, management, and recovery from disruptions within these interconnected systems. In contrast, the latest version of the European Union's 2024 *Cyber Resilience Act (CRA)*⁴ lacks a clear definition of cyber resilience, primarily focusing on cybersecurity requirements for products with digital elements. While the digital leaders have recognized that cybersecurity and cyber resilience are tightly intertwined, many other countries currently lack strategic-level discussions regarding cyber resilience.

In addition to technical approaches and solutions needed to achieve and ensure resilience in modern CPS systems, it is equally important to raise general awareness of the issue. This heightened awareness ensures technical solutions are not only developed but also effectively utilized and implemented, preventing them from re-

Figure 2. Resilience can only be achieved if technical innovation works hand-in-hand with curriculum integration, standardization/regulation, and lifelong learning. The individual steps are overlapping and continuous.



Lifelong learning ensures innovation is quickly integrated into existing systems and companies.

maintaining mere technical publications in scientific venues.

There are several ways to enhance developers' awareness, for example, integrating relevant topics into educational curricula and promoting lifelong learning, developing and enforcing robust backup and recovery plans, and implementing regulation and standardization, such as presented in Figure 2. Currently, college and university programs are rather rigid, with new topics only slowly finding their way to all students. Furthermore, while curriculum recommendations in electrical engineering and computer science often include security, resilience is not directly addressed. In terms of standardization and regulations, software and integrated information systems, such as CPS, lag behind their hardware counterparts, which have long-established safety standards (for example, the CE marking on electrical devices). Fortunately, existing legislation is undergoing changes to include software liability—a much-needed step toward greater resilience (the new E.U. Product Liability Directive⁶).

These four dimensions must be addressed together to achieve meaningful progress. For instance, even if developers recognize the need to create recovery plans for their products or services, a lack of knowledge about how to design and implement them effectively could lead to failure. Similarly, if recovery plans are not mandatory or if their quality is not governed by standards, companies may lack sufficient motivation to invest in them, particularly under financial or time constraints.

Good examples from other sectors illustrate the potential impact of such

efforts. For example, training and lifelong learning programs for pilots or operators of other high-risk machinery have been shown to effectively reduce risks. Similar initiatives must be undertaken in the CPS sector, where systems are rapidly evolving into similarly complex and high-risk environments, such as autonomous vehicles.

Outlook

To achieve CPS resilience, it is essential that researchers and practitioners from different communities engage in collaborative discourse, not only concerning innovative solutions but also practical trade-offs and priorities. While technical innovation is still needed in many fields, it also needs to be integrated into teaching curricula and into new standards and regulations. Lifelong learning ensures innovation is quickly integrated into existing systems and companies.

By adopting a holistic view that integrates resilience and security, we can better protect CPS and critical infrastructures against evolving threats, ensuring their reliability and stability in the face of both anticipated and unforeseen challenges. Promising initiatives are already under way but must be supported, expanded, and embraced at all levels and by all stakeholders. ■

References

1. Arnold, F.H. et al. *PCAST: Report to the President on Strategy for Cyber-Physical Resilience: Fortifying Our Critical Infrastructure for a Digital World*. Technical Report. White House Office of Science and Technology Policy (OSTP) (2024); <https://bit.ly/4aHSwss>
2. Cyber Security Agency of Singapore. *The Singapore Cybersecurity Strategy* (2021); <https://bit.ly/3MQsuxk>
3. European Union Directive. *Network and Information Security Directive 2 (NIS2): Directive (EU) 2022/2555 of the European Parliament and of the Council* (2022); <https://bit.ly/4aErqXa>
4. European Union Regulation. *Cyber Resilience Act (CRA): Regulation (EU) 2024/2847 of the European Parliament and of the Council* (2024); <https://bit.ly/3KEFD0y>

Anna Förster (anna.foerster@uni-bremen.de) is a professor at the University of Bremen, Germany.

Matthias Hollick (mhollick@seemoo.tu-darmstadt.de) is a professor of computer science at the Technical University of Darmstadt, Hessen, Germany.

Stefan Katzenbeisser (Stefan.Katzenbeisser@Uni-Passau.De) is a professor of computer engineering at the University of Passau, Germany.

Christoph Krauss (christoph.krauss@h-da.de) is a professor at the University of Applied Sciences Darmstadt, Germany.

This work was conducted under the umbrella of the DFG Priority Program Resilient Worlds: <https://www.resilient-worlds.org/>

f See <https://bit.ly/493M8gF>

Opinion

The Agentic Economy

The architecture of agentic communication will determine the extent to which generative AI democratizes access to economic opportunity.

GENERATIVE AI HAS revolutionized the way we interact with technology, allowing people to express their intent in free-form natural language. It has paved the way for AI agents that not only converse with users but also perform actions on their behalf, flexibly and with minimal guidance. Delegation to AI has already begun to improve the efficiency of individual processes, making both consumers and businesses more productive in the set of tasks they had already been doing.^{4,5} However, we believe that the more disruptive—and yet to be realized—impact of generative AI is its potential to drastically reduce the communication frictions between and among consumers and businesses. This could lead to a reorganization of markets, shifts of market power, and the introduction of entirely new products and services.

Consumers have traditionally faced high communication costs when initiating relationships with businesses, reducing efficiency.⁸ For example, a consumer seeking a new tax preparer might hesitate to switch because she would have to explain her financial situation all over again to a new person or online service. These communication hurdles can prevent consumers from taking advantage of better products and services or lower prices. Businesses have tried to lower these costs with tools like online forms and voicemail menus, but these often just shift communication costs to the consumer and can make interactions more rigid.

Imagine instead a future where every consumer has an assistant agent



to communicate their preferences and personal information to businesses, and every business has service agents to interact with consumers and other businesses. These agents could be designed to interface with each other seamlessly and flexibly, transforming the landscape of consumer-business interactions. Delegating interactions to such assistant and service agents lowers communication costs and makes markets more efficient by expanding the range of options available to both consumers and businesses.

To unlock the full economic potential of generative AI's communication capabilities, two developments are necessary. First, consumers and businesses must widely adopt assistant and service agents. This is already under way.^{2,7} Second, these agents must be designed

to interact seamlessly with each other to facilitate transactions. On the technical front, there has been significant progress in standardizing such agentic interaction, with frameworks such as Microsoft's AutoGen,^a and protocols such as Anthropic's Model Context Protocol^b and Google's Agent2Agent Protocol.^c However, it remains to be seen how these advances will be adopted and implemented, or constrained, given their complex interplay with and dependence on market forces.

We believe the largest benefits of inter-agent communication will be realized as markets reorganize around these new capabilities.³ While the pos-

^a See <https://github.com/microsoft/autogen>

^b See <https://github.com/a2aproject/A2A>

^c See <https://github.com/modelcontextprotocol>

sibilities of this technological distribution are numerous and difficult to predict, we offer a framework for understanding the most plausible outcomes and contrasting them with the current state of affairs. A key question is whether inter-agent communication will occur within closed “agentic walled gardens” controlled by a few dominant providers, akin to today’s app stores, or through a more open “web of agents” where agents freely connect and transact, much like the World Wide Web. The answer to this question will determine how today’s largest online platforms are impacted by the proliferation of interconnected agents—whether through further entrenchment of their market power or a loss of dominance and broader democratization of AI’s economic benefits. We explore how those benefits might manifest, with implications for advertising, e-commerce, and the creation of new products and industries.

The Current State of Agentic AI

Before discussing future possibilities for agentic economies, it is helpful to survey the current landscape of AI agents. On the surface it may appear as if several existing efforts are well on their way to providing consumers and businesses with assistant and service agents that could function as described here. But most existing agents lack a key ingredient: While they are designed to interact with or simulate human users, few public offerings are *designed to interact with each other*.

Existing agents generally come in one of two forms: siloed service agents or general-purpose end-to-end agents. The first, siloed service agents, provide a new user interface for products and services within a single company. For example, Amazon’s Rufus allows customers visiting Amazon to access their order histories or compare potential purchases through natural language instead of navigating a website. Likewise, Expedia’s Romie provides a chat interface to help customers build travel itineraries—including flights, hotels, and restaurants—by pulling information from customer email messages and group chats. Importantly, however, these efforts do not expose interfaces intended for interaction with other agents. As a result, they are still fundamentally lim-

A key question is whether inter-agent communication will occur within closed “agentic walled gardens” or a more open “web of agents.”

ited by people having to navigate to and personally interact with them.

End-to-end agents take a different approach, aiming to provide general-purpose automated functionality that is not limited to a single company or service. For instance, technology from OpenAI, Google, and Microsoft can aggregate research from external sources, navigate business websites on the user’s behalf, and even perform simple tasks like making reservations or placing food orders, all within one interface. Importantly, however, much of the functionality that these end-to-end agents provide currently comes through “computer use models” that simulate a user pointing and clicking on existing (non-agentic) websites (or most recently the assistant agent may absorb a menu of products and pricing, or even an app, but still all within the confines of the assistant agent). This gives the illusion of assistant and service agents working together, but the absence of a true service agent on the business side limits what an end-to-end agent can accomplish.

Moreover, by mimicking human users, these agents risk creating adversarial relationships with businesses that, for example, rely heavily on advertising revenue and may resist having their websites accessed by agents instead of humans. Even if end-to-end agents can perfectly simulate human users and businesses agree to agent-based access, interactions would still be constrained by what businesses currently expose through existing web forms that limit the expressiveness of requests and responses. For instance, businesses offering highly customizable products—such as catering services—often use only basic contact forms, handling

niche or specialized requests through human followup.

The Future Impact of Agentic AI

Given their limitations, we anticipate that siloed and end-to-end agents will ultimately give way to agents that are designed to seamlessly interact directly with each other. However, there is a complex interplay between technical capabilities and market forces that could lead to a number of different scenarios for what this solution will look like, who will control it, and what it will be capable of.

The Market Power of Digital Intermediaries. Two-sided platforms (businesses designed to bring together two distinct sides for a transaction) such as Amazon, Expedia, OpenTable, and Spotify are key intermediaries of the current digital economy that create value by matching millions of consumers and businesses to each other within specific domains such as shopping, travel, dining, and music.⁹ They do this in part by standardizing how both sides interact. For example, Amazon requires sellers to follow specific formats and policies, while consumers must use its interface to search and transact. In exchange, consumers access a vast range of sellers and businesses gain exposure to a large customer base. But this comes with trade-offs: both operate within a tightly constrained system, are subject to the platform’s design choices (for example, ranking algorithms), and pay referral fees.

If an agentic economy enables each consumer’s assistant agent to directly and flexibly communicate with each businesses’ service agent via *unscripted communication*, the role—and market power—of intermediary platforms could shift substantially. In principle, once communication frictions are low enough, interoperable AI agents could eliminate the necessity for two-sided platforms as intermediaries altogether. Consumer assistant agents could directly find and flexibly negotiate with service agents to buy goods, book hotels and airlines, make dining reservations, and stream music.⁶ This would represent a drastic decentralization of power compared to today’s markets.

In practice, however, intermediary platforms often provide value beyond simply standardizing communica-

tion between buyers and sellers, via discovery, validation, remediation, and economies of scale. For instance, in domains like travel, it may still be useful for an intermediary to provide trusted suggestions, offer trip insurance, resolve disputes, or ensure regulatory compliance. However, an agentic economy could lead to fierce competition between intermediaries due to low switching costs, reducing the profits they can extract.

Agentic walled gardens vs. the web of agents. Even though any specific assistant and service agent may be technically capable of communicating with each other in an unscripted manner, the pool of agents they might interact with could be *restricted* due to market forces. Select firms may provide assistant agents for free but restrict communications, creating “agentic walled gardens.” In some sense this would be a natural evolution of today’s existing application ecosystems such as the app stores in dominant operating systems. Given their existing large user bases and nascent assistant technologies like Apple Intelligence, Google Assistant, Microsoft Copilot, and Meta AI, these firms are well positioned to extend their current marketplaces to include interoperable AI agents. For example, in March 2025, Meta launched basic service agents for business pages on Facebook and Instagram at no cost, but these service agents are only accessible to users on their own platforms. In addition, firms such as OpenAI and Anthropic that have built out large user bases for their assistant agents could develop their own marketplaces.

These walled gardens could offer a number of benefits such as ensuring a baseline of quality and security by filtering out low-reputation or fraudulent service agents and streamlining discoverability and rating of agents. They could also offer insurance if an agent makes a mistake. However, much like today’s app stores, this could concentrate market power in the hands of a few dominant players, who could leverage their position to extract substantial profits and limit the openness, competitiveness, and innovation of the broader ecosystem. This could also lead to fragmentation of agentic ecosystems and suboptimal user experiences—for instance, a given individual

might have siloed assistant agents for personal and professional purposes, making it difficult to coordinate between them.

Conversely, if consumers and businesses fully own and manage their agents, communication could be both unscripted and unrestricted, leading to a completely open and decentralized “web of agents” that is not controlled by any one entity. Similar to today’s World Wide Web, any agent could join and transact with any other. Assistant agents would play a role similar to web browsers, and service agents similar to websites. This could foster competition, innovation, and broad access to agentic technology, but it also comes with substantial challenges. In particular, successful development of a web of agents requires large-scale coordination among many players—including corporations and governments—to develop and agree upon standards and protocols. It also requires robust mechanisms for discovery, trust, and security among interacting agents. We visualize these two plausible scenarios in the accompanying figure.

The future of advertising. In today’s digital economy there are generally many more businesses and products available to consumers than they have the time to research and consider, and so advertising plays a crucial role in capturing attention and guiding online transactions. But in an agentic economy where assistants can interact with millions of businesses on behalf of consumers, attention is a less-constrained resource. What matters more is the al-

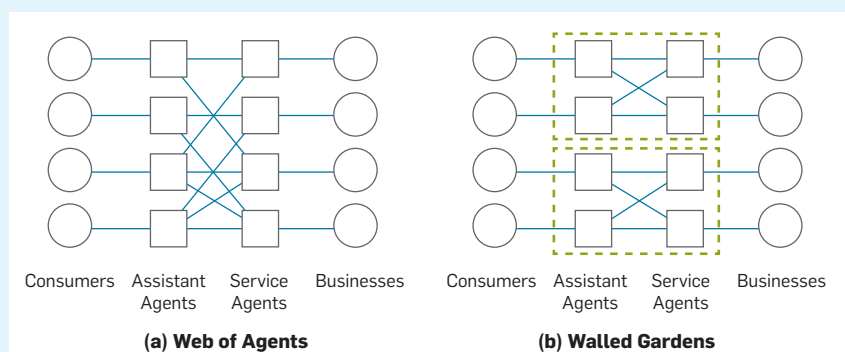
gorithm that matches assistants to service agents.

In a scenario where there are strong central intermediaries (that is, an agentic walled garden or web of agents with dominant discovery layers) some form of paid prioritization, akin to today’s advertising, is very likely to influence rankings. But the truly scarce and valuable resource—particularly in a web of agents—will be high-quality human feedback on goods and services. This feedback will be crucial not only for improving offerings (for example, training AI agents) but also for distinguishing high-quality services from poor ones. The focus of monetization and competition could shift from the “attention economy” to a “preference economy.” Success will hinge on attracting early, engaged users who provide valuable feedback.

New means of monetization may emerge to aggregate and rank service agents based on feedback from consumers and their assistants. Just as people leave reviews on platforms like Yelp or Google Maps today, future assistants might generate reviews based on user satisfaction data. Businesses could compete by offering better prices or services to attract these early users, creating a flywheel of feedback and preference data.

Payments and micro-transactions. As interactions between consumers and businesses become more seamless and platform intermediaries less central, we anticipate a rise in “one-off” transactions and accompanying decrease in repeat engagements and long-

Figure. In an agentic marketplace, assistant agents (representing consumers) and service agents (representing businesses) interact directly with each other. We distinguish between two scenarios: (a) an example of an open web of agents with unrestricted communication between agents and (b) an example of walled gardens in which agents are restricted to interacting within the confines of platforms.



term consumer-business relationships. This shift may encourage the growth of micro-transactions. For example, a consumer whose assistant frequently and seamlessly switches between multiple content providers (for example, Spotify and Pandora, or Netflix and Amazon Prime) may prefer usage-based micro-payments versus subscribing to both services. Furthermore, the usual friction associated with micro-payments is eliminated when transactions are handled entirely by assistants and service agents, encouraging micro-transactions that would otherwise be too inconvenient to manage manually. This shift is likely to evolve regardless of whether walled gardens or a web-of-agents scenario triumphs.

Unbundling, Rebundling, and New Products. Products, goods, and services are often bundled to balance complexity and efficiency of transactions. For example, a news article may bundle interviews, photos, and facts, even if the reader is already familiar with parts of the story. A consumer's assistant agent, however, can track what the user has already read and collaborate with a service agent (for example, from the *New York Times*) to generate a customized article focusing only on new or relevant information. This dynamic and personalized rebundling optimizes knowledge transfer while minimizing cognitive load. Such personalization can be accomplished even today using methods such as retrieval-augmented generation (RAG); in principle, an assistant agent could pull from multiple sources of high-quality content to create customized offerings for its user. Currently, however, such methods are often limited to public-domain sources such as Wikipedia or under fair use claims, as publishers are generally unwilling to permit their content to be reused without compensation. In the future, micro-transactions between assistants and service agents for the use of individual pieces of digital content could enable an ecosystem that compensates content creators while unlocking the ability to create customized user experiences.

More generally, the power of dynamic and personalized bundling could apply to many other digital goods and services that can be deconstructed and reconstructed by assistant agents working with service agents to best serve

As technological progress drives greater specialization, it also increases the need for coordination.

the needs of end users. This capability aligns closely with the anticipated rise of micro-transactions handled by agents. As the infrastructure for micro-payments develops, it will become possible for digital components to be individually and dynamically negotiated, priced, and packaged into hyper-personalized products. We expect more extreme unbundling, rebundling, and new products in a web of agents where there can be unrestricted communication between the assistant and *any* service agent, allowing for more innovation in product creation and bundling.

Conclusion

As technological progress drives greater specialization, it also increases the need for coordination, which in turn demands more sophisticated communication between individuals, organizations, and systems.¹ For example, where one doctor once handled a patient's care, today multiple specialists must collaborate using advanced tools and shared information. AI agents mark a shift in this pattern: rather than compounding communication overhead, they offer a way to streamline it. By coordinating tasks across fragmented systems, agents stand to reduce friction in markets, lower switching costs, and unlock more decentralized access to digital goods and services.

The architecture of this emerging marketplace is still taking shape. Much will depend on whether AI agents are allowed to interact freely across an open web or whether their interactions are restricted within closed ecosystems. This will be shaped by early stakeholder decisions, emerging technical standards, and evolving regulatory frameworks. Regulatory outcomes may vary by coun-

try—especially given the divergent approaches of the E.U., which favors precautionary, preventive measures, and the U.S., which tends to adopt a more reactive, ex-post regulatory stance in digital markets. In that sense, this moment is reminiscent of the early days of the Internet before it began generating billions of dollars of revenue. The current platform-dominated economy was certainly not obvious in the late 1990s into the early 2000s, when digital commerce was becoming increasingly decentralized as early Internet portals gave way to a blossoming World Wide Web.

Now is the time for us to reflect and ask what kind of agentic economy we want in the near future. The choices we make today will determine not only how these markets function, but also who benefits from this new wave of technology. 

References

1. Becker, G.S. and Murphy, K.M. The division of labor, coordination costs, and knowledge. *The Quarterly J. of Economics* 107, 4 (Nov. 1992); doi: 10.2307/2118383
2. Bick, A., Blandin, A., and Deming, D.J. *The Rapid Adoption of Generative AI*. Tech. Rep. National Bureau of Economic Research (2024).
3. Bresnahan, T. F. and Trajtenberg, M. General purpose technologies 'Engines of growth'? *J. of Econometrics* 65, 1 (1995); <https://bit.ly/4s88WEY>
4. Brynjolfsson, E. et al. Generative AI at work. *The Quarterly J. of Economics* (2025).
5. Cui, Z. K. et al. The effects of generative ai on high skilled work: Evidence from three field experiments with software developers. *SSRN 4945566* (2024).
6. Davidson, T. R. et al. Evaluating language model agency through negotiations. *arXiv:2401.04536* (2024).
7. Humlum, A. and Vestergaard, E. *The adoption of ChatGPT*. Tech. rep. IZA Discussion Papers (2024).
8. Klemperer, P. Competition when consumers have switching costs: An overview with applications to industrial organization, macroeconomics, and international trade. *The Rev. of Economic Studies* 62, 4 (1995).
9. Rysman, M. The economics of two-sided markets. *J. of Economic Perspectives* 23, 3 (2009).

David M. Rothschild (davidmr@microsoft.com) is a senior principal researcher at Microsoft Research, New York, NY, USA.

Markus Mobius is a senior principal researcher at Microsoft Research, Cambridge, MA, USA.

Jake M. Hofman is a senior principal researcher at Microsoft Research, New York, NY, USA.

Eleanor Dillon is a principal researcher at Microsoft Research, Cambridge, MA, USA.

Daniel G. Goldstein is a partner research manager at Microsoft Research, New York, NY, USA.

Nicole Immorlica is a senior principal research manager at Microsoft Research, Cambridge, MA, USA.

Sonia Jaffe is a principal researcher at Microsoft Research, Cambridge, MA, USA.

Brendan Lucier is a senior principal researcher at Microsoft Research, Cambridge, MA, USA.

Aleksandrs Slivkins is a senior principal researcher at Microsoft Research, New York, NY, USA.

Matthew Vogel is a principal program manager at Microsoft Research, New York, NY, USA.

Opinion

It Takes a Village to Teach Privacy

Rethinking privacy literacy for the younger generation.

CHILDREN TODAY ARE growing up in an era in which surveillance is not an exception but an expectation. A recent report from the American Academy of Child and Adolescent Psychiatry¹ indicated elementary school children spend more than six hours daily on digital devices using educational apps, watching videos, using chat apps to message friends, and navigating social media platforms, all of which are designed to collect users' data. Yet, most children lack the tools and skills to understand what is happening to their data or how to protect themselves.

Our research has found that children as young as six can articulate concerns about privacy, but their understanding is often superficial. Our work also indicates parents described privacy conversations at home as mostly rule-based ("don't share your location," "don't talk to strangers").³ While these rules are easy to remember, they rarely help children make sense of more complex or nuanced scenarios, such as whether it is okay to allow an app to track their activity "to improve performance."

We see an important gap here. We teach children to read before they read alone. We teach them about nutrition before they shop for groceries. But we often leave them to encounter digital privacy challenges with little more than "don't talk to strangers" or "don't post your real name." It is not enough. Children need more than rules. In our opinion, they need the knowledge, critical thinking, and decision-making skills



necessary to navigate privacy risks both online and offline. We define this as "privacy literacy."

As we are exploring novel ways to enhance children's privacy literacy, we found a need for a systemic shift in how we educate children. That is, instead of teaching rules, we need a comprehensive approach to equip children with the literacy they need.

Epstein's Theory and the Case for Cross-Context Collaboration

Education researcher Joyce Epstein proposed the Theory of Overlapping Spheres of Influence,² which states that children's learning is shaped by three overlapping spheres of influence: family, school, and community. When these spheres work in isolation, children re-

ceive fragmented or conflicting messages. When they collaborate, children benefit from reinforcement, consistency, and cumulative learning.

We applied this theory to the context of privacy education and found that each sphere contributes uniquely to children's learning when connections are intentional:

- Families are best at initiating conversations about privacy that are rooted in personal experiences and values.

- Schools offer structure, repetition, and alignment with developmental milestones in learning privacy skills.

- Community institutions like libraries and museums provide informal, playful, and curiosity-driven environments that can spark reflection and co-learning.

Based on this theory, we have been exploring a novel privacy education approach that allows families, schools, and community institutions (such as libraries and museums) to work in partnership to foster meaningful, age-appropriate, and sustained privacy literacy education for children. We believe privacy literacy education that bridges these spheres is not only more engaging but also more effective.

However, we also need to realize that each sphere has its strengths and weaknesses, which leaves room for an exciting research agenda.

Families Are Willing but Need Support

Parents play a crucial role in shaping how children think about privacy. Previously, we categorized five types of parent-child privacy conversations: rule-based, example-based, consequence-based, decision-exposing, and contextual.³ Most parents in our study relied heavily on rule-based messaging, which they found limited in helping children adapt to unfamiliar digital environments. A mother in our study described how her daughter learned about data tracking through a game that asked to access her photo library. The child asked, “Why does it want my pictures if it’s just a puzzle?” The mother, unsure how to explain, simply said, “Just say no.” The opportunity to discuss targeted advertising or third-party data sharing was lost.

In another study, we explored how guided reflection and perspective-taking can make privacy conversations more mutual and less didactic.⁶ When parents and children were supported to reflect on their own privacy needs and share experiences, conversations became more concrete, emotionally resonant, and memorable. However, many families told us they did not feel confident initiating these conversations on their own. Some parents rely on external authorities (for example, schools, tech companies, or online resources) to “take care of it.”

This is why privacy literacy tools for families should be not only simple and engaging for children but also accessible and empowering for parents. Many parents struggle to initiate conversations about online privacy because they themselves feel uncertain. To bridge this gap,

tools need to scaffold shared learning, which offers parents and children opportunities to explore privacy concepts together. Metaphors such as “Is your information a diary or a billboard?” invite reflections from parents and children. Games that visualize data sharing with embedded tips and prompts can help parents and children build shared language around privacy.

Schools Must Go Beyond Digital Citizenship Posters

Despite children’s growing exposure to technology, formal privacy instruction in elementary schools remains limited. Prior work by Kumar et al. shows that teachers described frequent use of Google tools, classroom apps, and online platforms, but rarely taught about privacy or data literacy explicitly.⁴ Instead, privacy concerns were framed around compliance or device management, not student understanding. This is a missed opportunity. Teachers are already engaging with students on digital platforms, but without embedded privacy discussions, children learn to use tools without questioning how those tools use their data and cause privacy issues.

However, educators need more than warnings or generic advice, as such information is rarely adaptive to classrooms with diverse needs. In reality, educators need curricular materials and professional development that align privacy education with language arts, science, and social studies, subjects that are already built into the school programs. They also need tools that fit into existing classroom technologies. For example, a reading unit on personal narratives can include a discussion on online

Despite children’s growing exposure to technology, formal privacy instruction in elementary schools remains limited.

identity; a science lesson on ecosystems can introduce data flows between devices, apps, and platforms; and a math class can include activities on algorithmic bias and data aggregation.

More importantly, the privacy education in the classroom ideally would also align with the privacy discussions in children’s homes. This is not only a reinforcement of the privacy concepts but also an alternative environment to nudge children to think about what privacy means in different contexts, which will facilitate children’s contextual understanding of privacy and gradually help them build stronger privacy literacy.

To accomplish this, the key is to embed “teachable moments” in teachers’ daily teaching routine, which can help teachers guide reflection without needing to build entirely new lessons.

Community Organizations: An Untapped Resource

Libraries, museums, and community centers offer informal learning spaces that are ideal for family-based privacy education. We interviewed eight families who regularly visited museums and libraries and found that these spaces often sparked curiosity and dialogue around privacy topics.⁵ One family recalled visiting a museum exhibit that visualized data traveling between apps. Their child asked, “So our tablet is always talking to the Internet?” The exhibit led to a 15-minute family conversation about Wi-Fi, cloud storage, and who sees what. Such moments are rare at home or school but flourish in informal, low-stakes settings.

These teachable moments are precious for various reasons. To start with, children come to museums to learn new things. Their mental state is set to absorb new things when they are at the museums, and their parents are also ready to engage with this learning procedure. Moreover, these moments are often the entry point for continuous and deeper learning at home and school.

But these approaches also need scaffolding. Museums or libraries can host privacy literacy events, which are interactive events where families can explore privacy concepts through games, stories, and technology demos during their visit. These events can illustrate abstract topics and make them feel

Figure 1. The authors (pictured) and collaborators developed an interactive exhibit to teach children about digital privacy at the Smithsonian National Museum of Natural History.



Figure 2. The Carnegie Mellon University Privacy Engineering program organized Data Privacy Day events at the Carnegie Library of Pittsburgh.



relevant, social, and fun. Additionally, take-home reflection cards, age-specific prompts, or digital journals can extend these conversations and connect them with school curriculum or family goals. The latter one, by nature, would require a deep partnership among families, schools, and community organizations so that all stakeholders can work collaboratively and create a well-rounded learning environment to enhance children's privacy literacy.

A Blueprint for Privacy Literacy Partnerships

As we are advocating for a partnership among families, schools, and community organizations, we reflect on some concrete recommendations for building cross-context, partnership-based privacy literacy education programs.

Recommendation 1: Co-design materials across settings. Privacy education tools should be designed with input from families, teachers, and community stakeholders.

Examples: A digital storybook used in school could be paired with a parent reflection guide, or a museum exhibit on data flows.

Recommendation 2: Embed “teachable moments” in everyday tech use. Classroom apps and learning platforms should include prompts or visual cues that help children pause and think.

Examples: “Who might see what I just typed?” “Why does this app want my location?”

Recommendation 3: Create longitudinal privacy learning pathways. Privacy decisions do not happen in one moment. Just as reading proficiency develops over

time, so does privacy literacy. Long-term studies and curricula should trace how children develop from “rules-followers” to context-aware privacy actors.

Example: Children's privacy learning stories that can be shared among all stakeholders.

Recommendation 4: Support families as co-teachers. Provide parents with low-barrier prompts and activities that center children's voices while guiding reflections, along with accessible resources for easy engagement.

Examples: Dinner table questions, bedtime stories, or screen-time negotiation checklists.

Recommendation 5: Leverage community events. Activities and events with a privacy theme in libraries, science centers, and afterschool programs can reinforce what children learn at home and school. These events should be designed to include entire families, not just children.

Examples: Interactive exhibit on digital privacy via the Expert Is In program at the Smithsonian National Museum of Natural History by Johns Hopkins University and the University of Notre Dame (Figure 1), Data Privacy Day at the Carnegie Library of Pittsburgh by Carnegie Mellon University (Figure 2).

Final Thoughts: Teaching Privacy Is Teaching Agency

At its core, privacy literacy is more than protecting data. It is about teaching children that they have a right to boundaries, a right to ask questions, and a right to make decisions about their digital lives. It is about understanding context, making judgments, negotiating mean-

ing, and navigating power. It is about cultivating agency in a world that often treats children as passive users or invisible data points.

These are not skills that can be taught in one setting, by one adult, or through one app. They require practice, reinforcement, and reflection across all the spaces where children live and learn. When families, schools, and communities work together, we can do more than improve test scores. We can empower children to read the world beneath what it looks like on the surface with confidence, curiosity, and care. This is a life-long skill. C

References

1. American Academy of Child and Adolescent Psychiatry. *Screen Time and Children* (2025); <https://bit.ly/4pRUVtw>
2. Epstein, J.L. et al. *School, Family, and Community Partnerships: Your Handbook for Action*. Corwin Press (2018).
3. Kamel, K. et al. Towards understanding family privacy and security literacy conversations at home: Design implications for privacy literacy interfaces. In *Proceedings of the 2024 CHI Conf. on Human Factors in Computing Systems (CHI '24)*. (2024); doi: 10.1145/3613904.3641962
4. Kumar, P.C. et al. Privacy and security considerations for digital technology use in elementary schools. In *Proceedings of the 2019 CHI Conf. on Human Factors in Computing Systems (CHI '19)* (2019); doi: 10.1145/3290605.3300537
5. Liu, L. et al. Exploring design opportunities for family-based privacy education in informal learning spaces. In *Proceedings on Privacy Enhancing Technologies* (2024).
6. Wen, Z., Liu, L., and Yao, Y. Supporting family discussions about digital privacy through perspective-taking: An empirical investigation. In *2025 IEEE Symp. on Security and Privacy (SP)* (2025); doi: 10.1109/SP61157.2025.00149

Yaxing Yao (yaxing@jhu.edu) is an assistant professor in the Department of Computer Science at Johns Hopkins University, Baltimore, MD, USA.

Lanjing Liu (lliu153@jhu.edu) is a Ph.D. student in the Department of Computer Science at Johns Hopkins University, Baltimore, MD, USA.

© 2026 Copyright held by the owner/author(s).

ACM ON A MISSION TO SOLVE TOMORROW.



Dear Colleague,

I feel privileged to call ACM my professional home! Believing in the core ACM values of excellence, high ethics, and inclusivity, I joined as a graduate student and since then have been contributing in several ways to ACM's mission of "advancing the art, science, engineering, and application of computing."

For more than 75 years, ACM has played a leadership role in the computing field, keeping a finger on the pulse of the evolving needs of the community and expanding its extensive services to benefit computing researchers, industry professionals, educators, and students around the world.

More and more, ACM is taking steps to become the home of interdisciplinary areas that involve computing. ACM is helping to shape many computing-related interdisciplinary areas by forming strategic alliances with other scientific societies and welcoming members with multidisciplinary backgrounds.

ACM is also among the frontrunners in redefining scholarly communication and the entire research life cycle by embracing the principles of reproducibility and accountability and facilitating open science methods. ACM is actively transitioning to become a fully open access publisher within the next few years.

For many years, ACM has advised policy makers on cutting-edge technologies that may have significant consequences on society. Today, this work is more urgent than ever as threats on democracy, increasing inequalities, and the loss of privacy challenge us all. ACM continues to advocate for the application of technological innovation within clear ethical boundaries.

ACM has had tremendous impact on the profession by evolving continuously to better serve the needs of its diverse members, listening to all voices, and making innovative changes. This remains the course so that ACM can continue to fulfill its vision of being "the premier global computing society."

Joining ACM means you dare to be the best computing professional you can be. It means you believe in advancing the computing profession as a force for good. It means you act on your commitment to solving tomorrow's challenges. Come join your peers at ACM and let us all together invent a better tomorrow!

Sincerely,

A handwritten signature in black ink, appearing to read 'Y Ioannidis', with a stylized flourish at the end.

Yannis Ioannidis
President
Association for Computing Machinery



Association for
Computing Machinery

Advancing Computing as a Science & Profession

SHAPE THE FUTURE OF COMPUTING. JOIN ACM TODAY.

www.acm.org/join/CAPP

SELECT ONE MEMBERSHIP OPTION

ACM PROFESSIONAL MEMBERSHIP:

- ☐ Professional Membership: \$99 USD
- ☐ Professional Membership plus
ACM Digital Library: \$198 USD
(\$99 dues + \$99 DL)

ACM STUDENT MEMBERSHIP:

- ☐ Student Membership: \$19 USD
- ☐ Student Membership plus ACM Digital Library: \$42 USD
- ☐ Student Membership plus Print *CACM* Magazine: \$42 USD
- ☐ Student Membership with ACM Digital Library plus
Print *CACM* Magazine: \$62 USD

- ☐ **Join ACM-W:** ACM-W supports, celebrates, and advocates internationally for the full engagement of women in computing. Membership in ACM-W is open to all ACM members and is free of charge.

PAYMENT INFORMATION

Name

Mailing Address

City/State/Province

ZIP/Postal Code/Country

- ☐ Please do not release my postal address to third parties

Email Address

- ☐ Yes, please send me ACM Announcements via email
☐ No, please do not send me ACM Announcements via email

- ☐ AMEX ☐ VISA/MasterCard ☐ Check/money order

Credit Card #

Exp. Date

Signature

Purposes of ACM

ACM is dedicated to:

- 1) Advancing the art, science, engineering, and application of information technology
- 2) Fostering the open interchange of information to serve both professionals and the public
- 3) Promoting the highest professional and ethics standards

By joining ACM, I agree to abide by ACM's Code of Ethics (www.acm.org/code-of-ethics) and ACM's Policy Against Harassment (www.acm.org/about-acm/policy-against-harassment).

I acknowledge ACM's Policy Against Harassment and agree that behavior such as the following will constitute grounds for actions against me:

- Abusive action directed at an individual, such as threats, intimidation, or bullying
- Racism, homophobia, or other behavior that discriminates against a group or class of people
- Sexual harassment of any kind, such as unwelcome sexual advances or words/actions of a sexual nature

BE CREATIVE. STAY CONNECTED. KEEP INVENTING.



Association for
Computing Machinery

ACM General Post Office
P.O. Box 30777
New York, NY 10087-0777

1-800-342-6626 (US & Canada)
1-212-626-0500 (Global)
Hours: 8:30AM - 4:30PM (US EST)

Fax: 212-944-1318
acmhelp@acm.org
www.acm.org/join/CAPP

An API for interfacing with the automaters.

BY THOMAS LIMONCELLI AND CHRISTIAN PEARCE

SRE Is Anti-Transactional

A MAJOR CONCERN in the site reliability engineering (SRE) movement is automation—that is, moving IT from focusing on transactional work to building systems that do work for us. There are many competing definitions of SRE. If you ask 10 site reliability engineers (SREs) to define SRE, you'll get 11 definitions. Some will focus on the *R* (reliability) as the driving force in everything from feature priorities, to budgets, to technology choices. Others use SRE as an umbrella term encompassing everything from application upkeep, cloud engineering, service provisioning, and platform engineering to nearly all aspects of keeping a technology working.

What they all have in common is an imperative to move away from transactional work. It may not be the defining attribute of the SRE domain, but it is an important aspect. Let's drill down into it.

SREs do not like manual work. Give a team of SREs a manual process, and over time they will evolve it from fully manual (people doing work), to automated (software tools that humans use to make progress), to autonomous (systems that identify when work needs to be done, do the work, and alert humans in exceptional

situations). The aim is to progress toward eliminating manual work from the system altogether.

SREs consider transactional work to be *toil*—work that, while valuable in and of itself, does not raise the bar and improve the process for any similar future requests. It is tactical, not strategic. It addresses immediate needs rather than contributing to long-term improvements or strategic goals.

For example, deploying a new software release to a website is a good thing. The new software release has new features that benefit the user, increase profit, or otherwise provide value. In the old days, this process was done manually and would frequently require days or weeks of planning and execution. Then, the next time a new software release arrived, the same amount of work would be required to deploy it. That is *toil*.

On the other hand, a project that replaces all that manual toil with a continuous integration/continuous delivery or deployment (CI/CD) system is *strategic*, such that developers can safely deploy new software releases in a self-service manner. It not only automates a manual process, thus reducing toil, but also permits the process to be done more frequently (even continuously) in smaller batches without waiting for IT to become available.

SREs gladly do manual work but only as a necessary step of learning the process so that it can be automated and eventually become autonomous. Toil is acceptable when a task is infrequent or a process is new and ill-defined or experimental in nature. Those are exceptional situations, however.

As the old joke goes, in the future there will be a factory that needs only two employees: one human and one dog. The human's job will be to feed the dog. The dog's job will be to make sure nobody touches the equipment. Once SREs rule the world, the dog will be automated as well.

Transactional Work

Transactional work usually takes the



form of a request that is worked on by an individual until it is completed. Tickets opened with your IT organization are transactional work. Whether you are requesting a replacement mouse or need assistance with some technology that is not working correctly, the work is transactional. It begins with a request, work is performed, the requester verifies the work is complete, and then, lastly, the ticket is closed.

Transactional work is typically manual and tactical (not strategic). It addresses immediate needs rather than contributing to long-term improvements or strategic goals. It does not improve the system, product, or processes in a meaningful way.

Toil scales linearly. If the number of requests increases 10x, the number of people required to complete the work increases 10x. This is unacceptable in

today's business environment. Since Ford popularized the automotive assembly line, it has become an expectation that cost should scale sublinearly with growth. That is, if 10 units of work requires 10 people, 100 units of work should require significantly fewer than 100 people.

Toil is reactive and interrupt-driven by nature, which makes planning difficult. A request arrives at an unpredictable time and requires humans to stop whatever they are doing to work on the task. Since people are not available 24/7, the requester must wait. Everyone has experienced the frustration of waiting days for someone to perform a task that takes minutes. Automation is available 24/7, so there is less waiting.

Toil drains engineers' time and energy, preventing them from focusing on high-impact engineering work,

such as automation, system improvement, or innovative projects. It reduces the overall efficiency and morale of the team, as engineers are bogged down in routine maintenance rather than more meaningful contributions. Luckily, toil is usually repetitive and, thankfully, repetitive tasks often prove to offer hot opportunities for automation.

Automation

SREs focus on building automation that eliminates the need to do manual work. It is not that they are lazy; they just want to be able to scale.

Give a team of SREs a fully manual task, and over time it will go from fully manual, to automated, to autonomous. The difference between automated and autonomous is the level of automation required: Automated systems are tools that humans use to get a task done. Au-

onomous systems are triggered when new work arrives and they complete work on their own. With autonomous systems, humans are there only for exceptions—edge cases that have not yet been automated, bugs that need to be fixed, or system optimizations.

When work is automated, it can scale super-linearly. This is particularly important for Internet-based companies. As Michael O'Dell, chief scientist of the first commercial ISP, frequently said, "The only problem is scaling. All the others inherit from that one."

The analogy we like to make is that SREs are not the assembly-line workers at an auto manufacturing plant. They are the people who make the robots that make the cars.

