

COMMUNICATIONS

CACM.ACM.ORG OF THE

ACM

09/2021 VOL.64 NO.09

The Future Is Big Graphs

Managing IT Professional Turnover

An Internet of Things Service Roadmap

Whose Smartphone Is It?

Q&A with ACM Computing Prize Winner David Silver

Association for
Computing Machinery





IPDPS
2022 Lyon, France

Lyon, France • May 30 - June 3, 2022

ipdps.org

36th International Parallel and Distributed Processing Symposium

CALL FOR PAPERS

The five-day IPDPS program includes three days of contributed papers, invited speakers, industry participation, and student programs, framed by two days of workshops with peer reviewed papers that complement and broaden the main program. In 2022, IPDPS will return to meeting in person and will do so in the European metropolis of Lyon. For more details, visit the website at ipdps.org.

Authors for the main conference are invited to submit manuscripts that present original unpublished research in all areas of parallel and distributed processing. Work focusing on emerging technologies and interdisciplinary work covering multiple IPDPS focus areas is especially welcome. Topics of interest include:

- **Parallel and distributed computing theory and algorithms (Algorithms)**
- **Experiments and practice in parallel and distributed computing (Experiments)**
- **Programming models, compilers and runtimes for parallel applications and systems (Programming Models & Compilers)**
- **System software and middleware for parallel and distributed systems (System Software)**
- **Architecture**
- **Multidisciplinary**

Abstracts due	October 1, 2021
Submissions due	October 8, 2021
Author response/rebuttal	November 23-24, 2021
First round decisions	December 3, 2021
Revised submissions due	January 5, 2022
Final notification	January 15, 2022

Sponsored by



IEEE COMPUTER SOCIETY
TCPP
Technical Committee on Parallel Processing



GENERAL CO-CHAIRS

Anne Benoit (ENS Lyon, France)
Laurent Lefèvre (Inria & ENS Lyon, France)

WORKSHOPS CHAIR AND VICE-CHAIR

Ananth Kalyanaram (Washington State University, USA)
Suren Byna (Lawrence Berkeley National Laboratory, USA)

PROGRAM CO-CHAIRS

Yves Robert (ENS Lyon, France & Univ. of Tennessee, USA)
Bora Uçar (CNRS, Laboratoire LIP, Lyon, France)

PROGRAM AREA CHAIRS & VICE-CHAIRS

• Algorithms:

Gagan Agrawal (Augusta University, USA)
Sherry Li (Lawrence Berkeley National Laboratory, USA)

• Experiments:

Kengo Nakajima (University of Tokyo/RIKEN R-CCS, Japan)
Keita Teranishi (Sandia National Laboratories, USA)

• Programming Models & Compilers:

Didem Unat (Koç University, Turkey)
Mohamed Wahib (AIST, Japan)

• System Software:

Thomas Herault (University of Tennessee, Knoxville, USA)
Ana Gainaru (Oak Ridge National Lab, USA)

• Architecture:

Radu Teodorescu (Ohio State University, USA)
Mengjia Yan (MIT, USA)

• Multidisciplinary:

Alba Cristina Melo (University of Brasilia, Brazil)
Sunita Chandrasekaran (University of Delaware, USA)

IPDPS 2022 VENUE

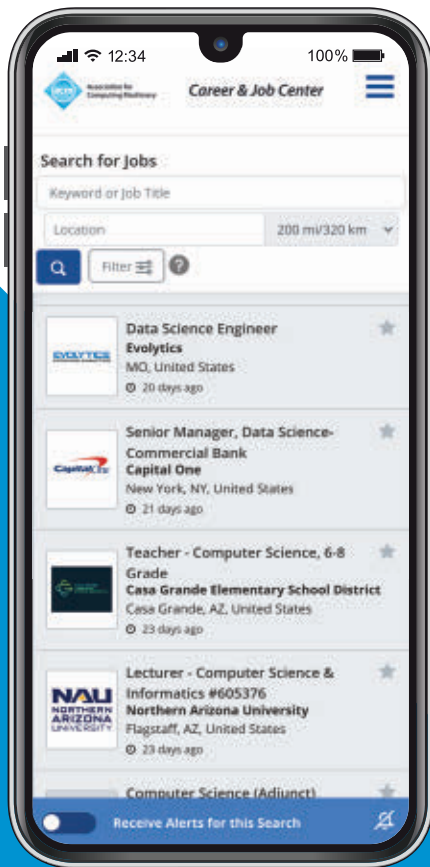
Lyon, France is a major science and technology center where Joseph Marie Jacquard designed the first mechanical weaving machine - a precursor of the first computer! Lyon still bears witness to 2,000 years of history, enriched by the signatures of modern architects and the dynamics of a full menu of cultural attractions, parks and rivers, and green city life. Located in the South-East of France, Lyon is 2 hours by TGV from Paris and less than 2 hours from the Mediterranean coast and the airport connects to one hundred plus international destinations.

The #1 Career Destination
to Find Computing Jobs.

Connecting you with top
industry employers.



The new ACM Career & Job Center offers job seekers a host of career-enhancing benefits, including:



Access to new and exclusive career resources, articles, job searching tips and tools.



Gain insights and detailed data on the computing industry, including salary, job outlook, 'day in the life' videos, education, and more with our new Career Insights.



Redesigned job search page allows you to view jobs with improved search filtering such as salary, location radius searching and more without ever having to leave the search results.



Receive the latest jobs delivered straight to your inbox with **new exclusive Job Flash™ emails**.



Get a free resume review from an expert writer listing your strengths, weaknesses, and suggestions to give you the best chance of landing an interview.



Receive an alert every time a job becomes available that matches your personal profile, skills, interests, and preferred location(s).

Your next job is right at
your fingertips.
Get started today!

Visit <https://jobs.acm.org/>

Departments

- 5 **Vardi's Insights**
The Sand-Heap Paradox of Privacy and Influence
By Moshe Y. Vardi
-
- 7 **Career Paths in Computing**
Computing Enabled Me To ...
How Music and Programming Led Me to Build Digital Microworlds
By Andrew Sorensen
-
- 9 **Letters to the Editor**
Turing Reaction
-
- 10 **BLOG@CACM**
Finding the Art in Systems Conversions, Naming
Doug Meil considers a third distinct type of development, while Mario Antoine Aoun ponders alternate names for ACM.
-
- 117 **Careers**

Last Byte

- 120 **Q&A**
Playing With, and Against, Computers
2019 ACM Computing Prize recipient David Silver on developing the AlphaGo algorithm, his fascination with Go, and on teaching computers to play.
By Leah Hoffmann

News



- 13 **A Model Restoration**
The architect of the Sagrada Familia appears to have done parametric modeling in his head; software is helping to complete the structure a century later.
By Marina Krakovsky
-
- 16 **Photonic Processors Light the Way**
Highly efficient light-based processors can overcome the bottlenecks of today's electronics.
By Samuel Greengard
-
- 19 **Non-Fungible Tokens and the Future of Art**
A new blockchain-based technology is changing how the art world works, and changing how we think about asset ownership in the process.
By Logan Kugler

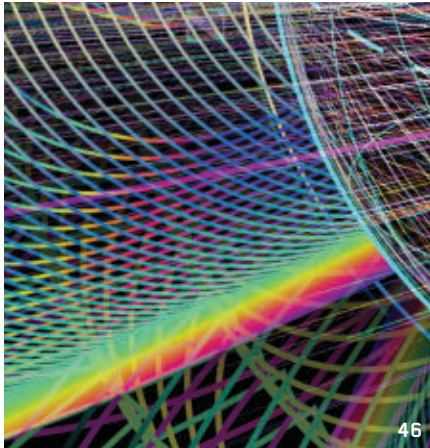
Viewpoints

- 22 **Law and Technology**
Protecting the Global Internet from Technology Cold Wars
Considering the perceived dangers of the global information flow.
By Anupam Chander
-
- 25 **Security**
Security Done Right Can Make Smart Cities Wise
Seeking security improvements for smart cities.
By Marjory S. Blumenthal
-
- 28 **Historical Reflections**
Women's Lives in Code
Exploring Ellen Ullman's 'Close to the Machine' and AMC's 'Halt and Catch Fire.'
By Thomas Haigh
-
- 35 **The Profession of IT**
Back of the Envelope
Back-of-the-envelope calculations are a powerful professional practice.
By Peter J. Denning
-
- 38 **Viewpoint**
Testing Educational Digital Games
Diversifying usability studies using rapid application development.
By Lamont A. Flowers
-
- 41 **Viewpoint**
Whose Smartphone Is It?
Should two private companies have complete control over the world's cellphones?
By James R. Larus
-
- 43 **Viewpoint**
AI Ethics: A Call to Faculty
Integrating ethics into artificial intelligence education and development.
By Illah Reza Nourbakhsh



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

Practice



46 **The Complex Path to Quantum Resistance**
Is your organization prepared?
By Atefeh Mashatan and Douglas Heintzman

54 **Quantum-Safe Trust for Vehicles: The Race Is Already On**
A discussion with Michael Gardiner, Alexander Truskovsky, George Neville-Neil, and Atefeh Mashatan.

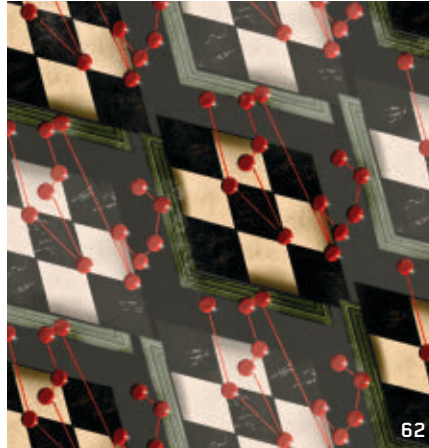
Q Articles' development led by **acmqueue** queue.acm.org

About the Cover:

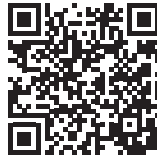
The record growth of interconnected data accentuates the vital role of graph processing today. There is no one-stop solution able to manage such diverse and voluminous data. This month's cover story—a joint effort of the computer systems and data management communities—explores the budding array of big graph processing systems. Cover illustration by Alli Torban.