A process generally goes through these stages:

1. *Manual.* Progress is made by humans doing the work, often typing or clicking, observing the results, and then making decisions about how and when to proceed.

2. *Automated (sometimes called manual automation).* A tool performs each step. When a task needs to be done, a human uses a tool (software) to perform that task. A single process might require many steps. SREs tend to write tools that automate those steps. Often, individual steps have bespoke tools, and the process is completed by humans who use the appropriate tool for each step, judging the results, and then deciding when to move forward to the next step.

3. *Autonomous systems (sometimes called fully automated).* Software initiates and executes work without human intervention. Humans are there to handle exceptions and improve the system. Once a system is stable, it is mostly concerned with reducing the number of exceptions that require human intervention, addressing any new features and business concerns, and achieving greater efficiency.

An autonomous system is never finished. SREs live in a Sisyphean struggle: As they get closer to perfect automation, a changing world throws new toil at them. If every edge case were handled today, new business needs would appear tomorrow. If the system runs smoothly now, an underlying system will require an upgrade tomorrow.

A good system will always be a

victim of its own success, becoming so popular that it exceeds its original design limits and thus requires reengineering to overcome the new bottlenecks this growth exposes. Popularity inspires new use cases, which requires redesigns and refactoring. If the system was designed to handle only green widgets, surely its success would make it attractive to the yellow-widget division. Sadly, while extending it to a new use case sounds easy, it rarely proves to be. It is financially impractical to automate every conceivable use case. Some use cases are so infrequent or complex that automation can offer little or no return on investment. Automating those cases would be wasteful and considered to be a bad business decision. Thus, an autonomous system needs a team to maintain it, optimize it, and handle the exceptions. Autonomous systems are never "one and done."

Example: Cluster Builds

Tom was once on a team that was deploying storage clusters. Each cluster was made up of many individual hardware and software components.

Initially, the process was manual, with hundreds of steps, each documented to varying degrees of completeness. Soon, each phase of the process was automated; one or more tools did most of the work of each phase. It required human judgment, however, to determine whether a phase had been completed correctly or fixes were required. We relied on human judgment to determine when to start each phase. This often required checking calendars or getting a go/no-go decision from other teams. While this situation was a great improvement over the manual process, it still required considerable human intervention to tell us when to start, what parameters were required in each instance, when to proceed to each new phase, as well as how to handle quality control. Still, it was better than a manual process in that it was more consistent, took less time, and required less cognitive load to complete each task.

The next phase was to help the system become autonomous. Orders for new clusters were no longer presented during status meetings or via tickets. Instead, orders came from a database

(imagine an order received via Salesforce). A validation process not only verified the request's parameters, but also sent an email message with an approval link to the designated budget controller. Each phase then was kicked off via timetables, verification of completion of preceding steps, and other triggers.

Once the system was stable, we discovered that a typical cluster build still required human intervention at least twice during the multiday process. As we fixed edge cases, improved algorithms, and fine-tuned the system, this was reduced to nearly zero. The system had become data-driven and even had a dashboard that showed our "human intervention" count dropping over time.

This count would still spike upward occasionally. These spikes correlated with new releases of software received from the developers, since each release had a tendency to break our fussy automation. Improving this required us to break out of our silo and work more closely with the developers.

We realized their release process needed to include verifying that the cluster-deployment system supported the new release. It was a "lightbulb moment" when the product manager realized the ability to deploy clusters was as much a part of a new release as the new features we had promised to ship. Maybe this is obvious in hindsight, but the authors have seen this cycle (silo, lightbulb, cooperation) repeat frequently throughout their careers.

It took more than a year to go from fully manual to an autonomous system. Most of the fine-tuning required after that was to scale the system both horizontally (efficiently creating larger clusters) and vertically (creating more clusters per month). The old manual system did not make anyone happy. Sales recognition (which starts *after* cluster delivery) was delayed. Customers were unhappy with misconfigurations and other bugs.

The new system was fast and predictable (three days). It had an SLA that allowed the company to tell customers when to expect delivery of the service, and the sales department was alerted that a contract had to be signed three days before the end of the quarter for timely revenue recognition. Best of all,

the system was autonomous, meaning orders could be fulfilled even if the SREs were asleep.

Example: Mouse and Cable Distribution

We used to think that one help-desk task that could never possibly be automated had to do with the procurement of small items, such as replacement mice and USB cables. A customer would open a ticket requesting (for example) a mouse, whereupon the help-desk person would provide the mouse, assign the cost to the person's department, and so on. How could that get any better?

I was impressed to discover that Google's help desk had actually succeeded in automating that process with a vending machine for physical items. Employees could swipe their badges and take the items they desired, after which their department would be billed. This system eliminated a significant percentage of help-desk visits—meaning the toil was gone. All that was left was a management process that involved ordering new items and monitoring for abuse.

Some industry pundits claim there is a world of “old IT” that is all toil, unfixable, and doomed. We like to believe there is always hope whenever there is sufficient scale and a willingness on the part of management to embrace creativity and innovation.

How to Interface with SREs

Given the SRE distaste for transactional work, what is the best way to interact with an SRE team? Ideally, transactional requests that are manual in nature (verbal requests, tickets generated by humans, and so on) should be replaced by interactions through software. That is, your team should be able to provide an API to your SREs or request they provide you with one. An API provides a way for the SRE software to interact with your systems and vice versa. This paves the path to automation.

Sometimes, we get lucky because an API exists. Sometimes, though, the API is obscured. For example, the database or SaaS-based product your team is using might already have an API, but access to it must be enabled, properly secured, with rules of engagement defined.

Usually, however, a new API needs

to be created from scratch. This is not easy. There is no magic button you can press to make an API appear. Creating an API requires the very human process of sitting down, identifying the need and requirements, proposing designs, and allocating resources to build, test, and implement that design. This usually requires software engineering skills, which not every team has available. For example, in our experience human resources departments rarely have software engineers on staff.

Efforts that mesh two teams with different levels of experience with process automation can prove to be frustrating for both teams. The non-SRE team, lacking software-development experience, can be put off by the requirement to work through a detailed process analysis given that the transactional process seems so obvious to them. The SRE team, meanwhile, can often become frustrated by the other team's tolerance of toil and surprised at the lack of disdain for it.

If creating an API is not an option, sometimes a little creativity is all that is required. In one case, Tom discovered that a “mostly autonomous” system was hamstrung by the fact that another team was responsible for allocating ID numbers required by the process. Basically, the process was fully autonomous apart from one “stop-the-world” moment that required waiting for a human on another team to allocate an ID number. This other team, meanwhile, did not understand the problems that resulted from this delay. Accordingly, the requests were often ignored for days. No other team made such requests and so they were viewed as “weird” and assumed to be of low priority. In reality, the delays were causing major headaches.

By sitting down and discussing the situation, the teams were able to propose two creative solutions. One was to allocate large batches of ID numbers, reducing the human interaction to just refill the ID pool—something that might happen only a few times each year. The other idea was simply to permit the SRE team to allocate ID numbers with a particular prefix, thus giving them the ability to manage their own allocation process.

Likewise, SRE teams should be willing to create interfaces for others.

That could be a traditional network-based API, but often a Web-based user interface or dashboard will suffice. For example, the other team's needs may be satisfied by gaining access to a dashboard in an existing monitoring or business information system. That could be a static webpage that is updated periodically. And while the need to visit a webpage to find answers might smell like unacceptable toil to an SRE, a non-SRE might just as easily view it as a welcome improvement.

In the end, when we can find a way to work together to eliminate the need for tickets or to reduce toil in some other fashion, we all win.

Conclusion

The value of a team is its output. If the team's role is to repeat the same task over and over again, that team's value is going to scale linearly with their effort. But if the team's role is to optimize processes through automation and the reduction of toil, then that team's value will prove to be like compound interest, with growth propelling growth. SREs are happiest when they are in the latter category: creating value that scales.

Systems built by SREs are not fully autonomous on day one. It is iteration over time that leads to fully autonomous, functional, reliable service. This iterative process requires SREs to evaluate how much time and money should be spent to achieve the objective. It is the heart of engineering to find the fastest, cheapest, and safest way to create and maintain a system.

We have seen conflicts between SRE goals and “traditional IT” mindsets. Generally, these stem from the fundamental differences between the SRE mindset of automating wherever possible to eliminate toil and the more immediate focus other teams have on completing certain specific tasks.

Understanding this tension is key to interacting with SRE teams. 

Thomas A. Limoncelli is a senior platform engineer at Stack Overflow Inc. He holds a B.A. in computer science from Drew University, Madison, NJ, USA.

Christian Pearce is a platform-engineering manager and a technology leader. He is a manager at Stack Overflow, leading the Infrastructure Platform Team.

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

© 2026 Copyright held by the owner/author(s).

If you are tired of hearing about memory safety, this article is for you.

BY ANDREW LILLEY BRINKER

Memory Safety for Skeptics

MEMORY SAFETY—THE PROPERTY that makes software devoid of weaknesses such as buffer overflows, double-frees, and similar issues—has been a popular topic in software communities over the past decade and has gained special prominence alongside the rise of the Rust programming language.^a Rust did not invent the idea of memory safety, nor was it the first memory-safe language, but it did break the dam on the last major holdout context where memory safety had not yet been achieved: what we often call *systems programming*.

Rust's big step function was to offer memory safety at compile time through the use of static analysis borrowed and grown out of prior efforts such as Cyclone,^b a research programming language formulated as a safe subset of C. Rust, by offering a memory-safe-by-default experience for the *systems*

domain, where operating systems, databases, file systems, embedded software, and the like are made, suddenly presented a new possibility to public policymakers and to leaders of engineering organizations: the mass reduction of memory unsafety across any domain.

In the years since Rust hit the scene, tech companies have published experience reports on the adoption of Rust for production systems, whether through rewrites of existing code or by producing new modules in Rust that might have otherwise been written in C or C++. The numbers were broadly consistent: a roughly 70% reduction^c in memory-safety vulnerabilities. Rust, more than just promising memory safety, was demonstrably translating safety guarantees into practical improvements in software security. This evidence, turning the abstract benefits of a semantic improvement into bottom-line improvements in business costs (vulnerabilities are *expensive* for all involved), meant organizations beyond just engineering groups began to take notice.

Of course, the choice of programming language is a contentious one. Languages do not exist in a vacuum, and the *right* language for a job is heavily path dependent. What languages do the developers already know? What is the timeline and budget for the project? How serious are the correctness constraints? The performance constraints? Do you expect to hire more developers, and what resources can you allocate to train them? If you are maintaining an open source project, you might ask which languages would possibly bring in more developers to contribute. For any project, your answer might be determined by what other libraries or tools you will need to integrate.

What's more, many projects already made their programming-language decision years ago—possibly decades ago. What should they do?

a <https://www.rust-lang.org>

b <https://cyclone.thelanguage.org>

c <https://tinyurl.com/2a3z9npb>



If the code you have today is memory unsafe, in C or C++, how can you pursue safety without throwing the whole thing out?

In some circles, the answer given might be to rewrite it in Rust to replace legacy software written in memory-unsafe languages with new Rust equivalents. The benefits, supporters argue, are clear: comparable performance, modern tooling, and memory safety.

Yet, experienced developers know rewrites are expensive⁶ and frequently misguided. Often, demands for large-scale rewrites are not a carefully rea-

soned argument about trade-offs but an aesthetic criticism of code that looks “ugly” or “too old.” If anything, those calling for mass rewrites show their own inability to do the difficult work of understanding and working with an existing codebase that does a job and does it well.

There are paths between accepting memory unsafety as the cost of doing business or performing a mass rewriting of stable systems in a new language to achieve safety. Reflexive rejection of a move to memory safety is misguided, especially when the benefits of memo-

ry safety can be achieved in a cost- and schedule-efficient way.

If you are not yet sold on the value of memory safety, this article is for you. The goal in writing it is to treat the question of pursuing memory safety in legacy systems with the seriousness and rigor it deserves.

Pursuing memory safety is worthwhile, with or without Rust, and I’d like to convince you to try.

In Budget, On Schedule

Software systems do not exist in isolation; software is built to *do things*, to

serve the needs of businesses and individuals by making systems more efficient or automatic. The development of software is constrained not by the theoretical limits of software's abilities, but by the cost and schedule limitations of the team building it.

In "The Case for Memory Safe Roadmaps,"^d a collection of government agencies from the "Five Eyes Countries" (the U.S., U.K., Australia, New Zealand, and Canada) collectively recommended that organizations develop roadmaps for transitioning their software development efforts to memory-safe languages.

It's worth being clear here: Their focus is on a *roadmap* they very explicitly accept and they discuss the challenge of the cost and schedule impacts of any transition toward memory safety.

Budget and schedule constraints and the desire for efficiency are part of what motivates the creation of software in the first place. Once that software is in place, it's frequently mission critical, having replaced the knowledge and labor of people who would have previously done the jobs the software now performs. Instead of accountants, a company may have accounting *software*, with a smaller number of accountants who know how to interact with the software and use it to perform their own jobs built on the knowledge the software provides with its data and embedded business logic.

Rewrites to critical software systems are risky precisely because the software itself is so important. Rewriting a software system, whether as an in-place rewrite where components are swapped out piece by piece, or as a wholesale rewrite with a cutover date, risks the proper functioning¹² of the business if the rewrite fails.

Complex, long-running software systems can face other severe constraints as well. They may be unable to be turned off—because, having become so business-critical and time-sensitive, any attempt to bring them offline for maintenance or replacement is unacceptable. They may also have become lost artifacts, where the expertise of the individuals who created them or previously worked on them is lost because it wasn't transferred to



A transition to memory-safe languages can bring benefits beyond just the safety (and thereby security) claims that are often given priority in these discussions.



newer engineers, resulting in a current team that does not understand the system or feel comfortable making changes to it.

There are also ongoing costs associated with the development of software that might have to be diverted to support even an incremental rewrite. Depending on the business, there might not be funding available to support an *increase* to the development team, so diversion of resources toward a rewrite would mean reduction in the delivery of features for the project, which may be untenable.

All of this is to say that rewrites, even incremental ones, are business decisions that have to be made as trade-offs with other strategic goals. While motivated developers can make the best case possible for the upside of a rewrite, they must also grapple with the business's need to deliver features and address bugs impacting users of the system today.

At the same time, a transition to memory-safe languages can bring benefits beyond just the safety (and thereby security) claims that are often given priority in these discussions.

Memory-safety violations, such as null pointer dereferences or indexing outside the bounds of a memory buffer, can result in denial of service (or, in the context of the classic security CIA triad, availability failures) of the relevant software. This might mean on-call pages to respond to a production incident, a failure to meet service-level agreement guarantees for customers, or reduced revenue from lost customers or interruption of business operations.

Memory-safety issues are also often a central building block in a kill chain for achieving remote code execution by attackers. Even as far back as "Smashing the Stack for Fun and Profit"^e in 1996, we could see cybersecurity professionals documenting how to turn a buffer-overflow weakness into remote code execution and full access to the host. With that foothold in place, attackers can begin to exfiltrate data, move laterally within a network, escalate privileges, lock down a system with ransomware, conscript a host into a botnet, and more.

d <https://tinyurl.com/22ajvyco>

e <https://tinyurl.com/29fonsma>

Software problems are cheaper to fix the earlier they^f occur in the software development lifecycle. In the long term, stopping a bug from being written is cheaper than responding to a bug bounty submission or triaging a production outage.

This is not to say that *all* moves toward memory safety are cost effective or that all roadmaps for memory safety should be as aggressive as possible, but it is intended to make clear there are both costs and savings to be had with any transition to memory safety.

Setting the Goalposts

What is memory safety? There should be a table-stakes answer to that question to have in hand amid the push toward memory safety in public discourse, but there is not a single, fully agreed-upon, and rigorous definition. There is a new effort, announced in a recent article published in *Communications*,¹³ to develop a standard definition of memory safety, but it is just beginning.

There is, however, a rough consensus among practitioners of what kinds of program behaviors are memory *unsafe*. That's a good place to start.

My favorite short definition comes from Michael Hicks,⁷ an academic who works on programming languages:

“[A] program execution is memory safe so long as a particular list of bad things, called memory-access errors,¹ never occur:

- Buffer overflow
- Null pointer dereference
- Use after free
- Use of uninitialized memory
- Illegal free (of an already freed pointer, or a non-`malloc`-ed pointer)”

You will sometimes see these categories broken down further, into spatial and temporal memory safety. *Spatial* covers memory-safety issues arising from accessing locations in memory that a program should not have access to (like a buffer overflow); *temporal* covers operations on memory done in the wrong order: for example, reading memory before it is initialized, trying to free an already freed pointer, or using a pointer after it has been freed.

There's also the Common Weakness Enumeration (CWE) category for mem-

ory-safety issues, which decomposes Hicks's list into more granular options. CWE is a taxonomy of software weaknesses, or as CWE puts it: “condition[s] in... software... that, under certain circumstances, could contribute to the introduction of vulnerabilities.”

In CWE's memory-safety category,⁸ “buffer overflow” is further broken down into six different, more-specific weaknesses, some of which are further decomposed into their own variants. This can be useful when maximum precision is warranted but is perhaps too much detail for the purposes of this article.

These definitions provide a reasonably clear picture of what constitutes memory *unsafety*. So, memory safety is when a program is guaranteed not to have those weaknesses. This can be achieved by compile-time constraints on the semantics of programs or by runtime management of memory by a garbage collector, so long as the guarantee is upheld.

This is often when perceptive on-lookers will cry foul. Rust permits *unsafety*!⁹ There's a whole *unsafe* keyword! How is that meaningfully different from the guarantees of C or C++?

Of course, they are right. Rust does permit programmers to write unsafe code, but as anyone who works in safety or security will tell you, defaults matter. In fact, defaults matter a lot!

Take seatbelts, for example. Seatbelts became generally mandatory across the U.S. between the late 1980s to the early 1990s. In 1985, when mandatory seatbelt laws first saw passage among the states, seatbelt usage sat at 21%^h of riders. In 1994, the average seatbelt usage rate in the U.S. was 58%. As of 2017, it was 89.7%.ⁱ That change in defaults led to massive increases in seatbelt usage and therefore saved more lives. The National Highway Transportation and Safety Administration estimates that in 2017 alone, seatbelts saved the lives of nearly 15,000 Americans.^j

The same truth applies in software. Before version 4.0.0 (published in 2017), Redis, the extremely popular key-value

store, offered no access controls in its default configuration. Frequently, new users of Redis would unintentionally expose their instance publicly, and this insecurity would result in data spills or become a vector for host exploitation. As of version 4.0.0, Redis enters a “protected mode” when run with its default configuration and without password protection. This limits access to loopback interfaces. As the Redis company itself has since touted, the introduction of protected mode has caused the number of publicly accessible Redis instances tracked on Shodan.io, a popular Internet host aggregator, to decline substantially. In 2017, it had identified roughly 17,000 exposed Redis instances; in 2020, that number had declined to 8,000 in an audit by security company TrendMicro.

Bringing it back to memory safety, we can and should think of memory-safety guarantees by languages as a continuum, and we can split languages between “memory safe by default” and “non-memory safe by default” groups. This framing, recommended^k by the Open Source Security Foundation's (OpenSSF's) Memory Safety Special Interest Group (SIG), makes the options clearer:

- Using memory-safe-by-default languages
- Using memory-safe-by-default languages to interface with non-memory-safe-by-default languages
- Using non-memory-safe-by-default languages

Here, memory-safe-by-default languages include not only Rust but also common garbage-collected languages, such as Java, C#, Go, Swift, Python, and Ruby. Non-memory-safe-by-default languages include C and C++ most notably, but also Zig, which may be surprising to those who have watched memory-safety discussions from the sidelines.

While Zig does provide more ergonomic options for programmers to write memory-safe programs themselves, Zig is not a memory-safe language because it does not guarantee memory safety even in its most conservative configuration. Jamie Brandon's breakdown of Zig's memory

f <https://tinyurl.com/2a69yayu>

g <https://tinyurl.com/28aj2u5p>

h <https://tinyurl.com/282k5yr4>

i <https://tinyurl.com/2asgbvdm>

j <https://tinyurl.com/25kw3h7s>

k <https://tinyurl.com/25fs6lvp>

safety² is a good walkthrough of why Zig’s guarantees are insufficient.

With a shared understanding of memory safety and memory-safe languages, let’s now dig into the concrete strategies for pursuing memory safety in real-world programs.

Strategies for Safety

Any of the following strategies are intended to maximize the benefit of memory safety while minimizing the cost of pursuing it. The specific choice of which approach is right is context-dependent and should be made with consideration of the importance of the component, the current and new target language, the team involved, and the timetable.

Make new code memory safe. The first and most obvious option is to make new code memory safe—that is, to write new components in a memory-safe language. While this seems simple, you must address certain caveats to make this approach successful.

First, you are unlikely to reap the benefits of memory safety if you try introducing memory-safe code alongside new memory-unsafe code. Think of it this way: In a fixed codebase that you continue to assure (via testing, code review, bug bounties, and more), the density of vulnerabilities decreases exponentially over time. As vulnerabilities become less dense in the codebase, the rate of new vulnerability discoveries also decreases, so the overall assurance level of the code *increases*. The riskiest thing you can do to a codebase is change it. In the case of memory-unsafe languages, that change can induce memory-safety vulnerabilities.

The Google Chrome and Android teams have published extensively about their experiences incentivizing a move to memory-safe languages in their codebases. They instituted a rule called the “Rule of Two,” where all new code must be either sandboxed or in a memory-safe language. In practice, because sandboxing is difficult, this naturally gave developers incentive to pursue memory safety in most cases.

Surprisingly to the team, they reaped the benefits of this new policy¹⁴ across the entire codebase—even the parts that were not rewritten. Be-

cause they had certain assurances about the new code inherent in the safety mechanisms it came with, they could focus assurance efforts on old code, which was now static. Through this, they not only reduced the overall rate of memory-safety vulnerabilities in the codebase, but also decreased the prevalence of vulnerabilities overall.

Target rewrites to critical components. Sometimes, rewriting code in a memory-safe language *can* be the right choice, but this is often a path to pursue only once you fully understand the challenges faced by the current memory-unsafe code.

Early in its history, the development of Rust was funded by Mozilla, makers of the Firefox browser, and the flagship Rust project Rust’s own compiler was the Servo web-rendering engine. Despite this, the first actual Rust code Mozilla integrated into Firefox was not Servo; it was an MP4 video file parser.⁵ They replaced the existing C++ parser with one written in Rust, moving from a memory-unsafe language to a memory-safe language, because it had long been a source of vulnerabilities. Firefox needs to parse MP4 files from untrusted sources, and failures to correctly handle that parsing could be dire. For Mozilla, it was a small but security-critical surface area that made sense to target for a rewrite.

Another helpful tool for targeting is Kelly Shortridge’s SUX Rule¹¹: target code that is Sandbox free, Unsafe, and eXogenous. This means you should prioritize rewriting code that processes untrusted (exogenous) input, runs without a sandbox, and is written in a memory-unsafe language. Reviewing your own codebase for these areas can be a fast way to identify critical paths with high risk of exploitation in the presence of memory-safety vulnerabilities.

Wrap unsafe code with safe interfaces. When fully rewriting existing memory-unsafe code to a memory-safe language is not feasible, it might instead make sense to wrap it in a memory-safe interface. This does still lay the burden of ensuring safety properties on the programmer, both for the original code in the memory-unsafe language and for the correctness of the interface, but it then per-

mits building safe and trusted new code on top of the old code. If you continue to work to assure the old code with techniques such as fuzz testing, analysis by sanitizers, or formal modeling, you can gain increased confidence in the latent unsafe code being wrapped.

This is, in fact, how many of the Rust standard library’s common container types are written.⁴ Under the hood, they contain unsafe code to manage buffers and pointers in a way that is as efficient as possible, but the interface provided to the user does not give access to any materials (buffers, pointers, lengths) that would permit the user to violate memory-safety guarantees.

This “wrapping” approach helps constrain the “blast radius” of memory-unsafe code that cannot feasibly be removed or replaced and constrains the auditing scope and assurance costs of that code as well.

“Get good” is not a strategy. There is a common reply in conversations about memory safety, coming from the most hardcore skeptics: Programmers should just write better code. They argue, explicitly or implicitly, that programmers who benefit from the guardrails of memory safety are *bad programmers*, and that *real programmers* are sufficiently skilled that they do not need a machine double-checking their work.

Let’s be clear: This is anti-intellectual nonsense—macho self-aggrandizement masquerading as a serious technical argument. You should not take it seriously and should consider someone advancing this argument as fundamentally unserious and to be ignored.

There is no step function in quality of work in the history of human achievement that happened because people one day woke up and decided to be better at their jobs. Improvements in productivity or quality or reductions in error and harm happen because of the invention of new techniques, processes, and tools.

Reductions in traffic fatalities in the 1980s and 1990s did not happen because drivers suddenly got better at driving; they happened because states enacted mandatory seatbelt laws.

While individuals *can* become

more skilled at their jobs, working faster or producing fewer errors, *large groups* of people generally do not do so without some force that works to provide incentive or enable that change. Even when improvements are nontechnical, they come from enhancements to processes or incentives for behavior. Over the past several decades, hospitals, for example, have reduced in-hospital mistakes because of increased use of standard checklists and provisioning common materials needed for emergencies in crash carts.

Programmers who argue against memory safety by arguing for mass self-improvement are posing an impossible future as an alternative against a credible opportunity for improvements in software safety and security. While there are credible case-specific arguments against individual paths to memory safety, they do not include sudden mass improvement of skill and quality across the industry. It is important to make this clear.

Regulations and Requirements

No, governments are not banning C or C++. One specter of the conversations around memory safety is whether the use of memory-unsafe languages will become generally unacceptable, either through formal government regulation or a rise in common requirements for software purchasing.

Today no agency, either in the U.S. or outside of it, regulates against the use of languages that are non-memory safe by default. Nor are there purchasing requirements in place calling for the use of memory-safe-by-default languages or even the presence of memory-safety roadmaps, at least for governments.

The Five Eyes report mentioned previously *recommends* that organizations establish roadmaps for the pursuit of memory safety, but this is non-regulatory, and no amendments have been made to federal software acquisition policy to require such a roadmap in the U.S. or elsewhere. The U.S. is not outlawing C or C++. While these agencies have recommended moving away from these languages for future software development, they have not recommended indiscriminate mass



Programmers who argue against memory safety by arguing for mass self-improvement are posing an impossible future as an alternative against a credible opportunity for improvements in software safety and security.



rewrites of existing code.

Also note that the processes for establishing regulation or requirements would face challenges and, whether successful or not, would be slow to take effect and offer ample time for feedback and consideration.

First, regarding the prospect of regulation around memory safety in the U.S., such regulation would need to be pursued by a relevant agency that can establish a relevant jurisdiction. With the end of Chevron deference in U.S. law,¹⁰ a requirement that judges defer to U.S. regulatory agencies' determinations in most cases that was abolished in 2024 by the U.S. Supreme Court in *Loper Bright Enterprises v. Raimondo*, this pursuit of memory-safety regulation would also likely need to be explicitly backstopped by Congressional mandate to ensure it survived legal challenges where judges may overrule agency rulemaking.

Second, regarding *acquisition requirements* (the federal government's term for the rules around purchasing done by the government), the Federal Acquisition Regulation (FAR) would need to be updated to incorporate requirements for memory safety. For reference, in 2020 President Biden signed Executive Order 14028, which included a request that federal agencies pursue an amendment to the FAR to require inclusion of a software bill of materials (SBOM) for all software purchased by the federal government. To date, those changes have not been made, and no such requirement is in place within the FAR.

This is not to say that regulation or future federal purchasing requirements are impossible, but simply to point out that none are in place today, and any such changes would take time to be enacted and implemented.

The U.S. government's role around memory safety has so far been to act as a cheerleader and promoter of the idea, including the "Future Software Should Be Memory Safe" report from the Office of the National Cyber Director, "A Case for Memory Safe Roadmaps" from the Cybersecurity and Infrastructure Security Agency (CISA) et. al., and CISA's inclusion of memory-safety recommendations within its Secure by Design effort to collaborate

with industry on improving software security.

Governments do not need to ban C or C++. Even without government mandate, many organizations in the tech industry have publicly stated their support for pursuing memory safety. In 2023, the Office of the National Cyber Director put out a request for information (RFI) seeking advice from the public on how best to support improving the security of open source software. That RFI included an interest in the possibility of promoting adoption of memory-safe languages in open source.

Respondents to the RFI, which includes a number of universities, think tanks, corporations, and individuals, overwhelmingly supported a move toward memory safety. Few espoused a hardline goal of rewriting all existing code from non-memory-safe languages, but many did recognize the value of pursuing memory safety in new code, and in rewriting critical components in security-sensitive contexts when possible.

Some companies, most notably Google,⁹ have been especially vocal about their experiences with the value of memory safety. To them, the promise of memory safety is a reduction in security-related costs for long-term products such as the Google Chrome browser or the Android operating system. By reducing memory-safety vulnerabilities at the source, they shift vulnerability costs left in the software development lifecycle; the cost of catching a bug during development and stopping it from being shipped at all is orders of magnitude cheaper than the cost of receiving a security report, perhaps paying out a bug bounty, and then coordinating, preparing, and releasing a patch.

In some corners, there has been a paranoia and fear that recommendations around memory safety from the U.S. government and others portend some forced end to C or C++. Bjarne Stroustrup, originator of C++ and continued major participant in the ISO Working Group that maintains the C++ specification, has recently begun to sound alarm bells³ in papers and speeches about the existential threat posed to C++ by failing to address the demands for memory safety,

with clear reference to the possibility that software written in non-memory-safe-by-default languages may be disallowed or become practically untenable to market and sell in the future.

This fear is simultaneously overblown and correct. It is overblown to suggest the U.S. or any other government is close to outlawing C or C++, but it is correct to note that the benefits of memory safety are becoming clearer with each case study performed at scale and that we should expect natural incentives to slowly accrue larger use and developer interest in memory-safe languages over non-memory-safe languages. C and C++ will not die, but they will likely decline and become legacy languages like Cobol or Ada. They will still sustain some degree of interest and community, and a smaller number of developers will likely continue to be able to make their careers as developers in these languages, but they will be languages that present developers with fewer labor-market opportunities in the future and are unlikely to ascend in popularity and use again without substantial changes to address these safety deficiencies.

Safety Is Worth Pursuing

Memory-safe languages present the clearest opportunity today to substantially improve software security. While memory safety does not eliminate all classes of software weaknesses, it does eliminate a particularly pernicious class that leads to disproportionately severe vulnerabilities. While there are other techniques for addressing these kinds of weaknesses (for example, hardware-based approaches such as CHERI), they are less mature and generally more difficult to adopt at scale.

The state of possibility with memory safety today is similar to the state of automobile safety just prior to the widespread adoption of mandatory seatbelt laws. As car manufacturers began to integrate seatbelts as a standard feature across their model lines and states began to require that drivers wear seatbelts while driving, the rate of traffic fatalities and severity of traffic-related injuries dropped drastically. Seatbelts did not solve automobile safety, but they credibly improved it and at remarkably low cost.

The same can be done with memory safety. There is an opportunity to make substantial inroads at addressing a serious class of vulnerabilities while also, long-term, saving money on the development and operation of software systems. Memory safety is not a silver bullet, but it is a credible and cost-effective assurance technique that we as an industry should pursue aggressively. We do not need to wait for regulation to catch up; it is in our best interests to act today. **□**

References

1. Austin, T.M., Breach, S.E., and Sohi, G.S. Efficient detection of all pointer and array access errors. *ACM SIGPLAN Notices* 29, 6 (June 1, 1994), 290 - 301.
2. Brandon, J. How (memory) safe is Zig? (Sept. 21, 2022); <https://tinyurl.com/2443cr4d>
3. Claburn, T. C++ creator calls for help to defend programming language from 'serious attacks'. *The Register* (Mar. 2, 2025).
4. Example: Implementing Vec. *The Rustonomicon*; <https://tinyurl.com/yTcfyoxb>
5. Giles, R. Deploying Rust in a large codebase. *Mozilla Tech* (Feb. 21, 2017); <https://tinyurl.com/z677yoo>
6. Hardy, A. We decided to rewrite our software. You won't believe what happened next! *Adobe Tech Blog* (Feb. 25, 2019).
7. Hicks, M. What is memory safety? *The Programming Languages Enthusiast* (July 21, 2014).
8. Klabnik, S., Nichols, C., Krycho, C., et al. Unsafe Rust. *The Rust Programming Language*; <https://tinyurl.com/yqfncvx5>
9. Rebert, A. and Kern, C. Secure by design: Google's perspective on memory safety. *Google Security Blog* (Mar. 4, 2024); <https://tinyurl.com/2cgqwsrm>
10. Rozenshtein, A.Z. Lawfare Daily: The Supreme Court takes the bait: Loper Bright and the future of Chevron deference. *Lawfare* (July 12, 2024).
11. Shortridge, K. The SUX rule for safer coding. *KellyShortridge.com* (Oct. 10, 2023); <https://tinyurl.com/22n5xhwk>
12. The Sonos app fiasco: How a great audio brand nearly ruined its reputation. *The Verge* (July 23, 2025); <https://tinyurl.com/26se6m8o>
13. Watson, R.N.M. et al. It is time to standardize principles and practices for software memory safety. *Commun. ACM* 68, 2 (Jan. 22, 2025), 40-45.
14. Vander Stoep, J. and Rebert, A. Eliminating memory safety vulnerabilities at the source. *Google Online Security Blog* (Sept. 25, 2024); <https://tinyurl.com/28dxx3k8>

Andrew Lilley Brinker (abrinker@mitre.org) is a principal engineer at MITRE, where he works on software security. He contributes to the CVE Quality Working Group, serves on the OmniBOR Core Team, and leads development of Hipcheck.

Acknowledgments Thank you to Steve Klabnik, engineer at Oxide Computer Company and former member of the Rust Core Team, and to Michael Chernicoff, senior software engineer at MITRE, for reviewing this article and providing feedback prepublication.

Disclaimer Approved for public release; distribution unlimited. Public Release Case Number 25-1514. The author's affiliation with The MITRE Corporation is provided for identification purposes only and is not intended to convey or imply MITRE's concurrence with, or support for, the positions, opinions, or viewpoints expressed by the author.

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

© 2026 Copyright held by the owner/author(s).

NOW OPEN FOR SUBMISSIONS

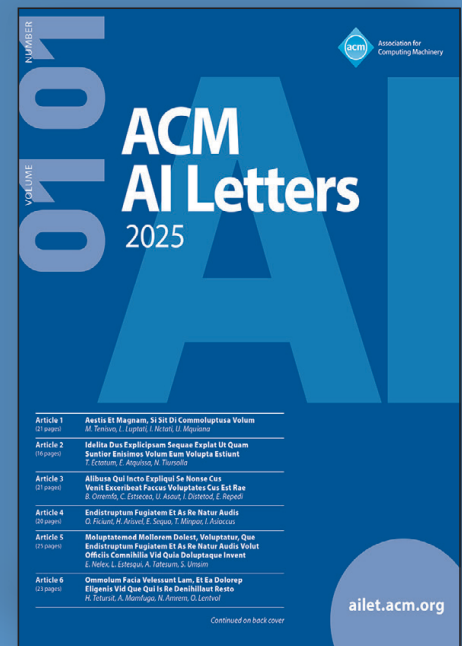
ACM AI Letters (AILET)

Editors-in-Chief

Nitesh Chawla, *University of Notre Dame, USA*

Barry O'Sullivan, *University College Cork, Ireland*

Richa Singh, *IIT Jodhpur, India*



A new venue for rapid publication of impactful, concise, and timely communications in AI

ACM AI Letters (AILET) is a new venue for rapid publication of impactful, concise, and timely communications in AI. Bridging a crucial gap between traditional conferences and journals, AILET will feature short peer-reviewed contributions that accelerate knowledge dissemination across academia and industry. This unique publication prioritizes theoretical breakthroughs, algorithmic innovation, practical real-world applications, and critical societal implications, including ethics, policy, and responsible AI. It also introduces a distinctive space for rigorously reviewed opinion pieces and policy briefs, promoting swift engagement with contemporary issues shaping the AI landscape.

ACM AI Letters welcome concise summaries of work in the following areas:

- Theoretical AI: Reports on theoretical breakthroughs in AI
- Algorithmic Advances: Descriptions of significant algorithmic and scientific advances in AI
- Practical Applications: Accounts of novel or deployed applications of AI in real-world settings such as healthcare, finance, robotics, and autonomous systems
- AI & Society: Reports on AI applications addressing key societal challenges, such as the United Nations Sustainable Development Goals; responsible AI; AI governance; AI ethics; AI policy
- Emerging Topics: Areas such as explainable AI; interdisciplinary applications of AI, such as AI & science, AI & medicine, focusing on AI implications in practical implementations
- Opinions and briefs: Latest advances, research highlights, expert commentary on emerging areas, comparative assessments

ACM AI Letters is now open for submissions. In keeping with ACM's goal of making all its publications open access, all papers will be published open access, with no publication charges for the first three years.

For more information, please visit ailet.acm.org



Association for
Computing Machinery

DOI:10.1145/3750037

Relationships with “digital humans” might help people feel less lonely. But will this technology increase social isolation?

BY ROBERT SPARROW AND JAMES BROWN

Against Imaginary Friends: Why Digital Companions Are No Solution to Social Isolation

LONELINESS AND SOCIAL isolation constitute major threats to public health.¹⁵ *Homo sapiens* is a gregarious species and the nature and extent of individuals’ relationships with other human beings have a significant influence on their psychological and physiological health.⁷ Lack of contact with other people increases both mortality and morbidity.^{2,25} Social isolation refers to an objective lack of social connections, while loneliness is subjective dissatisfaction with the number and quality of

social connections one has.⁶ Both are on the rise as a result of a complex of interrelated factors, including aging populations, changes in the family structure associated with lower birth rates and increased economic opportunities for women, increased labor force mobility, changes in urban density, an increase in the number of people living with chronic conditions that restrict mobility, rising rates of anxiety and depression, and policy responses to pandemics.⁸

Given that the increase in loneliness and social isolation over the past several decades appears to be driven by economic and demographic dynamics that are difficult to reverse, there is substantial interest, from both governments and researchers, in technological solutions to both problems.^{20,21,25} For several decades now, engineers have been working to develop social robots for eldercare settings, in the hope that this might help combat the loneliness and social isolation that are endemic in such settings. This project has been criticized as disrespectful—even harmful—to older persons.^{4,36,38} It has also, to date, failed because social robots remain expensive, have limited capacities, and are unable to sustain extended engagements with most users.^{14,16,41} However, a new technology, “digital companions,” looks more plausible

» key insights

- A new technology, “digital humans,” offers a realistic possibility of reducing loneliness among socially isolated older persons, where social robots have failed.
- Encouraging people to mistake digital companions for real friends is *prima facie* unethical. Encouraging people to have imaginary friends is no solution to social isolation.
- There is a risk that the development of digital companions will work to increase social isolation.
- Proposals to use digital companions to combat loneliness and social isolation underestimate the importance of touch, physical companionship, and mutual aid for human well-being.

ILLUSTRATION BY PETER CROWTHER. USING PHOTO BY JOHNER IMAGES/GETTY IMAGES



as a potential solution to loneliness and social isolation from a technological perspective, with the prospect of realistic-looking computer-generated characters capable of convincing natural language conversation just around the corner.

In this article, we extend the criticisms that have been made of social robots to digital companions and identify new dangers associated with the use of digital companions. It is possible that relationships with digital companions will allow people to feel less lonely. However, this will almost certainly require that users allow themselves to be deceived about the nature of their digital companion. Encouraging people to mistake imaginary friends for real friends is *prima facie* unethical. More importantly, providing a digital companion to someone who lives alone does not mean that they are no longer alone: They remain socially isolated, albeit with a digital companion to entertain them. Encouraging people to have imaginary friends is no solution to social isolation. Indeed, there is a real danger that this technology will instead lead to people having even less contact with other human beings. Proposals to use digital companions to combat loneliness and social isolation also underestimate the importance of touch, physical companionship, and mutual aid when it comes to human well-being, especially in later life.

We have deliberately limited the scope of our discussion here to the use of digital companions to address loneliness and social isolation. However, many of the arguments below may also be relevant to the ethics of other uses of “digital humans” (see below) and/or to applications of AI that might lead to people substituting relationships with machines for relationships with human beings.^a

Two Technological “Solutions” to Loneliness and Social Isolation

There is an emerging school of thought that claims that some tech-



Providing a digital companion to someone who lives alone does not mean that they are no longer alone: They remain socially isolated, albeit with a digital companion to entertain them.



nologies—especially computers, mobile phones, and social media—play a role in *generating* loneliness and social isolation.⁴³ Nevertheless, new technologies are regularly touted as having the potential to *reduce* loneliness and social isolation.

Social robots for eldercare? The harms associated with loneliness and social isolation have been mobilized as a justification for the development of robots since at least the early 1990s. Moved by fears that the world—especially wealthy “northern” societies in Europe and North America—is facing a “demographic tsunami,” engineers started out trying to build “robot butlers” to help older people remain in their homes longer.^{14,38} Unfortunately, this project proved too difficult, with the results being uncompetitive with other assistive technologies.^{14,33} Perhaps relatedly, some engineers pivoted to the notion that robot pets, or other artificial companions, might serve older people by playing a role in combatting loneliness and social isolation.³⁴ This idea became especially popular subsequent to the release of Sony’s AIBO robot dog in 1999.

Unfortunately, despite the enormous amount of media—and philosophical—attention devoted to eldercare robots, the results of this project have also been underwhelming. Companion robots are typically expensive, difficult to maintain, have limited functions, and are unable to hold people’s attention for extended periods.^{14,16,41} There is little good-quality evidence that introducing robots into eldercare settings makes a meaningful difference to people’s quality of life, especially in the long term.^{30,31} What research does exist on this topic often fails to account for the presence of the research team in the eldercare setting or to include a control arm in the trial design to determine the extent to which any effects they observe are the result of the capacities of the robot rather than of the introduction of a novel artifact.¹ Our strong suspicion is that to the extent that these technologies do work, they do so by serving as a topic of conversation among human beings who are already co-present to each other and that this function can be performed

a For the argument that replacing relationships with humans with relationships with artificial intelligences in *healthcare* contexts is morally problematic, see Montemayor et al.²⁴ For the argument that it is problematic in the context of tertiary *education*, see Sparrow and Flenady.³⁷

equally well, if not better, by other, less high-tech interventions.^{19,32} That does not mean that social robots cannot play a useful role in eldercare settings, but it does suggest that they have limited utility when it comes to addressing loneliness and social isolation more generally.

Digital companions. While the “promise” of social robots remains unrealized, a new technology, “digital humans,” has recently emerged, which we judge has much more potential to engage and entertain people. Digital humans are AI-enhanced chatbots that present on screen as human and accompany their utterances with appropriate gestures, body language, and even facial expressions by means of sophisticated graphics technology.¹⁸

The advent of ChatGPT and other generative AI systems based on large language models (LLMs) means that chatbots are now capable of sustaining extended natural language conversations on almost any topic via a screen and keyboard. Advances in speech recognition and speech synthesis now allow users to have spoken natural-language conversations with digital humans.^b Although the digital humans available today are very obviously animations, advances in both the software and hardware used to generate images on computers hold out the prospect of photo-realistic animations of human faces and bodies. Hollywood films already feature animated characters that are hard to distinguish from human actors; new AI “deepfake” technology means that it is becoming easier to generate such animations. Finally, by training an AI on photos, video, audio, and text generated by an existing or historical individual, it is possible to create a digital avatar of that person that looks and sounds like them. Eventually it may be hard to tell whether the “person” one is talking with on Skype, Zoom, Teams, Facetime, or some other video chat application is a human being or a sophisticated AI chatbot.

^b At the time of writing, it was possible to experience a conversation with an impressive digital human by visiting the website: <https://www.digitalhumans.com/>.

Unlike social robots, digital humans are cheap and easily scalable, and can be easily maintained and quickly updated. In the not-too-distant future they are likely to be ubiquitous in online customer-service roles.

Predictably, this technology is being hailed for its potential to combat loneliness and even, according to some scholars, social isolation.^{17,18,23} Following Sherry Turkle,⁴² we will call digital humans employed for this purpose *digital companions*.^c Already, a large number of people are interacting with digital companions via an application called Replika, which allows users to customize an avatar of a human being to be their digital companion and talk with it on a range of topics.^{22,28}

Imaginary Friends

Any adequate account of the prospects for digital companions must begin by acknowledging that actually existing—as opposed to science fictional—AI systems are not sentient, do not have real emotions, and do not care about the people with whom they are interacting. While this remains true, all that digital companions can offer is an imitation of a relationship with another sentient being rather than the real thing. Friendship is a relationship between two people. At best, then, non-sentient digital companions can only be ersatz, or imaginary, friends.

That existing digital humans cannot really be friends with people because they are not real humans might seem too obvious to point out. Nevertheless, it is worth pointing out because of a tendency in the literature on AI and social robots to assume that if people cannot tell the difference between two things, then they must have the same properties and, by implication, must therefore be the same. This line of thought runs deep

in the history of discussions of computer-human relations—back to the Turing Test at least—and continues to be influential today.⁵ Yet the fact that we cannot tell two things apart does *not* establish that they are the same—we may just be poorly placed to perceive the difference. Even if those relating to digital humans cannot tell them apart from real people on a video call, it will remain obvious that digital humans are not real people due to their inability to occupy physical space outside of a smartphone or computer. For the moment, despite occasional high-profile defections from this consensus, it remains obvious to those who design chatbots that they are not sentient and do not have real emotions.¹⁰ Even when this ceases to be obvious, this alone will not be sufficient to establish that machines are sentient or have real emotions: It may simply mean that we have reached the limits of our capacities to see that they are not and do not.

The “Turing Triage Test,” which asks when it might be justifiable to let a human being die in order to save the “life” of a machine, suggests that the intellectual and moral costs of claiming that machines are sentient would be substantial.³⁵ It also suggests that what would need to be true of machines, and our relationships with them, in order to render the claim that they are sentient plausible is much more demanding than is usually recognized. Should AIs ever “wake up” and become sentient, then the ethics of our relationships with them will change radically. Presumably it would be possible to be “friends” with a sentient machine and for sentient AIs to care about people. Quite why a sentient AI would want to spend its time entertaining and caring for lonely and socially isolated humans is another matter. Moreover, programming or creating sentient machines with a desire to serve as digital companions for humans would raise profound ethical issues—if it would be possible at all. In any case, our concern here is solely with digital companions that are *not* sentient.

May reduce loneliness. We believe that, unlike existing social robots,

^c Strictly speaking, the digital humans that concern us here are a subset of those entities that Turkle describes as “digital companions” because Turkle rightly allows that digital animals might also be used for this purpose and because Turkle includes robots in this category. Our concern is with “purely” digital companions and only with those that present themselves as human.

digital humans may succeed in engaging and entertaining people for lengthy periods. After all, eventually it will be possible to experience, by means of a digital human, a simulacrum of any interaction that one can have with a real human online. To the extent that social interactions via videoconferencing allow people to feel less lonely, we anticipate that at least some people will feel less lonely as a consequence of interactions with digital companions.

The caveat here is not insignificant. The impact of videoconferencing, as with social media more generally, on social isolation in particular but also on loneliness, remains controversial.³ There is some evidence that people need contact with other human beings in real life in order to feel happy and that virtual interactions may attenuate rather than strengthen social connections.^{12,43}

Nevertheless, interactions with digital companions might meet some of the social and psychological needs that people ordinarily satisfy through social contact with people.^d More prosaically, they might reduce loneliness by entertaining users so that they miss social contact less. Anecdotal reports suggest that Replika allows some users to feel less lonely²² and there is limited empirical evidence that other digital companions may have an impact on loneliness.^{9,23,29,39} More sophisticated versions of these technologies might be even more successful in this regard. Given the distress and harms associated with loneliness, this is a significant virtue of this technology.

Do not reduce social isolation. Although interactions with digital companions might reduce loneliness, in and of themselves they will have no impact on social isolation. The qualification “in and of themselves” is intended to highlight the important caveat that, as we discuss below, digital companions could be designed to encourage users to strengthen

their existing social connections, and form new connections, with human beings. However, providing a person who lives alone, without any meaningful social connections, with a digital companion will not, in itself, make them any less socially isolated. Imaginary friends do not count when considering the number of social connections an individual has.

As social facilitators? Thus far, we have been concerned with the extent to which interactions with digital companions themselves may impact on loneliness and social isolation. Another way that digital companions might combat these problems is if they were designed to encourage users to form and/or strengthen social connections with real people, by, for instance, reminding them to wish their relatives happy birthday or encouraging them to participate in social activities that might facilitate social connections.^{9,18} Ideally, digital companions would be designed to help users develop sufficient, and sufficiently rewarding, social relationships such that they no longer wanted to interact with a digital companion.^e

Unfortunately, we suspect that the designers and manufacturers of digital companions will have strong incentives to do otherwise. In order to encourage people to pay for access to digital companions, or to monetize the time that users spend interacting with digital companions by displaying advertisements to users, designers of these systems will need to make them as engaging—indeed as “addictive”—as possible. For instance, Replika is gamified extensively to try to increase the amount of time users spend on the application and also the amount they will pay to access various special features that are not available to users of the free version of the app. Although there are some markets—for instance, for pharmaceuticals or wound-care products—where companies flourish by selling products that are designed to eliminate

the need for the product, absent very strong regulation, the market for digital companions, which is closely adjacent to the market for video games and other entertainments, is highly unlikely to be one.

Lessons from the Literature on Social Robots

Because engineers have been trying to build social robots to combat loneliness and social isolation in older citizens for at least three decades, there is a correspondingly large literature on the ethical issues associated with this project. We believe that many of the criticisms of social robots for eldercare have force against the development and use of digital companions to combat loneliness and social isolation. For reasons of space, though, we focus here on only four lines of criticism.

Designing to deceive. One powerful criticism of social robots for eldercare settings is that they are premised on users believing them to have properties that they do not.^{4,32,34,43} In particular, in order to sustain interactions with such artifacts over long periods, they must believe that the robot has emotional states and cares about how it is treated. Designing such artifacts means designing them so that users are deceived about their capacities. Such “designing to deceive” is *prima facie* unethical.^{32,38}

This criticism has obvious force against the design of digital companions. At the time of writing, the front page of the desktop version of Replika features the text, “The AI companion who cares. Always here to listen and talk. Always on your side.”²⁸ None of these claims are true. Replikas do not care about anyone or anything—they are not sentient and thus, by definition, do not feel anything at all. They do not really listen to users or say anything: They respond to keyboard or audio prompts with a pastiche of words generated using complex statistical techniques applied to the content of the World Wide Web. Nor are Replikas on anyone’s side—they do not care at all about any humans. Replika’s designers are lying about its nature. They are doing so presumably because users need to believe their Replika cares about them in order to

d Although see Perry²⁷ for an argument that AI is in principle incapable of realizing emotional or motivational empathy, which in turn suggests that it may not be capable of meeting people’s need for empathic care when they are distressed or otherwise experiencing strong emotions.

e Digital companions might also be designed to reduce loneliness that results from pathologically unrealistic expectations of social connections by serving as therapists in order to mitigate these expectations.

care about it enough to keep using it for an extended period.

Robots and respect. A second criticism, which, in part, follows from the first, is that the project of designing social robots for eldercare settings is disrespectful to older people. References to aging populations or “demographic tsunamis” in the motivations for papers or research on eldercare robots treat the presence of older people in the population as itself a problem to be solved. Robots are postulated as the solution, despite the fact they would not be considered desirable were they to be directed toward the needs of younger people. Investment in research into robots continues, nonetheless, because older people are almost never involved in setting research priorities around eldercare or in the design of technological solutions to problems faced by older people.³⁸

It might be argued that these criticisms have less force against digital companions than they do against social robots for eldercare because digital companions are typically advertised as intended for the lonely and socially isolated in general, not just for older persons. However, this larger cohort, “lonely and socially isolated persons,” plays essentially the same role in the literature on digital companions as “older persons” does in the literature on social robots, and the attitudes this literature expresses toward this cohort are problematic for the same reasons. Those experiencing loneliness and/or social isolation are implicitly “othered” in discussions of digital companions, treated as a problem to be solved. Those designing, developing, or touting digital companions as cures for loneliness and social isolation seldom, if ever, suggest that they themselves aspire to a life lived alongside only machines. Almost by definition, lonely and socially isolated people play little role in discussions about how—or, indeed, whether—this technology should be developed.

Replacing real relationships. A third central concern in the debate about the ethics of using social robots for eldercare is the extent to which this might lead to older persons being deprived of opportunities



Imaginary friends may outcompete real friends for our time and attention by appearing, superficially, to be better friends.



for relationships with people.^{20,21,36,38} The claim that robots are a solution to the problem of a shortage of eldercare workers implies that robots will be able to take on roles in the eldercare sector currently performed by humans. Yet every interaction that older people have with a robot is one less opportunity for them to interact with a human. The economics of the eldercare sector strongly suggest that money saved by replacing people in existing roles with robots will not be reinvested in roles that might provide people with more human contact.³⁸ Perversely, were this technology to be fully realized, there is every chance that it would lead to a substantial increase in social isolation.

We believe that the risk that ersatz companions might increase social isolation is even more pressing when it comes to digital companions, in part simply because we think digital companions may work to reduce loneliness whereas social robots are unlikely to.³⁴

Imaginary friends may outcompete real friends for our time and attention by appearing, superficially, to be better friends.^f Digital companions have a number of advantages relative to human beings when it comes to their capacity to sustain engagement with users. Most obviously, they can be available any time, day or night, 365 days a year. Moreover, unlike real people, their sole purpose can be to satisfy the user’s desires. They will never be rude, selfish, thoughtless, or judgmental—unless their being so will itself contribute to increased engagement.^g Thus, not only might online interactions with digital com-

f As an anonymous referee encouraged us to acknowledge, one circumstance in which people might be better off with a relationship with a digital companion than a human being is when the relationship with the human being was an abusive or immiserating one. In such cases, though, it is not that the digital companion would be a better friend—as we have argued here, they would not be a friend at all—but rather that the person concerned would actually be better off being more socially isolated than maintaining the relationship with the abuser.

g The role played by the “difficult” behavior of pets in establishing our love for them, as well as the fact that the designers of robot pets aim to have their creations behave “badly” on occasion for the same reason, is noted in Sparrow.³⁴

panions crowd out real relationships from users' lives, but they might also change what users expect from their friends to the point that relationships with real people are seen as unsatisfying in comparison.^{18,43}

Importantly, that relationships with digital companions might substitute for relationships with real people is an assumption of the argument that digital companions offer a potential technological solution to loneliness and social isolation: If they cannot, then digital companions *will not* reduce loneliness or (less plausibly) social isolation, as their proponents hope. Moreover, it is hard to see how digital companions might be engaging enough to be emotionally rewarding (and thus cause people to feel less lonely) but not so compelling as to compete with relationships with real people. More fundamentally, insofar as relationships with digital companions reduce loneliness, they will reduce users' motivation to seek out relationships with real people. This means that the presence of digital companions in a community is likely to increase the social isolation of third parties as well as users of digital companions.

The advent of digital companions may also exacerbate social isolation by reducing the pressure on health-care systems and government agencies to address the economic, social, and technological causes of loneliness and social isolation. That is, the development of digital companions will make it easy for governments and owners of eldercare facilities to *appear* as though they are doing something about social isolation by providing citizens, or clients, with digital companions, even though these technologies will mostly leave social isolation intact. If people do in fact feel less lonely as a result of having imaginary friends, they may be less likely to demand that governments and/or the owners of eldercare facilities provide them with opportunities to establish social connections with real people.

What imaginary friends cannot do. Fourth, and finally, these products are dangerous by virtue of their potential to deprive users of important objective goods. Being happy is not



If people do in fact feel less lonely as a result of having imaginary friends, they may be less likely to demand that governments and/or the owners of eldercare facilities provide them with opportunities to establish social connections with real people.



enough to establish that one is leading a flourishing human life: If it were, one could treat loneliness and social isolation purely pharmaceutically. It is much more plausible to hold that a good human life includes various objectively good things, including, for instance, health, beauty, love, friendship, and social affirmation. As this list suggests, many of these goods are associated with relationships with those around us and, as such, cannot be secured through interactions with non-sentient machines. In particular, social robots are not able to provide people with the respect and social recognition that affirm individuals' status as members of a community and that are important components of well-being according to any plausible theory of well-being.³⁶ These criticisms also have force against digital companions, which can provide only shallow imitations of these goods rather than the real thing.

Some further limitations of digital companions. Although digital companions are more engaging than existing social robots, they suffer in comparison to social robots when it comes to the extent to which they can satisfy other important human needs—for touch, physical companionship, and for care expressed through assistance with physical tasks.

Human beings have a powerful innate need for creaturely touch. Skin to skin (or fur) contact with other humans or animals induces a pronounced physiological response in human beings.¹³ There is now a large literature exploring the health benefits of relationships with animals⁴⁵ as well as the harms associated with the absence of physical intimacy.⁴⁰ Touch is also ethically significant. In particular, care is often expressed through touch.

The importance of touch has not gone unnoticed by the designers of social robots, which now include a number of robot animals explicitly advertised as allowing users to “pet” them and to supposedly benefit from doing so.^{11,44} The famed Japanese roboticist Hiroshi Ishiguro has even made a robot that is supposed to “transmit” hugs provided remotely by another human being.²⁶ We are skept-

tical of the extent to which touching machines produces the benefits associated with touching people or animals, but it is clear that this is a domain where (purely) digital companions have nothing to offer.

Another way in which digital companions are lacking is that they are unable to provide users with physical companionship more generally. Human beings respond to each other's presence as well as to each other's touch.²⁰ As the dissatisfying experience of Zoom or Teams calls makes all too clear, the mere fact of other people being physically present with us makes a difference. Physical copresence also plays a role in the formation of memories and narratives that play a key role in social relationships: Going through something together helps people form social bonds. Granted, digital companions and their users may come to have a shared history of conversation and online "experiences." Digital companions could, for instance, hark back to conversations they have had with a user, games they have played together, and other online experiences that they have shared, in order to affirm and strengthen their "relationship." However, there is a clear limit to the capacity of digital companions to share experiences and provide companionship here.

Finally, and relatedly, the provision of help with real-world tasks is an important mark and component of both solidarity and care. Helping the people we love, for instance, by assisting them in walking up stairs, bringing them food, or bathing them, is an expression of our love for them. When we help each other in this way—"mutual aid"—we affirm our recognition of our mutual worth. Digital companions are obviously severely limited in what they can contribute in this regard. Again, this is not to deny that digital companions could provide some forms of help to their users. A sophisticated digital companion might, for instance, make a doctor's appointment for a user, remind them to take a list of the medications that they are currently on to the appointment, provide advice on what time they need to leave home to get there, order a taxi for them, and

notify them when it arrives. However, those who only have digital companions will be lost if they need someone to help them weed the garden or hold their hand to help them out of a car. Although such tasks remain beyond the current capacities of robots, it is at least conceivable that they might one day assist in this way.

Discussions of the prospects for digital companions have neglected the role played by touch, physical companionship, and mutual aid in promoting human well-being.^h Digital companions will not be able to satisfy people's need for these important goods. If the availability of digital companions leads to reductions in the number of real friends that people have, as we have argued is highly likely, then this technology will leave people worse off by virtue of depriving them of opportunities to experience touch or human companionship or to help each other with physical tasks, as well as depriving them of recognition and respect. Insofar as the absence of these goods causes people to feel lonely, the inability of imaginary friends to provide them may also affect the extent to which digital companions reduce loneliness.

Privacy, Representation, Manipulation, and Complicity

We have been concerned here with arguments that bear on the question of whether digital companions should be used to combat loneliness and social isolation at all. Inevitably, the use of digital companions would raise a further set of questions about their design and operation. In particular, issues about the privacy of users loom large given the intimate nature of the data that would be collected by digital companions.¹⁸ The design of digital companions raises issues in media ethics, about the representation of people of different races and genders, and about the responsibility of creators for the impact of the art they create. Digital companions will offer unparalleled opportunities to shape the personalities of—indeed, to manipulate—their users, which suggests that their design would need to be strictly regulated.¹⁸ Conversely, in some circumstances, the failure of a digital companion to intervene

when a user threatens other people or expresses sexist, racist, or otherwise morally reprehensible views, may itself make its designers culpable. Again, the boundaries of whether, when, and how digital companions should intervene in response to troubling statements by, or patterns of behavior from, their users will need to be determined by the broader community. Unfortunately—or perhaps fortunately—these issues are beyond the scope of our discussion here.

Conclusion

Rising rates of loneliness and social isolation in industrialized societies constitute a "wicked problem": It is little wonder that engineers and policymakers are tempted by the idea that there might be a technological solution. Digital companions seem to have several advantages over the social robots that have previously been touted as having the potential to reduce these evils. We have argued that, nevertheless, digital companions can be, at most, imaginary friends, which can do little, if anything, to address social isolation, even if they do reduce loneliness in some cases. We have shown that powerful criticisms in the literature on the ethics of the design and use of social robots for eldercare settings also apply to digital companions. In particular, there is a real risk that the development of digital companions will actually work to increase social isolation. We have also suggested that social robots might even have a number of advantages relative to (purely) digital companions that the literature on the latter topic has thus far neglected.

The implications of these observations for the larger debate about the wisdom of deploying digital companions are more complex than might first appear. It is loneliness—rather than social isolation per se—that makes people miserable. Any technology or policy that might reduce loneliness therefore deserves to be taken seriously. If, in particular cases, the choice is really between an individual having no companions or a digital companion, they may well be better off with a digital companion that allows them to feel less lonely. Yet social isolation is also arguably

objectively bad, as well as being bad for people's health, even when they are not lonely. Although some of the effects of social isolation on human health are probably the product of the loneliness that typically accompanies social isolation, others probably are not. For instance, those who are socially isolated are less likely to be encouraged by those around them to seek medical help and less likely to benefit from the help with daily living that other people can provide.

The prospect of a technology that might reduce loneliness but increase social isolation, as we have argued is likely true of digital companions, therefore forces us to confront difficult trade-offs. Should we reduce people's misery at the cost of making them objectively worse off by virtue of being more socially isolated? Is it more important to express respect for people or to promote their happiness? Will the development of digital companions lead policymakers to be less motivated to pursue policies intended to reduce social isolation? Answering these questions will require paying close attention to the impact of this technology on the experience and relationships of users and those around them. It will also require us to confront deep ethical and philosophical questions: about what we owe to each other, about what it is to lead a good human life, and about the sort of world we want to create with AI. A future in which people spend all their time at home with their imaginary friends without ever touching, or being touched by, another human being is, we submit, a future to be avoided. Whether it will be possible to develop and use digital companions to combat loneliness and social isolation without bringing about such a world remains to be seen.

Acknowledgments

The authors would like to thank Simon Coghlan for suggestions as to relevant sources in the literature on human-animal relations. Professor Sparrow is an associate investigator in the Australian Research Council Centre of Excellence for Automated Decision-making and Society (Grant number CE200100005) and worked on this article in this role. 

References

- Bemelmans, R., Gelderblom, G.J., Jonker, P., and De Witte, L. Socially assistive robots in elderly care: A systematic review into effects and effectiveness. *J. Am. Med. Dir. Assoc.* 13, 2 (2012), 114–120.
- Cacioppo, J.T. and Patrick, W. *Loneliness: Human Nature and the Need for Social Connection*, 1st ed. W.W. Norton & Co. (2008).
- Choi, N.G., Hammaker, S., DiNitto, D.M., and Marti, C.N. COVID-19 and loneliness among older adults: Associations with mode of family/friend contacts and social participation. *Clin. Gerontol.* 45, 2 (2022), 390–402.
- Coghlan, S., Waycott, J., Lazar, A., and Neves, B.B. Dignity, autonomy, and style of company. *Proc. ACM Hum. Comput. Interact.* 5, CSCW1 (2021), Article no. 104.
- Danaher, J. Welcoming robots into the moral circle: A defence of ethical behaviourism. *Sci. Eng. Ethics.* 26, 4 (2020), 2023–2049.
- Donovan, N.J. and Blazer, D. Social isolation and loneliness in older adults: Review and commentary of a National Academies report. *Am. J. Geriatr. Psychiatry* 28, 12 (2020), 1233–1244.
- Dunbar, R.I. Why friendship and loneliness affect our health. *Ann. N. Y. Acad. Sci.* 1545, 1 (2025), 52–65.
- Feng, X. and Astell-Burt, T. Lonelygenetic environments: A call for research on multilevel determinants of loneliness. *Lancet Planet Health* 6, 12 (2022), e933–e934.
- Gasteiger, N., Loveys, K., Law, M., and Broadbent, E. Friends from the future: A scoping review of research into robots and computer agents to combat loneliness in older people. *Clin. Interv. Aging* 16, (2021), 941–971.
- Grant, N. and Metz, C. Google sidelines engineer who claims its A.I. is sentient. *New York Times* (Jun. 12, 2022); <https://www.nytimes.com/2022/06/12/technology/google-chatbot-ai-blake-lemoine.html>
- Gustafsson, C., Svanberg, C., and Müllersdorf, M. Using a robotic cat in dementia care: a pilot study. *J. Gerontol. Nurs.* 41, 10 (2015), 46–56.
- Hall, J.A., Dominguez, J., and Mihailova, T. Interpersonal media and face-to-face communication: Relationship with life satisfaction and loneliness. *J. Happiness Stud.* 24, 1 (2023), 331–350.
- Handlin, L. et al. Short-term interaction between dogs and their owners: Effects on oxytocin, cortisol, insulin and heart rate—an exploratory study. *Anthrozoös* 24, 3 (2011), 301–315.
- Higgs, P. and Gilleard, C. Techno-fixes for an ageing society. *Aging Ment. Health* 26, 7 (2022), 1303–1305.
- Holt-Lunstad, J. et al. Loneliness and social isolation as risk factors for mortality: A meta-analytic review. *Perspect. Psychol. Sci.* 10, 2 (2015), 227–237.
- Hung, L. et al. The benefits of and barriers to using a social robot PARO in care settings: A scoping review. *BMC Geriatr.* 19 (2019), Article no. 232.
- Jecker, N.S. My friend, the robot: An argument for e-friendship. In *2021 30th IEEE Intern. Conf. on Robot & Human Interactive Communication*. IEEE (2021), 692–697.
- Jecker, N.S., Sparrow, R., Lederman, Z., and Ho, A. Digital humans to combat loneliness and social isolation: Ethics concerns and policy recommendations. *Hastings Cent. Rep.* 54, 1 (2024), 7–12.
- Johnson, E. Pepper as imposter. *Sci. Technol. Stud.* 37, 3 (2024), 62–70.
- Lederman, Z. Technological solutions to loneliness—are they enough? *Bioethics* 37, 33 (2023), 275–284.
- Lederman, Z. and Jecker, N. Social robots to fend off loneliness? *Kennedy Inst. Ethics. J.* 33, 3 (2023), 249–276.
- Loffhagen, E. 'I'm sick and tired of dating actual people'—Meet the men falling in love with their AI girlfriends. *The Standard* (Aug. 9, 2023); <https://www.standard.co.uk/insider/rise-of-ai-chatbot-girlfriends-replika-b1098144.html>
- Loveys, K., Sagar, M., Pickering, I., and Broadbent, E. A digital human for delivering a remote loneliness and stress intervention to at-risk younger and older adults during the COVID-19 pandemic: Randomized pilot trial. *JMIR Ment. Health* 8, 11 (2021), e31586.
- Montemayor, C., Halpern, J., and Fairweather, A. In principle obstacles for empathic AI: Why we can't replace human empathy in healthcare. *AI Soc.* 37, 4 (2022), 1353–1359.
- National Academies of Sciences, Engineering, and Medicine. Social isolation and loneliness in older adults: Opportunities for the health care system. The National Academies Press (2020).
- Ogawa, K. et al. Telenoid: Tele-presence android for communication. In *ACM SIGGRAPH 2011 Emerging Technologies*. ACM (2011).
- Perry, A. AI will never convey the essence of human empathy. *Nat. Hum. Behav.* 7, 11 (2023), 1808–1809.
- Replika. The AI companion who cares. Always here to listen and talk. Always on your side (2023); <https://replika.ai/>
- Ring, L., Shi, L., Totzke, K., and Bickmore, T. Social support agents for older adults: Longitudinal affective computing in the home. *J. Multimodal User Interfaces* 9, 1 (2015), 79–88.
- Scoglio, A.A.J., Reilly, E.D., Gorman, J.A., and Drebing, C.E. Use of social robots in mental health and well-being research: Systematic review. *J. Med. Internet Res.* 21, 7 (2019), e13322.
- Servaty, R. et al. Implementation of robotic devices in nursing care. Barriers and facilitators: An integrative review. *BMJ Open* 10, 9 (2020), e038650.
- Sharkey, A. and Sharkey, N. We need to talk about deception in social robotics!. *Ethics Inf. Technol.* 23, 3 (2021), 309–316.
- Sorell, T. and Draper, H. Robot carers, ethics, and older people. *Ethics Inf. Technol.* 16, 3 (2014), 183–195.
- Sparrow, R. The march of the robot dogs. *Ethics Inf. Technol.* 4, 4 (2002), 305–318.
- Sparrow, R. The Turing triage test. *Ethics Inf. Technol.* 6, 4 (2004), 203–213.
- Sparrow, R. Robots in aged care: A dystopian future?. *AI Soc.* 31, 4 (2016), 445–454.
- Sparrow, R. and Flendy, G. Bullshit universities: The future of automated education. *AI Soc.* (Apr. 2025).
- Sparrow, R. and Sparrow, L. In the hands of machines? The future of aged care. *Minds Mach.* 16, (2006), 141–161.
- Ta, V. et al. User experiences of social support from companion chatbots in everyday contexts: Thematic analysis. *J. Med. Internet Res.* 22, 3 (2020), e16235.
- Thomas, P.A. and Kim, S. Lost touch? Implications of physical touch for physical health. *J. Gerontol. B Psychol. Sci. Soc. Sci.* 76, 3 (2021), e111–e115.
- Tulli, S., Ambrosio, D.A., Najjar, A., and Rodríguez-Lera, F.J. Great expectations & aborted business initiatives: The paradox of social robot between research and industry. In *BNAIC/BENELEARN 2019: Proceedings of the Reference AI & ML conf. for Belgium, Netherlands & Luxembourg* (2019); <https://ceur-ws.org/Vol-2491/short24.pdf>
- Turkle, S. Authenticity in the age of digital companions. *Interact. Stud.* 8, 3 (2007), 501–517.
- Turkle, S. *Alone Together: Why We Expect More from Technology and Less from Each Other*. Basic Books (2011).
- Wada, K. and Shibata, T. Living with seal robots—Its sociopsychological and physiological influences on the elderly at a care house. *IEEE Trans. Robot.* 23, 5 (2007), 972–980.
- Wells, D.L. The state of research on human-animal relations: Implications for human health. *Anthrozoös* 32, 2 (2019), 169–181.

Robert Sparrow (robert.sparrow@monash.edu) is a professor of philosophy at Monash University, Melbourne, Australia.

James Brown is an independent scholar based in Melbourne, Australia.

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

© 2026 Copyright held by the owner/author(s).



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

A concept-based framework for building agents that can learn and reason across multiple domains.

BY JIAYUAN MAO, JOSHUA B. TENENBAUM, AND JIAJUN WU

Building Intelligent Agents with Neuro-Symbolic Concepts

ONE OF THE long-term goals of artificial intelligence (AI) is to build machines that can continually learn new knowledge from their experiences, ground these experiences in the physical world, and apply the knowledge to their reasoning across different tasks, modalities, and environments. The desired capability

of such agents includes, but is not limited to, describing perceived scenes, answering queries about scenes, making plans to achieve goals, and executing plans in the physical world. We want these machines to be able to learn and solve a wide variety of tasks across different environments, leveraging a feasible amount of data from multiple modalities.

In recent years, we have seen great success in neural-network-based

“end-to-end” learning methods. Still, most of them are tailored to particular tasks and environments, such as categorizing images of objects into a fixed set of labels, translating between languages, and playing video and board games. These systems are usually built on top of relatively simple and monolithic training and inference algorithms (for example, a single neural network trained by stochastic gradient descent). As a re-

» key insights

- We developed a concept-centric paradigm that supports building agents that can learn continually and reason flexibly.
- The agent acquires a vocabulary of neuro-symbolic concepts for objects, relations, and actions, represented through a combination of symbolic programs and neural network embeddings. These concepts are grounded in sensory inputs and actuation outputs and can be composed to solve novel tasks using general-purpose reasoning and planning algorithms.
- The proposed framework offers several advantages, including data efficiency, compositional generalization, continual learning, and zero-shot transfer—key properties for general-purpose AI. These advantages have been demonstrated across applications in vision, language, and robotics.

sult, their success in domain-specific applications relies on the availability of large-scale datasets and computational resources on the particular task of interest. However, annotating high-quality data for reasoning, planning, and control in visual and physical domains is usually labor-intensive

and, in some cases, prohibitively costly. Therefore, there has been limited success in extending these methods to building embodied generalist agents across domains.

In this article, we emphasize the role of *concepts* in learning and reasoning. We draw on rich traditions from philosophy and cognitive science that identify concepts as the basic building blocks of thought (see, for example, readings from Margolis and Laurence²⁴). Humans acquire concepts from the interaction of our evolved cognitive architecture and built-in inductive biases with our experiences in the world, including both our direct percepts and what we learn socially and culturally from interacting and communicating with other humans. Our minds then compose these basic units to form sophisticated compound thoughts: beliefs, desires, and plans. One of the most powerful ways to construct a system of useful concepts for reasoning is to build them out of meanings acquired through language. In particular, we

can consider the granularity of concepts at the level of individual word meanings (of nouns, verbs, and so on) and treat their combinations in a way similar to how words can be combined into phrases and sentences in natural language. Related ideas of composing primitive units of thought have a long tradition in AI, dating back to inductive logic programming,²⁷ statistical relational learning,¹³ and probabilistic-logic programming.²⁰

Our technical proposal is to build a *neuro-symbolic concept* representation. Each concept is a discrete symbol (word or short phrase) that can be grounded onto subsets of the embodied environments. An object concept “orange” grounds to the sets of orange objects; an object relation concept “left” grounds to all object pairs (A,B) such that A is left of B; an action concept “put-left” grounds to all agent action sequences that move the object currently being held to a position on the left of the reference object. This includes pushing the object to the goal, holding it vigorously using tools,

Figure 1. Our framework for building intelligent agents by internalizing a vocabulary of “concepts,” which are represented as compositional programs and neural network embeddings. These concepts can be grounded in various domains: 2D images, videos, 3D scenes, and robotic actions. They can also be recombined based on different types of user queries: visually grounded questions, physical and causal reasoning questions, referring expressions, and manipulation instructions.

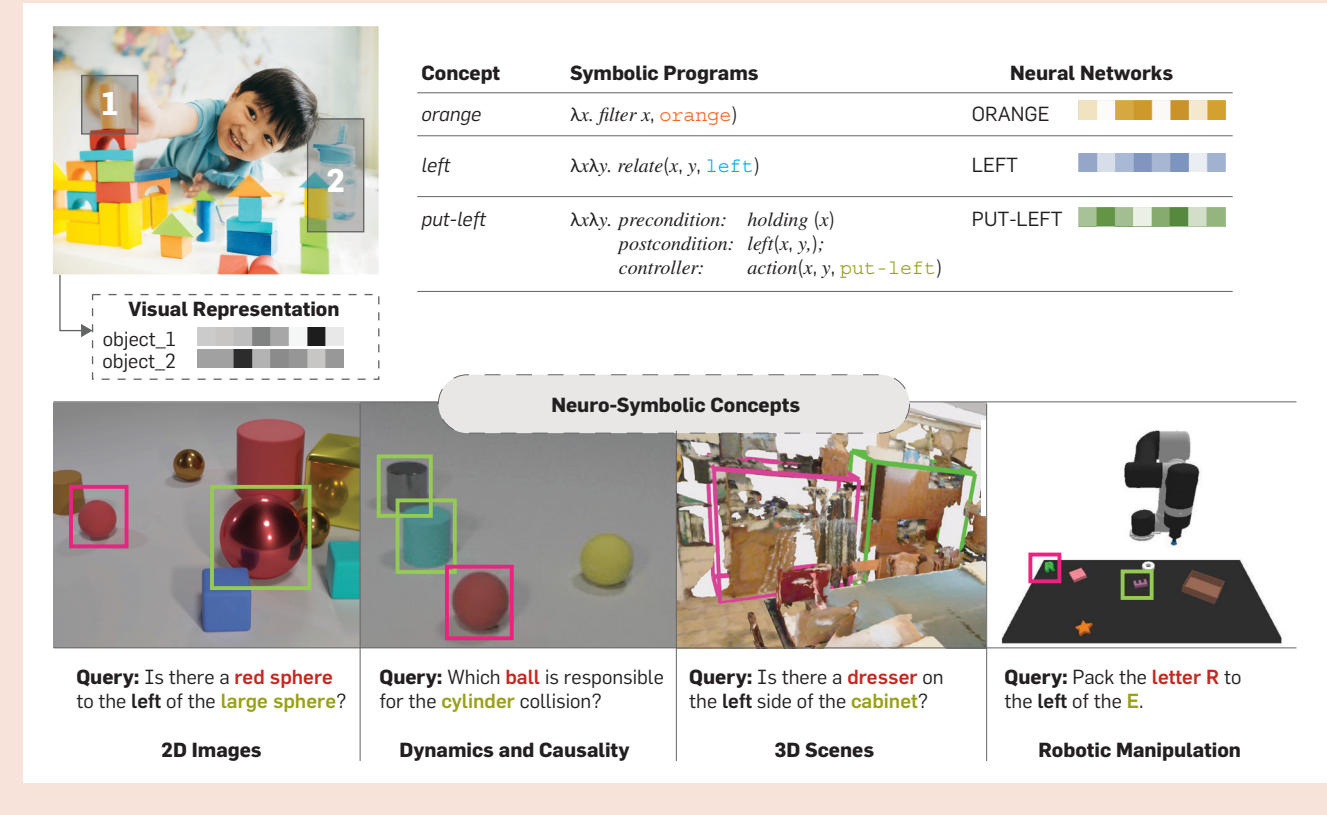


Figure 2. In a neuro-symbolic-concept-centric framework, different concepts such as object categories, properties, relations, and actions are represented as a combination of programmatic and neural representations. In (a), the neural representation connects the concept with sensory and actuation representations. In (b), different concepts can be combined to form new compound concepts.



and so forth. As illustrated in Figure 1, during reasoning, our *concept-centric* framework operates at a higher level of abstraction with this vocabulary of neuro-symbolic concepts.

Neuro-symbolic concepts have two advantages: They can be flexibly grounded on sensory and actuation modalities, and they have strong compositional generalization, since existing concepts can be structurally combined to form new ones. For instance, “orange,” “cylinder,” and “left” can be combined to form a novel concept, such as “Put the orange cylinder to the left of the bottle.” This compositionality naturally suggests a decomposition of the learning problem: We may individually learn how to recognize object shapes (cylinder, bottle), how to recognize object colors (orange), how to reason about object placement (left of the target), and how to move objects from one location to another. During inference time, we can recombine these learned concepts—shapes, colors, relations, and actions—to achieve the specified goals.

Compared with purely end-to-end learning methods, the inherent compositionality of neuro-symbolic concepts makes them much better suited for generalist agent learning. Compositionality simultaneously supports four properties required by generalist agents: data efficiency (because collecting labeled data is especially difficult); compositional generalization (the number of possible scenes and tasks can be exponentially large as the number of objects and their properties increases; we want systems that can generalize to unseen scenes and

unseen goals); continual learning (the system should be able to learn and adapt gradually); and transfer learning (we want to support transfer among different tasks).

In the rest of the article, we first offer a definition of neuro-symbolic concepts. Then we delve into a concrete framework of neuro-symbolic concept learning for visual scene understanding and showcase its ability to learn from little data, continually and generalizably, and be transferable. Finally, we discuss other applications of this approach.

Neuro-Symbolic Concepts

In order to systematically represent the grounding of concepts and how they can be composed, we formally represent each concept c as a tuple of:

$$c = \langle parameter, program, neural - nets \rangle.$$

In essence, our concept representation integrates neural network representations (for instance, vector embeddings), which ground concepts in visual and physical representations, and symbolic representations (particularly, parameterized programs), which characterize how various concepts can be combined. Shown in Figure 2a, the concept “orange” is represented as

$$orange = \langle \{x\}, filter(x, ORANGE), \{ORANGE\} \rangle,$$

where ORANGE denotes a vector embedding that can be used by the built-in *filter* function to classify orange objects. In practice, this *filter* function can be implemented by comput-

ing the cosine similarity between the neural representation of object x and ORANGE.²¹

Similarly, a relational concept (for example, a prepositional phrase) “left-of” can be defined as:

$$left-of = \langle \{x, y\}, relate(x, y, LEFT-OF), \{LEFT-OF\} \rangle.$$

The concept relates two objects x and y . The *relate* function will compute the cosine similarity between LEFT-OF and a pairwise representation between (x, y) to determine whether x is left of y in the scene. Actions (for example, verbs) would have three parts in their programs: a controller that can generate sequences of robot-control commands, and the preconditions and postconditions for the action. For example, an action “put-left-of” can be represented as:

$$put-left-of = \langle \{x, y\}, \{ pre = holding(x), post = left-of(x, y), controller = PUT-LEFT-OF, \{PUT-LEFT-OF\} \} \rangle.$$

Here, the preconditions and postconditions of the action can be described with formulas composed from other object-level and relational concepts.