Contributed Articles



62 **The Future Is Big Graphs: A Community View on Graph Processing Systems**
Ensuring the success of big graph processing for the next decade and beyond.
By Sherif Sakr, Angela Bonifati, Hannes Voigt, and Alexandru Iosup



Watch the authors discuss this work in the exclusive *Communications* video. <https://cacm.acm.org/videos/the-future-is-big-graphs>

72 **Managing IT Professional Turnover**
Organizational distrust, not compensation, is more likely to send IT pros packing.
By Kelly Idell, David Gefen, and Arik Ragowsky

78 **Formalizing and Guaranteeing Human-Robot Interaction**
As robots begin to interact closely with humans, we need to build systems worthy of trust regarding the safety and quality of the interaction.
By H. Kress-Gazit, K. Eder, G. Hoffman, H. Admoni, B. Argall, R. Ehlers, C. Heckman, N. Jansen, R. Knepper, J. Křetínský, S. Levy-Tzedek, J. Li, T. Murphey, L. Riek, and D. Sadigh

Review Articles



86 **An Internet of Things Service Roadmap**
A blueprint for leveraging the tremendous opportunities the IoT has to offer.
By Athman Bouguettaya, Quan Z. Sheng, Boualem Benatallah, Azadeh Ghari Neiat, Sajib Mistry, Aditya Ghose, Surya Nepal, and Lina Yao

Research Highlights

98 **Technical Perspective**
The Importance of WINOGRANDE
By Leora Morgenstern

99 **WinoGrande: An Adversarial Winograd Schema Challenge at Scale**
By Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi

107 **Technical Perspective**
Does Your Experiment Smell?
By Stefano Balietti

108 **Planalyzer: Assessing Threats to the Validity of Online Experiments**
By Emma Tosch, Eytan Bakshy, Emery D. Berger, David D. Jensen, and J. Eliot B. Moss



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

Executive Director and CEO
Vicki L. Hanson
Deputy Executive Director and COO
Patricia Ryan
Director, Office of Information Systems
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
Gabriele Kotsis
Vice-President
Joan Feigenbaum
Secretary/Treasurer
Elisa Bertino
Past President
Cherri M. Pancake
Chair, SGB Board
Jeff Jortner
Co-Chairs, Publications Board
Joseph Konstan and Divesh Srivastava
Members-at-Large
Nancy M. Amato; Tom Crick;
Susan Dumais; Mehran Sahami;
Alejandro Saucedo
SGB Council Representatives
Sarita Adve and Jeanna Neefe Matthews

BOARD CHAIRS

Education Board
Elizabeth Hawthorne and Chris Stephenson
Practitioners Board
Terry Coatta

REGIONAL COUNCIL CHAIRS

ACM Europe Council
Chris Hankin
ACM India Council
Abhiram Ranade
ACM China Council
Wenguang Chen

PUBLICATIONS BOARD

Co-Chairs
Joseph Konstan and Divesh Srivastava
Board Members
Jonathan Aldrich; Jack Davidson;
Chris Hankin; Mike Heroux; James Larus;
Marc Najork; Michael L. Nelson;
Holly Rushmeier; Eugene H. Spafford;
Bhavani Thuraisingham;
Julie R. Williamson

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

Computer Science Teachers Association
Jake Baskin
Executive Director

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
Diane Crawford
Managing Editor
Thomas E. Lambert
Senior Editor
Ralph Raiola
Senior Editor/News
Lawrence M. Fisher
Web Editor
David Roman
Editorial Assistant
Danbi Yu

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

Columnists
David Anderson; Michael Cusumano;
Peter J. Denning; Mark Guzdial;
Thomas Haigh; Leah Hoffmann; Mari Sako;
Pamela Samuelson; Marshall Van Alstyne

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

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

WEBSITE
<http://cacm.acm.org>

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

AUTHOR GUIDELINES
<http://cacm.acm.org/about-communications/author-center>

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

EDITORIAL BOARD

EDITOR-IN-CHIEF
Andrew A. Chien
aie@cacm.acm.org
Deputy to the Editor-in-Chief
Morgan Denlow
cacm.deputy.to.eic@gmail.com
SENIOR EDITOR
Moshe Y. Vardi

NEWS

Co-Chairs
Marc Snir and Alain Chesnais
Board Members
Tom Conte; Monica Divitini; Mei Kobayashi;
Rajeev Rastogi; François Sillion

VIEWPOINTS

Co-Chairs
Tim Finin; Susanne E. Hambrusch;
John Leslie King
Board Members
Virgilio Almeida; Terry Benzel; Michael L. Best;
Judith Bishop; Lorrie Cranor; Boi Falting;
James Gimmelmann; Mark Guzdial;
Haym B. Hirsch; Anupam Joshi; Richard Ladner;
Carl Landwehr; Beng Chin Ooi; Francesca Rossi;
Len Shustek; Loren Terveen; Marshall Van Alstyne; Jeannette Wing; Susan J. Winter

PRACTICE

Co-Chairs
Stephen Bourne and Theo Schlossnagle
Board Members
Eric Allman; Samy Bahra; Peter Bailis;
Betsy Beyer; Terry Coatta; Stuart Feldman;
Nicole Forsgren; Camille Fournier;
Jessie Frazelle; Benjamin Fried; Tom Killalea;
Tom Limoncelli; Kate Matsudaira;
Marshall Kirk McKusick; Erik Meijer;
George Neville-Neil; Jim Waldo;
Meredith Whittaker

CONTRIBUTED ARTICLES

Co-Chairs
James Larus and Gail Murphy
Board Members
Robert Austin; Nathan Baker; Kim Bruce;
Alan Bundy; Peter Buneman;
Premkumar T. Devanbu; Jane Cleland-Huang;
Yannis Ioannidis; Rebecca Isaacs;
Trent Jaeger; Somesh Jha; Gal A. Kaminka;
Ben C. Lee; Igor Markov; m.c. schraefel;
Hannes Werthner; Reinhard Wilhelm;
Rich Wolksi

RESEARCH HIGHLIGHTS

Co-Chairs
Shriram Krishnamurthi and Orna Kupferman
Board Members
Martin Abadi; Amr El Abbadi;
Animashree Anandkumar; Sanjeev Arora;
Michael Backes; Maria-Florina Balcan;
Azer Bestavros; David Brooks; Stuart K. Card;
Jon Crowcroft; Lieven Eeckhout;
Alexei Efros; Bryan Ford; Alon Halevy;
Gernot Heiser; Takeo Igarashi;
Srinivasan Keshav; Sven Koenig;
Ran Libeskind-Hadas; Karen Liu;
Joanna McGrenere; Tim Roughgarden;
Guy Steele, Jr.; Robert Williamson;
Margaret H. Wright; Nicholai Zeldovich;
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 © 2021 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.

ACM Media Advertising Policy

Communications of the ACM and other ACM Media publications accept advertising in both print and electronic formats. All advertising in ACM Media publications is at the discretion of ACM and is intended to provide financial support for the various activities and services for ACM members. Current advertising rates can be found by visiting <http://www.acm-media.org> or by contacting ACM Media Sales at (212) 626-0686.

Single Copies

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





Moshe Y. Vardi

DOI:10.1145/3477583

The Sand-Heap Paradox of Privacy and Influence

I LOOK IN THE mirror every day. I look the same as the day before. No change. Then I have a Zoom call with a person I have not seen in many years, I wonder how they got so old, and I realize that the other person must be thinking the same! The human mind always struggled with comprehending the cumulative impact of a large number of very small changes. This phenomenon has already been the subject of two classical Greek paradoxes: The Sand-Heap paradox and Zeno's paradoxes. When I was an elementary-school pupil, a favorite brain-twister was "How much is infinity times zero?"

We are facing the same paradox with respect to privacy and influence on the Internet. There are information items that we clearly want to protect, such as credit-card numbers. When such sensitive information is stolen via a cybersecurity breach, we clearly feel our privacy has been violated. But it is harder to feel a loss of privacy when we reveal a tiny bit of information at a time: a link clicked or a social-media posting "liked." Yet Internet companies have mastered the art of harvesting the grains of information we share with them, knowingly or unknowingly, and using them to construct sand heaps of information about us. Shoshana Zuboff, of Harvard University, named this business model of Internet companies "Surveillance Capitalism" in a 2019 book.

Zuboff called surveillance capitalism "an assault on human autonomy" and "a threat to freedom and democracy." We all realize that Internet companies persuaded us to give up some privacy for the sake of convenience,

but how much privacy have we given away? This is opaque to us. We see each grain of information given away, but not the heap of information. It is also opaque to us how this heap of information has been used by others not only to predict our behavior but also to influence and modify it. After the January 6, 2021 Capitol Insurrection in Washington, D.C., Zuboff wrote that "We can have democracy, or we can have a surveillance society, but we cannot have both."

The core issue, I believe, is that of human agency. Enlightenment thinkers downplayed divine authority and emphasized human agency. Rousseau wrote that "in the depths of my heart, traced by nature in characters which nothing can efface. I need only consult myself with regard to what I wish to do." Of course, we all know that the poet John Donne was right when he wrote "No man is an island entire of itself; every man is a piece of the continent, a part of the main." Our decisions and actions are clearly influenced by the social context. Yet, unless we feel coerced, we do not feel a loss of agency due to such social context.

Advertising, which originated in antiquity but emerged as a major commercial activity in the 19th century, expanded our social context, yet we still felt in control. After all, you can always go to the bathroom during a television commercial. Subliminal advertising, invented in the 1950s, uses sensory stimuli below an individual's threshold for conscious perception. While there is some controversy about its effectiveness, most people find subliminal advertising offensive, because it robs us from our sense of agency: we are being influenced without our

awareness. Indeed, many countries ban subliminal advertising.

The Internet has become a subliminal influence machinery. The days where the results of a Google search are ranked by the page algorithm are long gone. Google search results are now customized for each user individually by an opaque algorithm. The argument in favor of such customization is that it is aimed at maximizing user benefit, but it could also be aimed at maximizing advertising revenues. Analogously, the stream of posting on a Facebook user's wall is algorithmically customized, with the goal of "maximizing user engagement." Just like the grains of information we reveal about ourselves result in a heap of information about us, the grains of information that Internet companies give us result in a heap of influence we are not aware of.

Marc Rotenberg raised privacy concerns in a U.S. Senate testimony^a in 2000, but no action was taken then by the U.S. on Internet privacy. In 2018, Arnold King wrote a famous blog article,^b "How the Internet turned bad." The loss of privacy is at the core of that. It is time for us, as a community, to ask now: "How do we turn the Internet good?"

Follow me on Facebook and Twitter. 