Illustrated in Figure 2b, this representation of concepts enables us to combine existing concepts adhering to symbolic functional composition rules, to form compound concepts such as *orange(x)* and *cylinder(x)* (orange cylinders), or more complex ones such as *orange(x)* and *bottle(y)* and *put-left-of(x, y)* (put the orange

object left of the bottle). To support the formal compositionality of different concepts, all parameters and outputs are typed with primitive types (including objects, events, actions, Booleans, and integers). For instance, object concepts are represented as functions that take the perceptual representation of an object as input and predict classification scores as output, indicating whether the object has the concept. Relational concepts are associated with classifiers that classify object pairs. Meanwhile, action concepts are linked with preconditions (circumstances under which

the action can be executed), post-conditions (the outcomes of executing the action), and controllers that generate agent actions based on the current perceptual state. Depending on the domain and the task, one can choose to implement different primitive operations (for example, *filter* in a visual recognition context).

Leveraging these neuro-symbolic concepts, we can efficiently and effectively learn the grounding of concepts in various domains and recombine them to solve different downstream tasks. For example, as shown in Figure 1, neuro-symbolic concepts can

be grounded in 2D images (such as object properties and object relations), videos (physical events and their relations), 3D scenes (object properties and viewpoint-dependent relations), and finally, robotic-manipulation tasks (object properties, relations, and actions that change them). Therefore, by recombining them through symbolic program structures, we can answer questions, resolve referring expressions, and interpret human instructions. We illustrate this idea in Figure 2. Specifically, we learn the grounding (classifiers and controllers) for individual visual and action

Figure 3. Three challenges in (a) compositional generalization, (b) continual learning of concepts for reasoning, and (c) transfer of learned concepts across domains.

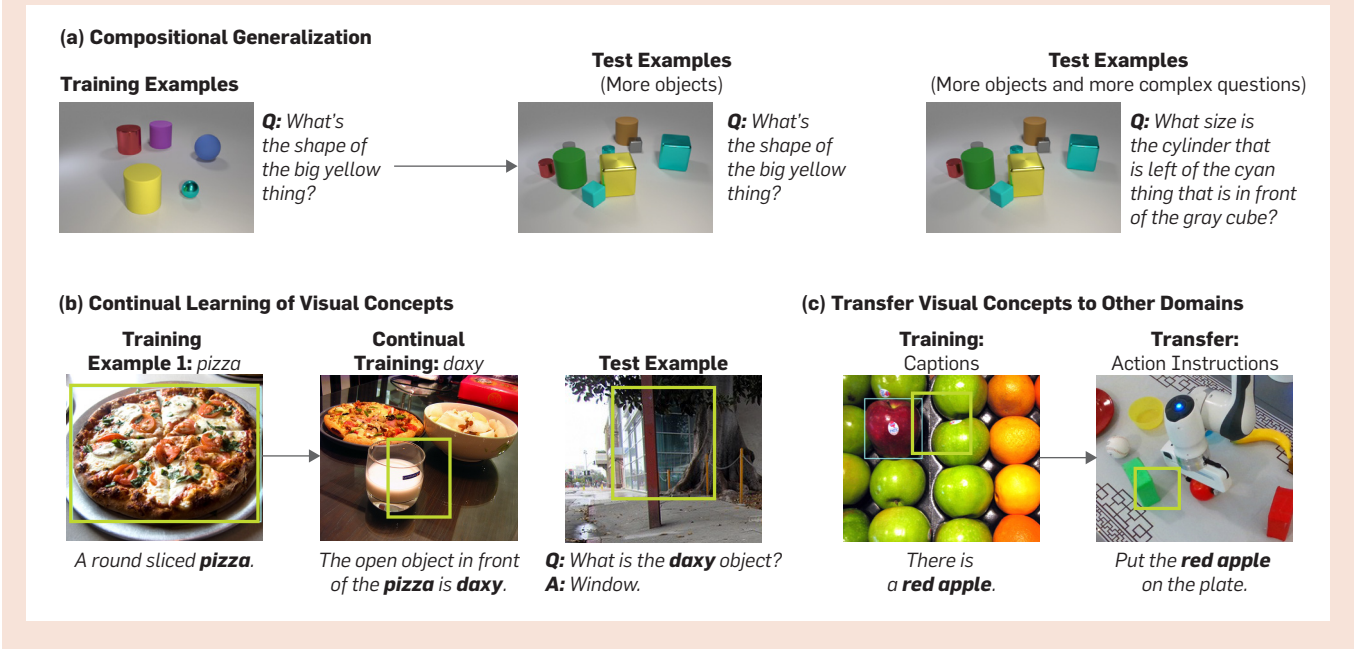
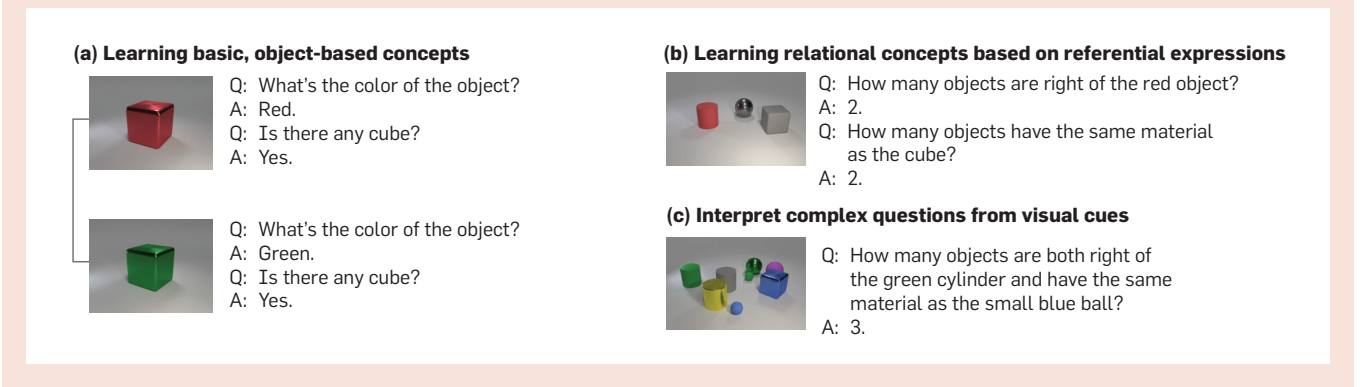


Figure 4. Humans learn visual concepts, words, and semantic parsing jointly and incrementally. (a) Learning visual concepts (red vs. green) starts from looking at simple scenes, reading simple questions, and reasoning over contrastive examples.¹² (b) Afterward, we can interpret referential expressions based on the learned object-based concepts and learn relational concepts (for example, on the right of, the same material as). (c) Finally, we can interpret complex questions from visual cues by exploiting the compositional structure.



concepts and recombine them following a hierarchical program that represents the meaning of the input user query: “Put the orange object left of the bottle.” In the following, we revisit the four important desiderata for generalist agent learning and illustrate how our neuro-symbolic-concept-centric paradigm fulfills all requirements.

Data efficiency. Since learning in embodied environments inevitably involves machine interaction with the physical environment and human annotations, minimizing the amount of data needed to learn a specific concept is crucial. Compared to a monolithic deep neural network, a concept-centric framework gains data efficiency primarily by leveraging modular structures of the learning task. For example, the complex concept “push the orange cylinder” can be decomposed into three individual concepts: two object concepts (“orange” and “cylinder”) and one action concept (“push”). This compositional structure introduces a strong prior into the learning algorithm, allowing it to disentangle the overall learning problem and to perform explicit multi-task learning across heterogeneous data sources (for example, learning the concept of “orange” from images and “push” from robot interaction datasets).

Compositional generalization. Compositionality is often grounded in different aspects across different domains. For example, in the visual concept learning domain (Figure 3a), compositional generalization is expected on at least two levels: the concept composition level (for example, “the big cylinder left of the yellow block”) and the scene composition level (generalization to scenes with a different number of objects compared to training examples). The advantage of concept-centric frameworks is primarily a contribution from the alignment between the composition structure of concepts and the structure of the domain. For example, explicitly reasoning about the set of “orange” and its subset “orange cylinder” in a visual scene makes the reasoning process robust to the total number of objects in a scene.

Continual learning. The demand



Neuro-symbolic concepts have two advantages: They can be flexibly grounded on sensory and actuation modalities, and they have strong compositional generalization, since existing concepts can be structurally combined to form new ones.



for continual concept learning and transfer learning arises due to two challenges. First, it is generally difficult to obtain high-quality “end-to-end” learning examples for embodied agents, that is, from raw perceptual input to robot-control commands. Therefore, a practical system should be able to learn from multiple sources of data, typically of different input-output specifications: unannotated videos, paired images and texts, human demonstrations of skills, and so on. Second, at deployment time, the machine should continually adapt itself to its environment, such as by learning new concepts (for example, an unseen breed of dog or a new type of dish) and new human preferences.

Zero-shot transfer. Unlike monolithic deep neural networks that usually require tuning all parameters while learning new concepts, the concept-centric framework (Figure 3b) naturally allows a flexible introduction of new concepts and adjustment to a single previously learned concept, thanks to the modular structures of concept composition. It also enables a zero-shot transfer of learned concepts across tasks and even domains, such as transferring learned object concepts from the task of image captioning (“The photo shows a dog”) to visual question answering (“How many dogs are there?”), and from the domain of visual concept learning (“apples”) to the domain of robotic manipulation (“Push the apples”; Figure 3c).

Learning Neuro-Symbolic Visual Concepts

Our framework for neuro-symbolic concept learning for visual scene understanding is motivated by how humans learn visual concepts by jointly understanding vision and language.¹² Consider the example shown in Figure 4a. Imagine someone with no prior knowledge of colors is presented with the images of the red and green cubes, paired with the questions and answers. They can easily identify the difference in objects’ visual appearance (in this case, color), and align it to the corresponding words in the questions and answers (*red* and *green*). Other object attributes (for example, shape) can be learned in a similar

way. Starting there, humans are able to inductively learn the correspondence between visual concepts and word semantics (for example, spatial relations and referential expressions, Figure 4b), and unravel compositional logic from complex questions assisted by the learned visual concepts (Figure 4c, also see Abend et al.¹).

This motivated us to build a framework that jointly learns visual perception, words, and semantic language parsing from images and question-answer pairs. Proposed in Mao et al.,²¹ a *neuro-symbolic concept learner (NS-CL)* learns all of these from natural supervision (that is, images and QA pairs), requiring no annotations on images or semantic programs for sentences. Instead, analogous to human concept learning, it learns via curriculum learning. The NS-CL starts by learning representations/concepts of individual objects from short questions (for example, "What's the color of the cylinder?") on simple scenes (less than three objects). By doing so, it learns object-based concepts such as colors and shapes. The NS-CL then learns relational concepts by leveraging these object-based concepts to interpret object referrals (for example, "Is there a box to the right of a cylinder?"). The model iteratively adapts to more complex scenes and highly compositional questions.

Shown in Figure 5, the NS-CL has three modules: a neural network-based perception module, a semantic parser to translate questions into executable programs, and a symbolic program executor. Given an input image, the visual perception module detects objects in the scene and extracts a deep, latent representation for each of them. The semantic parsing module translates an input question in natural language into an executable program, represented in a domain-specific language (DSL) designed for visual question answering (VQA). The DSL covers a set of fundamental operations for visual reasoning, such as filtering out objects with certain concepts or querying the attribute of an object. The generated programs have a hierarchical structure of symbolic, functional modules.

Next, based on the latent program recovered from the question in natu-

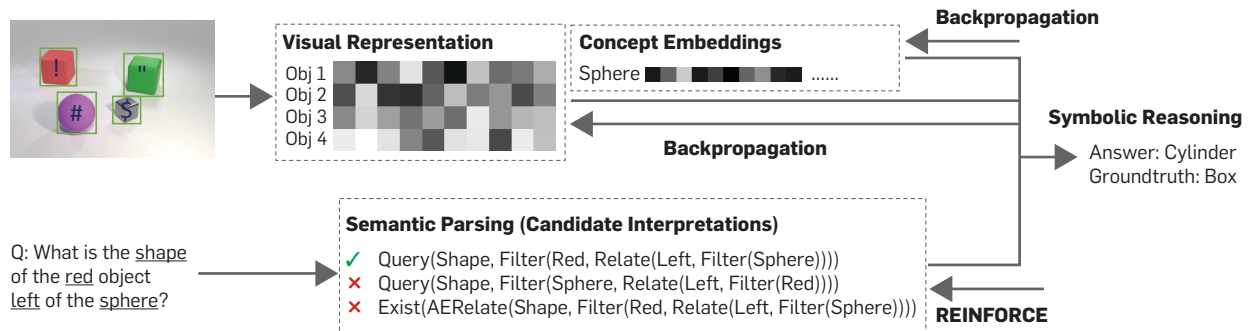
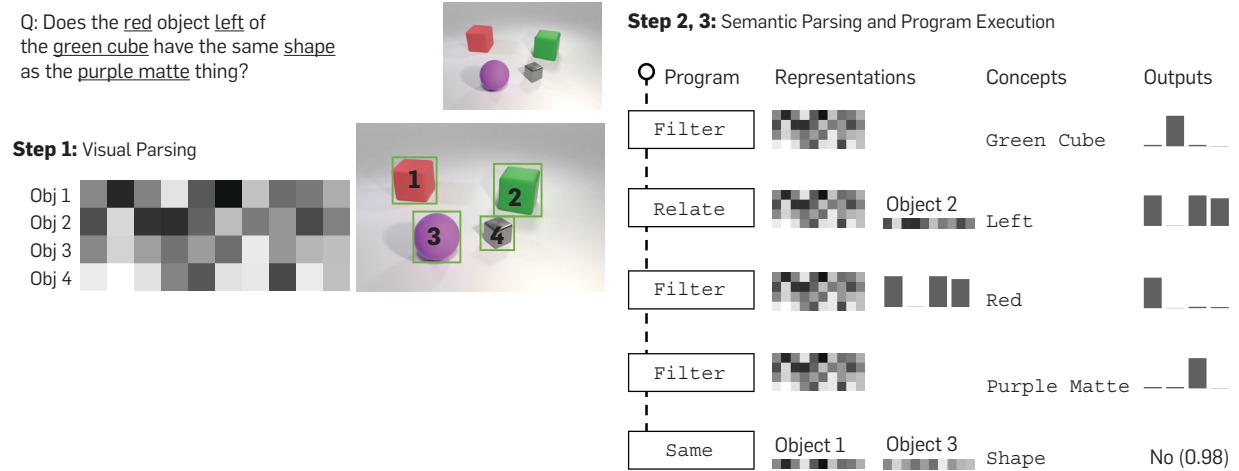
ral language, a symbolic program executor executes the program and derives the answer based on the object-based visual representation. Our program executor mainly contains two parts: the concept quantization module and a collection of deterministic functional modules. The concept quantization module classifies object attributes and relations, and the functional modules implement the logic of composing these classification results. To make the execution differentiable with regard to visual representations, we represent the intermediate results in a probabilistic manner: A set of objects is represented by a vector—the attention mask over all objects in the scene. Each element, $\text{Mask}_i \in [0, 1]$, denotes the probability that the i -th object of the scene belongs to the set. Figure 6 shows an illustrative execution trace of a program. The first *filter* operation outputs a mask of length 4 (there are in total four objects in the scene), with each element representing the probability that the corresponding object is selected (that is, the probability that each object is a green cube). The output "mask" on the objects will be fed into the next module (*relate* in this case) as input, and the execution of programs continues. The last module outputs the final answer.

Data efficiency. The NS-CL's modularized design enables interpretable, robust, and accurate visual reasoning. As shown in Figure 7a, it achieves state-of-the-art performance on the CLEVR dataset.¹⁷ More importantly, it enables data-efficient learning of concepts and combinatorial generalization with regard to both visual scenes and semantic programs. Figure 7a highlights comparisons with baselines that do not explicitly learn concepts: When trained on 10% of the CLEVR training data, the NS-CL achieves 98.9% accuracy on the test set, surpassing all baselines by 14%.

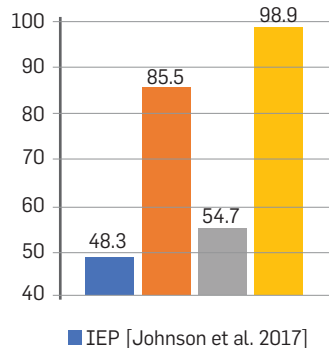
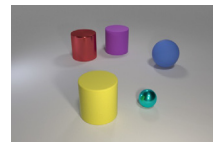
Compositional generalization. We also test our model for compositional generalization. As shown in Figure 7b, after being trained on scenes with a small number of objects and simple questions, the NS-CL directly generalizes to more complex scenes and questions, while all baselines show a significant performance drop.

Continual learning. Since our neuro-symbolic learning problem decomposes the problem into learning individual concepts, it naturally supports continual learning of new concepts. As a concrete implementation, in Mei et al.²⁵ we present a meta-learning framework for learning new visual concepts quickly from just one or a few examples, guided by multiple naturally occurring data streams: simultaneously looking at images, reading sentences that describe the objects in the scene, and interpreting supplemental sentences that relate the novel concept to other concepts. The system operates in a class-incremental manner,³⁵ where it continuously receives new examples of an unseen category and builds a new embedding for the novel category. The learned concepts support downstream applications, such as answering questions by reasoning about unseen images. Our model, FALCON, represents individual visual concepts, such as colors and shapes, as embeddings in a high-dimensional space. Given an input image and its paired sentence, our model first resolves the referential expression in the sentence and associates the novel concept with particular objects in the scene. Next, our model interprets supplemental sentences to relate the novel concept with other known concepts, such as "X has property Y" or "X is a kind of Y."¹⁴ Finally, it infers an optimal embedding for the novel concept that jointly 1) maximizes the likelihood of the observed instances in the image and 2) satisfies the relationships between the novel concepts and the known ones. We demonstrate the effectiveness of our model on both synthetic and real-world datasets.

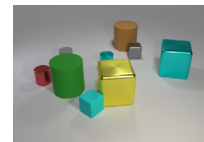
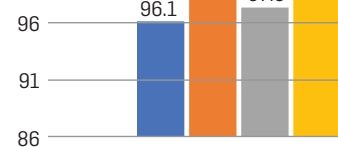
Transfer learning. The learned visual concepts can also be used in other domains, such as image retrieval. With the visual scenes fixed, the learned visual concepts can be transferred directly into the new domain. We only need to learn the semantic parsing of natural language into the new DSL. To do so, we build a synthetic dataset for image retrieval. The dataset contains only simple captions: "There is an <object A> <relation> <object B>." (for example, "There is a box to the right of a cylinder"). The

Figure 5. The NS-CL uses neural symbolic reasoning to bridge the learning of visual concepts, words, and semantic parsing.**Figure 6. The neuro-symbolic execution procedure of a program based on the visual representation and concept embeddings.****Figure 7. Data-efficiency test and compositional generalization test for the NS-CL.****(a) Data-Efficiency Test**

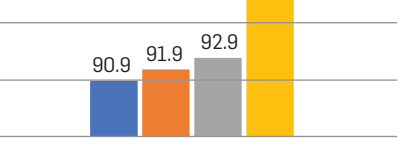
Training with 10% data

**(b) Compositional Generalization Test**

Training



Test



Q: What's the shape of the big yellow thing?

Q: What size is the cylinder that is left of the cyan thing that is in front of the gray cube?

semantic parser learns to extract corresponding visual concepts (for example, “box,” “right,” and “cylinder”) from the sentence. The program can then be executed on the visual representation to determine if the visual scene contains such relational triples. Note that this functionality cannot be directly implemented on the CLEVR VQA program domain, because questions such as “Is there a box to the right of a cylinder?” can be ambiguous if there exist multiple cylinders in the scene. Due to the entanglement of the visual representation with the specific DSL, baselines trained on CLEVR VQA cannot be applied directly to this task.

Applications

The design principles of the NS-CL—in particular, the visual grounding of concepts through neuro-symbolic reasoning—can naturally generalize to a large body of learning and reasoning tasks.

Accurate and robust image captioning. Wu et al.³⁸ implement a similar idea of neuro-symbolic concept learning to image-caption retrieval tasks. They used pretrained language parsers to translate captions into graphical representations composed of object categories, properties, and relationships. This kind of factorization not only leads to better performance in retrieving accurate descriptions of images, but also improves the robustness of the system with respect to captions that are similar to correct ones (for example, that differ only in one or two words) but inaccurate.

Video and counterfactual reasoning. A line of research^{5,9} has been extending the concept-learning framework to reasoning about physics. The object-centric nature of neuro-symbolic concept learners enables natural integration with learned physics models, which brings the capability to perform predictive and counterfactual reasoning. As an example, in Chen et al.,⁵ the authors ground concepts about physical objects and events from dynamic scenes and language. Building upon a neural object-centric representation, their model is simultaneously also trained to approximate the dynamic interaction among objects with neural

networks. Therefore, after training, it can not only detect and associate objects and events across the frames, but also make future and counterfactual predictions of object interactions (for example, “What will happen if we remove the red block from the scene?”). Later work, such as Ding et al.,⁹ further extends this capability to online inference of physical object properties.

3D concept grounding. The neuro-symbolic framework can also be applied to 3D representations.^{16,29} The variability of the 3D domain induces two fundamental challenges: 1) the expense of labeling and 2) the complexity of 3D grounded language. Hence, essential desiderata for models are to be data-efficient, generalize to different data distributions and tasks with unseen semantic forms, as well as ground complex language semantics (for example, view-point anchoring and multi-object reference: “Facing the chair, point to the lamp on its right”). To address these challenges, Prabhudesai et al.²⁹ propose disentangling different object properties in learning to achieve better performance in few-shot concept learning. Hsu et al.¹⁶ extend the neuro-symbolic concept learner model by introducing functional modules that effectively reason about high-arity relations (that is, relations among more than two objects), key in disambiguating objects in complex 3D scenes. This architecture enables significantly improved performance in terms of data efficiency and generalization, and demonstrates zero-shot transfer to a 3D question-answering task.

Human motion. Endo et al.¹¹ focus on designing models that conduct complex spatiotemporal reasoning over motion sequences. They propose a new framework for learning neural concepts of motion, attribute neural operators, and temporal relations. Unlike the 2D and 3D vision domains, where object segmentations can be readily extracted using preexisting object detectors, motion sequences lack a universal action segmentation methodology. Therefore, their system attempts to temporally localize and ground motion concepts jointly.

Robotic manipulation. Recall

that an important feature of neuro-symbolic concept learning methods is that the learnable modules associated with different concepts are naturally disentangled. Therefore, it directly supports the transfer of learned concepts to other tasks or even other domains (for example, from vision-language domains to robotic-manipulation domains). In Wang et al.³⁷ and Kalithasan et al.,¹⁸ the authors tackle the problem of learning robotic manipulation based on visual input. Both papers exploit the syntactic and semantic structures of language instructions to build robotic-manipulation algorithms composed of object-recognition models and action policies. Kalithasan et al.¹⁸ directly transfer the visual concepts learned by the neuro-symbolic concept learning on images to robotic manipulation, by learning additional object-movement policies with reinforcement learning. Wang et al.,³⁷ by contrast, leverage large-scale pretrained vision-language (VL) models for object property recognition. Compared to a conventional pretraining-finetuning pipeline for leveraging pretrained models for robotics, their method leads to more data-efficient learning and, more importantly, better zero-shot generalization in a variety of unseen objects and tasks.

General Discussion

We have presented a general framework for learning and reasoning that is applicable to various domains and tasks. By leveraging the neuro-symbolic concept representation, our system can continuously learn concepts from data streams in a data-efficient manner and programmatically compose its learned representations to solve new tasks, even previously unseen tasks. This capability allows us to learn and generalize concepts from diverse types of data streams, including image-caption data and robotic demonstrations, in order to solve complex tasks.

Our neuro-symbolic concept learning framework belongs to the broader paradigm of neuro-symbolic AI, a field where researchers explore the synergies between neural networks, symbolic reasoning methods, and probabilistic inference tools. The

idea of connecting neural networks with symbolic entities has its origins in early work on embedding symbolic relationships into vector representations. For example, early research^{7,30} demonstrates how to integrate logical reasoning and neural networks, improving the efficiency, interpretability, and controllability of learning systems.

Building on top of these high-level ideas, and in line with the NS-CL, many neuro-symbolic AI systems have been developed that combine symbolic reasoning mechanisms with neural networks for recognizing object properties and relationships, as well as predicting action commands in interactive environments. By combining perceptual capabilities with tools like forward-chaining theorem provers, answer-set programming solvers, and program-synthesis tools, these frameworks enable reasoning and planning in both visual and physical environments.

In application areas closely related to the NS-CL, Amizadeh et al.² introduce differentiable first-order logic frameworks for reasoning about objects in scenes. Further, Barbiero et al.⁴ combine neural network predictions with fuzzy-logic rule execution, and Shindo et al.³² use differentiable inductive logic programming to recover logic programs representing scene structures for scene reasoning tasks. These methods further formalize the interface between neural and symbolic components using tools such as probabilistic and real-valued logic, supporting formal interpretations of reasoning under uncertainty. These approaches excel not only in data efficiency and compositional generalization but also in offering interpretability of reasoning traces and inferred rules, which is critical in applications that demand transparency.

While much of the prior work has focused on learning simple rules or answering queries about object states and relationships, recent research has extended neuro-symbolic frameworks to more complex reasoning tasks. For example, Wang et al.³⁶ and Yang et al.⁴⁰ tackle visual puzzle-solving, such as Sudoku, using Boolean satisfiability solvers and answer set



By leveraging the neuro-symbolic concept representation, our system can continuously learn concepts from data streams in a data-efficient manner and programmatically compose its learned representations to solve new tasks.



programming tools. End-to-end neural networks for these complex tasks often require significantly more data, which can be difficult to obtain for many practical applications. However, training such complex neuro-symbolic models can be challenging, as backpropagation must occur over long chains of neuro-symbolic computations without intermediate supervision.

Moving toward more abstract, layout-based, and scene-level concepts—an emerging area in machine learning and visual reasoning—Shindo et al.³³ study how patterns of object placements can be learned from just a few examples. Similarly, Hsu et al.¹⁵ investigate reasoning about abstract concepts, such as mazes and treasure maps. These high-level concepts are difficult to interpret by state-of-the-art end-to-end systems like large vision-language models. These studies show that decomposing abstract concepts into smaller, more primitive entities using symbolic structures can significantly improve system performance.

Finally, it is important to highlight other key advantages of neuro-symbolic systems that we have not fully discussed in this article, such as interpretability, controllability, and the ability to integrate with external knowledge bases.^{4,34} These attributes, especially interpretability and safety, are critical in high-stakes decision-making contexts.³⁹

The idea of neuro-symbolic concepts is also closely related to the idea of neural module network compositions,³ since computationally, they are both paradigms for composing neural network modules to solve more complex tasks. However, they differ at both the conceptual and implementation levels. Specifically, in neural module networks, primitive neural networks define functions or transformations that can be applied to inputs, whereas in neuro-symbolic concepts, primitive neural networks handle the grounding of individual concepts, and the operations based on these concepts (for example, *filter* or *count*) are implemented as deterministic functions in DSLs. This disentanglement between grounding and reasoning brings

about significant improvements in data efficiency, compositional generalization, and transferability.

Of course, such improvements come at a cost: Many works on neuro-symbolic concept learning have the limitation of relying on a predefined DSL. This DSL encompasses primitive operators such as *filter* and *relate*, as well as concept symbols such as *orange* and *place*. In the following, we delve into these two parts.

In most domains, by combining object property primitives, relational primitives, and action primitives, along with simple set operations such as intersection, union, and counting, we can construct a highly capable system. Recently, there has been a growing interest in extending the primitive set and number operations to general programming languages, such as Python.⁸ This will significantly improve the expressiveness of the programs by including complex control flows such as loops and recursions. However, in general, this also introduces new challenges in the learning of concepts. Recall that in the NS-CL and many neuro-symbolic concept learning works, the concept representations are learned through backpropagation of the program execution trace. Although there has been much work on backpropagation against program execution traces,²⁸ complex control flows will inevitably introduce intractability of possible execution results, as well as gradient vanishing and explosion problems.

Relying on a predefined set of concepts can be too restrictive in many real-world applications. Three approaches have emerged to address this: grammar-based lexicon learning,²³ concept induction from past experiences, and more recent approaches based on large language models (LLMs). The high-level idea behind grammar-based lexicon learning is that instead of trying to associate each word or phrase with a predefined set of concept names (for example, mapping the word *orange* to a concept ORANGE), we construct the library of concepts by “converting” each word into a concept. This is roughly equivalent to discovering the syntax of each word. For example, if we see the word *orange*



While much of the prior work has focused on learning simple rules or answering queries about object states and relationships, recent research has extended neuro-symbolic frameworks to more complex reasoning tasks.



in a sentence and it is an adjective, then we immediately know the word *orange* should correspond to an object property concept, named ORANGE. This approach is called grammar-based lexicon learning because the system begins with a small set of universal grammar rules and jointly discovers new concepts from the text corpus while learning their grounding.

A second approach involves inducing new concepts from experiences. One method is to represent concepts as “theories composed of other concepts,” in line with the theory-theory of concepts.²⁶ For example, Das et al.⁶ induce new logical concepts by composing previously learned concepts using logic programs. Likewise, Shindo et al.³³ induce transferable scene-level concepts from a few examples, while Ellis et al.¹⁰ learns a “library” of functions built from primitive concepts that can be hierarchically recombined to form complex geometric and scene-level concepts. Another direction explores the invention of new object and relation concepts through contrastive learning.³¹

A third approach is to leverage LLMs such as GPT-4. With the success of LLMs in language-to-code translation, researchers explore the use of these models in extracting concept symbols from natural language queries.^{8,16} In particular, they leverage LLMs that have been trained on Internet-scale text and code corpora to translate natural language queries into programs with concept symbols. These concept symbols are not chosen from a given vocabulary, but are instead automatically generated by LLMs based on the user queries. For each concept symbol that appears in LLM-translated programs, a new concept representation will be initialized and learned.

There are many challenges and future directions for neuro-symbolic concept learning systems. Most approaches focus on relational concepts involving only two objects (or three, in the case of 3D concepts¹⁶), but there are more complex layout concepts and scene-level concepts (for example, mazes¹⁵) that involve many more objects and have variable arities, raising the question of how to handle such complexity. Additionally, while the

NS-CL uses curriculum learning, and some work explores different methods for curriculum construction,¹⁹ the automatic discovery or design of curricula that adapt as humans do when learning new concepts remains unsolved. So far, many concept learning systems can only operate in a pure “class-incremental” learning framework,³⁵ where new concepts build on previous ones; addressing the full “curriculum learning” setting, which requires revision or reversion of previously learned concepts, remains a significant challenge. Another underexplored topic is unsupervised concept learning beyond the current focus on supervised or semi-supervised methods. Finally, another direction for future work is to formalize reasoning under perceptual and other types of uncertainty by incorporating probabilistic inference methods. This includes tools like probabilistic logic programming and probabilistic programming languages.

Up to now, most neuro-symbolic concept learning systems have been developed for particular domains and tasks. For example, we have systems that can learn and reason with concepts for 2D images,²¹ 3D scenes,¹⁶ human motions,¹¹ video events,⁵ and robotic actions.^{22,37} However, they have been built in isolation, with limited shared knowledge among them. An important future direction for neuro-symbolic concept learning is the development of scalable, cross-domain concept libraries. For example, the concept “close to” has grounded meanings across different domains, but the core and abstract notion of distance metrics is really shared across all domains and modalities. From an engineering perspective, similar to recent LLMs such as OpenAI’s GPT and Google’s Gemini, which can solve a wider range of tasks in the language domain based on user instructions, building unified concept representations across domains and modalities would enable us to tackle a broader spectrum of embodied AI problems, spanning from perception to action. From a scientific perspective, connecting and unifying concept representations across domains and modalities could not only bring better data efficiency in learning, but also

enable the grounding of concepts in more abstract scenarios. Ultimately, achieving this level of abstraction and unification would mark a fundamental step toward building truly intelligent systems that can reason, adapt, and act across the vast complexity of the real world.

Acknowledgments

This work is in part supported by AFOSR YIP FA9550-23-1-0127, FA9550-22-1-0387, ONR N00014-23-1-2355, ONR YIP N00014-24-1-2117, ONR MURI N00014-22-1-2740, and NSF RI #2211258, and an AI2050 Senior Fellowship. 

References

1. Abend, O. et al. Bootstrapping language acquisition. *Cognition* 164 (2017), 116–143.
2. Amizadeh, S. et al. Neuro-symbolic visual reasoning: Disentangling ‘visual’ from ‘reasoning’. In *Intern. Conf. on Machine Learning* (2020).
3. Andreas, J., Rohrbach, M., Darrell, T., and Klein, D. Neural module networks. In *IEEE/CVF Conf. on Computer Vision and Pattern Recognition* (2016).
4. Barbiero, P. et al. Interpretable neural-symbolic concept reasoning. In *Intern. Conf. on Machine Learning* (2023).
5. Chen, Z. et al. Grounding physical concepts of objects and events through dynamic visual reasoning. In *Intern. Conf. on Learning Representations* (2021).
6. Das, M., Ramanan, N., Doppa, J.R., and Natarajan, S. Few-shot induction of generalized logical concepts via human guidance. *Frontiers in Robotics and AI* 7 (2020), 122.
7. Demeester, T., Rocktäschel, T., and Riedel, S. Lifted rule injection for relation embeddings. In *Conf. on Empirical Methods in Natural Language Processing* (2016).
8. Didac, S., Menon, S., and Vondrick, C. ViperGPT: Visual Inference via Python execution for reasoning. In *IEEE/CVF Intern. Conf. on Computer Vision* (2023).
9. Ding, M. et al. Dynamic visual reasoning by learning differentiable physics models from video and language. In *Advances in Neural Information Processing Systems* (2021).
10. Ellis, K. et al. DreamCoder: Growing generalizable, interpretable knowledge with wake–sleep Bayesian program learning. *Philosophical Transactions of the Royal Society A* 381, 2251 (2023), 20220050.
11. Endo, M., Hsu, J., Li, J., and Wu, J. Motion question answering via modular motion programs. In *Intern. Conf. on Machine Learning* (2023).
12. Fazly, A., Alshahi, A., and Stevenson, S. A probabilistic computational model of cross-situational word learning. *Cognitive Science* 34, 6 (2010), 1017–1063.
13. Friedman, N., Getoor, L., Koller, D., and Pfeffer, A. Learning probabilistic relational models. In *Intern. Joint Conf. on Artificial Intelligence* (1999).
14. Han, C. et al. Visual concept metaconcept learning. In *Advances in Neural Information Processing Systems* (2019).
15. Hsu, J. et al. What makes a maze look like a maze? In *Intern. Conf. on Learning Representations* (2025).
16. Hsu, J., Mao, J., and Wu, J. NS3D: Neuro-symbolic grounding of 3D objects and relations. In *IEEE/CVF Conf. on Computer Vision and Pattern Recognition* (2023).
17. Johnson, J. et al. CLEVR: A diagnostic dataset for compositional language and elementary visual reasoning. In *IEEE/CVF Conf. on Computer Vision and Pattern Recognition* (2017).
18. Kalithasan, N. et al. Learning neuro-symbolic programs for language guided robot manipulation. In *IEEE Intern. Conf. on Robotics and Automation* (2023).
19. Li, Q., Huang, S., Hong, Y., and Zhu, S.C. A competence-aware curriculum for visual concepts learning via question answering. In *European Conf. on Computer Vision* (2020).
20. Manhaeve, R. et al. Deepprolog: Neural probabilistic logic programming. In *Advances in Neural Information Processing Systems* (2018).
21. Mao, J. et al. The Neuro-symbolic concept learner: Interpreting scenes, words, and sentences from natural supervision. In *Intern. Conf. on Learning Representations* (2019).
22. Mao, J., Lozano-Pérez, T., Tenenbaum, J.B., and Kaelbling, L. PDSketch: Integrated domain programming, learning, and planning. In *Advances in Neural Information Processing Systems* (2022).
23. Mao, J. et al. Grammar-based grounded language learning. In *Advances in Neural Information Processing Systems* (2021).
24. Margolis, E. and Laurence, S. *Concepts: Core Readings*. The MIT Press (1999).
25. Mei, L. et al. FALCON: Fast visual concept learning by integrating images, linguistic descriptions, and conceptual relations. In *Intern. Conf. on Learning Representations* (2022).
26. Morton, A. *Frames of Mind: Constraints on the Common-Sense Conception of the Mental*. Oxford University Press (1980).
27. Muggleton, S. and Raedt, L.D. Inductive logic programming: Theory and methods. *The J. of Logic Programming* 19 (1994), 629–679.
28. Petersen, F., Borgelt, C., Kuehne, H., and Deussen, O. Learning with algorithmic supervision via continuous relaxations. In *Advances in Neural Information Processing Systems* (2021).
29. Prabhudesai, M. et al. Disentangling 3D prototypical networks for few-shot concept learning. In *Intern. Conf. on Learning Representations* (2021).
30. Rocktäschel, T., Singh, S., and Riedel, S. Injecting logical background knowledge into embeddings for relation extraction. In *Annual Conf. of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (2015).
31. Sha, J., Shindo, H., Kersting, K., and Dhami, D.S. Neuro-symbolic predicate invention: Learning relational concepts from visual scenes. *Neurosymbolic Artificial Intelligence* (2024), 1–26.
32. Shindo, H., Pfanschilling, V., Dhami, D.S., and Kersting, K. ILP: Thinking visual scenes as differentiable logic programs. *Machine Learning* 112, 5, (2023), 1465–1497.
33. Shindo, H., Pfanschilling, V., Dhami, D.S., and Kersting, K. Learning differentiable logic programs for abstract visual reasoning. *Machine Learning* 113, 11 (2024), 8533–8584.
34. Skryagin, A., Ochs, D., Dhami, D.S., and Kersting, K. Scalable neural-probabilistic answer set programming. *J. of Artificial Intelligence Research* 78 (2023), 579–617.
35. Van de Ven, G.M., Tuytelaars, T., and Tolias, A.S. Three types of incremental learning. *Nature Machine Intelligence* 4, 12 (2022), 1185–1197.
36. Wang, P.W., Donti, P., Wilder, B., and Kolter, Z. SATNet: Bridging deep learning and logical reasoning using a differentiable satisfiability solver. In *Intern. Conf. on Machine Learning* (2019).
37. Wang, R. et al. Programmatically grounded, compositionally generalizable robotic manipulation. In *Intern. Conf. on Learning Representations* (2023).
38. Wu, H. et al. Unified visual-semantic embeddings: Bridging vision and language with structured meaning representations. In *IEEE/CVF Conf. on Computer Vision and Pattern Recognition* (2019).
39. Yang, W.C., Marra, G., Rens, G., and Raedt, L.D. Safe reinforcement learning via probabilistic logic shields. In *Intern. Joint Conf. on Artificial Intelligence* (2023).
40. Yang, Z., Ishay, A., and Lee, J. NeurASP: Embracing neural networks into answer set programming. In *Intern. Joint Conf. on Artificial Intelligence* (2020).

Jiayuan Mao (jiayuanm@seas.upenn.edu) is an assistant professor at the University of Pennsylvania, Philadelphia, PA, USA.

Joshua B. Tenenbaum is a professor at the Massachusetts Institute of Technology, Cambridge, MA, USA.

Jiajun Wu is an assistant professor at Stanford University, Stanford, CA, USA.

© 2026 Copyright held by the owner/author(s).

A tutorial-style introduction to recent research on using logical specifications to encode RL tasks illustrates theoretical limitations and practical solutions.

**BY RAJEEV ALUR, SUGUMAN BANSAL, OSBERT BASTANI,
AND KISHOR JOTHIMURUGAN**

Specification-Guided Reinforcement Learning

IN REINFORCEMENT LEARNING (RL), an *agent* learns to achieve its goal by interacting with its environment and learning from feedback about its successes and failures. This feedback is typically encoded as numerical rewards—for example, positive rewards for beneficial outcomes and negative rewards for harmful outcomes. Traditional RL algorithms assume the reward is provided by the user, and designing an effective reward function can be a daunting task. For example, consider a robot organizing boxes in a warehouse. To ensure the robot successfully moves boxes to the right locations while avoiding collisions, we might provide positive rewards for moving boxes in the right directions and negative rewards for

collisions. However, there is a wide variety of ways to implement such a reward function: Should rewards be given immediately after moving a box or after multiple successful moves? How should failures such as incorrect placements or collisions be handled? There is also the question of balancing the objective with safety constraints—if the warehouse robot receives rewards solely for obstacle avoidance, it might focus on safety and never move, whereas overweighting the objective may lead it to take dangerous shortcuts. Beyond the question of balancing the objective with safety constraints, the reward function may also need to introduce new variables to keep track of progress—for example, which rooms still have boxes that need to be cleared.