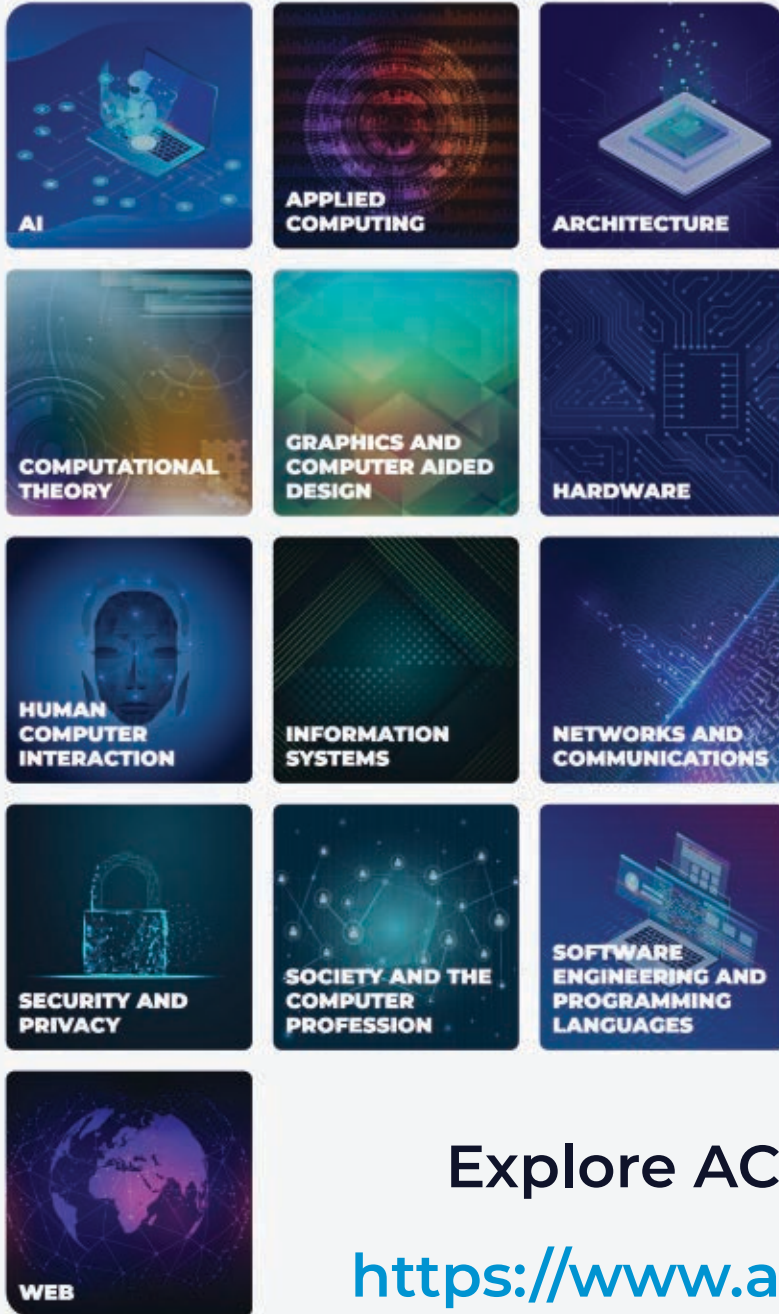
^a <https://epic.org/privacy/internet/senate-testimony.html>

^b <https://hackernoon.com/how-the-internet-turned-bad-bf348cdb99e7>

Moshe Y. Vardi (vardi@cs.rice.edu) is University Professor and the Karen Ostrum George Distinguished Service Professor in Computational Engineering at Rice University, Houston, TX, USA. He is the former Editor-in-Chief of *Communications*.

Copyright held by author.

Introducing **ACM Focus**



A New Way to Experience the Breadth and Variety of ACM Content

ACM Focus consists of a set of AI-curated custom feeds by subject, each serving up a tailored set of the latest relevant ACM content from papers to blog posts to proceedings to tweets to videos and more. The feeds are built in an automated fashion and are refined as you interact with them.

Explore ACM Focus today!

<https://www.acm.org/acm-focus>



Association for
Computing Machinery



SCITRUS

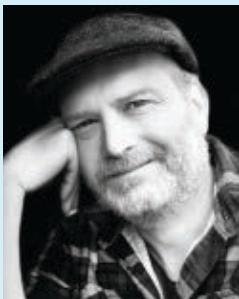


CAREER PATHS IN COMPUTING

DOI:10.1145/3476462

Computing enabled me to . . .

How Music and Programming Led Me to Build Digital Microworlds



NAME

Andrew Sorensen

BACKGROUND

**Australian, loving living
in Tasmania**

CURRENT JOB TITLE/EMPLOYER

Co-Founder MOSO

EDUCATION

**Ph.D. in Computer Science,
Australian National University**

I'VE ALWAYS THOUGHT of programming as a vehicle for exploration and discovery. My first experience with a computer, at the age of 10, was an interactive programming experience with Seymour Papert's Logo language. In a special math class, held in a touring computer lab (our school didn't own computers!) we "discovered" the properties of various geometric primitives by interactively driving a little turtle around a monochrome screen through issued commands in the Logo programming language. I was hooked!

The ability to discover and experience world building is a relatively unique privilege afforded to computer programmers. Programming my little turtle literally helped me 'experience' geometry. How fortunate we are to be

able to conceive of some microworld and then to attempt to create it, without requiring anything more than a computer and some power. In my early 20s, I remember thinking that with time and effort, I really can build anything. While clearly naive, such a mindset is something I truly hope to never outgrow.

Alongside my early love of programming was a love of music, and inevitably these two worlds collided. My first degree was in jazz trumpet performance—but since my early 20s, my primary musical instrument has been a programming environment. The environments I built were able to modify the sonic and visual landscape of the real world in real time when explored by a programmer. To paraphrase Stephen Holtzman, computers are of our time, and we are compelled to explore them.

My interest in programming music and sound has never been with the intention of automating its production. Instead, I create each new musical work as its own program, and I have spent inordinate amounts of time developing the programming languages that I use to help me construct these bespoke musical microworlds. At its extreme, the construction of this music becomes a real-time process where a complete performance is "livecoded" on the fly in front of an audience. What fascinates me about creating music in this way is that, at its best, there is an "ideal" point where the code becomes an essential component of the artwork.

An unintended consequence of building microworlds has been the number and depth of rabbit holes I've dived down. My younger self would never have imagined that my computer

music exploration would lead to interests in compiler design, type theory, analog simulation, time predictable architectures, digital signal processing, cyber-physical systems, distributed systems, and more. All of these technical interests have developed in the pursuit of real-world goals.

Another way I like to explore the world is through the creation of programming languages. My languages have been used to operate robotic telescopes, steer physics simulations on supercomputers, build massive distributed interactive computer-graphics installations and to play robotically controlled acoustic pianos in concert halls worldwide. What ties these disparate ideas together is that my languages allow for these systems to be modeled, built, modified, and controlled through a programming language interface in real time. Code in these interactive programming systems is often ephemeral, being an incomplete representation of the state of the system and the world in which it runs. This is a style of programming that focuses on code as a means of exploration first and of production second. A means to orchestrate complex systems at the meta level.

Throughout my career in computing, I have been fortunate to have roles in both industry and academia, as the CTO of a public company, Senior Research Fellow, and company founder. I have thoroughly enjoyed every position I've had. As the world becomes ever more programmable, I continue to look for increasing opportunities to orchestrate the world at the meta level.

Copyright held by author/owner.

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.



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

Turing Reaction

THE COVER PHOTO of the June 2021 edition of *Communications of the ACM* depicts Jeffrey Ullman and Alfred Aho, winners of ACM's 2020 Turing Award, and editorial content in the issue celebrates this selection. As ACM and *Communications* are well aware, for many Iranian members of the computing community, Ullman is the face of discrimination in academia. For more than 15 years, Ullman maintained a Web page that denigrates Iranian students using demeaning and derogatory language and told them they are not welcome in the computing community because of political reasons. The text is so clear and direct that it is hard *not* to see it as violating ACM's Policy Against Harassment, but of course it was "just" published on Ullman's page for more than a decade rather than being delivered at an ACM event, so in ACM's eyes it does not count.

For context: shortly after the award announcement, a public letter¹ signed by more than 1,200 individuals from academia and industry including over 450 ACM members condemned the decision and shared details of Ullman's discriminatory correspondence over the years. On April 19, ACM officially confirmed receiving the letter and published a response,² in which they did not even recognize the victims and that any harm was done. In parallel to publishing celebratory content about the winners, ACM's Executive Committee (EC), upon the request of *Communications* Editor-in-Chief to intervene, decided to reject an Op-Ed submitted by myself and five colleagues in which we criticized ACM leadership's, in our opinion, weak and inadequate response to the "CS for inclusion" letter and proposed concrete steps to be taken to help the cause.

To summarize:

1. Despite the claim that ACM EC and *Communications* were unaware of Ullman's long history of discriminatory rhetoric, they were officially made aware of this since mid-April;

2. Yet *Communications* published

celebratory content about Ullman two months after the award;

3. In parallel, ACM EC and *Communications* reject (as far as we know the only) "communication" done by the community through "Communications of ACM" about such an important issue, based on the reason that "enough has been said about it already;"

4. ACM still claims to hold D&I as core value,³ and that it "cannot accept any conduct that discriminates or denigrates an individual on the basis of citizenship or nationality."

Of course, *Communications* can come up with logistic excuses, but they cannot be relevant when it comes to any issue of such importance with more than a month left until the publication date, and evidently *Communications* and ACM's EC could in fact coordinate quickly when they choose to. I hope the ACM EC and *Communications* either will have the courage to accept the harm they furthered by Ullman's selection and *Communications*' celebratory coverage of it and correct their stance now that it matters most, or alternatively not claim to truly care about diversity and inclusion as core values or to be really against discrimination based on citizenship or nationality.

Resources

1. <https://csforinclusion.wordpress.com/>
2. <https://www.acm.org/response-to-letter>
3. <https://www.acm.org/about-acm/mission-vision-values-goals>

Mohammad Mahmoody,
Charlottesville, VA, USA

Editor-in-Chief's Response:

To air community concerns raised that the June issue of Communications did not present a balanced view, we are publishing the above letter.

The ACM's official response to the letter from CSforinclusion can be found at <https://www.acm.org/response-to-letter> and addresses the current criteria and process for Turing award selection, and the efforts under way to improve them to live up to the values ACM has articulated.

With respect to the actions of Communications, here is a bit of context

for how the production logistics of the June issue operate. For each month, final production for ~120 pages of editorial content begins approximately 10 weeks before the issue date. This content production was well under way weeks before the Turing selection was made. Shortly after the award selection is announced, an exceptional process is triggered to collect photographic and interview information to add to the "Turing issue." So this overload process, for the Turing issue and consistent with monthly publication, produced final layouts for the issue by May 1. With our physical publishers and mail delays, this is a hard constraint. The claim that logistics cannot be relevant belies the reality they are a constraint in any practical endeavor.

We regret any upset that inclusion of traditional content for an ACM A.M. Turing Award in the June 2021 issue caused but chose to do so based on the judgment to proceed in usual fashion until greater clarity emerged.

Andrew A. Chien, Chicago, IL, USA

Credit Where It Is Due

In Neil Savage's June 2021 news article "Getting Down to Basics" (p. 12) describing the work of 2020 ACM A.M. Turing Award recipients Alfred Aho and Jeffrey Ullman, I was surprised to read they were credited with the development of LEX and YACC. LEX was the work of Mike Lesk and Eric Schmidt, and YACC was the work of Stephen (Steve) Johnson (Aho provided some insight and came up with the name)—these tools were released as Unix utilities, and all three scientists were employees of Bell Labs.

David M. Abrahamson, Dublin, Ireland

Editor-in-Chief's Response:

Thanks, David.

Andrew A. Chien, Chicago, IL, USA

Communications welcomes your opinion. To submit a Letter to the Editor, please limit your comments to 500 words or less, and send to letters@cacm.acm.org

© 2021 ACM 0001-0782/21/9 \$15.00

The *Communications* website, <http://cacm.acm.org>, features more than a dozen bloggers in the BLOG@CACM community. In each issue of *Communications*, we'll publish selected posts or excerpts.



Follow us on Twitter at <http://twitter.com/blogCACM>

DOI:10.1145/3474351

<http://cacm.acm.org/blogs/blog-cacm>

Finding the Art in Systems Conversions, Naming

Doug Meil considers a third distinct type of development, while Mario Antoine Aoun ponders alternate names for ACM.



Doug Meil The Art of Speedy Systems Conversions

<https://bit.ly/2TpmcsG>

June 1, 2021

Building a software system de novo is the baseline way that software engineering is taught and understood. Use cases are identified, architectures and patterns are designed, and then software is implemented and deployed. Users are onboarded. This kind of green-field development can be exhilarating opportunity to create anew. Upgrading an existing system is a second and more frequent type of development as for any given system there is only one initial release but many subsequent releases. While upgrades primarily focus on incremental improvements, it is arguably a more complex case as upgrades are primary risks of outages and functional regressions, whereas with the baseline case there is nothing else in place at the time of initial release.

But what if there is a prior operational system in place? Specifically, one that is being replaced.

System conversions represent a third distinct type of development. The project scope now includes all of the effort of an initial software release plus an entirely new set of complexities. The prior system is often caught in a downward spiral: technical constraints may exist that make upgrades difficult, which in turn can diminish the organizational will to improve the system, which in turn reduces system performance and viability. The prior system, however, must be kept alive long enough to transition the functionality as well as support the data conversion to a new platform. This can become an anxiety-inducing software “race against time.” As an example of life imitating art, the 1994 action movie *Speed* with Keanu Reeves offers some surprisingly insightful lessons and how this situation can be managed.

Lesson #1: The Bus Couldn't Slow Down

In the movie, a transit bus is wired with a bomb and cannot go below 50 MPH without dire consequences. From a

software standpoint, if an existing system is highly utilized and still running critical functions but not well maintained, it can feel like this. There may be multiple factors all pulling on the existing system to slow it down: an outdated and non-scalable architecture, an outdated codebase, and perhaps even a lack of developers to support the aforementioned items. Ignoring the current system, though, only makes the problem worse.

Lesson #2: A Second Bus Was Required

To save the initial bus, a second bus had to be obtained. In the software world, the “second bus” represents the new system and the development team to create that system. This could either be managed as one team with two major responsibilities (support old system, build new system) or two teams, but one thing is clear: there is effectively twice as much work. A key mistake of system conversion development is only budgeting for the “new” development.

Lesson #3: The Second Bus Was Accelerated to Catch the First Bus

Achieving functional parity is one of the most difficult aspects of system conversions especially when the first bus has a 100-mile head start, metaphorically speaking. The “chasing system” needs a long-enough runway both in terms of time and budget, complicated by the fact that the prior system may still continue to evolve at the same time and not be a static target. Even the most well-intentioned projects can get tripped up on this. This type of development could take multiple fiscal quarters or years, and one of the biggest issues is executive expectation management.

Lesson #4: The Passengers Are Rescued

In the movie, the passengers are rescued in dramatic fashion, and anyone that has lived through a large system conversion will recognize this is pretty much what it feels like. To rescue the passengers, both buses must be operating not just at high speed, but also close proximity, re-emphasizing the importance of feature parity. Having a second bus running 50 MPH but 5 miles distant and receding doesn't help.

Additionally, software to assist in conversions—particularly large-scale data migrations—is required and is a special art. Such software still needs to adhere to software engineering best practices, but also needs to be fast (as conversion windows are always under a time crunch), explainable (as conversions are always being asked to explain exactly what happened), and automatable (as the best conversions are always heavily practiced).

The management of conversions is an important aspect of software engineering and not for the faint of heart. The process represents the bridge from the old to the new.

Lesson #5: The First Bus Was Retired

In the movie, the first bus exploded spectacularly after the passengers were rescued. In real life, such kinetic outcomes are not generally desirable. Shutdown processes informed by contractual or regulatory provisions are important considerations, such as saving the existing system state for a required period of time and potentially

leaving the system online in a read-only state. If a system state is saved as a backup, confirming that the backup can actually be restored is advised.

Conclusion

System conversions are a hard problem and will be ever-present in the software world as today's blue-sky development efforts become tomorrow's legacy code. Reasons for system-rot are myriad: technological obsolescence of frameworks or languages are one set of causes, but more than a few systems with reasonably current architectures have been undercut by boom-and-bust budgeting behaviors as systems are deployed with an enthusiastic initial release and then lay fallow. Technology leaders must actively manage every system in a portfolio. It's a lot of work to do this, but the alternative is worse.



Mario Antoine Aoun The Name Game

<https://bit.ly/3eYRBKN>

May 19, 2021

There was a recent discussion in *Communications'* Letters to the Editor section regarding a name change for ACM. Editor-in-Chief Andrew A. Chien even encouraged sending him ideas or suggestions for new ways to rethink the letters A-C-M. I, too, thought of an interesting name change for ACM, but after careful consideration, I realized I adore the current name for its longstanding value and history.

Concerning previous suggestions made by others (see *Communications* June 2020 and September 2020), we must be careful that our association is not dedicated only to its registered members. That is, it is dedicated for advancing computing machinery as science and profession, and not just for members contributing to its mission or benefitting from it. For instance, articles are published in *Communications* or other ACM periodicals by authors who are not ACM members. Also, people (like my lovely wife, for instance) may read ACM proceedings or attend ACM conferences with attendees who are not members of the association.

I liked Andrew Chien's comment concerning name change and the idea of recursion. For that reason, I suggest the following list of poten-

tial substitutions for Association for Computing Machinery:

- ▶ Association of Computing Minds
- ▶ Association for Computing Minds
- ▶ Association for Computing Minds and Machinery
- ▶ Association of Computing Minds for Computing Machines
- ▶ Association of Computing Minds for All Computing Machines

My personal favorite is *Association for Computing Minds* because it encapsulates many meanings and its hold on ACM's mission is twofold: it works toward the advancement of computing in terms of machinery, and it works toward the advancement of computing for scientists and professionals (as per ACM's motto, “Advancing Computing as a Science & Profession”). Besides, it reminds us of Turing's paper “Computing Machinery and Intelligence,” thus it implicitly offers tribute to him and explicitly to the evolution of computers while highlighting the mind and intelligence (natural and artificial).

What is interesting is ‘Computing Minds’ can refer to both a human and a computing machine. On one side, it gives legacy to the evolution of computers from their early invention as pure mechanical programmable calculators, as well as today's intelligent decision makers and knowledge discoverers. On the other side, it inspires programmers, software engineers, database designers, and computer scientists by calling them ‘computing minds’ as they create computing solutions by transforming thoughts into computing codes. In this way, we elevate ‘machinery’ to ‘mind,’ and at the same time we considered every person interested in this stuff as a computing mind, too.

Moreover, ‘Association for Computing Minds’ is new, novel, and unusual!

Still, as I said at the outset, I still adore ‘Association for Computing Machinery’ for its originality, value, and history.

What do you think?

Doug Meil is a software architect at Ontada. He also founded the Cleveland Big Data Meetup in 2010.

Mario Antoine Aoun is an ACM Professional member who has been a Reviewer for *ACM Computing Reviews* since 2006. He has 25 years of computer programming experience and holds a Ph.D. in Cognitive Informatics from the Université du Québec à Montréal. His main research interest is memory modelling based on chaos theory and spiking neurons.

© 2021 ACM 0001-0782/21/9 \$15.00

A New Journal from ACM

Co-published with SAGE



Collective Intelligence, co-published by ACM and SAGE, with the collaboration of Nesta, is a global, peer-reviewed, open access journal devoted to advancing the theoretical and empirical understanding of collective performance in diverse systems, such as:

- human organizations
- hybrid AI-human teams
- computer networks
- adaptive matter
- cellular systems
- neural circuits
- animal societies
- nanobot swarms

The journal embraces a policy of creative rigor and encourages a broad-minded approach to collective performance. It welcomes perspectives that emphasize traditional views of intelligence as well as optimality, satisficing, robustness, adaptability, and wisdom.

Accepted articles will be available for free online under a Creative Commons license. Thanks to a generous sponsorship from Nesta, Article Processing Charges will be waived in the first year of publication.

For more information and to submit your work,
please visit <https://cola.acm.org>

A Model Restoration

The architect of the Sagrada Família appears to have done parametric modeling in his head; software is helping to complete the structure a century later.

GLANCING AT BARCELONA'S still-unfinished Sagrada Família Roman Catholic basilica, with its famous sandcastle-like exterior, it is easy to get the wrong idea about its architect, Antoni Gaudí, as a carefree, loosey-goosey artist. The whimsical exterior hides a geometrically sophisticated, structurally advanced design—a big part of the reason this grand basilica, begun in 1882, has taken so many decades to build, remaining the world's longest-running ongoing architectural project.

This complexity required an utterly different approach to modeling than what architects had typically deployed. Instead of using two-dimensional drawings to guide builders, Gaudí relied heavily on large, high-fidelity plaster models—models that needed to be reverse engineered and rebuilt after extensive damage during the Spanish Civil War. In a separate project, Gaudí pioneered the use of hanging-chain models that enable changes in real time; though he did not use these interactive models on the Sagrada Família, they guided his thinking and prefigured the so-called parametric design software that has been instrumental to the acceleration of the project's pace in recent years.