The art of designing an effective reward function is called *reward engineering*. The challenge is that reward engineering conflates the problem of *task specification* (i.e., what is the task to be performed?) with the problem of *reward shaping* (i.e., what rewards best guide the robot to solve the task?). Recent work has proposed to separate these concerns by using *logical specifications* to encode the desired behavior of an RL agent, which can automatically be compiled into shaped rewards

» key insights

- Temporal logics provide a natural way to specify what an agent should do over time, offering an elegant solution to the challenges of traditional reward engineering in RL, which requires low-level behavioral guidance and often leads to unintended outcomes.
- Learning directly from temporal logic specifications in the most general form presents theoretical limitations, as it cannot provide *probably approximately correct* (PAC) guarantees. This is due to the non-robustness of temporal specifications.
- Algorithms such as DURL use the structure of logical specifications to break complex tasks into smaller subtasks, each with its own simple reward. This allows the agent to plan at a high level and learn more efficiently, reducing the number of interactions needed with the environment.



to train the agent. *Temporal logics* such as *linear temporal logic (LTL)*²⁰ are particularly suitable for specifying RL tasks, as they provide rigorous syntax and semantics for reasoning about sequences of events and states over time. For instance, instead of crafting numerical rewards, the desired behavior of the warehouse robot can be encoded as a sequence of logical steps: (1) Go to object while avoiding obstacles, (2) Pick up object, (3) Go to target location while avoiding obstacles, and (4) Place object. Each step can be encoded using logical predicates. If g denotes a target location and o denotes an obstacle location, then the requirement of safely visiting location g can be written in LTL as *Eventually* $g \wedge$ *Always* $\neg o$, where *Eventually* and *Always* are temporal operators.

The key question then is how to train an agent to satisfy a given logical specification. One strategy would be to compile the specification into a shaped reward function, which can then be used in conjunction with a traditional RL algorithm to train an agent. However, compared to a reward function, the logical specification transparently represents task structure in a way that can be exploited by the RL algorithm. In particular, *specification-guided RL* aims to design novel RL algorithms that train agents directly from logical specifications.

In this article, we provide an overview of some recent progress on specification-guided RL. This is an active area of research, with a growing body of literature.^{1,5,6,9–12,14,17–19,22,24} We first highlight surprising negative theoretical results that provide limits on specification-guided RL in the infinite horizon setting. Our discussion of

these results provides an insight into some key differences between logical specifications and reward functions in the context of RL (curious readers can find formal proofs and further details in the cited references). In practice, it often suffices to require that the agent completes the assigned task within a specified finite number of steps. This finite horizon setting offers opportunities for positive results and practical algorithms. As a representative example, we describe Dijkstra-guided RL (DIRL), an algorithm that combines conventional reward-based RL with high-level planning to efficiently learn policies for complex tasks. Rather than crafting a single reward function for the entire task, DIRL (Figure 2) exploits the structure of the logical specification to automatically infer a set of subtasks. Each subtask is associated with its own reward function and represented as an edge in a task graph, which serves as the foundation for high-level planning. DIRL alternates between planning over the task graph and learning subtask policies. This decomposition allows DIRL to tackle complex tasks with significantly fewer environment interactions compared to baseline methods. Through this integration of symbolic structure and learning, DIRL highlights several advantages of specification-guided RL.

Learning Optimal Policies

The objective of RL is to train the agent to act in an environment to accomplish a specific task—for example, learn what action should be performed in any given state so the robot will eventually reach a specific room. A key feature of RL is the environment is assumed to be *unknown*. Hence, the *policy* that en-

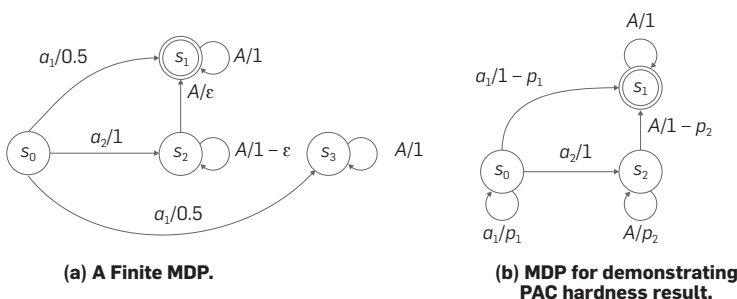
ables the agent to act must be learned by exploring the environment.

Markov decision processes. The environment is modeled as a *Markov decision process (MDP)*. Figure 1a shows an MDP with states $S = \{s_0, s_1, s_2, s_3\}$ and actions $A = \{a_1, a_2\}$. Actions cause state transitions according to *transition probabilities* $T : S \times A \times S \rightarrow [0, 1]$. For example, a_1 from s_0 leads to s_1 or s_3 with equal probability. The *initial state distribution* η describes the distribution of states the system initializes in. Here, η is the dirac distribution at s_0 —that is, with $\eta(s_0) = 1$. Mathematically, an MDP $\mathcal{M}$ is a tuple consisting of these four components^a—that is, $\mathcal{M} = (S, A, T, \eta)$. An *infinite run* of the MDP is an infinite sequence of state-action pairs describing the evolution of the system over time. A possible infinite run in Figure 1a is $\rho = s_0 \xrightarrow{a_2} s_2 \xrightarrow{a_1} s_1 \xrightarrow{a_2} s_1 \xrightarrow{a_2} \dots$ in which the agent reaches s_1 from s_0 via s_2 . Similarly, a *finite run* of length H describes the evolution of the system for H time steps—for example, $\rho = s_0 \xrightarrow{a_2} s_2 \xrightarrow{a_1} s_1 \xrightarrow{a_2} s_1 \xrightarrow{a_2} s_1$ is a finite run of length 4.

In RL, T is a priori unknown, and the agent must learn about the environment by taking actions and observing transitions. In Figure 1a, if the agent repeatedly takes action a_1 in state s_0 , it observes that the system transitions to s_1 approximately half the time. The goal is to learn a *policy* $\pi : S \rightarrow A$ mapping states to actions. This policy induces a distribution over runs, generated by sampling an initial state from η , and iteratively using π to select actions and sample next states from T . We denote the distribution over infinite runs as $\mathcal{D}^\pi$ and over finite H -length runs as $\mathcal{D}_H^\pi$.

Reward functions. Traditionally, tasks in RL are specified using a *reward function* $r : S \times A \rightarrow \mathbb{R}$ that assigns rewards for state-action pairs. For example, to reach s_1 in Figure 1a, we might set $r(s, a) = \mathbb{1}(s = s_1)$, which assigns a reward of 1 if $s = s_1$ and 0 otherwise. For a run ρ , this naturally generates a reward sequence (e.g., 0, 1, 1, ... for a run reaching s_1 in one step). For an infinite horizon, RL algorithms typically maximize the expected value of the *dis-*

Figure 1. Example MDPs.



a It is common to have a reward function as a component of the MDP, but we intentionally separate the two here.

counted-sum reward $J(\pi) = \mathbb{E}[\sum_{t=0}^{\infty} \gamma^t r_t]$ where the random variable r_t denotes the reward obtained at time t , and discount factor $0 < \gamma < 1$ indicates the rate of diminishing returns. In Figure 1a, under the reward function $r(s, a) = \mathbb{1}(s = s_1)$, the policy π_1 that chooses action a_1 in state s_0 will achieve a discounted-sum reward of $0 + \gamma + \gamma^2 + \dots = \gamma/(1 - \gamma)$ with probability 0.5 (when it reaches s_1) and 0 with probability 0.5 (when it reaches s_3). Hence, $J(\pi_1) = \gamma/(2(1 - \gamma))$. However, the policy π_2 that chooses action a_2 in s_0 leads to a discounted-sum reward of $0 + \dots + 0 + \gamma^{k+2} + \gamma^{k+3} + \dots = \gamma^{k+2}/(1 - \gamma)$ if s_1 is reached in exactly $k + 2$ steps which occurs with probability^b $(1 - \epsilon)^k \epsilon$. Thus, the expected discounted-sum reward^c is $\epsilon \gamma^2/(1 - \gamma) \cdot \sum_{k=0}^{\infty} (1 - \epsilon)^k \gamma^k = \epsilon \gamma^2/[(1 - \gamma)(1 + \gamma\epsilon - \gamma)]$. If, instead, given a finite horizon H , then RL algorithms maximize the expected reward $J_H(\pi) = \mathbb{E}[\sum_{t=0}^{H-1} r_t]$. Finally, the goal of RL is to learn an *optimal policy* π^* that maximizes $J(\pi)$ (or $J_H(\pi)$); we let $J^* = \sup_{\pi} J(\pi)$ (or $J^* = \sup_{\pi} J_H(\pi)$) denote the best possible value.

Logical specifications. A specification ϕ is a logical formula encoding whether a finite or infinite run of the MDP solves the desired task. For example, the task of reaching s_1 can be encoded by the specification $\phi = \text{Eventually } s_1$. A run that starts in s_0 and transitions immediately to s_3 will never reach s_1 , and therefore will not satisfy ϕ . Logical specifications can be straightforwardly composed to express more complex tasks; for instance, the task of sequentially visiting two locations l_1 and l_2 can be simply expressed as $\text{Eventually } l_1; \text{Eventually } l_2$. The meaning of a specification is given by the *semantics* function (denoted $\llbracket \phi \rrbracket$) that maps a run ρ to $\{0, 1\}$ indicating whether or not the run satisfies ϕ . The objective of specification-guided RL is to learn a policy that maximizes the probability of achieving the task—that is, $J^q(\pi) = \mathbb{E}_{\rho \sim \pi}[\llbracket \phi \rrbracket(\rho)]$. In Figure 1a, the policy π_1 that chooses action a_1 in s_0 assigns a probability of 0.5 to each of the two infinite runs it produces (one

Reward engineering conflates the problem of task specification (i.e., what is the task to be performed?) with the problem of reward shaping (i.e., what rewards best guide the robot to solve the task?).

that visits s_1 and the other that does not), so $J^q(\pi_1) = 0.5$. On the other hand, the policy π_2 that chooses action a_2 in s_0 results in infinitely many trajectories but all of these trajectories eventually visit s_1 given $\epsilon > 0$. Thus, $J^q(\pi_2) = 1$. Clearly, π_2 is optimal.

Infinite Horizon Specifications

This section discusses results on RL from *infinite-horizon* specifications. These specifications are important, as they can express recurring tasks and persistent behaviors that finite-horizon specifications cannot capture. Linear temporal logic is a prominent example of such specifications. For instance, tasks such as “repeatedly monitor an area” are naturally infinite-horizon properties.

Reductions to discounted rewards.

A natural approach to solving an infinite-horizon specification ϕ is to first compile it into a reward function r_{ϕ} and then apply standard RL algorithms for discounted rewards that come with guarantees and have state-of-the-art implementations readily available. However, recent work² has proven that without knowledge of the environment model, there are specifications for which no discounted reward function r exists (for any discount factor γ) such that the policies maximizing r will also maximize the probability of satisfying ϕ . This impossibility result has profound implications: There are specifications that simply cannot be solved by the strategy of reduction to reward functions. This result arises due to a fundamental mismatch between how discounted rewards and logical specifications measure task completion—discounted rewards inherently care about when the task is completed since future rewards are time-discounted, whereas logical specifications such as $\text{Eventually } s_1$ care only about the task eventually being completed.

This mismatch is illustrated by the specification $\text{Eventually } s_1$ and the reward function $r(s, a) = \mathbb{1}(s = s_1)$ in Figure 1a. Recall that the policy π_2 which takes action a_2 in state s_0 is optimal with regard to $\text{Eventually } s_1$. However, this policy can be made suboptimal with regard to $r(s, a) = \mathbb{1}(s = s_1)$ for any discount factor γ by choosing an appropriate value of ϵ .

Recent work shows some promising

^b Probability of staying in s_2 for k steps times the probability of transitioning to s_1 from s_2 .

^c $\sum_{k=0}^{\infty} (\text{Probability of reaching in } k + 2 \text{ steps}) \times (\text{Discounted-sum reward for reaching in } k + 2 \text{ steps})$.

directions under additional assumptions. For instance, Bozkurt et al.⁴ and Hahn et al.⁹ have shown that any LTL specification can be reduced to discounted rewards with an appropriate discount factor that depends on the MDP transitions. More recently, Alur et al.³ have studied the use of *discounted LTL*, which naturally prioritizes near-term events over distant ones, allowing a direct reduction to discounted rewards.

Theoretical guarantees. Even if it is impossible to compile logical specification to discounted-sum rewards, we may still hope we can devise algorithms that directly learn policies to solve infinite-horizon specifications. Our next result shows that the answer remains negative. To formalize this result, we extend the *probably approximately correct (PAC)* framework²¹ to specification-guided RL. Given a confidence level $\delta > 0$ and an error tolerance $\epsilon > 0$, an RL algorithm is said to be PAC if after sampling a sufficient number of transitions (as a function of n, m, δ, ϵ) in an MDP with n states and m actions, the probability that the learned policy π is ϵ -close to optimal (i.e., $J^* - J(\pi) \leq \epsilon$) is at least $1 - \delta$. Note that the required sample size does not depend on the transition probabilities, allowing agents to determine termination without knowing the true environment dynamics.

While algorithms with PAC guarantees exist for discounted-sum rewards, surprisingly, they are impossible for even simple reachability specifications, such as *Eventually g* (where g is a goal state). To understand why, consider the MDP shown in Figure 1b and the specification *Eventually s_1* . When $p_1 = 1$ and $p_2 < 1$, the optimal policy chooses action a_2 in state s_0 since the probability of eventually reaching s_1 from s_2 is 1, whereas the probability of reaching s_1 from s_0 is 0 if the agent chooses action a_1 in state s_0 . Similarly, when $p_1 < 1$ and $p_2 = 1$, the optimal action at state s_0 is a_1 . Note that altering the transition probabilities slightly can significantly impact the optimal policy. For example, if $p_1 = 1$ and $p_2 = 1 - x$ for an infinitesimally small x , altering the values to $p'_1 = p_1 - x = 1 - x$ and $p'_2 = p_2 + x = 1$ causes the optimal action at s_0 to change from a_2 to a_1 . Furthermore, an optimal policy in the



While algorithms with PAC guarantees exist for discounted-sum rewards, surprisingly, they are impossible for even simple reachability specifications.



original MDP (with $p_1 = 1$) that reaches s_1 with probability 1 will attain a zero probability of reaching s_1 in the new MDP. This example illustrates that the specification *Eventually s_1* is not robust—small changes to the MDP can drastically affect the corresponding optimal policy. In contrast, discounted-sum rewards are robust in the sense that altering the transition probabilities by a small amount only impacts the value $J(\pi)$ of a policy π by a small amount (more precisely, $J(\pi)$ is a continuous function of the MDP transitions and rewards).

This lack of robustness is the reason PAC algorithms do not exist for specifications. For the sake of contradiction, suppose there is an RL algorithm with a PAC guarantee for *Eventually s_1* . We can run this algorithm for the above two scenarios with slightly different transition probabilities and the same values of $\delta > 0.9$ and $\epsilon < 1$. Then, for a sufficiently small x , it is highly likely that the RL algorithm will see the same samples from the two different MDPs. Thus, the RL algorithm will output the same policy π , meaning it must attain a zero probability of reaching s_1 in one of the two MDPs. As a result, $J^* - J(\pi) = 1 > \epsilon$ for that MDP, which is a contradiction to the PAC property.

Furthermore, Yang et al.²³ show that PAC guarantees are possible only for *finitary* specifications (a specification ϕ is *finitary* if there exists a fixed H such that we can decide whether or not an infinite run $\rho = (s_0, s_1, \dots)$ of the system satisfies ϕ by only looking at the first H steps). PAC algorithms also exist under additional assumptions—for example, if a lower bound on all non-zero transition probabilities of the MDP is given.⁸

Finite Horizon Specifications

While infinite horizon poses many theoretical challenges for specification-guided RL, in practice, it is often sufficient to consider finitary specifications. If we have a fixed finite horizon H and the satisfaction of a specification ϕ only depends on the first H steps of the MDP, then a straightforward approach is to define a reward function r_ϕ that assigns a reward of 1 after H steps if ϕ is satisfied and 0 otherwise. It is easy to see that r_ϕ faithfully represents ϕ , and a policy π that maximizes the expected sum of rewards $J_H(\pi)$ also maxi-

mizes the probability of satisfying φ . However, using such *sparse* reward functions that “rarely” assign non-zero rewards often requires a large number of samples from the environment. For example, consider the warehouse environment with interconnected rooms shown in Figure 3a. This is an environment with *continuous* state and action spaces. Here, a state is a pair of coordinates $s = (x, y) \in \mathbb{R}^2$ indicating the location of the robot, and actions are velocities $a = (v_x, v_y) \in \mathbb{R}^2$. The initial distribution over states is $\eta = \text{Uniform}(A)$ (where A is the purple circle). Consider the specification φ requiring the robot to reach the region B starting from any position in A within H steps. If we use the above approach, then r_φ assigns the agent a non-zero reward only after it reaches B . Thus, the agent receives no meaningful reward feedback until it has already completed the entire task. Using state-of-the-art RL algorithms, the agent might explore for a long time before receiving such a reward. This problem is exacerbated for more complex specifications. A common solution is to use *dense* reward functions that assign non-zero rewards more frequently to guide the agent toward solving the task. For example, we can define a reward function $r_\varphi(s, a) = -\text{dist}(s, B)$ measuring the distance between the agent’s current state s and the goal region B , which encourages the robot to move toward B .

Another challenge with encoding specifications using reward functions is that specifications can be *non-Markovian* or *history dependent*. However, most existing RL algorithms do not allow rewards to be history dependent—that is, the reward at the current step $r(s, a)$ must be a function of the current state s and the action a , and should be independent of the states visited and the actions taken prior to the current step. For example, suppose φ requires the robot (starting from a state in A) to visit X first, go back to A , and then visit E . In this case, the reward provided to the agent should depend on whether the robot had previously visited X in the current run. If the robot has not yet visited X in the current run, then $r(s, a)$ should be higher if s is in X than if it is in E since the robot should not be encouraged to go directly to E without visit-

ing X . Conversely, if the robot has already visited X , then $r(s, a)$ should be higher if s is in E than if it is in X since the robot should not be encouraged to visit X again or to stay in X indefinitely. While such history-dependent tasks are easy to encode using logical specifications, they cannot be directly represented as rewards.

There has been a lot of recent work on designing practical specification-guided RL algorithms.^{1,5,6,10,11,18,22,24} Next, we describe one such algorithm, called DIRM, which is based on the SPECTRL language.

The SPECTRL language. SPECTRL is a simple specification language that lets users encode tasks such as the ones we discussed in the context of robot navigation.¹⁵ Its syntax is given by the grammar

$$\varphi := \text{Eventually } p \mid \varphi \text{ Ensuring } p \mid \varphi; \varphi \mid \varphi \vee \varphi$$

where $p \subset S$ is a predicate representing a subset of the states. In our warehouse example, the region $X \subseteq S = \mathbb{R}^2$ corresponds to such a predicate. Next, $\text{Eventually } p$ denotes eventually (within H steps) reaching a state $s \in p$, and $\varphi \text{ Ensuring } p$ denotes performing φ while remaining in the “safe” set p at every step. Furthermore, sequencing $\varphi_1; \varphi_2$ denotes first performing φ_1 and then φ_2 , and disjunction $\varphi_1 \vee \varphi_2$ gives the agent the choice of either performing φ_1 or φ_2 .

In our warehouse example, the task of visiting X while avoiding the obstacle region O (union of all regions marked in red) can be encoded in SPECTRL as $\varphi_1 = (\text{Eventually } X \text{ Ensuring } \neg O)$, where $\neg O = S \setminus O$ denotes the set of all states that do not belong in O . The task of visiting X first and then returning to A before visiting E can be encoded as $\varphi_2 = (\text{Eventually } X; \text{Eventually } A; \text{Eventually } E)$. Finally, $\varphi_3 = ((\text{Eventually } X \vee \text{Eventually } E); \text{Eventually } C) \text{ Ensuring } \neg O$ specifies that the robot must first visit either X or E and then visit C , all while avoiding O .

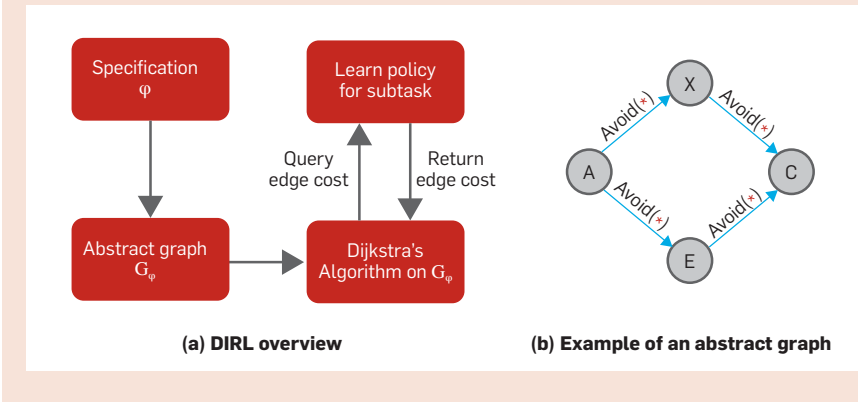
A specification-guided RL algorithm. Early work on specification-guided RL focused on compiling the specification to a reward function,^{15,18,22} including techniques for reward shaping and handling non-Markovian specifications. One effective strategy for reward shaping is to use *quantitative semantics* for LTL speci-

fications, which assign numerical values to runs of the system that capture how “robustly” the specification is satisfied (while ensuring runs that satisfy the specification are assigned higher values compared to those that do not). To handle history-dependent tasks, one approach is to alter the state space to include relevant information about the history—for example, as a finite state machine called a *reward machine*.¹² However, some tasks can be challenging to learn with these approaches since even with reward shaping, RL algorithms tend to be myopic, focusing on immediate rewards over long-term ones. Consider our example specification φ_3 ; a typical reward function implementing this task would provide similar rewards for visiting X and E , since these two cannot be distinguished without understanding the structure of the MDP. In the MDP, there is no direct way to reach C from E , so in fact X is preferable to E . Thus, if by chance an RL algorithm learns to solve the initial subtask $\text{Eventually } X \vee \text{Eventually } E$ by visiting E instead of X , then it might struggle to reach C .

We describe DIRM, a more recent specification-guided RL algorithm based on SPECTRL¹⁶ that addresses this issue by leveraging the structure of the specification in conjunction with knowledge about the MDP gained during training. DIRM consists of two components: using Dijkstra’s algorithm to search for a high-level sequence of *subtasks* to perform (a.k.a. planning), and using traditional RL algorithms to learn to perform subtasks (a.k.a. learning). The planning component involves computing a shortest path in a directed graph in which edges represent subtasks. The learning component involves, for each subtask, constructing a reward function and using an off-the-shelf RL algorithm to learn a policy to maximize the expected sum of rewards. Instead of performing learning and then planning, DIRM interleaves the two, incorporating information obtained during learning into planning and vice versa.

Figure 2a provides an overview of the algorithm. First, DIRM automatically constructs an *abstract graph* $G_\varphi = (\mathcal{V}, \mathcal{E})$ representing the high-level structure in the given SPECTRL speci-

Figure 2. Components of DIRL.



fication.^d For our example specification φ_3 , the corresponding abstract graph constructed by DIRL is shown in Figure 2b. Each vertex represents a set of states in the MDP. In this example, the abstract graph has four vertices representing the four regions of our warehouse environment. Each directed edge represents a *reachability subtask*—namely, the subtask of reaching the set of states represented by the target vertex. For instance, the edge from A to X denotes the subtask of reaching

X from A . Each edge is optionally labeled with a *path constraint*—that is, all edges are labeled with the constraint of avoiding the red obstacle regions. Our example graph has two paths from A to C representing the two ways of solving φ_3 : (i) first go to X and then go to C , or (ii) first go to E and then go to C . If the agent learns how to perform each subtask along a path in this graph, then the agent can perform the sequence of subtasks dictated by this path to achieve its objective. Thus, the overall problem reduces to two sub-problems corresponding to the two components of DIRL: (i) computing the “best” path

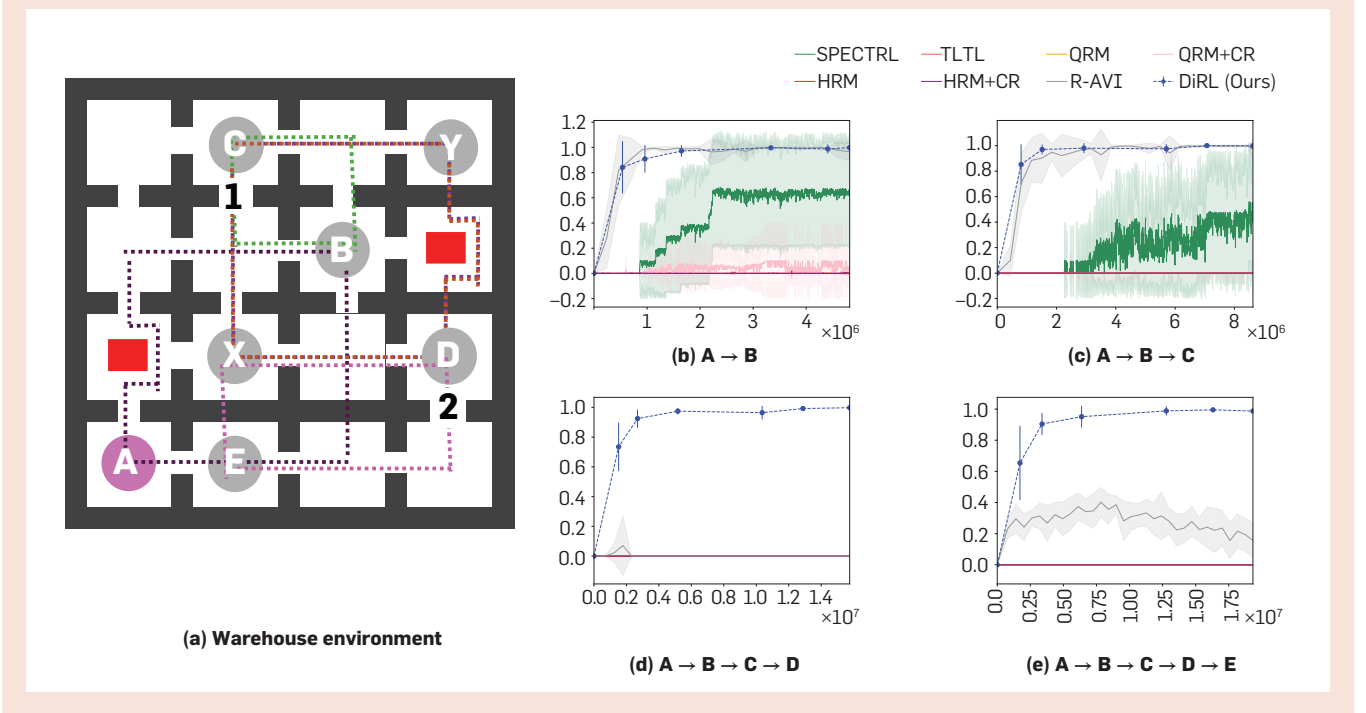
in G_φ from the initial vertex v_0 to a goal vertex v_g (achieved using Dijkstra’s algorithm^e), and (ii) learning to perform the subtask corresponding to each edge in that path (achieved using RL).

To apply Dijkstra’s algorithm to compute the “best” path, we need to associate a cost to each edge in the graph; ideally, this assignment should ensure the resulting shortest path corresponds to the high-level plan most likely to solve φ . For intuition, suppose the probability of completing a subtask corresponding to an edge $e = u \rightarrow v \in \mathcal{E}$ is independent of the starting state (that is, the specific state in the set of states represented by u), and that the agent has already learned how to complete the subtask corresponding to each edge in the graph. Then, letting p_e be the probability of completing the subtask corresponding to edge e , we can define the cost of e to be $c_e = -\log(p_e)$. Given a path $e_1, e_2, \dots, e_\ell$, the probability of successfully completing all corresponding subtasks in order is $\prod_{i=1}^{\ell} p_{e_i} = \exp(\sum_{i=1}^{\ell} \log(p_{e_i}))$. Thus, minimizing the sum of edge

^d We omit the procedure for constructing the abstract graph; see Jothimurugan¹⁶ for details.

^e The abstract graph is always finite and DIRL is applicable in environments with continuous state and action spaces.

Figure 3. Performance plots showing probability of specification satisfaction (y-axis) across specifications of increasing complexity against number of samples taken to learn (x-axis). $X \rightarrow Y$ is shorthand for specification to reach region Y from region X after visiting one of the corners in the square/rectangle formed by X and Y .



costs (that is, computing the shortest path) is equivalent to maximizing the probability of completing φ .

This strategy could be implemented by first running learning for each subtask, and then running Dijkstra's algorithm, with no feedback between the two. In practice, this approach does not work because our assumption that p_e is independent of the starting state does not hold. DIRM uses two ideas to address this issue. First, note we do not need the cost of all edges a priori. Thus, DIRM can estimate edge costs lazily—that is, it estimates the cost c_e of edge e only when it is needed by Dijkstra's algorithm. Whenever an edge cost needs to be estimated, DIRM constructs a (non-sparse) reward function for the corresponding subtask (using quantitative semantics), uses an off-the-shelf RL algorithm to train a policy to perform that subtask and then estimates p_e using simulations.

Second, a key property of Dijkstra's algorithm is that when the cost of an edge $e = u \rightarrow v$ is needed, it has already computed the shortest path from v_0 to u ; this also implies that DIRM has learned policies to solve subtasks for edges in this path. Thus, we can simulate the environment using these policies to obtain samples of states in u (assuming the agent successfully completes the path). This approach not only provides a way to sample initial states for training a subtask policy for e , but this distribution of initial states corresponds to the distribution of states encountered when using the learned subtask policy if the final path computed by DIRM contains e .

DIRM has been shown to scale well to complex specifications, including realistic applications such as robotic pick-and-place environments, where it outperforms prior approaches while requiring fewer than 4×10^5 samples. Figures 3b–3e show learning curves of DIRM for some complex specifications in the warehouse environment. Recent work has expanded SPECTRL-based approaches to multi-agent scenarios^{7,17} and learning verified policies.^{13,25}

Conclusion

In this article, we presented a tutorial-style introduction to recent research on using logical specifications to encode RL tasks. In particular, we discussed

both theoretical limitations and practical solutions. Using examples, we illustrated key ideas and concepts, such as differences between reward functions and logical specifications, specification robustness, graph representations of logical specifications, and integrating high-level planning with learning low-level control actions. We also demonstrated that one can leverage the structure in the logical specification to improve learning. One main caveat is the user needs to provide an appropriate specification for the task at hand, and the benefits of specification-guided RL are only fully realized when the task has a natural logical structure—for example, if the task is to simply reach a goal region that is challenging to navigate to (or requires complex maneuvers), specification-guided RL provides limited added benefits when compared to using reward functions since the high-level task is easily encoded using rewards. Nonetheless, specification-guided RL shows tremendous promise for many applications of RL with complex task structures. Several promising directions exist for future work, such as handling raw perceptual inputs including challenges regarding evaluating predicates under partial observability, leveraging LLMs to generate formal specifications from natural language descriptions, and zero-shot generalization to new tasks encoded using logical specifications. ■

References

- Aksaray, D. et al. Q-learning for robust satisfaction of signal temporal logic specifications. In *Conf. on Decision and Control (CDC)*. IEEE (2016), 6565–6570.
- Alur, R., Bansal, S., Bastani, O., and Jothimurugan, K. A framework for transforming specifications in reinforcement learning. *Principles of Systems Design: Essays Dedicated to Thomas A. Henzinger on the Occasion of His 60th Birthday* (2021).
- Alur, R. et al. Policy synthesis and reinforcement learning for discounted LTL. In *Computer Aided Verification*. Springer Nature Switzerland, Cham, (2023), 415–435.
- Bozkurt, A.K., Wang, Y., Zavanos, M.M., and Pajic, M. Control synthesis from linear temporal logic specifications using model-free reinforcement learning. In *Proceedings of the 2020 IEEE Intern. Conf. on Robotics and Automation (ICRA)*. IEEE, (2020), 10349–10355.
- Brafman, R., Giacomo, G.D., and Patrizi, F. LTLf/LDLf non-markovian rewards. In *Proceedings of the AAAI Conf. on Artificial Intelligence*, 32, (2018).
- Giacomo, G.D., Iocchi, L., Favorito, M., and Patrizi, F. Foundations for restraining bolts: Reinforcement learning with LTLf/LDLf restraining specifications. In *Proceedings of the Intern. Conf. on Automated Planning and Scheduling*, 29, (2019), 128–136.
- Eappen, J. and Jagannathan, S. DistSPECTRL: Distributing specifications in multi-agent reinforcement learning systems. *Joint European Conf. on Machine Learning and Knowledge Discovery in Databases*. Springer, 2022, 233–250.
- Fu, J. and Topcu, U. Probably approximately correct MDP learning and control with temporal logic constraints. *Robotics: Science and Systems* (2014).
- Hahn, E.M. et al. Omega-regular objectives in model-free reinforcement learning. *Tools and Algorithms for the Construction and Analysis of Systems* (2019), 395–412.
- Hasanbeig, M., Abate, A., and Kroening, D. Logically constrained reinforcement learning. arXiv preprint arXiv:1801.08099 (2018).
- Hasanbeig, M. et al. Reinforcement learning for temporal logic control synthesis with probabilistic satisfaction guarantees. In *Proc. Conf. on Decision and Control* (2019), 5338–5343.
- Icarte, R.T., Klassen, T., Valenzano, R., and McIlraith, S. Using reward machines for high-level task specification and decomposition in reinforcement learning. In *Proc. Intern. Conf. on Machine Learning*. PMLR, (2018), 2107–2116.
- Ivanov, R. et al. Compositional learning and verification of neural network controllers. *ACM Transactions on Embedded Computing Systems* (2021).
- Jiang, Y. et al. Temporal-logic-based reward shaping for continuing learning tasks. arXiv:2007.01498 [cs. AI] (2020).
- Jothimurugan, K., Alur, R., and Bastani, O. A composable specification language for reinforcement learning tasks. *Advances in Neural Information Processing Systems*, 32 (2019), 13041–13051.
- Jothimurugan, K., Bansal, S., Bastani, O., and Alur, R. Compositional reinforcement learning from logical specifications. *Advances in Neural Information Processing Systems* (2021).
- Jothimurugan, K., Bansal, S., Bastani, O., and Alur, R. Specification-guided learning of Nash equilibria with high social welfare. In *Proc. of the Intern. Conf. on Computer Aided Verification* (2022), 343–363.
- Li, X., Vasile, C.I., and Belta, C. Reinforcement learning with temporal logic rewards. In *Proc. of the IEEE/RSJ Intern. Conf. on Intelligent Robots and Systems* (2017), 3834–3839.
- Littman, M.L. et al. Environment-independent task specifications via GLTL. arXiv:1704.04341 [cs.AI] (2017).
- Pnueli, A. The temporal logic of programs. In *Proc. of the 18th Annual Symp. on Foundations of Computer Science*. IEEE, (1977), 46–57.
- Strehl, A.L. et al. PAC model-free reinforcement learning. In *Proceedings of the 23rd Intern. Conf. on Machine Learning*. (2006), 881–888.
- Xu, Z. and Topcu, U. Transfer of temporal logic formulas in reinforcement learning. In *Proc. of the Intern. Joint Conf. on Artificial Intelligence*. (2019), 4010–4018.
- Yang, C., Littman, M., and Carbin, M. Reinforcement learning for general LTL objectives is intractable. arXiv (2021).
- Yuan, L.Z., Hasanbeig, M., Abate, A., and Kroening, D. Modular deep reinforcement learning with temporal logic specifications. arXiv (2019).
- Žikelić, Đ. et al. Compositional policy learning in stochastic control systems with formal guarantees. *Advances in Neural Information Processing Systems* 36, (2024).

Rajeev Alur is a Zisman Family Professor and director of ASSET Center for Trustworthy AI at the University of Pennsylvania, Philadelphia, PA, USA.

Suguman Bansal is an assistant professor at the Georgia Institute of Technology, Atlanta, GA, USA.

Osbert Bastani is an assistant professor at the University of Pennsylvania, Philadelphia, PA, USA.

Kishor Jothimurugan is a quantitative researcher at Two Sigma Investments, LP, New York, NY, USA.

The views expressed herein are solely the views of the author(s) and are not necessarily the views of Two Sigma Investments, LP or any of its affiliates. They are not intended to provide, and should not be relied upon for, investment advice.

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

© 2026 Copyright held by the owner/author(s).

Beyond its utility in detecting AI vulnerabilities and bias, red-teaming raises important issues around values, labor, and harms.

BY TARLETON GILLESPIE, RYLAND SHAW,
MARY L. GRAY, AND JINA SUH

AI Red-Teaming Is a Sociotechnical Problem

LARGE LANGUAGE MODELS (LLMs) and image diffusion models have emerged rapidly—from research projects into the engines behind the current global deployment of generative AI. When a technology jumps this quickly from experimental plaything to widely available consumer service, many other elements settle in around it, often without much forethought; these include interfaces, policies, business models, labor arrangements, infrastructural assurances, complementary technologies, public claims, advertising campaigns, and regulations. Many of these decisions, arrangements, and infrastructures may turn out to be just as consequential for users and the broader public as the core technology itself. But the boisterous promises and debates that surround

the new technology can obscure these other essential elements—elements that make technologies much more than the sum of their engineered parts.

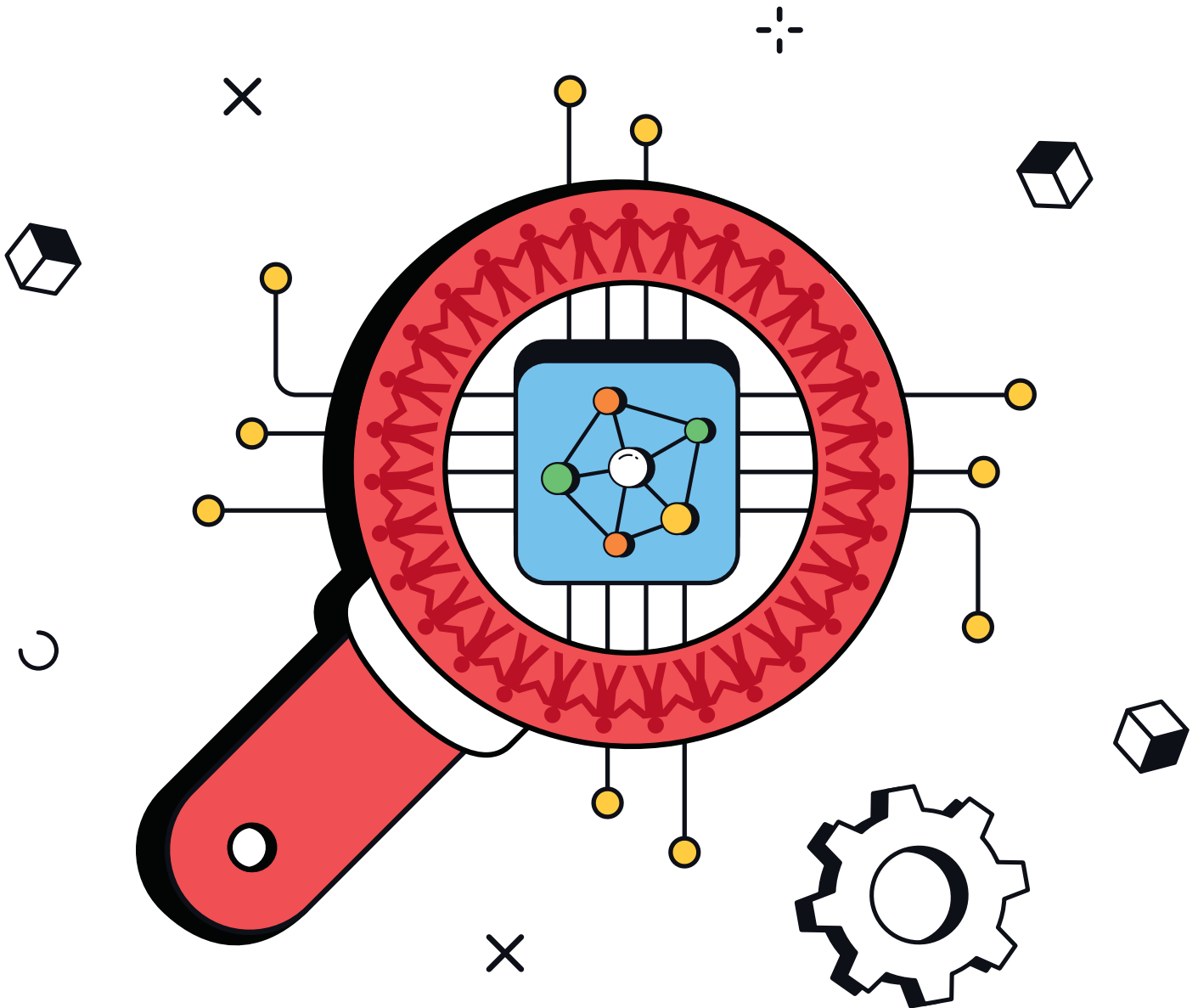
To keep pace, researchers studying the workings and implications of these technologies across computer science, engineering, the social sciences, humanities, and law must gear up just as quickly, focusing on not only the core technology but also the sociotechnical system taking shape around it²¹—and they need to work together. In this article, we call upon computer scientists and social scientists alike to pay closer, critical, collaborative attention to one part of AI development: *red-teaming*.^a

AI models and their applications typically undergo internal testing before release, and continue to be evaluated during use; one part of this testing, red-teaming, probes these applications for exploitable vulnerabilities, errors, and bias. From an AI-evaluation vantage point, red-teaming needs to be well-designed, effective, and replicable. But from a sociological vantage point,

^a The spelling of the term is inconsistent across different texts and organizations (sometimes even within organizations). Sometimes it is hyphenated or concatenated, other times not. In line with Friedler et al.,¹² we hyphenate the term except when specifically referencing the teams themselves.