The exterior of the Sagrada Família basilica, whose construction reflects the use of parametric design.

“Gaudí didn’t like to draw; he liked to model,” says Rafael Gomez-Moriana, a Barcelona-based architect and architecture critic who leads architectural tours of the city. “When you draw in two dimensions, you tend to limit yourself to those two dimensions,” to the kinds of simple angles that can be drawn easily with a T-square and a triangle (set square). Gaudí resisted such limitations, preferring organic-looking shapes that took inspiration from trees

and other natural creations. “Because Gaudí wanted to work with a vocabulary that was free from two-dimensional drawing conventions, it made sense that he would work with models,” Gomez-Moriana explains.

To precisely communicate the intricacy of his vision to the stone masons working alongside him, Gaudí made extraordinarily large plaster models, usually on a scale of 1:25 and, for his master model, 1:10. “The stone

masons were often going just straight from his model,” says Mark Foster Gage, a Manhattan-based architect and a professor at the Yale University School of Architecture who has written about Gaudí. “The larger the model, the more accurate it would be.” As a result, some of these models were an astounding 16 feet (about five meters) tall.

A New Arrival

In one of the project’s many setbacks, the Spanish Civil War (1936–1939) destroyed parts of the church and shattered the plaster models. Some effort went into restoring the models, but it was not until 1979 that information needed to be extracted from the model fragments for building purposes. That was also the year a young New Zealander named Mark Burry, a newly minted architecture graduate of the U.K.’s University of Cambridge, arrived to study the way Gaudí mixed architectural styles. He met with two elderly project leaders who had worked with Gaudí himself decades earlier, before the architect’s death in 1926. When Burry asked them how Gaudí had communicated his vision to the people actually building the church, they showed him boxes and boxes of the broken, dust-covered plaster models.

Intrigued, Burry switched his focus to reconstructing these complex models. Before committing to a year-long contract, the elders gave him a week as a trial period to show that he had a working method. The good news was that he understood the problem: he needed to somehow reverse engineer the models’ underlying geometries, which meant figuring out the points at which three adjacent three-dimensional surfaces intersected at a single point, called a triple point. “If you have a broken model, and you can at least find where those triple points are, you can then begin to unpack the constituent geometry,” says Burry, who is now a professor at Australia’s Swinburne University of Technology, and founding director of its Smart Cities Research Institute.

However, none of his education had taught him how to intersect these geometries. The project leaders had their eye on building the church’s nave—the

“If you have a broken model, and you can at least find where those triple points are, you can then begin to unpack the constituent geometry.”

central part of the main building—and the nave’s elaborate, kaleidoscopic ceiling is made up entirely of intersecting hyperboloids. “Every single opening in the ceiling and the windows of the nave is a hyperboloid of revolution—every single one,” Burry says. A hyperboloid of revolution itself is not terribly complicated: it is the surface of the three-dimensional (3D) shape that is formed from rotating a hyperbola around its axis (and can be formed by lathing wet plaster spun on a model-maker’s turntable). Once you start intersecting these surfaces, especially when the hyperboloids are not similar (say, a circular hyperboloid and an elliptical one), the resulting curves can become bewilderingly sinuous.

So finding the triple point of three different hyperboloids is a real challenge. Gaudí himself had managed to figure out how to design these complex models through sheer genius, Burry says. Working through trial and error, the master achieved these extraordinary results in just a few iterations. “I am absolutely astonished that he was able to zero in on particular combinations with particular outcomes without having to do thousands of models,” Burry says.

To begin reverse engineering Gaudí’s models, Burry spent three days in a fruitless search for a solution to the triple-point problem—until, one morning, he woke up with a game-changing insight. “I don’t

know where it came from, but I knew I’d seen this problem before,” he recalls. What dawned on him was that a mountain peak is, like any of Gaudí’s triple points, a point at which the mountain’s ridges intersect. Burry realized that Gaudí was, in a way, building a landscape.

“That’s where it got easy for me,” he says, because seeing the models as a landscape meant Burry could use the same tool geographers use to represent mountains. They use contours—topographic maps—in which all the points on a contour line represent the same height, with distances between contours corresponding to distances in elevation. “So if I could produce contours of these surfaces, I would be able to figure out where the intersections were and where the triple points were.”

This insight enabled Burry to model the components of Gaudí’s three-dimensional models on paper. With no software to aid him, it was slow, painstaking work, but in 1979, such software simply didn’t exist. In fact, Burry recalls having traded in his slide rule for a calculator just a couple of years earlier.

A Revolution in Software

By 1989, though, when his colleagues at the Sagrada Família invited him back to resume the reverse-engineering effort, some architects had begun using software, and Burry thought he’d be able to do the same. “Computer experts said this will be a breeze using AutoCAD,” Burry says. However, the experts he had spoken to were wrong. “I discovered that all the [available] architectural software was completely useless” for the task at hand. “I had this contract to get results, and I couldn’t.”

It was again time for Burry to take a page from another field. Designing boat hulls, he realized, poses the same spatial challenges, so he looked for boat-design software. Finding Intergraph VDS (for vehicle design system), he spent two weeks proving to himself that it could do the job. Yet, at about \$75,000 per license, Intergraph was prohibitively expensive for the Sagrada Família. He found a suitable alternative from another company, Computervision, that granted him free access. This software ran only on pricey Sun work-

stations, but Burry got access to the requisite hardware through a collaboration with the Polytechnic University of Catalonia, in Barcelona, Spain, where he had secured a visiting professorship. The collaboration with the local university, which lasted for 26 years, went fantastically well, Burry says, with them responsible for the horizontal surfaces and him taking on the vertical ones—the columns and the windows.

Soon after Burry and his colleagues adopted Computervision's software, the company released a major upgrade that replaced explicit modeling (which merely mimics what an architect would normally do on paper) with parametric modeling, which gives designers an entirely different way of working. Using parametric (also called constraint-based) modeling software, an architect can change the value of one parameter and see the software immediately adjust the associated geometries according to pre-defined constraints. "That's the quantum leap in stepping away from Gaudí's world to the contemporary design world," Burry says.

While Gaudí worked a century before this software transformed architecture, he had actually done something conceptually similar to parametric modeling when he used his famous hanging-chain models to experiment with complex arch configurations in designing the Church of Colònia Güell, another unfinished project which, in Burry's view, would most likely have been Gaudí's best building had he been able to complete it.

Gaudí was drawn to catenary arches for their natural elegance. Unlike more traditional arches, which require support from either massive walls or flying buttresses, a catenary arch (mathematically, a hyperbolic cosine) supports its own weight. It can do so by minimizing the forces in compression much the way its upside-down analogue—a chain or rope suspended from two points—minimizes the forces in tension, perfectly distributing the natural force of gravity along the curve. Extending this basic idea, Gaudí created chandelier-like webs of chains hanging from other chains, pulling and deforming the chains (and thus their corresponding arches) in interesting ways. These catenary

models are parametric, explains Gomez-Moriana, because whenever Gaudí made a tweak—lengthening a chain, say, or hanging a bag of lead shot to represent an additional load—the whole model would automatically adjust itself. "That's what a digital [parametric] model of a building allows you to do."

In the decades since beginning to go digital, the Sagrada Família has embraced other technologies. Three-dimensional (3D) scanners and printers have enabled additional types of model-making. And these days, stone-cutting robots are helping stone masons bring these designs to life.

Now in its final phases of construction, the Sagrada Família is scheduled for completion by 2026, for the centennial of Gaudí's death. "But," says Yale's Gage, "Just when you think they'll finish it, COVID comes," slowing construction once again.

To the fervently devout Gaudí, such delays might be a reminder of who's ultimately in charge. Known as God's architect, Gaudí once said, "My client can wait." Gage has a somewhat less-reverent take on the delays: "You almost have to wonder if God really wants this church," he says. Maybe, Gage adds, "He just really likes watching it be built." **□**

Further Reading

Burr, M.

Scripting Cultures:
Architectural Design and Programming
<https://bit.ly/3tUpxxo>

Burry, J.R. and Burry, M.C.

Gaudí and CAD
ITcon Vol. 11 (2006)
<https://bit.ly/37ckBdz>

Herta, S.

Structural Design in the Work of Gaudí
Architectural Science Review,
Volume 494, pp 324,
<https://bit.ly/3l6Xrxn>

Burry, J., Felicetti, P., Tang, J., Burry, M., and Xie, M.

Dynamical structural modeling A
collaborative design exploration
*International Journal of Architectural
Computing, Issue 01, Volume 03*
<https://bit.ly/2TKQ8QI>

Based in the San Francisco Bay area, **Marina Krakovsky** is the author of *The Middleman Economy: How Brokers, Agents, Dealers, and Everyday Matchmakers Create Value and Profit* (Palgrave Macmillan).

© 2021 ACM 0001-0782/21/9 \$15.00

ACM Member News

COMBATING ONLINE MISINFORMATION



Filippo Menczer is Distinguished Professor in the Luddy School of Informatics, Computing, and

Engineering at Indiana University, where he also serves as director of the Observatory on Social Media, located at the Luddy School.

Menczer received his undergraduate degree in physics from the University of Rome, in Italy. He earned his master's degree in computer science and his Ph.D. in computer science and cognitive science at the University of California, San Diego.

After obtaining his doctorate, Menczer became an assistant professor of management sciences at the University of Iowa, before joining the faculty of Indiana University, where he has remained ever since.

Menczer's current research interests focus on analyzing and modeling the spread of information and misinformation in social networks, and detecting and countering the manipulation of social media.

"I study all aspects of information diffusion, especially misinformation and manipulation of social media," Menczer says. This includes the detection of bots and the coordination of misinformation campaigns, as well as various kinds of abuse on social media by bad actors.

One tangent of his work is developing machine learning tools that are available to the public to detect and understand online manipulation.

Menczer does not think the challenges of online manipulation will go away anytime soon. He hopes his research will have an impact on computing, policy, and education to increase the quality of information shared on social media, but without censorship or hindering free speech.

"I think that's going to be one of our top priorities, to create a healthier information ecosystem and be less vulnerable to manipulation," Menczer says.

—*John Delaney*

Photonic Processors Light the Way

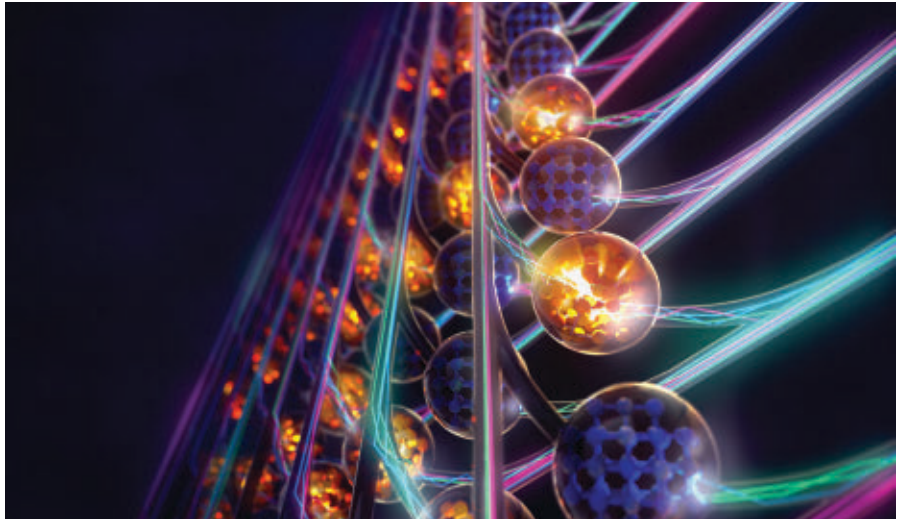
Highly efficient light-based processors can overcome the bottlenecks of today's electronics.

ONGOING ADVANCES IN electronics and computing have introduced opportunities to achieve things that once seemed inconceivable: build autonomous machines, solve complex deep learning problems, and communicate instantaneously across the planet. Yet, for all the advances, today's systems—which rely on electronic processors—are grounded in a frustrating reality: the sheer physics of electrons limits their bandwidth and forces them to produce enormous heat, which means they draw vast amounts of energy.

As demand for fast and low-energy artificial intelligence (AI) grows, researchers are exploring ways to push beyond electrons and into the world of photons. They are replacing electronic processors with photonic designs that incorporate lasers and other light components. While there is skepticism among some observers that the technology can transform analog computing, researchers in the optical space are now building systems demonstrating significant benefits in AI and deep learning.

“Photonic processors can resolve the bottlenecks associated with today's electrical computing systems. They are highly efficient from an energy perspective and they can overtake the standard clock rates of electronic systems by almost two orders of magnitude,” according to Maxim Karpov, a researcher at Switzerland's École Polytechnique Fédérale de Lausanne (EPFL). However, considerable challenges remain in optimizing photonics in integrated circuits, including finding the right mix of materials to replace silicon, which does not perform well with optical, and improved packaging techniques.

Nevertheless, the technology is emerging from research labs and popping up in real-world systems, including from a handful of commercial startups.



Demonstrating the principle underlying the photonic processor by showing the spread of light in a matrix of phase-change materials.

The possibilities are particularly enticing in areas such as deep learning, machine learning, and quantum computing. Technical advancements, including miniaturization and better packaging, are pushing the field forward at a rapid clip. Says Karpov, “Photonic computing and especially the area of integrated photonic computing, which uses silicon-based chips for optical signal processing, is actively evolving and beginning to make an impact.”

Seeing the Light

The idea of using light to speed processing is rooted in research from the 1980s. Yet, until recently, the idea had mostly stalled out. For one thing, the level of miniaturization required for components did not exist. For another, lasers and other components were not ready for primetime. As a result, the focus has mostly remained on eking out performance gains from conventional computing frameworks.

Although challenges still exist in the optical space—for example, it is not clear whether researchers package

photonics in a way that actually delivers widespread benefits in general computing systems—the field is advancing. “With the rise of machine learning and artificial intelligence, photonic processors found a field in which it can shine. Increasing volumes of data bring current electronic technologies to their limits,” says Johannes Feldmann, a post-doctoral researcher at Oxford University in the U.K.

Optical technology, already widely used for cabling, communications, and increasingly, system interconnects, takes direct aim at the growing limitations surrounding Moore's Law and von Neumann architectures. When compute-intensive tasks such as deep learning are tossed at electronic processors, they choke on advanced tasks and they devour energy. Moreover, scaling up systems to handle increasingly complex tasks is cost prohibitive. On the other hand, optics excels with low-precision linear functions. “This is where photonic processors challenge electronic processors: as hardware accelerators for artificial intelligence,” Feldmann says.

It is important to recognize the key differences between electrical and optical systems, says Nathan Youngblood, an assistant professor in the Electrical Engineering Department at the University of Pittsburgh. At the most basic level, electrical systems constantly change the number of electrons in a line, thereby charging and discharging metal interconnects that span two logic gates on the chip. “Fundamentally, optics is not limited by the charging and discharging of the interconnect line, so you can transfer data at much higher speeds. You don’t have a trade-off between energy consumption and modulation.”

Photons are attractive for performing computations because, unlike electrons, they can occupy the same physical state as other photons. This makes them more efficient and ideally suited to handle matrix-vector multiplications (MVMs) and convolutions used for deep learning. Light signals, which are modulated to encode input data vectors, are sent to the optical chip. “The light then propagates through the mesh of photonic waveguides. It is passively attenuated and mixed to transform it so that it conforms to the data matrix we want to use for multiplications,” Karpov explains. As the chip generates output, the light bearing the result of the multiplication operation is detected.

Since photons can propagate within the chip in an ultra-efficient manner, the system puts the power it draws to maximum use. At the same time, the light modulation speed can easily hit tens of gigahertz, which radically boosts the throughput of the system in comparison to electronic components. Finally, multiple elements can be placed on a single chip, including modulators, detectors, and even light sources. This makes the technology ideal for a wide variety of other uses, including optical data transmission, spectroscopy, LiDAR, MRI scans, and even optical circuit switching in datacenters.

Fueling these advances are new photonic platforms that use materials such as silicon nitride and lithium niobate, and fabrication processes for extremely low-loss photonic waveguides based on these materials. Meanwhile, a growing number of commercial foundries are equipped to build photonic integrated circuits (PICs), and startup companies such as Lightelligence, Lightmatter,

“Many of the obstacles that have prevented the technology from advancing are now being solved.”

and Optalysys are introducing solutions that address various advanced computing and communications tasks. Says Youngblood, “Many of the obstacles that have prevented the technology from advancing are now being solved.”

Designs on Speed

Significant breakthroughs in photonic computing have appeared over the last few years. Some of the biggest advances revolve around the fundamental way these systems are designed and constructed. Packaging and interconnects are evolving rapidly and enabling even more sophisticated capabilities. Says Karpov, “We’re seeing a gradual transition from isolated chip-based photonic components to more complex photonic systems, where various technologies and material platforms are integrated on the same chip in a hybrid way.”

In fact, packaging is crucial. “Without the right packaging, you cannot fully take advantage of the functionality and performance of the integrated circuits,” says Paul Fortier, senior engineer for Photonic Packaging Development at IBM’s assembly and test facility in Bromont, Canada. “Legacy photonic assembly is too often manual, time consuming, and difficult to bring the technology into high-volume production.” IBM is focused on developing low-loss optical interconnects, thermal management packaging, integrating photonics with microelectronics, and further miniaturizing components. The goal is to “allow light to get on and off the chips with the highest bandwidth in the smallest space, all the while being low-cost, reliable, and scalable for automation,” he says.

In February 2021, the field took a significant leap forward when a group of

researchers—including Karpov, Youngblood, and Feldmann—introduced a new architecture for photonics that combines processing and data storage on a single chip. With this design, the group developed a hardware accelerator for MVMs that serves as the basis for neural networks. Relying on different light wavelengths that do not interfere with each other, they were able to build a processor that handles complex parallel calculations and produces highly accurate results on convolution operations.

The technology framework, which is built using microresonator-based optical frequency combs (microcombs), provides a simple and straightforward way to parallelize computing operations on photonic processors, Karpov explains. Microcombs create ultra-compact light sources providing multiple equidistantly spaced optical frequencies. They allow the photonic tensor core to accommodate simultaneous data transfer and computing at speeds comparable to those of fiber networks, while generating near zero heat.

Nevertheless, further advances are required to push the technology into the mainstream—particularly in areas such as machine vision, which require ultra-fast calculations, Youngblood says. One obstacle, for now, is that photonic devices are physically larger than electronic transistors—even if the computing density, speed, and output is considerably higher for photonics. This makes optical chips unsuitable for certain tasks and situations. Another factor is that certain types of optical processors, such as free-space designs that rely on diffractive optics, introduce barriers related to the stability of the setup and the slow modulation speeds of spatial light modulators, Feldmann explains.

Scaling can also be a problem because photonic architectures still rely on electronic control circuits, which create a bottleneck. “The photonic processor itself could easily handle much higher data rates,” Feldmann says. “The photonic chip operates at a very low power level. However, the electronic control circuit driving it introduces much higher power requirements.” This means that further improvements in electronics are necessary to drive better photonic performance.

Finally, photonic foundries remain relatively immature, and the fabrication

of photonic circuits must be more reproducible in the specifications of individual components such as multiplexers and light sources, he adds.

A Bright Future?

The jury is still out on whether photonics will deliver niche benefits to computing or revolutionize the space. “It is possible that photonics may play a role in some analog aspects of computing, such as in neural network systems. And there is great potential for photonics to help relieve the communications bottleneck at the edge of electronic chips,” says Rod Tucker, Melbourne Laureate Emeritus Professor at the University of Melbourne and former director of the Institute for a Broadband-Enabled Society (IBES).

However, Tucker believes formidable challenges remain for swapping out digital electronic processing with digital photonic processing. As a result, a general-purpose photonic computer is not likely to appear anytime soon. “There is no photonic device that come anywhere near a digital electronic gate in terms of miniaturization, low energy consumption per operation, logic level restoration, and noise suppression. And no photonic device can store a bit of digital data as efficiently or as long as an electronic memory cell,” he explains.

Moreover, Tucker says, “There have been recent examples of clever experiments that show how photonic devices can emulate digital electronics, but the challenges emerge when one tries to scale up to the processing capacity of a state-of-the-art electronic chip containing millions of ultra-low-energy devices.” He believes a focus on direct and

fair comparisons with state-of-the-art digital electronics is paramount.

Feldmann says critics often miss the mark on the role of photonics. “An optical general-purpose processor is not very close to reality, but photonic processors shine currently in accelerating AI workloads.” For example, Lightmatter—a photonics AI startup rooted in MIT—generates 1.2 million inferences per second on a ResNet50 deep learning architecture versus 300,000 on an Nvidia DGX GPU. “This is for a full hybrid electro-optic system,” he notes. “Other startups that focus on Theoretical Operations Per Second (TOPs/W) also beat electronics by a substantial margin.”