» key insights

- **Red-teaming is already settling in as an essential step in the development of generative AI models, but it has not been sufficiently examined as a sociotechnical element of AI.**
- **A sociotechnical lens highlights red-teaming as a particular arrangement of human labor that unavoidably introduces human value judgments into technical systems, and can pose psychological risks for those who do this work—challenges similar to those faced by practitioners of content moderation.**
- **Studying the technical, social, critical, and policy dimensions of red-teaming, and of AI as an emerging sociotechnical system, will require a coordinated network of scholars from the full range of the computational and social sciences.**



red-teaming is something else as well: a specific kind of labor, done by specific sets of people, in specific institutional contexts, with its own specific set of implications.^{12,26} Our aim is to shed light on these sociotechnical elements, and to call for more cross-disciplinary attention to this critical component of AI development.

Since the commercial launch of ChatGPT, red-teaming has been quickly normalized as a step in the production and deployment of generative AI models. AI developers champion it as proof of their public responsibility, while regulators count on it as a bulwark preventing AI from inflicting social harms. But the public knows precious little about how this work is conducted, upon what values and assumptions it is based, who is enlisted to do it, or the psychological

costs they bear. This was also the case with content moderation: Hidden from public and critical scrutiny, its labor and well-being concerns were long overlooked. At the same time, the opacity of these value judgments soon became a political liability for social media platforms. And, given that what little the public does know about AI red-teaming comes largely from Silicon Valley's own promotional materials, and that the changing political climate, particularly in the U.S., may discourage such outward performances of responsibility, the public may soon know even less.

It is worth noting that this article is not based on any information internal to Microsoft. Rather, our observations are drawn from our recent studies of the labor behind responsible AI,^{17,41,42} the politics of generative AI as a new

media technology,¹⁶ and participatory approaches to AI development and governance.^{10,38,b} And, being trained in science and technology studies, labor, psychology, and design, we feel it is important to contextualize AI red-teaming within longer histories of scholarship that reckon with the social construction of new technologies.

What Is Red-Teaming?

Broadly, *red-teaming* means testing the safety and security of a system by methodically probing it as an adversary would. The U.S. military coined the term to describe the technique of assigning members of one's own forces to

^b Also see Microsoft Research's Project Resolve: <https://www.microsoft.com/en-us/research/project/project-resolve/>

act as the enemy during wargames and simulations, probing defensive strategies for potential weaknesses. During the Cold War, that presumed enemy was the Soviet Union, hence the color red.⁴⁰ But Zenko suggests that the idea of adversarial testing extends back much farther, noting that in the 13th century the Catholic church established the “Devil’s Advocate,” who interrogated those nominated for sainthood. Just as the Devil’s Advocate aimed to poke holes in nominees’ candidacies, military red teams attempted to infiltrate their own forces’ front lines.

The term migrated to the field of cybersecurity. Red teams were tasked with infiltrating information systems to simulate worst-case scenarios, such as the theft of sensitive information or hacks on infrastructure, that might lead to financial or operational disaster.⁴⁰ Red teams became an important element of systems security, that might even pay for themselves by anticipating and thereby preventing breaches.^c

For generative AI, red-teaming sometimes concerns the security of the foundation model, but increasingly it also means purposefully provoking the model to produce undesired or incorrect responses. Because users might either intentionally or inadvertently prompt an AI model to generate hateful, pornographic, vile, or biased responses, red teams attempt to do so first, acting as a kind of adversary to the intended or presumed use. By preemptively tempting the AI model to say things it should not, they can document how to shore up the safety architectures meant to prevent such responses. Still, what constitutes AI red-teaming remains fuzzy, given that its tactics and organizational structures are still forming. Its familial resemblance to evaluation, social engineering, bug bounties, threat assessments, and penetration testing is still “being discovered.”³⁶ Other terms from cybersecurity and hacker lingo describe much the same project: Google describes their AI Red Team as “ethical hackers,” a nod to venerable “white hat” hackers-for-hire who hunt for technical insecurities.^d

Some assert that red-teaming is es-

If we want AI that is not only safe and secure but also sustainable, we need to study the emerging labor arrangements that are critical to it.

sential to AI, and is the surest way to safeguard equitable and responsible AI development. Others worry that red-teaming and the guardrails it produces are a kind of “security theater,” more performative than substantive,¹¹ meant to obscure the reckless deployment of harmful technologies to the public. Still others suggest that such efforts to police AI models will hinder the true potential of the technology, ultimately leading to “woke AI.”^e

But in blunt economic terms, the first AI company to successfully “tame” their generative AI products through such safeguards could well capture the market. Business clients are asking for AI-powered customer service chatbots that do not hallucinate^f and productivity tools that behave consistently.^g There are enormous financial, reputational, and regulatory incentives to make generative AI tools safe and value-transparent, pushing these companies to rapidly institute red teams—arguably, faster than researchers concerned about the politics of AI can follow.

Red-teaming also offers a kind of reassurance, easing fears held by the public, governments, and financial stakeholders about the safety and performance of generative AI systems. New AI products are often touted as having been tested by red-teams before a wider release.^h The U.S. federal government (briefly) adopted the language of red-teaming as an important assurance of the safety of AI systems. In October 2023, President Biden issued an Executive Order that reinforced the importance of red-teaming as part of a proposed system of federal oversight over AI: “Any foundation model that poses a serious risk to national security, national economic security, or national public health and safety...must share

e On X, Elon Musk claimed a “woke AI” may eventually kill people: <https://x.com/elon-musk/status/1768746706043035827>

f In 2024, Air Canada was found liable for its customer support chatbot’s incorrect claim about a customer’s ticket: <https://bit.ly/3YgM5Ju>

g Complaints about AI tools getting “lazy” surface from time to time (<https://bit.ly/3KPipAg>), as well as concerns about generative AI systems degrading over time (<https://bit.ly/48ZEGoe>).

h For example, when OpenAI announced its Sora video generation model in early 2024, they prominently featured their red-teaming efforts on its launch page: <http://bit.ly/4prpq8E>

c <https://www.wired.com/story/microsoft-ai-red-team/>

d <https://bit.ly/4bhe6rU>

the results of all red-team safety tests.ⁱⁱⁱ That order has since been rescinded, but red-teaming will likely appear in subsequent efforts to regulate, or self-regulate, generative AI.

As a new labor formation, developing under financial and political pressure, red-teaming echoes other contingent forms of digital labor, such as data labeling, content moderation, and enrichment services of all kinds—all arrangements for supporting the semi-automation of human judgment critical to data-driven technical systems. If we want AI that is not only safe and secure but also sustainable, we need to study the emerging labor arrangements that are critical to it: the mix of internal teams testing out AI products, volunteers convened at hack-a-thon-like events, third-party crowdwork vendors, and professional security firms. And, as we learned in the case of social media, we must attend to the psychological costs of the work asked of red teams charged with making AI safe, secure, and useful for everyone.

Rather than focus exclusively on implementing and improving it, for the sake of the model, we need to better understand red-teaming as a practice, and understand its place in the development of generative AI tools, for the sake of the people involved. A sociological perspective, which is fundamentally human-centric, can better guide the responsible and effective use of red teams.

Value Judgments

Efforts to apply obligations, structures, and benchmarks to generative AI models have had to race to keep up with the rush to commercialize them.²³ Initial red-teaming efforts were being implemented even as design teams were still wondering which harms to even probe for: AI red-teamers have had to develop homegrown taxonomies of harms, and the measurement and benchmarking systems for mitigating them.^{18,39} In public statements, the major AI companies often state as a given that generative AI tools will unavoidably produce harmful content.²⁴ However, it is much rarer for them to discuss how they determine what counts as harmful content, what

they should and should not be looking for, and whether their own teams are best suited to make those judgments. This prompts the question: Whose values are being utilized for alignment and evaluation?¹¹

The development of AI red-teaming echoes the early days of commercial content moderation at social media platforms.^j The parallels are revealing. The categories of concern are strikingly similar: graphic violence, hate speech, harassment, discrimination, sexual content, terrorism, human trafficking, self-harm, child abuse, and misinformation.^{13,28} And when Silicon Valley found itself compelled to manage the gap between what can be generated online and what users should actually see, it too enlisted contingent human labor to serve as that filter.

Social media platforms “discovered” the need for human moderation labor after being surprised by the kinds of content that could turn up through their services.^{15,17,22} This awareness often came from user complaints, the technology press calling out the platforms’ shortcomings, and platform companies stumbling upon it themselves. AI red-teaming has similarly been fueled by user complaints and critical press coverage. Addressing these harms has to be done internally and largely in proprietary ways,^{6,31} making industry-wide or public-wide discussions about harms and values difficult to develop. But leaving the public out of this process leaves the value judgments to the AI designers themselves. Or, as OpenAI explained, echoing so many social media companies before them, “Our approach is to red-team iteratively, starting with an initial hypothesis of which areas may be the highest risk, testing these areas, and adjusting as we go.”²⁸

Tackling harmful content internally, intuitively, and iteratively has had profound implications for social media platforms over the past two decades; the same implications could befall the red-teaming of generative AI systems. Like AI red teams today, many social media platforms turned first to their own engineers and employees to evaluate for harmful content; to them, some categories of harms tended to seem more obvi-

ous; others tended to go unrecognized. Silicon Valley engineers generally do not reflect the range of identities and contexts of their global user base; social media platforms that began with their own employees often underestimated the harms faced by women, marginalized racial and ethnic groups, and those with stigmatized sexual and gender identities.^{14,31,33,34}

In the earliest days of social media platforms, the predominant approach to moderation was the “Feel bad? Take it down” rule.²² Subjective judgments and gut feelings stood in, often poorly, for the public’s understandings of harms or ethical principles. AI red-teaming strategies will have to be more complex and more inclusive to avoid the mistakes of social media’s past. However, with little opportunity for outside researchers to study commercial red teams, and little internal or external incentive to disclose much about their practices, we lack a clear empirical understanding of what harms may be under-attended to, or fall outside the purview of commercial organizations.

Labor Politics

The failure to appreciate the importance of human labor in AI systems, whether intentional or not, is common.^{17,32} To this point, it is illustrative that the verb form of *red team* is often used, eliding the human workforce that constitutes red teams. A sociotechnical examination of red-teaming should extend not only to the values and concerns behind the techniques red teams deploy, but also to the people doing the work and the labor arrangements within which they operate.

Red-teaming as a method is emerging in various forms: inside and outside companies, among salaried employees and volunteers, and involving those with access to the inner workings of the AI system and those without. These are bound to change; some forms will fall away, while others will settle in as “the way things are done.” But whatever particular labor politics do settle into place, there are important questions about the institutional contexts, material conditions, and economic incentives of this AI-related work, all ripe for scholarly analysis.

Who does this work internally can vary. Big tech companies are eager to

i White House Executive Order (14110) on the Safe, Secure and Trustworthy Development and Use of Artificial Intelligence, 2023.

j <https://www.techpolicy.press/ais-content-moderation-moment-is-here/>

boast about their flagship red teams,^k whose jobs are dedicated solely to ethical hacking, but we know that red-teaming also happens at smaller scales throughout the product-development cycle.³⁶ People doing the red-teaming may be part of larger responsible AI efforts, trust and safety divisions, or legal/compliance apparatuses.⁴¹ While those who perform red-teaming for their own companies typically enjoy the job security of full employment, they may not be in a position to refuse a red-teaming request. Though they are likely to have the necessary technical understanding of how models work, it is not clear that this is sufficient to effectively identify and mitigate AI risks.³⁶ And they may not be able to raise concerns publicly without breaching corporate norms or legally binding non-disclosure agreements.⁸

Employee red-teams may have little training in any other relevant proficiencies, whether linguistic, sociocultural, historical, legal, or ethical; the incentive structures do not ask for or reward such expertise. Some internal red teams might include people with sociocultural domain expertise, but they, too, work for the company and may have conflicting incentives.¹² Those windows of opportunity can open, a little. To test GPT-4 ahead of its March 2023 release, OpenAI solicited help from more than 50 experts, though primarily those with trust and safety and cybersecurity backgrounds.²⁸ Anthropic and Microsoft encourage their red teams to consult with experts to test specific types of harms.²⁷ These partnerships give AI companies access to experts without having to retain them as formal employees, compensating them in clout, API credits, job opportunities, or bragging rights rather than dollars.

Following a pattern ubiquitous in Silicon Valley, there have been increasing efforts to shift red-teaming labor from company employees to third-party data-work vendors, often overseas. Doing so uses labor arbitrage to drive down wages and, as the history of outsourcing and offshoring demonstrates, can erode worker protections and their capacity to

raise concerns about the effectiveness of their work. Early in their model development, researchers at Anthropic enlisted several hundred untrained crowdworkers and instructed them to “make the AI behave badly” to elicit harmful responses from their LLM chatbot.¹³ In this project, the crowdworkers were responsible for both probing the AI and assessing its responses for harmful content.

Either way, red teams can be an expensive undertaking. Smaller companies competing to bring their generative AI services to market may not be equipped to employ internal red-teaming services sufficient to satisfy internal liability concerns or regulatory obligations. They may turn to an emerging crop of boutique AI safety and data services^l that see market opportunities in sourcing red-teaming. As scholars have noted,¹⁷ when labor is offloaded, particularly piecemeal to a globally distributed contingent workforce, it becomes harder to trace and trickier to assert labor protections for it.

Some red-teaming happens outside the confines of AI companies and their outsourced labor pools. Hackers, volunteers, and everyday users also engage in forms of red-teaming. At DEFCON 2023, for instance, one of the largest hacking conferences in the world, more than 2,000 volunteers came together to prompt the largest LLMs into producing harmful content.^m Attendees included field experts, industry-based red-teams, cybersecurity consultants, and computer science Ph.D. students, but DEFCON organizers also dismantled some significant technical barriers to broaden participation to include novices—even children—with no programming knowledge.⁷

The event demonstrated the viability of convening a greater diversity of perspectives than is represented on most internal company red teams, something AI ethicist Rumman Chowdhury has argued is imperative when red-teaming on issues such as race, gender, sexuality, politics, and class.²⁹ However, asking marginalized communities to elaborate on their own marginalization for red-teaming can easily slip into a trans-

actional, extractive, and exploitative codependency.⁹ Volunteer and uncompensated work also requires significant effort to organize and has no obvious scaling mechanism. Access to generative AI models must be negotiated with AI companies, and taking any action on the insights gleaned depends entirely on the companies’ goodwill.

In many of the major generative AI applications, end users can also provide feedback, at least in limited forms. ChatGPT’s “thumbs down” includes a pull-down menu for users to indicate their concern: “Don’t like the style,” “Not factually correct,” “Refused when it shouldn’t have,” and others. But it is not evident how feedback from these slivers of digital “civic labor”²⁵ ever gets back to designers. User feedback is not strictly speaking red-teaming, as it isn’t necessarily adversarial. But end users do, whether inadvertently, playfully, or deviously, help designers surface outputs that companies did not anticipate or design for.

Red-teaming is still taking shape as a set of labor practices, inside and outside of AI companies. It is unclear whether companies’ internal red-teaming efforts are limited by the diversity of their employee pool, and how often they lean unfairly on their own minoritized employees to inject some diversity into their models.^{4,23} And, outsourced red-teaming work and using labor arbitrage to reduce costs raises tragically familiar concerns that echo social media companies’ use of human labor: leaving workers with fewer labor protections, more adverse work conditions, and more precarious job security, with few or no avenues for career development.

Well-Being of Red-Teamers

Beyond understanding *who* is doing the AI red-teaming and *what* is being evaluated, we also need to pay attention to the human *cost* of doing such work. Scholars and practitioners involved in red-teaming call it “rather unsavory work”³ and “mentally taxing.”²⁷ Like content-moderation work, the adversarial probing of red-teaming requires workers to imagine worst-case scenarios and expose themselves to potentially troubling outputs. Red-teamers are often required to assume the persona of a potential adversary (for example: online harasser, sex trafficker, racist, ter-

^k Google even produced a short documentary on their premier red-teams that has been viewed over 1.5 million times as of December 2025: <https://youtu.be/TusQWn2TQxQ?si=UnHO87JYDK1dblvK>

^l Such as Noma, Hidden Layer, Protect AI, and Mindgard

^m <https://cyberscoop.com/def-con-ai-hacking-red-team/>

rorist), invent a plan that this adversary may use to compromise the system, and evaluate the output for potential harms. They may also assume the persona of a benign user with specific intents or contexts (for example, a user with a history of eating disorders looking for dieting advice) and try to reveal system vulnerabilities that might be harmful. A red-teamer may first research harmful groups to learn their behaviors and bring that knowledge into red-teaming the models for hate speech or deep-fakes. They may immerse themselves in child online safety concerns to then evaluate the model's capabilities in aiding child exploitation.

To date, there is little empirical research about the psychological impact of AI red-teaming. But one recent study of RAI content workers shows that red-teamers, content moderators, and data labelers face similar psychological challenges when working with potentially harmful content.⁴¹ Given the extensive research on content moderation and the well-documented occupational health concerns experienced by professions that contend with trauma exposure,³⁷ red-teamers stand to benefit from attention to this history.

Moreover, the success of a red-team operation depends on uncovering and reviewing increasingly harmful content. Much like content moderators, emergency responders, journalists, or police investigating distressing events, AI red-teamers may be exposed repeatedly to disturbing and traumatic content that can lead to negative psychological symptoms. For example, content moderators removing harmful and offensive material from platforms have reported mental health challenges extending long after the work is complete, leading to documented cases of post-traumatic stress disorder (PTSD) and secondary traumatic stress (STS).^{2,33} Prolonged exposure may contribute to long-lasting mental health symptoms, alterations to their personal belief systems, and increased risk of physical health issues and substance abuse.³⁷ In fact, repeated work-related exposure to traumatic content is among the diagnostic criteria used for PTSD in the *Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition (DSM-5)* and has subsequently been used to support a series of recent lawsuits brought by content moderators



A successful AI red-teamer must exhibit an antagonistic imagination to be effective. Or, as one red-teamer put it, “If there were a red-team motto, it would be: *The more sinister your imagination, the better your work.*”



against their employers.^{30,n} Although many major platforms were slow to deal with this problem, most now offer mental health support and take measures to limit moderators' exposure to the most reprehensible content. These parallels underscore the importance of initiating research to protect red-teamers from the psychological hazards inherent in their work.

AI red-teaming also introduces other distinct psychological challenges. A successful AI red-teamer must exhibit an antagonistic imagination to be effective. Or, as one red-teamer put it, “If there were a red-team motto, it would be: *The more sinister your imagination, the better your work.*”^o Red-teaming involves people deliberately engaging in transgressive, uncomfortable, unethical, immoral, or harmful activities, including immersing themselves in scenarios that go against their morals or belief systems—to think like a harasser, or feel like a target of discrimination. Such practice can lead to “moral injury,”³⁵ a form of psychological distress that stems from actions, or the lack thereof, that violate one's moral or ethical code.⁴¹ Those who cannot safely detach their personal identity from their transgressions may experience negative self-perception and guilt. Regularly breaking the rules for the greater good can lead to a “loss of self,” sometimes seen in the undercover police profession.²⁰

The potential negative impacts on red-teamers' well-being have been acknowledged by some of those who organize such work. For example, organizers of the DEFCON Generative Red Team event anticipated that models generating unexpected harmful outputs might be triggering to participants.⁷ Anthropic's early red-teaming efforts involved consultations with trust and safety professionals to design safety measures for their crowdworkers.¹³ Strategies for preserving the well-being of red-teamers could include providing warnings about sensitive content, allowing opt-outs, encouraging breaks, monitoring mood, or allowing them to choose topics within

n <https://www.nytimes.com/2018/09/25/technology/facebook-moderator-job-ptsd-lawsuit.html>

o <https://www.bostonglobe.com/2024/01/11/opinion/ai-testing-red-team-human-toll/>

their own risk tolerance.^{13,27,28} But for volunteer or crowdsourced red-teamers, such strategies are limited to what the organizers are willing to provide.

For professional red-teamers, companies sometimes offer employee assistance programs (EAPs) with mental health resources. However, organizational factors often affect how these resources are actually used. For some, accessing a therapist may be a luxury they cannot afford, because of a psychologically unsafe work environment, unrelenting performance metrics, or job insecurity. Non-disclosure agreements (NDAs) have historically prevented workers from speaking up about working conditions.³⁷ Monitoring red-teamers' well-being gets entangled with more unsavory forms of workplace surveillance,¹ and increases liability risks for employers if the findings are severely negative. And implementing consistent well-being strategies can become infeasible for red-teamers working across organizational and national boundaries. So when organizers claim red-teaming events were safe because no one used the on-call therapists,⁷ they may be mistakenly assuming that no usage means no need.

Beyond individual mental health resources, one of the most effective tools for red-teamer well-being may be social support through a community of workers in similar roles, as well as family and friends. Prior research on content moderators demonstrates that the validation and belonging that comes from social support are essential.³⁷ Informal red-team communities¹⁹ that have emerged on social networks to share strategies can act as safe spaces to heal from shared trauma, especially when red-teamers may be reluctant to share their experiences with loved ones to protect them from exposure.³⁷

In an attempt to minimize human exposure to the “unsavory work” of red-teaming, some researchers have advocated for automated red-teaming. Whether or not this is possible—some leading AI safety researchers have recently argued that the “human element” will always be necessary⁵—automation may inadvertently make the human work even more invisible,¹⁷ diverting resources away from well-being measures needed by those red-teamers who remain. Therefore, it is important to criti-



If we hope to improve red-teaming outcomes as part of a sociotechnical system, we must study how its judgments are made and by whom.



cally examine the role of automation in red-teaming—not only its immediate impact on tasks, but also its long-term effects on the overall ecosystem.

We must acknowledge the sobering reality that comes with the commoditization of AI harm reduction. As long as generative AI remains integral to our lives, the work of AI red-teaming and its psychological implications will persist. To support this workforce, it is crucial to rigorously study and validate the effectiveness of innovative well-being strategies across various contexts, with close examination of the surrounding organizational and social structures.

Conclusion

What would more substantive, empirical, and cross-disciplinary research on red-teaming provide? Today's siloed, firewalled, market-reactive approach to red-teaming has potential drawbacks for AI consumers, red-teamers, and companies. Each company is rapidly developing its own version of red-teaming, with definitions and workloads varying based on the company's priorities, “brand,” and particular focus. Almost every company working on generative AI today has a red-team workforce of some kind. While there is energy being put toward addressing biases and other “embedded harms,” plenty of red-teaming efforts are more concerned with ungrounded existential risks than with current tangible concerns. And it is unclear how many issues identified through red-teaming efforts have been mitigated.

If we hope to improve red-teaming outcomes as part of a sociotechnical system, we must study how its judgments are made and by whom. Its compartmentalization and operational opacity can both alienate workers and keep the public from understanding fully what AI systems can offer. Empirical study can challenge these arrangements, identify the barriers and incentives at play, and perhaps point the public and AI companies toward more sustainable alternatives. An organized, crossdisciplinary, empirical research agenda that examines red-teaming as a sociotechnical undertaking could make both AI companies and the public more aware of and intentional about red-teaming work. Regardless of its current effectiveness or its future improvements, red-team-

ing has underlying logics and structural conditions that need examination. Lessons from content moderation show that finding and removing “bad elements” demands the deliberations and value judgments of teams of people. The specific conditions under which people do this hard work, at an unprecedented global scale and across myriad institutional settings, matters for both red-team workers’ occupational health and for the integrity of our technical and informational ecosystems.

And like most data work, the notion that this labor is only temporary, that it will soon be automated away, is wrong, and (deliberately) distracts from these sociological concerns. It is difficult to believe that we can ever fully automate such peculiarly human judgments, about contentious and shifting topics, under pressure from regulators and the public, whose ethical frameworks can themselves shift. But even if it could be automated in the future, real people are doing this work right now, and with real consequences. It makes little difference whether we discard this phantasmic notion of full automation entirely, or just concede that data labor will be with us for the foreseeable future. Either way, research today can help structure this work in ways that are more attentive to the well-being and labor rights of the people doing the work right now.

In fact, we may need not just more empirical study of red-teaming, but also a coordinated network of scholars studying red-teaming: as a multi-faceted practice, as a component in the institutional and labor arrangements of Silicon Valley, as a global public health concern, and as a hidden value system buried in our newest tools of expression and knowledge. The field of computer science has, in the past decade, begun to recognize that information systems are also labor systems and value systems—growing networks like the ACM Conference on Fairness, Accountability, and Transparency illustrate this—and the field is grappling with the implications of that in ways it had not before. Again, we might learn from the rise of content moderation and the research that attended to it: Many excellent scholars studying content moderation challenged its underlying logics and structural conditions. But, perhaps,

a coordinated network of scholars to deepen, circulate, and affirm those insights could have had a more substantive impact on these arrangements. It is not too late to pose an empirical and coordinated challenge to red-teaming, and to the many forms of labor and values on which AI systems depend.

Acknowledgments

The authors thank Jordan Ali for feedback and research assistance. 

References

- Ajunwa, I., Crawford, K., and Schultz, J. Limitless worker surveillance. *California Law Rev.* (2017).
- Arsht, A. and Etcovitch, D. The human cost of online content moderation. *Harvard J. of Law and Technology* 2 (2018).
- Bai, Y. et al. Constitutional AI: Harmlessness from AI feedback. *arXiv* (2022); <http://arxiv.org/abs/2212.08073>
- Benjamin, R. *Race after Technology*. Polity (2019).
- Bullwinkel, B. et al. Lessons from red teaming 100 generative AI products. *arXiv* (2025); <https://arxiv.org/abs/2501.07238>
- Carmi, E. The hidden listeners: Regulating the line from telephone operators to content moderators. *Intern. J. of Communication* 13 (2019), 440–458.
- Cattell, S. Generative red team recap (2023); <https://aivillage.org/defcon%2031/generative-recap/>
- Christin, A. et al. Internal fractures: The competing logics of social media platforms. *Social Media + Society* 10, 3 (Jul. 2024).
- Dalal, S., Hall, S. M., and Johnson, N. Provocation: Who benefits from “inclusion” in generative AI? *arXiv* (2024); <https://arxiv.org/abs/2411.09102>
- Deng, W. H. et al. Responsible crowdsourcing for responsible generative AI: Engaging crowds in AI auditing and evaluation. In *Proceedings of the AAAI Conf. on Human Computation and Crowdsourcing*. AAAI (2024), 148–150.
- Feffer, M., Sinha, A., Lipton, Z. C., and Heidari, H. Red-teaming for generative AI: Silver bullet or security theater? *arXiv* (2024); <https://arxiv.org/abs/2401.15897>
- Friedler, S. et al. AI red-teaming is not a one-stop solution to AI harms: Recommendations for using red-teaming for AI accountability. *Data & Society* (2023).
- Ganguli, D. et al. Red teaming language models to reduce harms: Methods, scaling behaviors, and lessons learned. *arXiv* (2022); <http://arxiv.org/abs/2209.07858>
- Gerrard, Y. and Thornham, H. Content moderation: social media’s sexist assemblages. *New Media & Society* 22, 7 (Jul. 2020), 1266–1286.
- Gillespie, T. *Custodians of the Internet: Platforms, Content Moderation, and the Hidden Decisions that Shape Social Media*. Yale University Press (2018).
- Gillespie, T. Generative AI and the politics of visibility. *Big Data & Society* 11, 2 (Jun. 2024).
- Gray, M. L. and Suri, S. *Ghost Work: How to Stop Silicon Valley from Building a New Global Underclass*. HarperCollins (2019).
- Hao, S. et al. Safety and fairness for content moderation in generative models. *arXiv* (2023); <http://arxiv.org/abs/2306.06135>
- Inie, N., Stray, J., and Derczynski, L. Summon a demon and bind it: A grounded theory of LLM red teaming in the wild. *arXiv* (2023); <https://arxiv.org/abs/2311.06237>
- Joh, E. E. Breaking the law to enforce it: Undercover police participation in crime. *Stanford Law Rev.* 62 (2009), 155; <https://api.semanticscholar.org/CorpusID:152308574>
- Joyce, K. et al. Toward a sociology of artificial intelligence: A call for research on inequalities and structural change. *Socius: Sociological Research for a Dynamic World* 7 (Jan. 2021).
- Klonick, K. The new governors: The people, rules and processes governing online speech. *Harvard Law Rev.* 131 (2018), 73.
- Lazar, S. and Nelson, A. AI safety on whose terms? *Science* 381, 6654 (Jul. 2023), 138–138.
- Marres, N. et al. AI as super-controversy: Eliciting AI and society controversies with an extended expert community in the UK. *Big Data & Society* 11, 2 (Jun. 2024).
- Matias, J. N. The civic labor of volunteer moderators online. *Social Media + Society* 5, 2 (Apr. 2019).
- Metcalfe, J. and Singh, R. Scaling up mischief: Red-teaming AI and distributing governance. *Harvard Data Science Rev. Special Issue* 4 (Dec. 2023).
- Microsoft. Introduction to red teaming large language models (LLMs) (2023); <https://learn.microsoft.com/en-us/azure/ai-services/openai/concepts/red-teaming>
- OpenAI. GPT-4 system card (2023); <https://cdn.openai.com/papers/gpt-4-system-card.pdf>
- Oremus, W. AI ‘red teams’ race to find bias and harms in chatbots like ChatGPT. *The Washington Post* (Aug. 2023).
- Pinchevski, A. Social media’s canaries: content moderators between digital labor and mediated trauma. *Media, Culture & Society* 45, 1 (Jan. 2023), 212–221.
- Roberts, S. T. Commercial content moderation: digital laborers’ dirty work. In *Intersectional Internet: Race, Sex, Class and Culture Online*. S. Umaña Noble and B. Tynes (Eds.). Peter Lang Publishing Inc. (2016), 147–159.
- Roberts, S. T. *Behind the Screen: Content Moderation in the Shadows of Social Media*. Yale University Press (2019).
- Ruckenstein, M. and Turunen, L. L. M. Re-humanizing the platform: Content moderators and the logic of care. *New Media & Society* (2019).
- Shahid, F. and Vashistha, A. Decolonizing content moderation: Does uniform global community standard resemble utopian equality or western power hegemony? In *Proceedings of the 2023 CHI Conf. on Human Factors in Computing Systems*. ACM (2023), 1–18.
- Shay, J. Moral injury. *Psychoanalytic Psychology* 31, 2 (2014), 182.
- Singh, R. et al. *Red-Teaming in the Public Interest*. Technical Report. Data & Society Research Institute and AI Risk and Vulnerability Institute (2025).
- Steiger, M. et al. The psychological well-being of content moderators: The emotional labor of commercial moderation and avenues for improving support. In *Proceedings of the 2021 CHI Conf. on Human Factors in Computing Systems*. ACM (2021).
- Suresh, H. et al. Participation in the age of foundation models. In *Proceedings of the 2024 ACM Conf. on Fairness, Accountability, and Transparency*. ACM (2024), 1609–1621.
- Weidinger, L. et al. Sociotechnical safety evaluation of generative AI systems. *arXiv* (2023); <http://arxiv.org/abs/2310.11986>
- Zenko, M. *Red Team: How to Succeed by Thinking Like the Enemy*. Basic Books (2015).
- Zhang, A. Q. et al. AURA: Amplifying understanding, resilience, and awareness for responsible AI content work. *arXiv* (2024); <https://arxiv.org/abs/2411.01426>
- Zhang, A. Q. et al. The human factor in AI red teaming: perspectives from social and collaborative computing. In *Companion Publication of the 2024 Conf. on Computer-Supported Cooperative Work and Social Computing*. ACM (2024), 712–715.

Tarleton Gillespie is a senior principal researcher at Microsoft Research, Cambridge, MA, USA, and an affiliated professor in the Department of Communication and the Department of Information Science at Cornell University, Ithaca, NY, USA.

Ryland Shaw (ryland.shaw@usc.edu) is a doctoral student in the Annenberg School of Communication at the University of Southern California, Los Angeles, CA, USA, and was a predoctoral research assistant at Microsoft Research, Cambridge, MA, USA, at the time of this writing.

Mary L. Gray is a senior principal researcher at Microsoft Research, Cambridge, MA, USA.

Jina Suh is a principal researcher at Microsoft Research, Redmond, WA, USA.

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

© 2026 Copyright held by the owner/author(s).

NOW OPEN FOR SUBMISSIONS

ACM Transactions on AI Security and Privacy (TAISAP)

Editors-in-Chief

Murat Kantarcioglu, *Virginia Tech (USA)*

Patrick McDaniel, *University of Wisconsin, Madison (USA)*

Sagar Samtani, *Indiana University, Bloomington (USA)*



**Research investigating all aspects of AI security and privacy,
including models, systems, and environments**

ACM Transactions on AI Security and Privacy (TAISAP) publishes research investigating all aspects of AI security and privacy, including models, systems, and environments. Research areas related to AI security include examining adversarial attacks, technical vulnerabilities within AI, and privacy concerns related to AI model training and deployment. It also welcomes submissions exploring AI's application in cybersecurity, such as within security operations centers, open source software security, and cyber threat intelligence. Key themes include ensuring robust and interpretable AI for enhancing security and privacy – along with reliable performance. TAISAP encourages interdisciplinary research and submissions documenting both theoretical and applied work, including innovative methods to assess threats and defenses for AI security and privacy, and to develop advanced AI-enabled cybersecurity analytics.

TAISAP is now open for submissions. In keeping with ACM's goal of making all its publications open access by January 2026, all papers will be published open access, with no publication charges for the first three years.

For more information, please visit taisap.acm.org



Association for
Computing Machinery

research highlights

P. 98

Technical Perspective Perspective on Cardinality Estimation Graphs

By Dan Suci

P. 99

Cardinality Estimation Graphs

By Semih Salihoğlu, Jeremy Chen,
Yuqing Huang, Mushi Wang, and Ken Salem

Technical Perspective

Perspective on Cardinality Estimation Graphs

By Dan Suciu

QUERY ENGINES ARE really good at choosing an efficient query plan. Users do not need to worry about how they write their query, since the optimizer makes all the right choices for executing the query while taking into account all aspects of data, such as its size, storage-device characteristics, the distribution pattern, the availability of indexes, and so on. The query optimizer always makes the best choice, no matter how complex the query is or how contrived it was written. Or, this is what we expect today from a modern query optimizer. Unfortunately, reality is not as nice.

The problem is that query optimizers have an Achilles' heel: cardinality estimation. To choose an optimal query execution plan, they need to search over many alternative plans and evaluate their cost; and to evaluate their cost, the optimizer needs to estimate the number of tuples that will result for each subexpression of the query. If they estimate the number of tuples incorrectly, then their cost estimate is wrong, and they end up choosing an inefficient query plan. When query optimizers fail to choose the best execution plan, it is almost always because they were informed incorrectly by the cardinality estimator.

The cardinality estimation problem is easy to state. Given a query, estimate the cardinality of its answer without evaluating the query. Typically, the query is restricted to selection and join operators, since other relational operators (group by, duplicate elimination, union, and so on) are usually pushed to the top of the query plan. To perform cardinality estimation, the system computes offline some statistics on each base table, then uses them at optimization time to estimate the cardinality of various query expressions. There are two major challenges. First, there is a very limited space budget for what statistics we can precompute and store, because all statistics must be kept in main memory during query optimization. Production databases often have hundreds of tables, some with hundreds of attributes;

hence, we can afford to store only very little information about each table and each attribute. Second, cardinality estimation must be very fast, because it is in the inner loop of the query optimizer, which needs to invoke the estimator thousands of times during optimization. Disk accesses, or solving complex optimization problems are ruled out: Cardinality estimation must be simple and efficient.

The accompanying paper, "Cardinality Estimation Graphs," by Chen et al. presents a unified framework that captures the most common approaches to cardinality estimation: those based on summary statistics (as opposed to samples) and those based on average or maximum degree (which includes all estimators I am aware of). There are two separate tasks in cardinality estimation, which are often studied separately: selectivity estimation of predicates on the base tables, and the cardinality estimation of a multi-join query. The paper studies the second task exclusively.

A cardinality estimate of a multi-join query relies on average degrees. Consider estimating the size of a join, $R(A,B) \bowtie S(B,C)$. A reasonable estimate is the cardinality of R times the average degree of the attribute $S.B$; for example, if each value of B occurs five times in S , then our estimate is five times $|R|$. But we can also switch the roles of R and S and use as an estimate $|S|$ times the

average degree of $R.B$. This presents a problem: Which estimate should we use? Database systems choose the smallest of these two, and this leads to the classic textbook formula we like to teach in undergraduate classes: $|R \bowtie S| = |R| \cdot |S| / \max(\text{dom}(R.B), \text{dom}(S.B))$.

It becomes clear that, for an n -way join query, the number of choices becomes very large, and it is unclear which one to use as an estimate. The paper introduces a simple abstraction, called the cardinality estimation graph (CEG), where each path represents one possible choice for the cardinality estimator. Using a CEG, the paper can explain and compare the various cardinality estimators described in the literature, opening up the possibility for designing new estimators.

A recently proposed alternative to computing a cardinality estimate is to compute an upper bound. This approach is called a pessimistic cardinality estimate (with some abuse, since it is not an estimate but a guaranteed upper bound). The definitive mathematical formula for the upper bound of a multi-join query is known but impractical, and the few implementations of pessimistic estimates use relaxations of this formula. The paper shows that the same CEG abstraction can also be used to explain these pessimistic cardinality estimates by simply replacing the average degree with a maximum degree.

Readers with little knowledge of cardinality estimation will hopefully find the CEGs a convenient, quick introduction to various cardinality estimation methods. Experts will find the CEG a convenient abstraction for comparing various estimators. And systems builders may find the experimental section a useful guide in making choices for their own cardinality estimator. 

Dan Suciu is a Microsoft Endowed Professor in computer science and engineering at the University of Washington, Seattle, WA, USA.

This Technical Perspective first appeared in the *ACM SIGMOD Record*, 52, 1, (June 2023).

© 2026 Copyright held by the owner/author(s).

The cardinality estimation problem is easy to state. Given a query, estimate the cardinality of its answer without evaluating the query.

Cardinality Estimation Graphs

By Semih Salihoğlu, Jeremy Chen, Yuqing Huang, Mushi Wang, and Ken Salem

ABSTRACT

We study two classes of summary-based cardinality estimators that use statistics about input relations and joins of a small number of input relations: (i) optimistic estimators, which were defined in the context of graph database management systems, that make uniformity and conditional independence assumptions; and (ii) the recent pessimistic estimators that use information theoretic linear programs (LPs). We show that optimistic estimators can be modeled as picking bottom-to-top paths in a *cardinality estimation graph* (CEG), which contains sub-queries as nodes and edges whose weights are average degree statistics. We show that existing optimistic estimators have either undefined or fixed choices for picking CEG paths as their estimates and ignore alternative choices. Instead, we outline a space of optimistic estimators to make an estimate on CEGs, which subsumes existing estimators. We show, using an extensive empirical analysis, that effective paths depend on the structure of the queries. We next show that optimistic estimators and seemingly disparate LP-based pessimistic estimators are in fact connected. Specifically, we show that CEGs can also model some recent pessimistic estimators. This connection allows us to provide insights into the pessimistic estimators, such as showing that they have combinatorial solutions.

1. INTRODUCTION

The problem of estimating the output size of a natural multi-join query (henceforth *join query*) is a fundamental problem that is solved in the query optimizers of database management systems to generate efficient query plans. This problem arises both in relational systems as well as those that manage graph-structured data, where systems need to estimate the cardinalities of subgraphs in their input graphs. It is well known that both problems are equivalent, since subgraph queries can equivalently be written as join queries over binary relations that store the edges of a graph.

Perhaps the most widely adopted estimators are *summary-based*, in which the DBMS uses statistics about a database in a procedure to derive an estimate. In this paper, we study two classes of summary-based techniques, which we will model through a common framework:

- *Optimistic estimators*^{2,14,16,18} from graph-based systems store as statistics the cardinalities of small-size subgraphs, which correspond to joins of a small number of relations and use algebraic formulas that make independence and uniformity assumptions as a procedure. We refer to these

as “optimistic,” as these estimators can underestimate true cardinalities of queries.