The benefits of this alone could be significant. For instance, AI-related computing already consumes a considerable chunk of global energy, and the trendline indicates that increased demand for computing resources is nearly certain in the future, particularly as autonomous vehicles, robotics, and other machines demand more data-intensive input and output. Global sustainability hangs in the balance. “Photonic processors could reduce power consumption substantially,” Feldmann points out.

The biggest gains, however, would likely center on radically higher clock rates and parallelization that take machine learning and deep learning to an entirely different level—and unlock previously unachievable results. Optical signals can be modulated at up to 100 GHz, which opens the door to new and different uses. “For now, photonic processing makes sense where both high throughput and a high level of parallelization is needed on linear operations,

such as matrix-vector multiplications,” Feldmann says.

Although electronic microprocessors will continue to serve as the backbone of computing for the foreseeable future, photonic systems could begin to change computing—and many aspects of life. As researchers learn how to fully integrate electronic and photonic components into single systems and package them effectively, Markov sees a bright future for the field. Ultimately, “The technology is likely to lead to a variety of application-specific photonic processors that will support ongoing advances in digital technology and the rise of quantum computing.” □

Further Reading

Patterson, D., De Sousa, I., and Archard, L. The future of packaging with silicon photonics. *Chip Scale Review*, January 2017, <https://www.ibm.com/downloads/cas/MONL8N85>

Feldmann, J., Youngblood, N. Karpov, M., Gehring, H., Li, X., Stappers, M., Le Gallo, X., Fu, A., Lukashchuk, A., Raja, A.S., Liu, J., Wright, C.D., Sebastian, A., Kippenberg, T.J., Pernice, W.H.P., and Bhaskaran, H. Parallel Convolutional Processing Using an Integrated Photonic Tensor Core. *Nature*, 589, pp.52–58 (2021), February 23, 2021.

Shastri, B.J., Tait, A.N. Ferreira de Lima, T., Pernice, H.P., Bhaskaran, H., Wright, C.D., and Prucnal, P.R. Photonics for Artificial Intelligence and Neuromorphic Computing, *Nature Photonics*, 15, pp. 102–114 (2021), January 29, 2021.

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

© 2021 ACM 0001-0782/21/9 \$15.00

Milestones

ACM SIGACT Announces 2021 Knuth, Gödel Prizes

The ACM Special Interest Group on Algorithms and Computation Theory (ACM SIGACT) named Moshe Vardi of Rice University recipient of the 2021 Donald E. Knuth Prize for his outstanding contributions that apply mathematical logic to multiple fundamental areas of computer science.

Vardi’s work increased understanding of myriad

computational systems, and led to practical applications such as industrial hardware and software verification. The major themes of Vardi’s contributions are the use of automata theory and logics of programs to algorithmically prove correctness of system designs; the analysis of database issues using finite-model theory; characterizations of complexity classes such as P in terms of

logical expressions; and the analysis of multi-agent systems such as distributed computation systems, via epistemic logic.

ACM SIGACT also awarded the 2021 Gödel Prize to researchers Andrei Bulatov of Canada’s Simon Fraser University, Martin E. Dyer of the U.K.’s University of Leeds, David Richerby of the U.K.’s University of Essex, Jin-Yi Cai of the University of Wisconsin, Madison;

and Xi Chen of Columbia University, for advancing understanding of constraint satisfaction, an area of study within theoretical computer science.

Papers published by the researchers support a Complexity Dichotomy Theorem for counting constraint satisfaction and similar problems that are expressible as a partition function.

Non-Fungible Tokens and the Future of Art

A new blockchain-based technology is changing how the art world works, and changing how we think about asset ownership in the process.

BEHIND JEFF KOONS and David Hockney, the most lucrative auction for a piece of art from a living artist happened in 2021—and it was for a work that existed in a JPEG file. The artist Beeple’s “Everydays—The First 5,000 Days,” a series of digital artworks, sold at Christie’s for the princely sum of \$69.3 million.

It was a stunning event made possible by a technology called non-fungible tokens (NFTs).

NFTs are cryptographic tokens built on the Ethereum blockchain. NFTs are “minted,” then sold, just like Bitcoin. The difference, though, is that Bitcoin is “fungible.” If you swap Bitcoin with someone, you both still have the same asset: some amount of Bitcoin. There’s no functional difference between one Bitcoin or another.

However, NFTs are “non-fungible.” Each token is unique, and that token proves that you, and only you, have ownership rights over a digital asset—like Beeple’s art. As a random Internet user, you can view Beeple’s “Everydays—The First 5,000 Days” online, but only the person who bought the NFT tied to the art owns it.

This dynamic creates a simple, but powerful, change in how digital art works: it makes digital art exclusive. Once minted on the Ethereum blockchain, the NFT is represented on a public ledger that can’t be changed. By owning the token, you are proven the owner of the art piece. There is nothing stopping someone online from viewing, copying, and sharing a digital art file, but thanks to NFTs, they cannot fake possession of the art. NFTs make it possible to have exclusive ownership of digital art—something that was previously impossible.

In some cases, artists like Beeple may structure the NFTs tied to their work in unique ways. They may retain



rights to reproduce the image. They also may require automatic royalty payouts every time the NFT is resold.

Think of an NFT like the documents that come with owning an original Picasso. Art experts verify your Picasso is, in fact, original; they verify your ownership and provide documentation. As a result, the world accepts that you own an original Picasso.

The only big difference here is that NFTs make it possible to verify ownership of digital assets. Unquestionably there exist plenty of fraudulent Picassos, but given the limited supply of his works, and the legions of experts evaluating paintings, it is possible to prove that an individual owns a specific, legitimate Picasso.

It used to be impossible to do this for digital art. You could create digital art and everyone would know *you* made it, but anyone could reproduce it and share it with the entire world at the click of a button. In a scenario in which you can duplicate art with perfect fidelity indefinitely, the artist has some legal recourse to protect against how reproductions are used in commercial ventures. But who owns the original piece? *What* is the original piece? Is it

the first file the artist created? Is it the first version of the finished piece?

Before NFTs, there was no widely accepted way to determine the “original” piece of a digital artwork. There was also no widely accepted way to prove or transfer its ownership. NFTs have changed that, and with it, they’re changing the world of art.

“We feel very confident that this is just the beginning for NFTs,” says Meghan Doyle, a cataloguer of post-war and contemporary art at auction house Christie’s. “There is tremendous potential for NFTs in the art market and beyond. As a mechanism, the potential that NFTs have to shift the way that we establish ownership has no bounds.”

A Better Way to Create

With this ability to mint ownership of digital assets, NFTs have transformed how artists and creators make a living while changing how we buy, sell, and relate to art. NFTs also have expanded interest in blockchain technology beyond investment in Bitcoin and Ethereum. Experts still debate whether NFTs are the future of art or just a fad, but the amount of money changing hands for art backed by NFTs has the art world, technologists, and financiers paying attention.

The biggest mainstream use of NFTs today is for artwork, thanks to Beeple’s big sale.

NFTs are so prevalent in art because digitally native creators can bestow scarcity on works that consist entirely of pixels, says Doyle at Christie’s. They enable creators to earn more than they would outside the restrictions of the fine art world. Today, creators typically only get paid when they initially sell a piece of artwork; should the artwork’s new owner sell it to someone else, they pocket any gains made—and the artist gets nothing. However, NFTs use smart contracts to verify ownership and terms.

Those terms can include paying the original artist royalties every time the artwork changes hands.

“The smart contracts and royalties offer an attractive proposition to artists all over the world,” says Seyi Awotunde, who runs the Opulence (www.opulence.com) marketplace for art backed by NFTs.

Awotunde says he’s seeing artists demanding 10% royalties on resales become the market standard for NFTs. That gives artists the possibility of earning residual lifetime income from each piece of art they create. In turn, that means full-time artists can spend more time making art rather than working or freelancing to pay the bills.

“The reason why artists are interested in NFTs is that they’re easier, faster, more democratic, and far more appealing than the traditional art world model,” says Leighanne Murray, a contemporary art specialist who represents artists using NFTs.

She says artists are expected to go to art school, then work their way up the ladder. If they are lucky, they get noticed after years by a mid-tier gallery. From there, they are totally dependent on gallery directors to exhibit and sell their work. Only a select few ever reach the pinnacle of the pyramid.

Even if they do, success comes at a price. Art galleries often take a 50% commission on each sale, she says. Many have exclusivity contracts.

NFTs present an alternative. Any artist can mint an NFT for a piece of work. They set their own prices on easy-to-use online NFT marketplaces like OpenSea or Foundation. They control the promotion of their work through social media, and they keep all the earnings outside of basic transaction fees for selling an NFT.

For NFTs that pay royalties, the functionality is hard-coded into the smart contract on the blockchain. Once you set up the smart contract on the backend, royalties pay out automatically. There are no complicated payment platforms, invoicing, or logistics that artists need to manage.

There are benefits for collectors, too, says Chester Spatt, a finance professor at Carnegie Mellon University (CMU). Unlike physical art, it is easy to store and protect NFTs, since ownership is verified and secured on the

blockchain. It also is simple to transfer an asset electronically between platforms and devices.

Uncertain Profits, Uncertain Future

There is no doubt NFTs are having a moment, but their long-term value is still unclear, says Deeksha Gupta, a finance professor at CMU.

While NFTs enable digital art ownership on the blockchain, digital art itself has not changed. Looking at a print of a physical painting is different from the experience of seeing the original. That is why almost everyone has seen a *Mona Lisa* reproduction at some point, but millions still travel to see the artwork in person.

“Since the experience of looking at the original cannot be replicated when looking at a [physical] print, the original should have greater value,” says Gupta. “This is not an argument that can be made with digital art because any print is identical to the original.” That calls into question how valuable NFTs actually are.

Bryan Routledge, a finance professor also at CMU, thinks the downsides of NFTs go beyond how we value art. The very way the tokens work could create problems. Setting up the smart contracts that deal with future royalties for artists is not always easy, he says. And they make it possible to design complicated ownership structures.

“Is that a good idea? That is not clear,” says Routledge. “Some of the initial prices an artist receives reflect future appreciation.” NFTs may bring sale prices down by enabling royalties. In expanding the overall financial pie for creators, they also may reduce the size of everyone’s slice.

The so-called “last mile” problem with blockchain is another issue, says Christian Catalini, a professor who studies blockchain at the Massachusetts Institute of Technology Sloan School of Management.