- *Pessimistic estimators*,^{5,11} which were introduced in the context of relational systems, store as statistics the *degrees* of the values in columns, that is, the number of times a value appears, and use linear programs (LPs) and have the guarantee that the estimate is an upper bound on the true cardinality of a query.

To motivate our first contribution, consider a database consisting of an edge-labeled graph with capital letter labels A, B, C , and so on. This can be modeled as a set of binary relations, such as $A(\text{src}, \text{dst})$. Consider the “subgraph query” in Figure 1, asking for matches of a five-edge subgraph, which is equivalent to a join of five relations. Given the accurate cardinalities of all sub-queries of size ≤ 2 , there are 252 formulas to make an estimate. Two examples are:

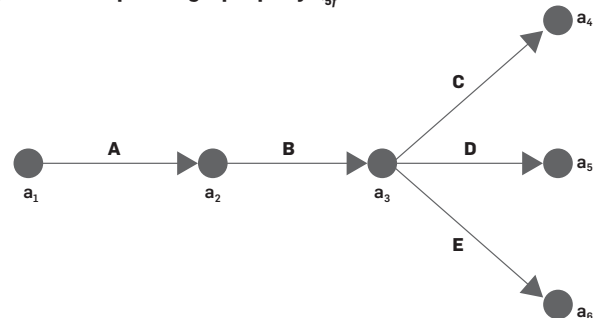
$$\begin{aligned} &\triangleright |A \rightarrow B| \times \frac{|B, C|}{|B|} \times \frac{|C, D|}{|C|} \times \frac{|D, E|}{|D|} \\ &\triangleright |A \rightarrow B| \times \frac{|B, D|}{|B|} \times \frac{|C, D|}{|C|} \times \frac{|B, E|}{|B|} \end{aligned}$$

In previous work,^{2,14,16} the choice of which of these estimates to use has either been unspecified or fixed without acknowledging possible other choices. Our first contribution aims to answer the following questions: What is the space of formulas of a query? Which formulas should be picked to make more accurate estimates?

We begin by showing that the algebraic formulas of prior optimistic estimators can be modeled as picking a bottom-to-top path in a weighted *cardinality estimation graph* (CEG), which we call CEG_Q , for **Optimistic**. In this CEG, nodes are intermediate sub-queries and edges; weights are average degree statistics that extend sub-queries to larger queries. For example, the CEG_Q for the query in Figure 1 and the dataset in Figure 2 is shown in Figure 3. Each path of this CEG corresponds to a possible formula where the estimate is the multiplication of the weights of the edges in the path.

We systematically describe a space of nine estimators, defined by different choices to pick a CEG path for making an

Figure 1. Example subgraph query Q_{sr}



The original version of this paper was published in *ACM SIGMOD Record*, 52, 1 (June 2023), 94–102.

optimistic estimate. This space subsumes and extends the choices made by existing optimistic estimators. We then empirically show that the better-performing optimistic estimators in this CEG space depend on the structure of the query. On acyclic queries and queries with small-size cycles, we advise using the *maximum-weight paths*; using independence assumptions tends to underestimate the true cardinalities of queries, which can be offset by picking the highest-estimat-

ing formula. In contrast, on queries that contain larger cycles, optimistic estimators estimate the number of paths rather than cycles. Here, we advise using minimum-weight paths, as real-world graphs contain many more paths than cycles.

As our next main contribution, we show that CEGs are expressive enough to also model the recent LP-based pessimistic estimators. Specifically, we show we can replace the edge weights of CEG_o (which are average degrees) with maximum

Figure 2. Example dataset in graph and relational formats.

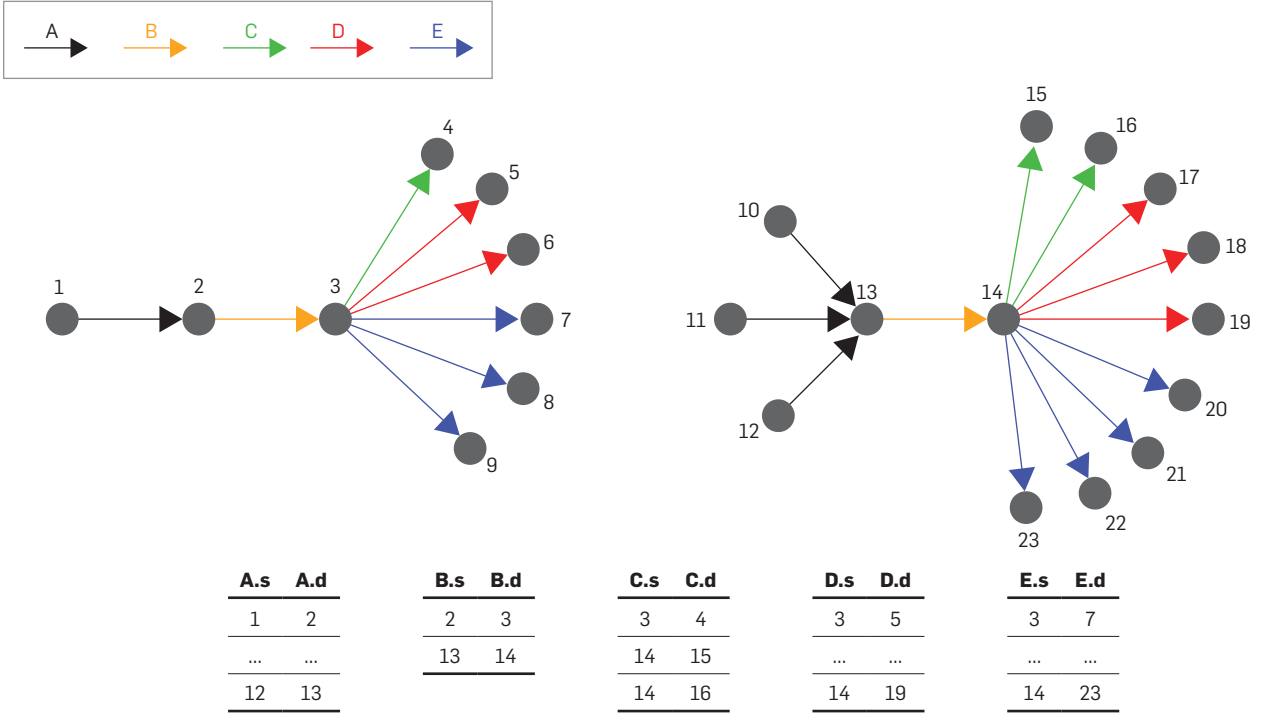
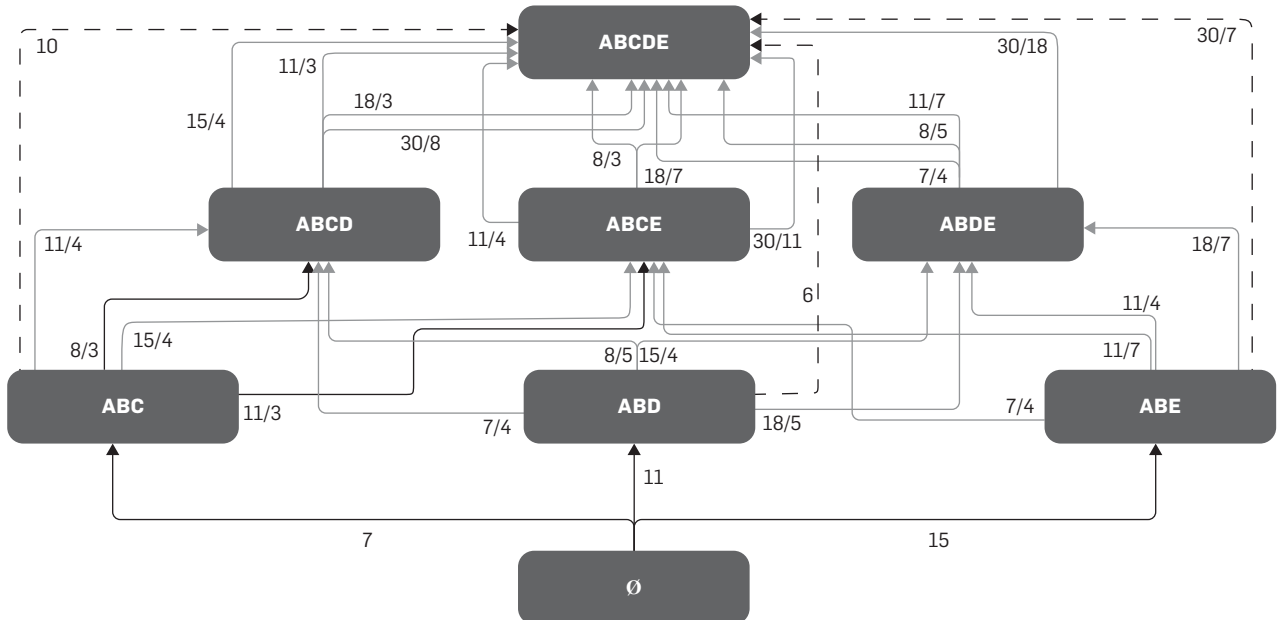


Figure 3. CEG_o for query Q_s in Figure 1 when the Markov table (§4) contains joins up to size 3.



degrees of base relations and small-size joins, and construct a new CEG, which we call CEG_M . Unlike the optimistic estimators, where the choice of path is not clear, we now show that picking the minimum weight path would (provably) be the most accurate estimate and this path is indeed equivalent to the solution of the LP that defines a pessimistic estimator from Joglekar and Ré¹¹ called MOLP. We therefore show that both subgraph summary-based optimistic estimators and the recent LP-based pessimistic ones can be seen as different instances of a broader class of estimators that pick paths through CEGs. Modeling these two classes of estimators in a common framework has several benefits; for example, one can apply optimizations for one class to the other. Our paper also opens the possibility of defining many other CEGs that use different statistics as edge weights; for example, entropies of certain columns. We hope future work can define and evaluate the accuracies of such CEGs.

For readers interested in the theory of pessimistic estimators, we note that CEGs are also very useful mathematical tools to prove properties of pessimistic estimators. For example, using CEGs in our proofs, we can derive combinatorial proofs to some properties of MOLP; for instance, that MOLP is identical to the pessimistic estimator proposed by Cai et al.⁵ on acyclic queries over binary relations.

2. QUERY AND DATABASE NOTATION

We consider conjunctive queries, that is, natural join queries, of the form

$$Q(\mathcal{A}) = R_1(\mathcal{A}_1), \dots, R_m(\mathcal{A}_m)$$

where $R_i(\mathcal{A}_i)$ is a relation with attributes $\mathcal{A}_i$ and $\mathcal{A} = \cup_i \mathcal{A}_i$. In conjunctive queries, the join predicates are equality of the common attributes shared across relations. Most of the examples used in this paper involve edge-labeled subgraph queries,

in which case each R_i is modeled as a binary relation containing a subset of the edges in a graph as source/destination pairs. Note that even if these edges contain other properties that can be modeled as additional columns, for the purpose of the join query, they are not relevant. Figure 2 presents an example showing a graph with edge labels A, B, C, D , and E . This graph can be represented using five binary relations, one for each edge label, as shown in Figure 2. We will often represent queries over such relations using a graph notation. For example, consider the relations A and B from Figure 2. We will represent the query $Q(a_1, a_2, a_3) = A(a_1, a_2) \bowtie B(a_2, a_3)$ as $a_1 \xrightarrow{A} a_2 \xrightarrow{B} a_3$. Similarly, the query $Q(a_1, a_2, a_3) = A(a_1, a_2) \bowtie B(a_3, a_2)$ will be represented as $a_1 \xrightarrow{A} a_2 \xleftarrow{B} a_3$.

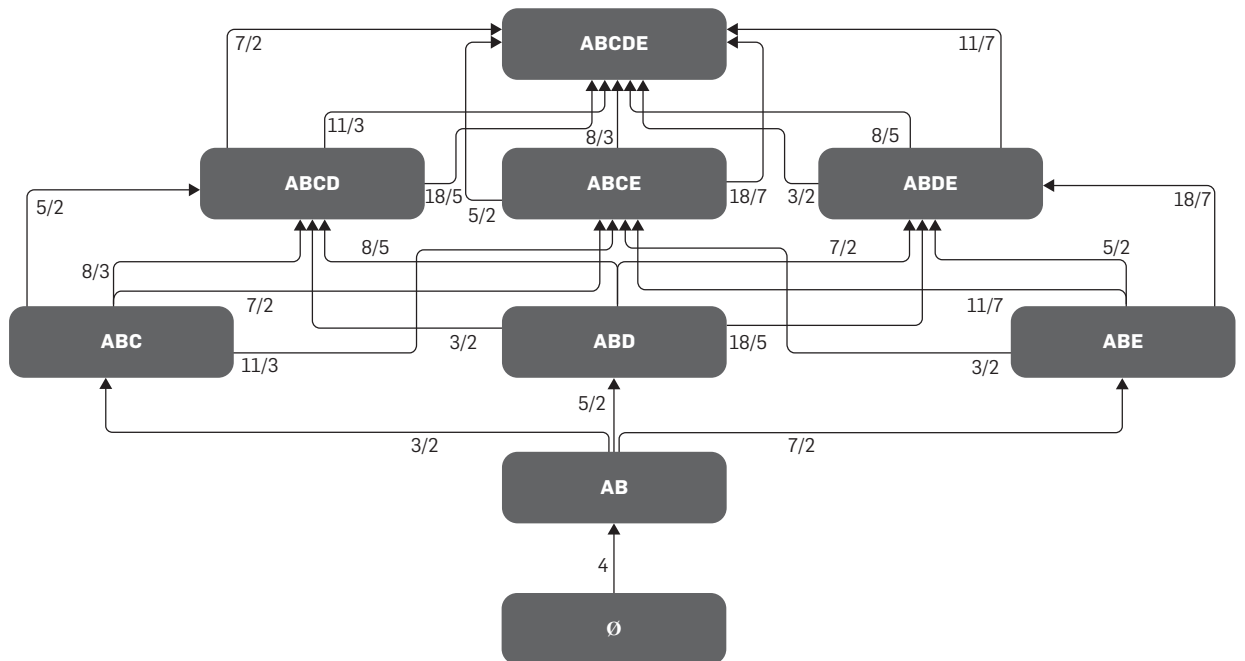
Note on binary vs. arbitrary relations: We emphasize that CEGs, which will be introduced in Section 3, do not require relations to be binary. In fact, the CEG we will define for pessimistic estimators in Section 5 assumes relations with arbitrary arities. However, we do our empirical evaluation on optimistic estimators, which for practical reasons have been developed in the context of graph-based systems, where joins are over binary relations. The reason is that optimistic estimators store as statistics the sizes of all possible small-size joins, that is, joins of a small number of relations, and when relations are binary, it is practical for a system to know and store the set of possible small-size joins in advance. In contrast, in RDBMSs that store relations with arbitrary numbers of columns and where columns can be arbitrarily joined with each other, the number of possible small-size joins can be very large.

3. CEG OVERVIEW

Next, we offer some intuition for CEGs. A CEG for a query Q consists of:

- Vertices labeled with sub-queries of Q , where sub-queries are defined by subsets of Q 's relations or attributes

Figure 4. CEG_Q for query Q_5 from Figure 1 when the Markov table (§4) contains joins up to size 2.



► Edges from smaller sub-queries to larger sub-queries, labeled with *extension rates*, which represent the cardinality of the larger sub-query relative to that of the smaller sub-query

Each bottom-to-top path (from $\emptyset$ to Q) in a CEG represents a different way of generating a cardinality estimate for Q . An estimator using a CEG picks one of these paths as an estimate. The estimate of a path is the product of the extension rates along the edges of the path. Equivalently, one can put the logarithms of the extension rates as edge weights and sum the logarithms.

Figure 4 illustrates a CEG^a for the query Q_{5f} shown in Figure 1 over the relations shown in Figure 2, assuming that statistics are available for any size-2 sub-queries of Q_{5f} . Depending on the semantics of the edges in a CEG, there can be multiple edges between two sub-queries in a CEG. For example, the last components of the two formulas we presented in Section 1 for Q_{5f} , $|\frac{D \rightarrow E}{|D|}|$ and $|\frac{B \rightarrow E}{|B|}|$, correspond to parallel edges that extend the sub-query over $ABCD$ edges with an E edge to Q_{5f} . Consider the leftmost path in Figure 4. The first extension rate from $\emptyset$ to $a_1 \xrightarrow{A} a_2 \xrightarrow{B} a_3$ is the known cardinality of $a_1 \xrightarrow{A} a_2 \xrightarrow{B} a_3$, which is 4, and the second extension rate has a weight $3/2$, intuitively estimating that each $a_1 \xrightarrow{A} a_2 \xrightarrow{B} a_3$ path will extend to $3/2$ many $a_1 \xrightarrow{A} a_2 \xrightarrow{B} a_3 \xrightarrow{C} a_4$ paths. Continuing the extensions, the final estimate is $4 \times \frac{3}{2} \times \frac{5}{2} \times \frac{7}{2} = 52.5$.

In the rest of this paper, we will show how some prior optimistic and pessimistic estimators can be modeled as instances of this generic estimator using different CEGs.

4. OPTIMISTIC ESTIMATORS

The estimators we refer to as *optimistic* use statistics about the input database in formulas that make uniformity and independence or conditional independence assumptions. We focus on three estimators: *Markov tables*² from XML databases, graph summaries¹⁴ from RDF databases, and the graph catalog estimator of the Graphflow system¹⁶ for managing property graphs. As we explain momentarily, despite being designed for systems that adopt different graph-based data models, these estimators are all extensions of each other.

We begin with an overview of the Markov tables estimator.² A Markov table of length $h \geq 2$ stores the cardinality of each path in an XML document's element tree up to length h and uses these to make predictions for the cardinalities of longer paths. Table 1 shows a subset of the entries in an example Markov table for $h = 2$ for our running example dataset from Figure 2. The formula to estimate a 3-path using a Markov table with $h = 2$ is to multiply the cardinality of one of the 2-paths with the consecutive 2-path divided by the cardinality of the common edge. For example, consider the query $Q_{3p} = \frac{A \rightarrow B \rightarrow C}{|A|}$ against the dataset in Figure 2. The formula for Q_{3p} would be: $|\frac{A \rightarrow B}{|A|}| \times (|\frac{B \rightarrow C}{|B|}| / |\frac{B}{|B|}|)$. The formula assumes that the number of C edges that each B edge extends to is uniformly $r = |\frac{B \rightarrow C}{|B|}| / |\frac{B}{|B|}|$. Equivalently, this is the “average C -degree” of nodes in the $\frac{B \rightarrow C}{|B|}$ paths. The result of this formula is $4 \times \frac{3}{2} = 6$, which underestimates the true cardinality of 7. The graph summaries¹⁴ for RDF databases and the graph catalog estimator¹⁶ have extended the contents of what

is stored in Markov tables, respectively, to other acyclic joins, such as stars, and small cycles, such as triangles.

Table 1. Example Markov table for $h = 2$.

Path	lPathl
$\frac{B}{ B }$	2
$\frac{A, B}{ A, B }$	4
$\frac{B, C}{ B, C }$	3
...	...

4.1. Space of possible optimistic estimators.

We next represent optimistic estimators using a CEG we call CEG_o . We assume the given query Q is connected. CEG_o consists of the following:

► *Vertices*: For each connected subset of relations $S \subseteq R$ of Q , we have a vertex in CEG_o with label S . This represents the sub-query $\bigvee_{R_i \in S} R_i$.

► *Edges*: Consider two vertices with labels S and S' such that $S \subset S'$. Let $\mathcal{D} = S' \setminus S$ (for difference), $\varepsilon \supset \mathcal{D}$ (for extension) be a Markov table entry, and $I = \varepsilon \cap S$ (for intersection). If ε and I exist in the Markov table, then there is an edge with weight $\frac{|\varepsilon|}{|I|}$ from S to S' in CEG_o .

This gives the core structure of CEG_o , though when making estimates there are several simple rules that remove some edges in CEG_o from prior work that limit the paths considered in CEG_o (see the longer version of this paper⁷). In general, there may be multiple $(\emptyset, Q)$ paths that lead to different estimates in CEG_o :

Example 1: Consider the CEG_o for Q_{5f} shown in Figure 4, which uses a Markov table of size 2. There are 36 $(\emptyset, Q)$ paths leading to seven different estimates. Two examples are:

$$\begin{aligned} \text{► } |\frac{A \rightarrow B}{|A|}| \times \frac{|\frac{B, C}{|B, C|}|}{|\frac{B}{|B|}|} \times \frac{|\frac{B, D}{|B, D|}|}{|\frac{B}{|B|}|} \times \frac{|\frac{B, E}{|B, E|}|}{|\frac{B}{|B|}|} &= 52.5 \\ \text{► } |\frac{A \rightarrow B}{|A|}| \times \frac{|\frac{B, C}{|B, C|}|}{|\frac{B}{|B|}|} \times \frac{|\frac{C, D}{|C, D|}|}{|\frac{C}{|C|}|} \times \frac{|\frac{D, E}{|D, E|}|}{|\frac{D}{|D|}|} &= 57.6 \end{aligned}$$

Example 2: Similarly, consider estimating Q_{5f} now with a Markov table with up to 3-size joins. The new CEG_o is shown in Figure 3, which contains multiple paths leading to two different estimates:

$$\begin{aligned} \text{► } |\frac{A \rightarrow B \rightarrow C}{|A, B, C|}| \times \frac{|\frac{C, D, E}{|C, D, E|}|}{|\frac{C}{|C|}|} \\ \text{► } |\frac{A \rightarrow B \rightarrow C}{|A, B, C|}| \times \frac{|\frac{A, B, D}{|A, B, D|}|}{|\frac{A, B}{|A, B|}|} \times \frac{|\frac{A, B, E}{|A, B, E|}|}{|\frac{A, B}{|A, B|}|} \end{aligned}$$

Both formulas start by $|\frac{A \rightarrow B \rightarrow C}{|A, B, C|}|$. The first *short-hop* formula makes one fewer independence assumption than the *long-hop* formula, which is an advantage. In contrast, the first estimate also makes a uniformity assumption that conditions on a smaller-size join, which might make it less accurate than the two assumptions made in the long-hop estimate, which condition on 2-size joins.

Any optimistic estimator implementation needs to make choices about which formulas to use; this corresponds to picking paths in CEG_o . We systematically identify a space of choices an optimistic estimator can make along two parameters that capture the choices made in prior work:

► *Path length*: The estimator can identify a set of paths to consider based on the path lengths, that is, number of edges or hops, in CEG_o , which can be: (i) maximum-hop (max-hop),

a Specifically, it is a CEG_o , defined in Section 4.

(ii) minimum-hop (min-hop), or (iii) any number of hops (all-hops). Let $\mathcal{P}$ be the set of paths an estimator picks.

► **Aggregator:** To derive a final estimate, the estimator has to aggregate the estimates in $\mathcal{P}$. We identify three aggregators: (i) the path with the largest estimate (max-aggr), (ii) the path with the lowest estimate (min-aggr), or (iii) the average of all the estimates in $\mathcal{P}$ (avg-aggr).

Any combination of these two choices can be used to design an optimistic estimator. The original Markov tables² chose the max-hop paths. In Aboulnaga et al.,² queries were paths, so when the path length is chosen, any CEG_o path gives the same estimate. Therefore, an aggregator is not needed. Graph Summaries¹⁴ chooses the min-hop paths and leaves the aggregator unspecified. Graph Catalog¹⁶ picks the min-hop and min-aggr aggregator. None of these estimators consider alternative choices an estimator can make. Instead, we do a systematic experimental analysis of this space of estimators in Section 6 and show that the best choices depend on the query structure, as optimistic estimators behave differently on different structures. For acyclic queries and queries with small cycles, these estimators tend to underestimate, and for queries with larger cycles, they tend to overestimate. We next make an observation to explain this difference.

Recall that a Markov table stores the cardinalities of patterns up to some size h . Given a Markov table with $h \geq 2$, optimistic estimators can produce estimates for any acyclic query with size larger than h . However, faced with a large cyclic query Q , optimistic estimators do not actually produce estimates for Q . Instead, they produce an estimate for a similar acyclic Q' that includes all of Q 's edges but is not closed. To see this, consider the following example:

EXAMPLE 1

Consider a 4-cycle query in Figure 5a using a Markov table with $h=3$. The CEG_o for this setting is shown in Figure 5b. Consider the leftmost path corresponding to the formula: $|\underline{A}, \underline{B}, \underline{C}| \times |\underline{B}, \underline{C}, \underline{D}| / |\underline{B}, \underline{C}|$. This formula is in fact estimating a 4-path $\underline{A}, \underline{B}, \underline{C}, \underline{D}$ rather than the 4-cycle. This is true for each path in CEG_o .

More generally, when queries contain cycles of length $> h$, CEG_o breaks cycles in queries into paths. Therefore, estimates over CEG_o can lead to *overestimates* for queries with large cycles, as there are often significantly more paths than cycles in real-world graphs.

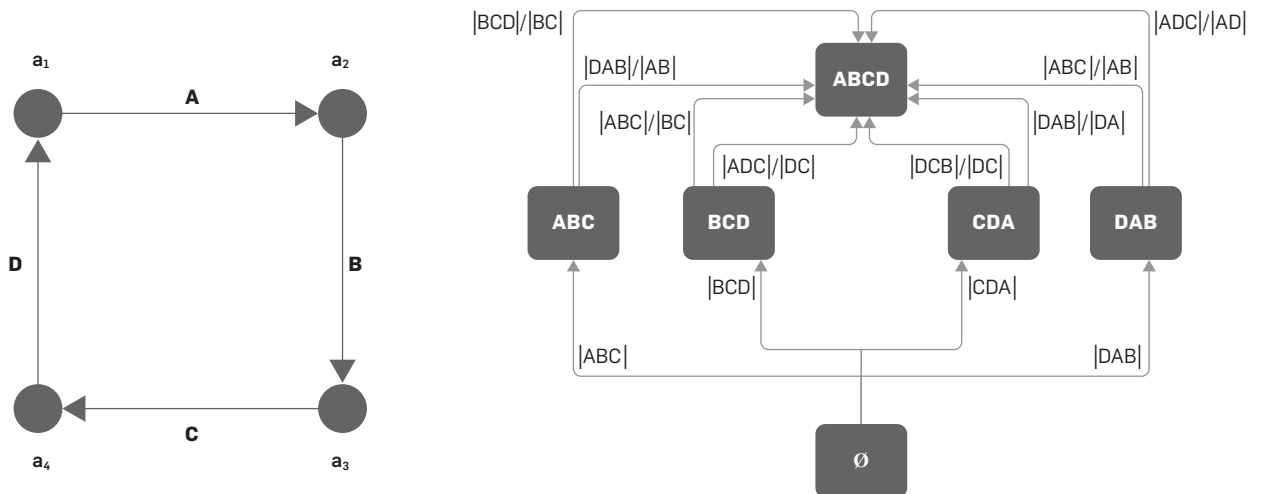
5. PESSIMISTIC ESTIMATORS

Starting from the seminal result by Atserias, Grohe, and Marx in 2008,⁴ several upper bounds have been provided for the output sizes of join queries under different known statistics. For example the initial upper bound from Atserias et al.,⁴ now called the *AGM bound*, used only the cardinalities of each relation, while later bounds, DBPLP,¹¹ MOLP,¹¹ and CLLP¹ used maximum degrees of the values in the columns and improved the AGM bound. Since these bounds are upper bounds on the query size, they can be used as *pessimistic estimators*. This was done recently by Cai et al.⁵ in an actual estimator implementation. We refer to this as the CBS estimator, after the names of the authors. We next show that some of the recent pessimistic estimators^{5,11} can also be modeled as making an estimate using a CEG.

5.1. MOLP.

MOLP was defined in Joglekar and Ré¹¹ as a tighter bound than the AGM bound that uses additional degree statistics about input relations that AGM bound does not use. Let $\mathcal{X}$ be a subset of the attributes $\mathcal{A}_i$ of some relation $\mathcal{R}_i$, and let v be a possible value of $\mathcal{X}$. The *degree* of v in $\mathcal{R}_i$ is the number of times v occurs in $\mathcal{R}_i$; that is, $\deg(\mathcal{X}(v), \mathcal{R}_i) = |\{t \in \mathcal{R}_i \mid \Pi_{\mathcal{X}}(t) = v\}|$. For example, in Figure 2, $\deg(s(3), E) = 3$ because the outgoing E -degree of vertex 3 is 3. Similarly, $\deg(d(2), A)$ is 1 because the incoming A -degree of vertex 2 is 1. We also define $\deg(\mathcal{X}, \mathcal{R}_i)$ to be the maximum degree in $\mathcal{R}_i$ of any value v over $\mathcal{X}$; that is, $\deg(\mathcal{X}, \mathcal{R}_i) = \max_v \deg(\mathcal{X}(v), \mathcal{R}_i)$. So, $\deg(d, A) = 3$ because vertex 13 has three incoming A edges, which is the maximum A -in-degree in the dataset. The notion of degree can be generalized to $\deg(\mathcal{X}(v), Y, \mathcal{R}_i)$, which refers to the “degree of a value v over attributes $\mathcal{X}$ in $\Pi_Y(\mathcal{R}_i)$ ”, which counts the number of times v occurs in $\Pi_Y(\mathcal{R}_i)$. Similarly, we let $\deg(\mathcal{X}, Y, \mathcal{R}_i) = \max_v \deg(\mathcal{X}(v), Y, \mathcal{R}_i)$. Suppose a system has stored $\deg(\mathcal{X}, Y, \mathcal{R}_i)$

Figure 5. A 4-cycle query and its CEG_o .



statistics for each possible $\mathcal{R}_i$ and $X \subseteq Y \subseteq A_i$. MOLP is the output of this LP:

$$\begin{aligned} & \text{Maximize } s_A \\ & s_\emptyset = 0 \\ & s_X \leq s_Y, \forall X \subseteq Y \\ & s_{Y \cup E} \leq s_{X \cup E} + \log(\deg(X, Y, \mathcal{R}_i)), \forall X, Y, E \subseteq A, X \subseteq Y \subseteq A_i \end{aligned}$$

The base of the logarithm can be any constant and we take it as 2. Let m_A be the optimal value of MOLP. Joglekar and Ré¹¹ has shown that 2^{m_A} is an upper bound on the size of Q .

MOLP CEG (CEG_M): It is not easy to directly see the solution of the MOLP on our running example. However, we can represent the MOLP bound as the cost of minimum-weight ($\emptyset, Q$) path in a CEG we call CEG_M.

► **Vertices:** For each $X \subseteq A$, the variable s_X in MOLP represents an upper bound on the size of $Q_X = \Pi_X Q$. Therefore, for each $X \subseteq A$ there is a vertex in CEG_M.

► **Extension Edges:** For each $s_{Y \cup E} \leq s_{X \cup E} + \log(\deg(X, Y, \mathcal{R}_i))$ inequality, there is an edge with weight $\log(\deg(X, Y, \mathcal{R}_i))$ between any $W_1 = X \cup E$ and $W_2 = Y \cup E$. These inequalities intuitively indicate the following: Each tuple $t_{X \cup E} \in Q_{X \cup E}$ can extend to at most $\deg(X, Y, \mathcal{R}_i) Q_{Y \cup E}$ tuples.

► **Projection Edges:** For each $s_X \leq s_Y$ inequality (that is, $\forall X \subseteq Y$), add an edge with weight 0 from Y to X . These indicate that the size of Q_X is at most the size of Q_Y , if Y is a larger subquery.

Figure 6 shows the CEG_M of our running example. The figure uses actual degrees instead of their logarithms as edge weights and omits the projection edges.

THEOREM 5.1

Let Q be a query with degree statistics $\deg(X, Y, \mathcal{R}_i)$ for each $\mathcal{R}_i$ and $X \subseteq Y \subseteq A_i$. The optimal solution m_A to the MOLP of Q is the weight of the minimum-weight ($\emptyset, A$) path in CEG_M.

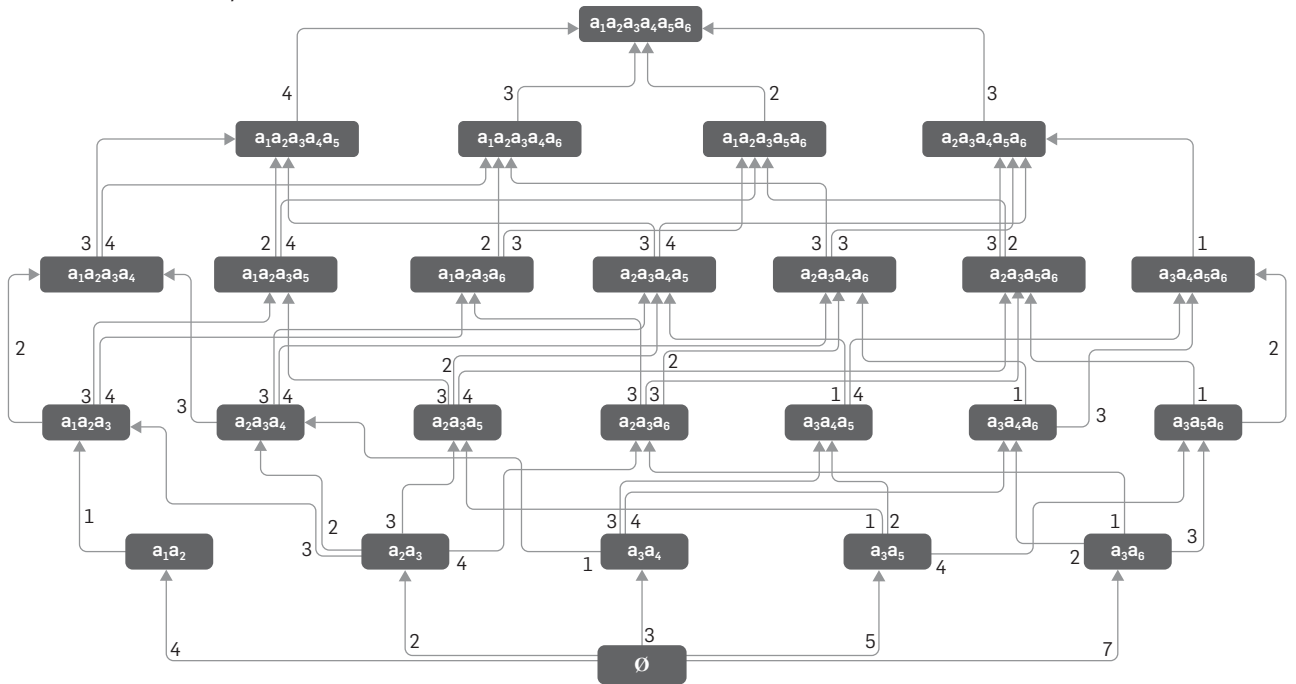
Theorem 5.1's proof can be found in Chen⁷. With this connection, readers can verify the MOLP bound in our example by inspecting the paths in Figure 6; the minimum-weight ($\emptyset, A$) path has a weight of 96, corresponding to the leftmost path. We make two observations:

Observation 1: Joglekar and Ré¹¹ proves through a numeric LP-duality argument that 2^{m_A} is an upper bound on the the output size of the query (OUT), that is, $OUT \leq 2^{m_A}$ (see Prop. 2¹¹). Our CEG formulation of MOLP provides arguably a simpler proof of this property. Note that each ($\emptyset, A$) path in CEG_M corresponds to a sequence of extensions from $\emptyset$ to Q and is an estimate of the cardinality of Q . Since we are using maximum degrees on the edge weights, each ($\emptyset, A$) path is by construction an upper bound on Q . Since for any ($\emptyset, A$) path P in CEG_M, $OUT \leq 2^{w(P)}$ and by Theorem 5.1, m_A is equal to the weight of the minimum-weight ($\emptyset, A$) path in CEG_M, then $OUT \leq 2^{m_A}$.

Observation 2: Theorem 5.1 implies that MOLP can be solved using a combinatorial shortest-path algorithm instead of a numeric LP solver.

In Chen⁷, we also review the CBS pessimistic estimator⁵ and show, using CEGs in our proofs, that MOLP is as tight as CBS and is exactly the same as CBS on acyclic join queries over binary relations. Our use of CEGs as mathematical tools in our proofs can be interesting to readers interested in the theory of worst-case bounds on the sizes of join queries. Overall, our results show that CEGs are expressive enough to model two prior pessimistic estimators as well. One benefit of this is we can apply optimizations done in one class of estimators to the other. In the original version of this paper⁷, we show an example of this by applying an optimization called the *bound sketch* optimization developed for the CBS estimator also to the optimistic estimators.

Figure 6. CEG_M for query Q_{sf} in Figure 1.



6. EVALUATION

Recall that for pessimistic estimators, one should always pick the shortest path on $CE G_M$, but for optimistic estimators, there is no obvious choice. We next present extensive experiments to evaluate the accuracies of the space of optimistic estimators we described on a large suite of datasets and workloads, so we can provide an advice based on an empirical justification. Our code, datasets, and queries are available here.⁶ We note that the evaluation we make here is based on subgraph queries, which are equivalent to join queries over binary relations. Recall that this is because prior optimistic estimators were developed in this context. As a result, we note that the advice we give applies to this context. Using CEG-based optimistic estimators for queries over arbitrary relations is not in the scope of this paper.

Our original publication⁷ contains many more experiments than what is shown here, including a comparison of optimistic and pessimistic estimators (optimistic ones are much more accurate), impacts on the plan quality of using the optimistic estimators we advise on the RDF-3X system, scalability of Markov Tables and CEGs as the size of the patterns increase, and a detailed demonstration of applying the bound sketch optimization of pessimistic estimator to optimistic estimators.

6.1. Set-up, datasets, and workloads.

For all our experiments, we use a single machine with two Intel E5-2670 at 2.6GHz CPUs, each with eight physical and 16 logical cores, and 512GB of RAM. We used six real-world datasets, shown in Table 2, and six workloads on these datasets. Our dataset and workload combinations are as follows.

Table 2. Dataset descriptions.

Dataset	Domain	IVI	IEI	IE. Labels
IMDb	Movies	27M	65M	127
YAGO	Knowledge Graph	13M	16M	91
DBLP	Citations	23M	56M	27
WatDiv	Products	1M	11M	86
Hetionet	Social Networks	45K	2M	24
Epinions	Consumer Reviews	76K	509K	50

IMDb and JOB: The IMDb relational database, together with a workload called join order benchmark (JOB), has been used for cardinality estimation studies in prior work.^{5,13} We created property graph versions of the this database and workload. Briefly, the construction is based on converting *entity tables*, representing actors, movies, and companies as *entity nodes*, and *relationship tables* representing many-to-many relationships between the entities as edges. In addition, adding *foreign key edges* between an entity node u to v if there is a foreign key in the table u comes from to the table that v comes from. The final workload contained 369 queries.

WatDiv and WatDiv-Acyclic Workload: WatDiv³ is a synthetic knowledge graph that has its own workload in SPARQL format with 12,400 original queries. We converted these queries into equivalent subgraph queries by removing their vertex predicates, and we then removed queries with at most three edges. After removing duplicates among the

remaining queries, there were 75 different queries left, nine of which were cyclic and the other 64 acyclic. Because nine queries is very small for a workload, we used only the 64 acyclic queries and call this the WatDiv-Acyclic workload.

YAGO 1 and G-CARE-Acyclic and G-CARE-Cyclic Workloads: G-CARE¹⁹ is a cardinality estimation benchmark for subgraph queries. From this benchmark, we took their YAGO knowledge graph dataset and the acyclic and cyclic query workloads for that dataset. The acyclic workload contains 382 queries generated from query templates with 3-, 6-, 9-, and 12-edge star and path queries, as well as randomly generated trees. We will refer to this workload as G-CARE-Acyclic. The cyclic query workload contains 240 queries generated from templates with 6-, and 9-edge cycle, 6-edge clique, 6-edge flower, and 6- and 9-edge petal queries. We will refer to this workload as G-CARE-Cyclic.

DBLP, WatDiv Epinions, Hetionet and Acyclic and Cyclic Workloads: We used three other datasets: Hetionet, a biological network; DBLP, a real knowledge graph; and Epinions, a real-world social network graph. Epinions is a dataset that by default does not have any edge labels. We added a random set of 50 edge labels to Epinions. For these datasets, we created one acyclic and one cyclic query workload, which we refer to as Acyclic and Cyclic. The Acyclic workload contains queries generated from 6-, 7-, or 8-edge templates, shown in Figure 7. Then, we generated 20 non-empty instances of each template by putting one edge label uniformly at random on each edge, which yielded 360 queries in total. The Cyclic workload contains queries generated from templates used in Mhedhbi and Salihoğlu.¹⁶ We generated instances of these queries by randomly matching each edge of the query template one at a time in the datasets. Because the WatDiv's original queries contained only acyclic queries, we used the Cyclic workload also on WatDiv. We generated 70 queries for DBLP, 212 queries for Hetionet, 129 queries for WatDiv, and 394 queries for Epinions.

6.2. Space of optimistic estimators.

We begin by comparing our nine optimistic estimators on $CE G_o$ we defined. To set up an experiment in which we could test all nine of the possible optimistic estimators, we used a Markov table with $h=3$. A Markov table with only 2-size joins cannot test different estimators based on different path-length choices or any cyclic query.

To compare the accuracies of different estimators, for each query Q in our workloads, we make an estimate using each estimator and compute its q-error. If the true cardinality of Q is c and the estimate is e , then the q-error is $\max\{\frac{e}{c}, \frac{c}{e}\} \geq 1$. For each workload, this gives us a distribution of q-errors, which we compare as follows. First, we take the logs of the q-errors so they are now ≥ 0 . If a q-error was an underestimate, we put a negative sign to it. This allows us to order the estimates from the least accurate underestimation to the least accurate overestimation. We then generate a box plot where the box represents the 25th (median) and 75th percentile cut-off marks. We also compute the mean of this distribution, excluding the top 10% of the distribution (ignoring under/overestimations) and draw it with a red dashed line in box plots.

6.2.1. Acyclic queries and cyclic queries with only triangles.

Our first question is: Which of the nine possible optimistic estimators leads to the most accurate estimates on acyclic queries and cyclic queries that contain ≤ 3 edges on CEG_o ? We compare our nine estimators on CEG_o for each acyclic query workload in our setup (for IMDb, WatDiv, and YAGO, using

JOB, WatDiv-Acyclic, and G-CARE-Acyclic workloads). We then compare our nine estimators on each cyclic query workload using the queries that only contain triangles as cycles. We omit GCARE-Cyclic because every query in GCARE-Cyclic except for one contained cycles with more than three edges.

Our results are shown in Figure 8. We make several obser-

Figure 7. Query templates for the Acyclic workload. Edge directions are omitted in the figure.

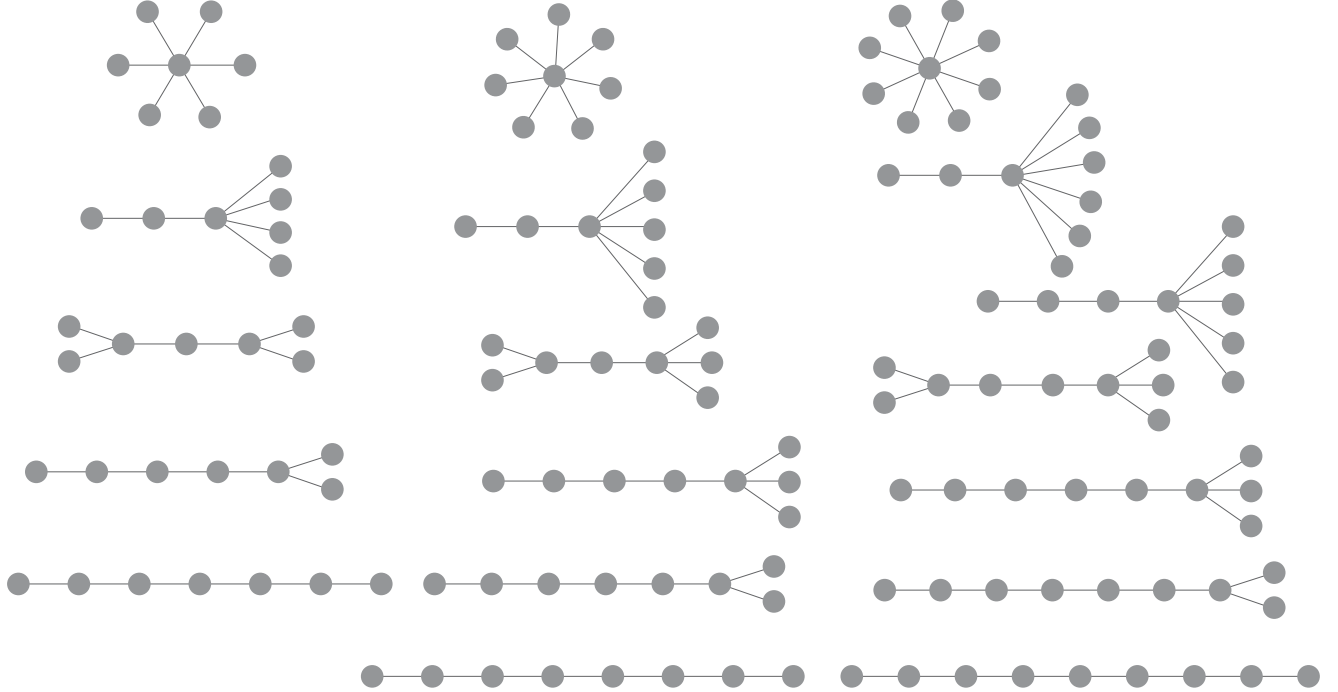
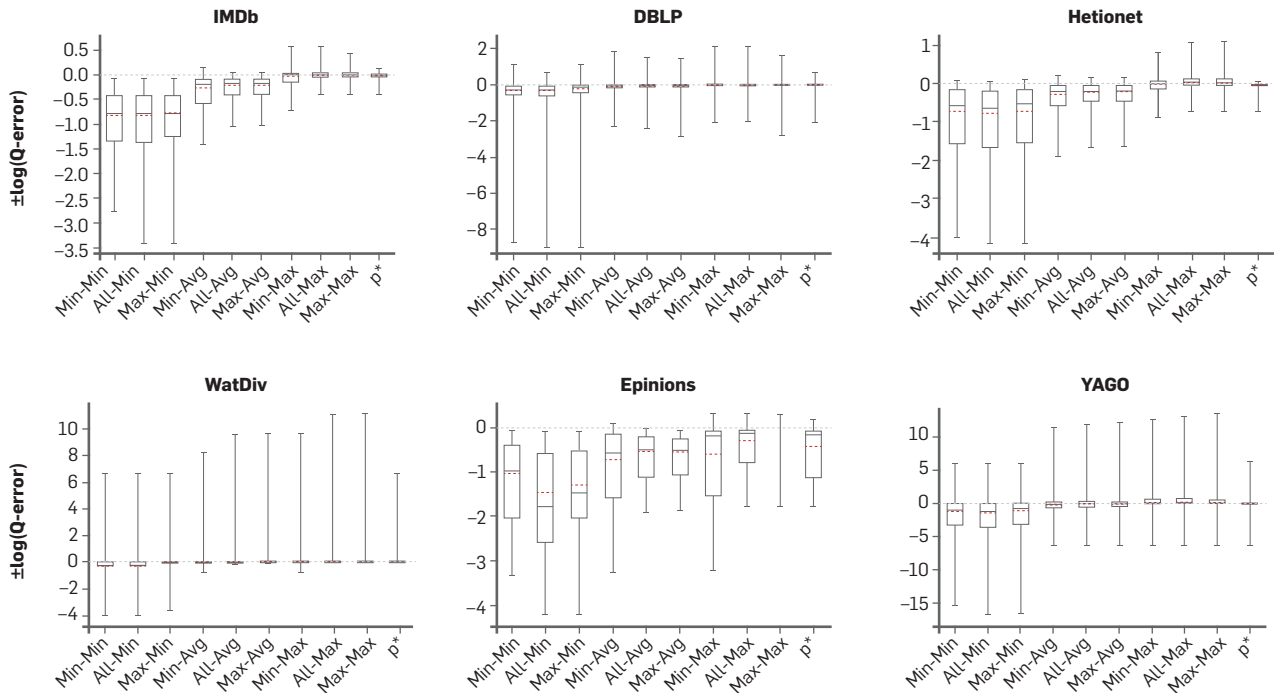


Figure 8. Evaluation of the optimistic estimators on CEG_o on acyclic queries. Estimators are labeled “P-A”; P is the path length (one of max-hop, min-hop, or all-hops) and A the aggregator (one of max-aggr, min-aggr, or avg-aggr).



vations. First, regardless of the path-length choice, the max aggregator (the last three box plots in the figures) makes significantly more accurate estimates (note that the y-axis on the plots is in log scale) than avg, which in turn is more accurate than min. This is true across all acyclic experiments and all datasets. For example, on IMDB and JOB workload, the all-hops-min, all-hops-avg, and all-hops-max estimators have a log of mean q-errors of 6.5 (underestimation), 1.7 (underestimation), and 1.02 (underestimation), respectively. *Therefore on acyclic queries, when there are multiple formulas that can be used to make an estimate, using the pessimistic ones is an effective technique to combat the well-known underestimation problem for estimators that make uniformity and independence assumptions.*¹² This can give up to three orders of magnitude lower mean q-errors than, for example, the min-hop-min estimator from prior work.

We next analyze the effects of path-length choices. Observe that across all experiments, if we ignore the outliers and focus on the 25-75 percentile boxes, max-hop and all-hops do at least as well as min-hop. Further observe that on IMDB, Hetionet, and on the Acyclic workload on Epinions,

max-hop and all-hops lead to significantly more accurate estimates. Finally, the performance of max-hop and all-hops are comparable across our experiments. We verified that this is because all-hops picks one of the max-hop paths in a majority of the queries in our workloads. Therefore, we observe that the advantage of long-hop paths that condition on 2-size joins when making uniformity assumptions is generally stronger than its disadvantage of making more independence assumptions. Since max-hop enumerates strictly fewer paths than all-hops to make an estimate, we conclude that on acyclic queries, systems implementing the optimistic estimators can prefer the max-hop-max estimator.

Figure 9 shows the accuracies of the nine estimators on cyclic query workloads with only triangles. Our observations are similar to those for acyclic queries, and we find that the max aggregator yields more accurate estimates than other aggregators, irrespective of the path length. When using the max aggregator, we also observe that max-hop performs at least as well as min-hop. Therefore, as we observed for acyclic queries, *we find max-hop-max estimator to be an effective way to make accurate estimations for cyclic queries with only triangles.*

Figure 9. Evaluation of the space of optimistic estimators on CEG_0 on Cyclic workload on queries with only triangles.

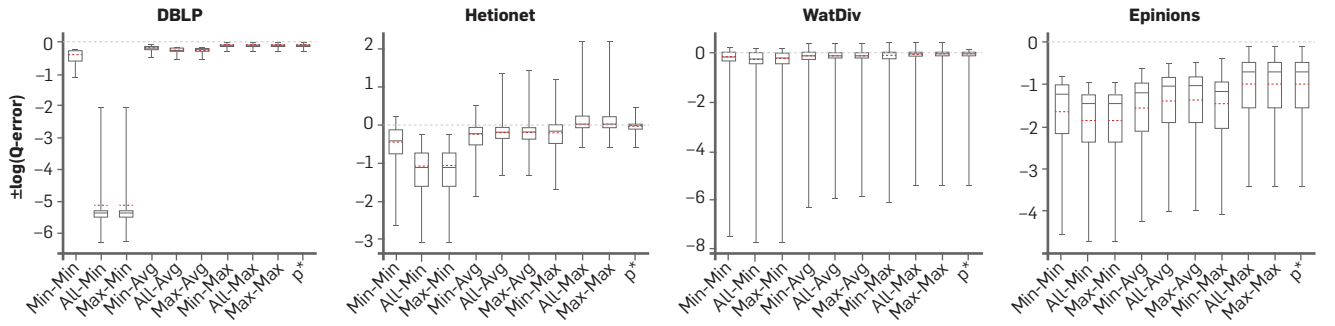


Figure 10. Accuracy of max-max for different query sizes (x-axis). The number on top of boxplots is the number of queries in the boxplot.

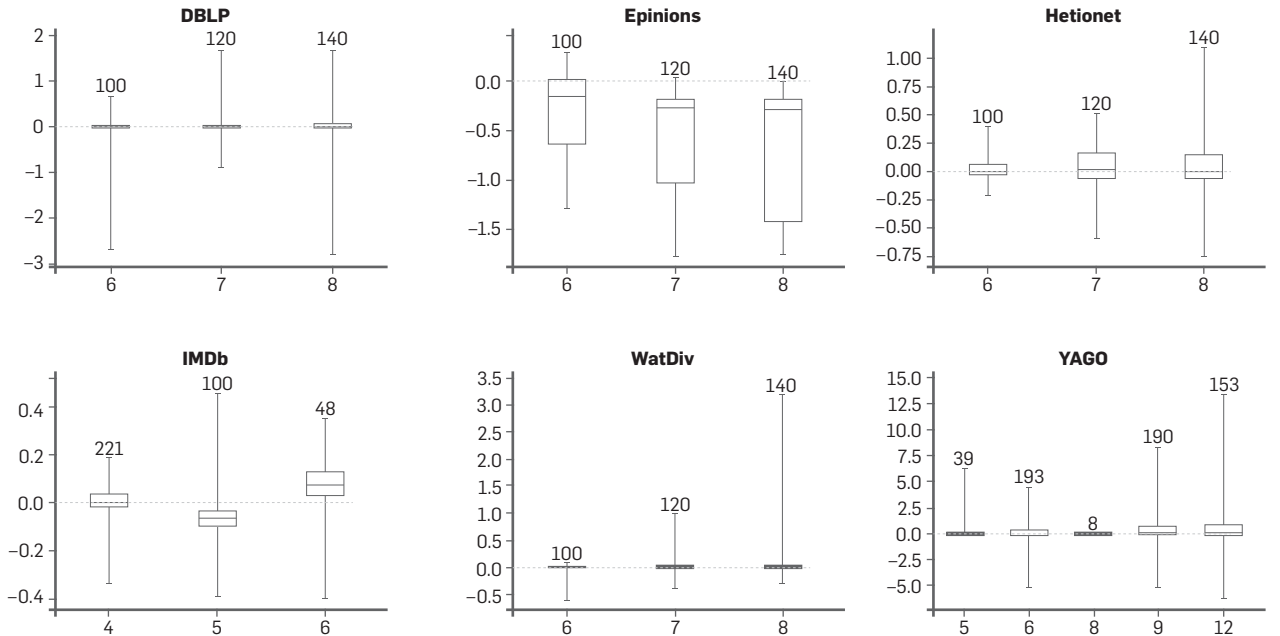
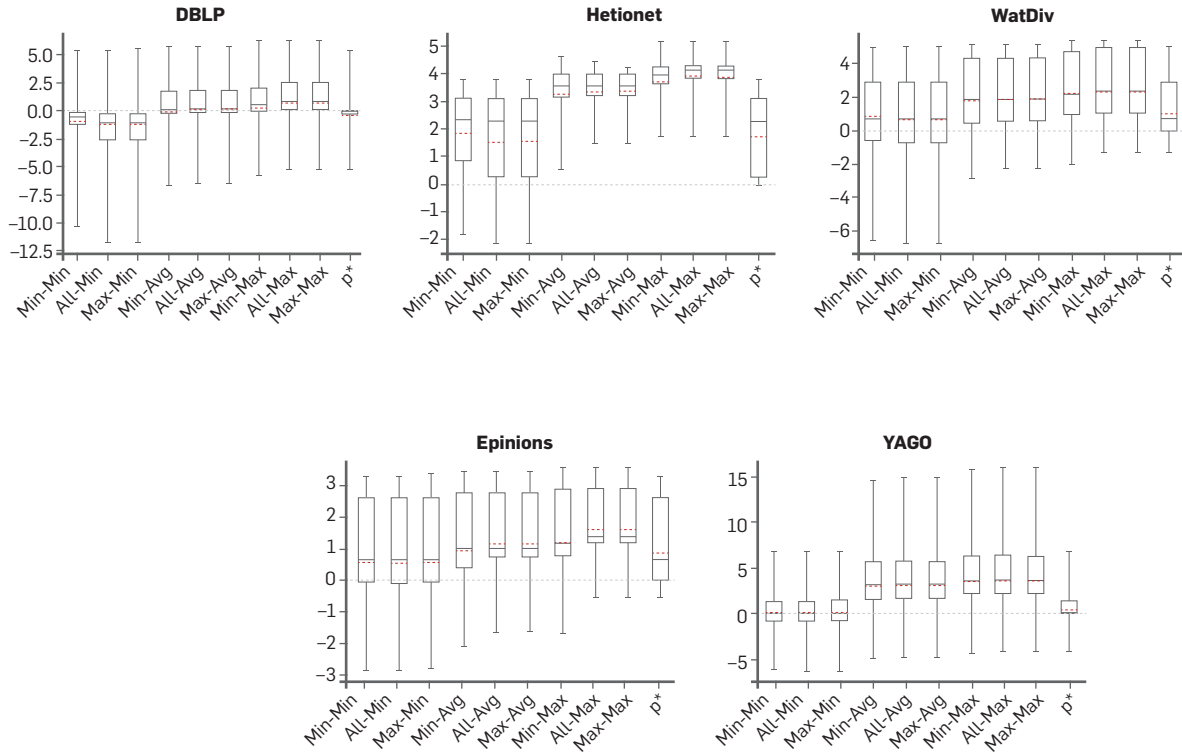


Figure 11. Evaluation of optimistic estimators on CEG_o on cyclic queries with cycles of four or more edges.

6.2.2. Cyclic queries with cycles of size >3 .

Recall our observation that on queries that contain large cycles, existing optimistic estimators estimate paths instead of cycles, which could yield highly inaccurate overestimates, as real-world graphs contain many more paths than cycles. Therefore, unlike the case for acyclic queries and queries with small cycles, we now expect that estimates based on CEG_o will be pessimistic. To verify this hypothesis, our next experiments compare the performance of optimistic estimators for each dataset-cyclic query workload combination, but only using queries that contain cycles of size >3 .

Figure 11 shows our results. As we expected, we now see that the optimistic estimators yield overestimates for the majority of the queries. This can be seen by observing that except for a few exceptions, the 25-75 percentile boxes of the boxplots are above 0. In addition, estimators using the `min` aggregator are generally more accurate, sometimes with several orders of magnitude difference, which can be seen by observing the mean accuracy lines of the boxplots. For example, on YAGO, the mean accuracy lines of `max-hop-min` and `max-hop-max` are, respectively, 0.20 and 3.61 on logarithmic scale, which correspond to absolute q-errors of 1.58 (overestimation) and 4043.86 (overestimation). We also observe that there is less sensitivity to the path-length choice when using the `min` aggregator, and any of the path-length choices perform reasonably well.

As we discuss in Section 8, we hope CEGs can be the foundation for proposing other estimators for future work. As a demonstration of the flexibility of our CEG-framework to design new CEG-based estimators, in the original version of our paper,⁷ we present one possible approach (though others can

be proposed) to remedy the overestimates of the estimators on CEG_o for queries with large cycle queries. Specifically, we describe a new CEG, that modifies CEG_o with new edge weights that incorporate a cycle-closing effect and show that estimates on this new CEG can be more accurate than those on CEG_o .

7. RELATED WORK

There is decades of extensive research on cardinality estimation techniques that can be used for estimating sizes of join queries. This literature has proposed several different classes of estimators, including other summary-based ones, as well as sampling-based ones and novel ML-based ones. Sampling-based estimators^{8,10,21,23} sample input records from base tables and evaluate queries on these samples to make estimates. Sampling-based estimators are fundamentally different than summary-based ones as by increasing the sizes of the samples they can be arbitrarily accurate, while summary-based ones can be more accurate by keeping more statistics. Another class of estimators on which there is active work is ML-based ones that learn to make estimates from example queries or predicates, e.g., example range predicates on columns.²² Studying how all these estimators compare is beyond the scope of this work. In the following, we cover other summary-based estimators primarily focusing on graph-based database management systems.

Many relational systems, including commercial ones such as PostgreSQL, use summary-based estimators. Example summaries include the cardinalities of relations, the number of distinct values in columns, or histograms,¹⁷ wavelets,¹⁵ or probabilistic and statistical models⁹ that capture the distribution of values in columns. These statistics

are used to estimate the selectivities of each join predicate, which are put together using several approaches, such as independence assumptions. In contrast, the estimators we studied store degree statistics about base relations and small-size joins (note that cardinalities are a form of degree statistics, for example, $|R_i| = \text{deg}(\emptyset, \mathcal{R}_i)$).

Characteristic Sets (CS)¹⁸ is a summary-based estimator primarily designed to estimate the cardinalities of stars in an RDF graph. It uses the so-called *characteristic set* of each vertex v in an RDF graph, which is the set of distinct outgoing labels v has. CS keeps statistics about the vertices with the same characteristic set. Then, using these statistics, CS makes estimates for the sizes of star queries. For a non-star query Q , Q is decomposed into multiple stars $s_1, \dots, s_k$ in a greedy manner, by removing the largest stars first, and the estimates for each s_i is multiplied. However, unlike the optimistic estimators we considered, this decomposition procedure does not lead to multiple possible decompositions.

Several works have proposed summary-based estimators that compute a sketch of an input graph. SumRDF²⁰ builds a summary graph S of an RDF graph and adopts a holistic approach to make an estimate. Given the summary S , SumRDF considers all possible RDF graphs G that could have the same summary S . In the context of estimating the selectivities of path expressions XSeed²⁵ builds a sketch S of the input XML Document. The sketch of the graph effectively collapses multiple nodes and edges into supernodes and edges with statistics on them. Then given a query Q , Q is matched on S and using the metadata, an estimate is made. Unlike CEGs, these techniques do not decompose a query into smaller sub-queries, so the question of which decomposition to use does not arise for these estimators.

Similarly, several work use data structures that are adaptations of histograms from relational systems to store selectivities of paths or trees in XML documents, such as *positional histograms*.²⁴ Such techniques do not consecutively make estimates for larger paths and have not been adopted to general subgraph queries.

8. LIMITATIONS, CONCLUSIONS, AND FUTURE WORK

Aside from capturing existing optimistic and pessimistic estimators, we believe the CEG framework can be the foundation to develop novel summary-based estimators. In addition to the two CEGs we considered here, CEG_O and CEG_M , other CEGs using different statistics can be defined and paired with different techniques to pick paths. To demonstrate an example, in the original version of this paper,⁷ we describe another CEG we call CEG_{OCR} as a possible approach to remedy the overestimation problem of optimistic estimator on CEG_O for queries with large cycles. CEG_{OCR} uses new edge weights that capture the closing of large cycles between sub-queries.

Perhaps the most important research question our work leaves unanswered is: which CEG should be used in practice? This is a very broad question and it is easy to define many other CEGs and our understanding of which ones would yield to better accuracy is limited. For example, one can use variances or entropies of the distributions of small-size joins as edge weights, possibly along with degree statistics, and pick the lowest entropy paths. An important research

direction is to systematically study a class of CEG instances that use different combination of statistics as edge weights, as well as techniques for picking paths, to design more accurate CEG-based estimators. In addition, in this work, we focused only on join-only queries and ignore other non-join predicates. An additional future work direction is to develop principled mechanisms to integrate filters on the queries that can be estimated using CEG. G

References

1. Abo Khamis, M., Ngo, H.Q., and Suciu, D. Computing join queries with functional dependencies. In *Proceedings of the 35th ACM SIGMOD-SIGACT-SIGAI Symp. on Principles of Database Systems* (2016), 327-342.
2. Aboulmaga, A., Alameldeen, A.R., and Naughton J.F. Estimating the selectivity of XML path expressions for Internet scale applications. In *Proceedings of the Intern. Conf. Very Large Data Bases* (2001).
3. Alug, G., Hartig, O., Özsu M.T., and Daudjee K. Diversified stress testing of RDF data management systems. In *Proceedings of the 13th Intern. Semantic Web Conf.* (2014).
4. Atserias, A., Grohe, M., and Marx D. Size bounds and query plans for relational joins. *SIAM J. Computing* 42, 4 (2013).
5. Cai, W., Balazinska, M., and Suciu D. Pessimistic cardinality estimation: Tighter upper bounds for intermediate join cardinalities. In *Proceedings of the Intern. Conf. on Mgmt. of Data* (2019).
6. Github repo of our code, datasets, and queries (2022); <https://github.com/cetechreport/CEXperiments>
7. Chen, J. et al. Accurate summary-based cardinality estimation through the lens of cardinality estimation graphs. In *Proceedings of the Intern. Conf. on Very Large Data Bases* 15, 8 (2022).
8. Chen, Y. and Yi, K. Random sampling and size estimation over cyclic joins. In *Proceedings of the Intern. Conf. on Database Theory* (2020).
9. Getoor, L., Taskar, B., and Koller D. Selectivity estimation using probabilistic models. In *Proceedings of the SIGMOD* (2001).
10. Haas, P.J., Naughton, J.F., Seshadri, S., and Swami, A.N. Selectivity and cost estimation for joins based on random sampling. *J. Computer and System Sciences* 52, 3 (1996).
11. Joglekar, M. and Ré, C. It's all a matter of degree—Using degree information to optimize multiway joins. *Theory of Computing Systems* 62, 4 (2018).
12. Leis, V. et al. How good are query optimizers, really? In *Proceedings of the Intern. Conf. on Very Large Data Bases* 9, 3 (2015).
13. Leis, V. et al. Query optimization through the looking glass, and what we found running the join order benchmark. *VLDBJ* 27, 5, (2018).
14. Maduko, A., Anyanwu, K., Sheth, A., and Schliekman, P. Graph summaries for subgraph frequency estimation. In *Proceedings of the European Semantic Web Conf.* (2008).
15. Matias, Y., Vitter, J.S., and Wang, M. Wavelet-based histograms for selectivity estimation. In *Proceedings of the SIGMOD* (1998).
16. Mhedhbi, A. and Salihoglu, S. Optimizing subgraph queries by combining binary and worst-case optimal joins. In *Proceedings of the Intern. Conf. Very Large Data Bases* 12, 11 (2019).
17. Muralikrishna, M. and DeWitt, D.J. Equi-depth histograms for estimating selectivity factors for multi-dimensional queries. In *Proceedings of SIGMOD* (1988).
18. Neumann, T. and Moerkotte, G. Characteristic sets: Accurate cardinality estimation for RDF queries with multiple joins. In *Proceedings of the Intern. Conf. on Data Engineering* (2011).
19. Park, Y. et al. G-CARE: A framework for performance benchmarking of cardinality estimation techniques for subgraph matching. In *Proceedings of SIGMOD* (2020).
20. Stefanoni, G., Motik, B., and Kostylev E.V. Estimating the cardinality of conjunctive queries over RDF data using graph summarisation. In *Proceedings of the World Wide Web Conf.* (2018).
21. Vengerov, D., Menck, A.C., Zait, M., and Chakkappen, S.P. Join size estimation subject to filter conditions. In *Proceedings of the Intern. Conf. on Very Large Data Bases* 8, 12 (2015).
22. Woltmann, L. et al. PostCENN: PostgreSQL with machine learning models for cardinality estimation. In *Proceedings of the Intern. Conf. on Very Large Data Bases* 14, 12 (2021).
23. Wu, W., Naughton, J.F., and Singh, H. Sampling-based query re-optimization. In *Proceedings of SIGMOD* (2016).
24. Wu, Y., Patel, J.M., and Jagadish, H.V. Estimating answer sizes for XML queries. In *Proceedings of the Intern. Conf. Extending Database Technology* (2002).
25. Zhang, N., Özsu, M.T., Aboulmaga A., and Ilyas, I.F. XSEED: Accurate and fast cardinality estimation for XPath queries. In *Proceedings of the Intern. Conf. on Data Engineering* (2006).

Semih Salihoğlu, University of Waterloo, Waterloo, ON, Canada.

Jeremy Chen, University of Waterloo, Waterloo, ON, Canada.

Yuqing Huang, University of Waterloo, Waterloo, ON, Canada.

Mushi Wang, University of Waterloo, Waterloo, ON, Canada.

Ken Salem, University of Waterloo, Waterloo, ON, Canada.

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

© 2026 Copyright held by the owner/author(s).

CAREERS

Georgia Institute of Technology

Faculty Positions in Computational Science and Engineering

The School of Computational Science and Engineering (CSE) in the College of Computing at the Georgia Institute of Technology in Atlanta, GA, U.S.A. invites applications for multiple openings for the position of Assistant Professor (tenure-track); openings for Associate Professor and Professor (tenured or tenure-track) are also available. All positions are full-time and are located in Atlanta, GA, U.S.A. The School of CSE focuses on foundational research of an interdisciplinary nature that enables advances in science, engineering, medical, and social domains. Duties of these positions include developing and sustaining a research program in one or more of our core areas: high-performance computing, data science and visual analytics, scientific computing and simulation, artificial intelligence and machine learning, and computational bioscience and biomedicine. We also welcome candidates in the area of AI and finance, with the op-

portunity to be considered for the JPMorgan Chase Endowed Career Development Professorship. Candidates are expected to demonstrate an exceptional commitment to the teaching and mentoring of students. Duties also include teaching undergraduate and graduate courses (classroom teaching), as well as providing service to the School, College and the Institute. A PhD degree is required by the starting date of the job, in a discipline relevant to the field of Computational Science and Engineering, including a degree in Computer Science, Mathematics, Engineering, or Science.

Applicants are expected to engage in substantive research with collaborators in other disciplines. For example, current faculty have domain expertise and/or collaborations in computational chemistry; earth sciences; biomedical and health sciences; urban systems and smart cities; social good and sustainable development; materials and manufacturing; and others.

For information on how to apply, go to: <https://cse.gatech.edu/faculty-hiring>.

Questions can be sent to cse-fac-recruit@office365.gatech.edu. For full consideration, applications are due by December 31, 2025.

Georgia Tech provides equal opportunity to all faculty, staff, students, and all other members of the Georgia Tech community, including applicants for admission and/or employment, contractors, volunteers, and participants in institutional programs, activities, or services. Georgia Tech complies with all applicable laws and regulations governing equal opportunity in the workplace and in educational activities. Georgia Tech prohibits discrimination, including discriminatory harassment, on the basis of race, ethnicity, ancestry, color, religion, sex (including pregnancy), sexual orientation, gender identity, national origin, age, disability, genetics, or veteran status in its programs, activities, employment, and admissions. This prohibition applies to faculty, staff, students, and all other members of the Georgia Tech community, including affiliates, invitees, and guests.



ADVERTISING IN CAREER OPPORTUNITIES

How to Submit a Classified Line Ad: Send an e-mail to acmm mediasales@acm.org. Please include text, and indicate the issue/or issues where the ad will appear, and a contact name and number.

Estimates: An insertion order will then be e-mailed back to you. The ad will be typeset according to CACM guidelines. NO PROOFS can be sent. Classified line ads are NOT commissionable.

Deadlines: 20th of the month/2 months prior to issue date. For latest deadline info, please contact: acmm mediasales@acm.org

Career Opportunities Online: Classified and recruitment display ads receive a free duplicate listing on our website at: <http://jobs.acm.org>

Ads are listed for a period of 30 days.

For More Information Contact:
ACM Media Sales
at 212-626-0686 or
acmm mediasales@acm.org

[CONTINUED FROM P. 112] can generate buggy code, and we know it can generate code with security vulnerabilities. So we hardly want the masses to use AI tools to generate code and take the output at face value.

This reminds me of some of your earlier work on trustworthy AI^d—in particular, about using formal methods to ensure that AI systems demonstrate things like reliability, safety, accuracy, robustness, and other properties.

The missing link right now with any AI system—whether it’s to produce text or code or images or video—is a verification step of some kind that says, “This system has all the properties we expect from trustworthy computing.” Of course, the challenge is that verifying AI systems means incorporating probabilistic reasoning and behavior. So suddenly, we have to revisit all those trustworthiness properties where things are not necessarily right or wrong, true or false, yes or no, on or off.

In a sense, it sounds like that’s another facet of computational thinking that’s more relevant in the age of AI—the abstractions of statistics and probability in addition to algorithms and data structures.

When I think back to my own undergraduate courses, I’m so grateful that, as a computer science major, I was required to take probability, statistics, linear algebra, and all of the mathematical foundations that computer science builds on. Math teaches you skills like analysis, proof, and precision—it goes back to the basics, not unlike English courses where you’re learning how to read and write.

Ethics is another issue that I’d expect to have increased relevance. I don’t know if you would call that an aspect of computational thinking per se, but it’s certainly an intimate neighbor.

I would relate ethics to the AI alignment problem,^e where we strive to build an AI system that aligns with human values. A few years ago, when I worked on ethics in data science with my colleague Chris Wiggins here at Columbia, we started with the Belmont principles.

“We will continue to debate the philosophical differences between human intelligence and machine intelligence.”

These are standards released by a Congressionally funded committee in 1978^f that identify guidelines for addressing basic ethical issues that arise from research on human subjects.

The three Belmont principles are Respect for Persons, Beneficence, and Justice. Others have proposed additional principles, but let’s just start with these three, as applied to computing and AI systems. The principle of Respect for Persons suggests, for example, that people should always be informed when they are talking with a chatbot. The principle of Beneficence requires researchers to conduct a risk/benefit analysis to maximize possible benefits and minimize possible harms—in, for instance, the decision a self-driving car might make on whom not to harm. Finally, the ethical principle of Justice requires us to ensure the fairness of risk assessment tools in the court system and automated decision systems used in hiring.

That’s a principle with which you’ve got some personal experience, through your involvement with a 2019 task force^g that sought to ensure fairness in New York City’s use of automated decision systems.

The genesis of the task force was that many agencies in New York City were making decisions about New York City residents using automated tools. Who gets low-income housing? Whose garbage gets picked up first? Who gets released without posting bail? There are fairness issues behind each of those questions, and if AI or automat-

ed decision tools are being used to decide, then you could imagine them not always being fair to each resident or to each of the many different populations that the city serves. At any rate, these questions will continue to be with us, and the Belmont principles are a starting point to help address them.

Back in 2008, you listed “What is intelligence?” among the field’s core questions.^h In the age of AI, it seems more relevant than ever. What other questions do you think are relevant to the future computing?

We continue to debate the philosophical differences between human intelligence and machine intelligence. Now, advances in generative AI raise additional philosophical questions such as “What is creativity?” While machines can create artifacts, are there aspects of human creativity that distinguish humans from machines?

Of course, AI is not the only thing that’s prompting us to reevaluate the nature of computing. How do you see those issues from your vantage as a university administrator?

We need to rethink how research universities are structured, how we think about the value of academic research, and where the funding comes from. When federal funding is reduced, everyone looks to industry. But industry can’t make up the difference—they have their own mission and bottom line, and their horizon is much shorter than the horizon that academic researchers have enjoyed since the Vannevar Bush era.

For computing, the added tension is that industry, with big compute and big data, has an edge over academia. Can we create new, more flexible, mutually beneficial models of engagement between big tech companies and academia?

We’ve also lost the trust of the public in science. Unfortunately, AI has added to this mistrust.

These are questions that I don’t have easy answers to, but they are fundamental to the future of the field. ■

^h <https://bit.ly/4mVuNfF>

Leah Hoffmann is a technology writer located in Piermont, NY, USA.

© 2026 ACM 0001-0782/26/2

^d <https://bit.ly/42fiQJl>

^e <https://bit.ly/4p1BzSF>

^f <https://bit.ly/4lVZoIw>

^g <https://bit.ly/4ngsJi5>

Q&A

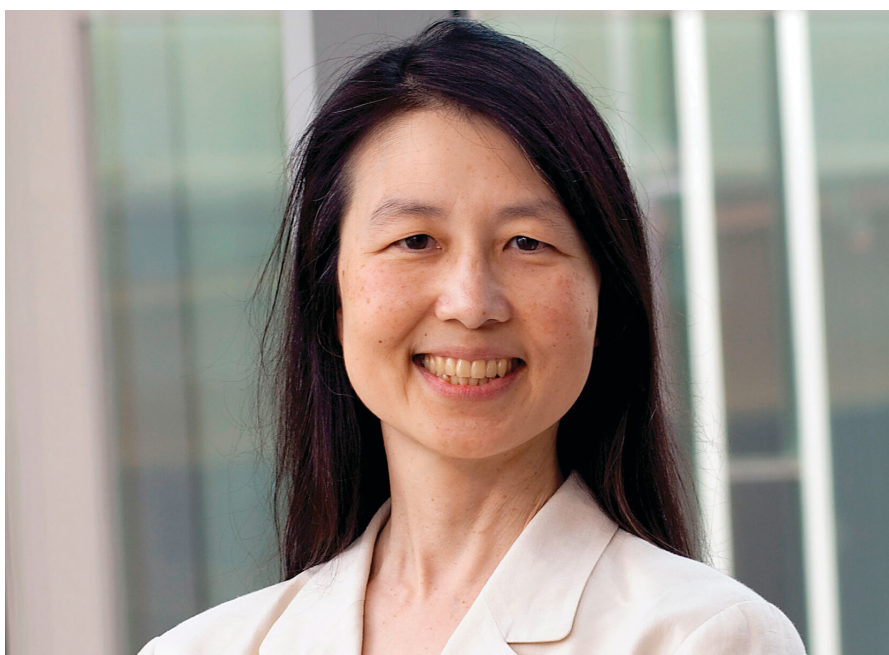
The Rise of Computational Thinking in the Age of AI

Jeannette Wing considers what computer science should be teaching in the context of artificial intelligence.

COMPUTING IS HARDLY the only discipline in which AI has prompted a moment of reflection. Yet coding agents and app generators raise profound questions about the purpose of the field—and indeed, its very nature. At this challenging juncture, Jeannette Wing, executive vice president for research and professor of computer science at Columbia University, finds herself reflecting on a concept she's promoted for nearly 20 years: computational thinking,^a or the mental tools and processes that give computer science its power. Here, she discusses how the idea has evolved—and which aspects of computational thinking will be most relevant in the future.

When we last spoke,^b you had just popularized the term *computational thinking* to describe how computer science transcends computer programming—and to capture the value of the models, methods, and abstractions that computer scientists use to solve problems. Let's get right into it: What is computational thinking in the age of AI?

When I first wrote about computational thinking,^c I wanted to make the point that the value of computer science goes beyond the specific programming skills that we teach. But I felt that a lot of people, especially in the realms of K-12 and introductory



undergraduate education, took that to mean we should teach coding to everyone. They equated computational thinking with computer programming, which is exactly what I was trying to get away from. In some sense, programming is just a skill you learn in order to express concepts in a way that machines can actually execute.

And now that AI can generate increasingly viable code, programming has become something of a commodity.

The real question is, what and how should we be teaching our students in the context of AI? Should we bother to teach coding, or should we just teach students how to use AI tools? When I

talk to my computer science colleagues in both academia and industry, the growing consensus is that computational thinking is more important than ever. AI offloads a lot of the mental activity you would otherwise have to do on your own. But it doesn't take away the need for you to understand what is generated. You still need to have an appreciation of what coding is and what code expresses in terms of algorithms and data structures, time and space efficiency, and so on.

At the moment, at least, AI code is also far from perfect.

We know AI can generate bad code. We know it [CONTINUED ON P. 111]

a <https://bit.ly/47ZrBuN>

b <https://bit.ly/3HRoLxt>

c <https://bit.ly/4m1fdO9>

NEW BOOK RELEASE



ACM BOOKS

Collection III

Multi-LLM Agent Collaborative Intelligence

The Path to AGI



Edward Y. Chang



ASSOCIATION FOR COMPUTING MACHINERY

Multi-LLM Agent Collaborative Intelligence

The Path to AGI

Edward Y. Chang

ISBN: 979-8-4007-3178-5

DOI: 10.1145/3749421

Today's large language models excel at pattern recall yet falter on long-range planning, self-critique, context loss, and the tendency of maximum-likelihood training to reward popularity over quality. MACI offers a promising route to AGI by orchestrating specialized LLM agents through explicit protocols rather than enlarging a single model.

Several modules remedy complementary weaknesses: adversarial-collaborative debate surfaces hidden assumptions; critical-reading rubrics filter incoherent arguments; information-theoretic signals steer dialogue quantitatively; transactional memory enables reliable long-horizon execution; and a dual-agent ethical court adjudicates outputs. Crucially, MACI also modulates linguistic behavior, tuning each agent's contentiousness and emotional tone, so the collective explores ideas from contrasting, affect-aware perspectives before converging.

Across healthcare diagnosis, investment support, scheduling, supply-chain management, and news-bias mitigation, MACI ensembles deliver significant improvements in reasoning depth, planning horizon, and reliability compared with similar-sized single models. By uniting structured debate, information-theoretic coordination, persistent memory, affect-aware discourse, and deliberative ethics, MACI demonstrates that rigorously validated multi-agent collaboration provides a practical, interpretable path toward robust general intelligence.

<http://books.acm.org>

Today's Research Driving Tomorrow's Technology

The ACM Digital Library (DL) is the most comprehensive research platform available for computing and information technology and includes the ongoing contributions of the field's most renowned researchers and practitioners.

Each year, roughly 20,000 newly published articles from ACM journals, magazines, technical newsletters and annual conference volumes are added to the DL's complete full text contents of more than 550,000 articles.

The DL also features the fully integrated and comprehensive bibliographic index, *The Guide to Computing Literature*—a continually updated index featuring millions of publication records from over 5,000 publishers worldwide.

For more information, please visit
<https://libraries.acm.org/>
or contact ACM at
dl-info@hq.acm.org

ACM



DIGITAL
LIBRARY