The last mile problem refers to how digital assets interact and interface with the offline world. For instance, you need a way to turn your offline money into Bitcoin. No matter how useful Bitcoin is from a technological perspective, the only way you can use it is with sites that act like on- and off-ramps that connect the offline world to the digital one. Bitcoin solves the

problem by being sufficiently well-established that there are plenty of exchanges on which to transact in the cryptocurrency. There also is a vast, passionate segment of users.

“What gives value and makes some of these NFTs ‘useful’ in the long run critically depends on what their link to the offline world is,” says Catalini. “This link could be as simple as a community supporting their creation and exchange. If that community loses engagement, so does much of the value associated with NFTs.”

If NFTs overcome these challenges, they could transform art and technology as we know it. Murray predicts traditional art institutions and major galleries will come to accept NFTs. Awotunde believes the incentives align well enough that we shall see an NFT-backed artwork sell for over \$100 million this year. Digital art itself is also getting its day in the sun thanks to NFTs, says Doyle at Christie’s.

“Before the introduction of NFTs and blockchain technology, it was impossible to assign value to works of purely digital means,” she says. “Digital art isn’t new; it’s only new to the traditional art market. And we are now seeing it make up for lost time.”

That’s to say nothing of the bump NFTs give to blockchain as a whole, says Catalini.

“NFTs will accelerate the growth of the cryptocurrency space outside of finance, and will bring novel ideas and approaches from new sets of creators, artists, collectors of digital items, developers and more.”

Further Reading

Clark, M.
NFTs, explained, *The Verge*, Mar. 2021,
<https://bit.ly/3ic4WkT>

Nguyen, T.
The value of NFTs, explained
by an expert, *Vox*, Mar. 2021
<https://bit.ly/3ye5hJj>

Tarmy, J.
Crypto Investor Moves On to Picasso
After \$69 Million Beeple NFT Miss,
Bloomberg, Apr. 2021,
<https://bloom.bg/3A3LbCb>

Logan Kugler is a freelance technology writer based in Tampa, FL, USA. He has written for over 60 major publications.



ACM BOOKS Collection II

Intelligent Computing for Interactive System Design provides a comprehensive resource on what has become the dominant paradigm in designing novel interaction methods, involving gestures, speech, text, touch and brain-controlled interaction, embedded in innovative and emerging human-computer interfaces. These interfaces support ubiquitous interaction with applications and services running on smartphones, wearables, in-vehicle systems, virtual and augmented reality, robotic systems, the Internet of Things (IoT), and many other domains that are now highly competitive, both in commercial and in research contexts.

This book presents the crucial theoretical foundations needed by any student, researcher, or practitioner working on novel interface design, with chapters on statistical methods, digital signal processing (DSP), and machine learning (ML). These foundations are followed by chapters that discuss case studies on smart cities, brain-computer interfaces, probabilistic mobile text entry, secure gestures, personal context from mobile phones, adaptive touch interfaces, and automotive user interfaces. The case studies chapters also highlight an in-depth look at the practical application of DSP and ML methods used for processing of touch, gesture, biometric, or embedded sensor inputs. A common theme throughout the case studies is ubiquitous support for humans in their daily professional or personal activities.

In addition, the book provides walk-through examples of different DSP and ML techniques and their use in interactive systems. Common terms are defined, and information on practical resources is provided (e.g., software tools, data resources) for hands-on project work to develop and evaluate multimodal and multi-sensor systems. In a series of short additions to each chapter, an expert on the legal and ethical issues explores the emergent deep concerns of the professional community, on how DSP and ML should be adopted and used in socially appropriate ways, to most effectively advance human performance during ubiquitous interaction with omnipresent computers.

This carefully edited collection is written by international experts and pioneers in the fields of DSP and ML. It provides a textbook for students and a reference and technology roadmap for developers and professionals working on interaction design on emerging platforms.

<http://books.acm.org>
<http://store.morganclaypool.com/acm>

Intelligent Computing for Interactive System Design

*Statistics, Digital Signal
Processing, and Machine
Learning in Practice*

Edited by
Parisa Eslambolchilar
Andreas Komninos
Mark Dunlop



ASSOCIATION FOR COMPUTING MACHINERY

Intelligent Computing for Interactive System Design

*Statistics, Digital Signal
Processing and Machine
Learning in Practice*

Edited by
Parisa Eslambolchilar
Mark Dunlop
Andreas Komninos

ISBN: 978-1-4503-9026-2
DOI: 10.1145/3447404

▶ James Grimmelmann, Column Editor

Law and Technology

Protecting the Global Internet from Technology Cold Wars

Considering the perceived dangers of the global information flow.

IN THE SUMMER of 2020, the global Internet suffered two setbacks in quick succession. First, the Court of Justice of the European Union struck down the principal mechanism for personal-data transfers from Europe to the U.S.^a Two weeks later, President Donald Trump announced the U.S. was banning TikTok, an app owned by a company headquartered in Beijing, China. Perhaps surprisingly, both of these actions shared a common justification: data flowing to a company with foreign ties might subject that data to foreign surveillance. Thus, not only is it unsafe to send data across the Atlantic, it is unsafe to send data across the Pacific. Call this the “dangerous waters” theory of the Internet.

Invocations of the dangerous waters theory are piling up. In March

2021, the Bavarian data protection authority banned the use of U.S.-based MailChimp because of the possibility of U.S. surveillance. The next month, Portugal’s data protection authority similarly barred national census data from being sent to U.S.-based Cloudflare. In May 2021, the European Data Protection Supervisor opened an inquiry into the public use of Amazon Web Services and Microsoft Office 365. Word, apparently, may be a weapon.

The dangerous waters theory threatens the foundations of the global Internet. Focusing on the TikTok ban, I argue in this column that app bans should be carefully scrutinized, lest they be used as cover for other political ends. I begin by describing how President Trump’s TikTok ban represented a major departure from a quarter-century of U.S. support for a global Internet, and then argue that the national security claims against TikTok proved overblown, and describe lessons from this experience.

President Trump’s About Face on the Global Internet

China started it. At the dawn of the Internet age, it adopted a “Golden Shield”—what we came to call the Great Firewall of China—the modern version of an ancient effort to keep barbarians at bay. As James Fallows describes, “In China, the Internet came with choke points built in.” American sites such as Facebook, Google’s search, Twitter, and Wikipedia would be banned, accessible only via virtual private networks that dodged the address blocks.

For decades, the U.S. deplored the Chinese efforts to erect barriers to cross-border information flows. In 2000, President Bill Clinton famously scoffed that these Chinese efforts were “like trying to nail Jell-O to the wall.” A decade later, Secretary of State Hillary Clinton added “the freedom to connect” to the four freedoms enunciated by President Franklin Delano Roosevelt—the freedom of expression, freedom of worship, freedom from want, and freedom from

^a Court of Justice of the European Union. *Data Protection Commissioner v. Facebook Ireland and Maximillian Schrems*, Case C-311/18 (2020).

fear. Secretary Clinton put the U.S. firmly on the side of the global Internet: “We stand for a single Internet where all of humanity has equal access to knowledge and ideas.” For decades, then, the U.S. advocated for an Internet where information could flow across borders relatively unencumbered, subject to a few limitations such as local hate-speech laws.

But in 2020, the U.S. retreated sharply from that vision. On July 31, 2020, President Trump surprised the country by declaring, “as far as TikTok is concerned, we’re banning them from the U.S.” An app designed to share short video clips with the world, TikTok now found itself in the middle of a geopolitical storm.

Is the TikTok ban merely turnabout as fair play? Or does it herald a dangerous turn—when the champion of a global Internet declares it too dangerous to tolerate?

TikTok as National Security Threat

On August 6, 2020, President Trump followed through on his threat and issued twin executive orders targeting TikTok, as well as another popular app originally from China, WeChat. The orders were based on the President’s powers under the International Emergency Economic Powers Act (IEEPA).^{b,c} The TikTok executive order provided that within 45 days, “any person..., subject to the jurisdiction of the United States” would be prohibited from transacting with ByteDance Ltd., the China-headquartered owner of TikTok, or any of its subsidiaries. The Department of Commerce implemented this order by making it illegal to provide hosting, peering, or mobile app store services to TikTok—services it would need to keep running in the U.S.

On August 14, 2020, President Trump followed up with a second order requiring ByteDance to sell or transfer TikTok within 90 days, based on a review by the Committee on Foreign Investment in the United States (CFIUS).^d

The executive orders made two central claims as to TikTok’s national-security threat, one about the collection of information and the other about its dis-



semination. First, the U.S. claimed the Chinese government would use TikTok to gather compromising data about Americans, which it could then use for “blackmail.” The Trump Administration seemed to be relying on a frighteningly broad provision of the Chinese National Intelligence Law, Article 7, which states “any organization or citizen shall support, assist, and cooperate with state intelligence work according to law.”

Second, the U.S. argued the Chinese government would use the app to censor American speech or to disseminate propaganda. For example, TikTok had indeed been caught suspending an American teenager who cleverly used an

eyelash tutorial to criticize the Chinese government’s treatment of Uyghur Muslims. When this act drew public attention, TikTok quickly apologized for what it described as an error and restored her account. Since that time, posts with the hashtag #uyghur have garnered 82.5 million views on the app.

Overblown Fears

The TikTok ban was an improbable mechanism to improve national security for a number of reasons. Indeed, it is not clear whether the national emergency posed by TikTok was the threat of China exfiltrating data or Sarah Cooper’s TikToks mercilessly mimicking

b Exec. Order No. 13,942, 85 Fed. Reg. 48,637 (Aug. 6, 2020).

c Exec. Order No. 13,943, 85 Fed. Reg. 48,641 (Aug. 6, 2020).

d Regarding the Acquisition of Musical.ly by ByteDance Ltd., Exec. Order, 85 Fed. Reg. 51297 (Aug. 14, 2020); <https://bit.ly/3wvglQp>

the president's own words or TikTok teens reserving tickets for his Tulsa rally they had no intention of using. TikTok, after all, was the largest social network the president or his supporters had failed to master. One could not have imagined the president targeting Twitter and jeopardizing his free platform to reach his millions of followers there.

First, the dangerous waters theory proves too much. China is hardly alone in having laws that compel the disclosure of data held overseas, though the standards to compel production will differ widely across the world. The CLOUD Act explicitly grants this authority to the U.S. government, subject to extensive procedural safeguards. Other countries with similar laws range from Australia to Serbia.^e

Second, the dangerous waters theory would forbid even apps from domestic enterprises if they had operations in foreign jurisdictions that could compel them to produce data wherever it is held. Under this reasoning, even Apple might pose a national security risk to Americans because it is subject to Chinese jurisdiction.

Third, the TikTok ban undermines U.S. efforts against data localization. The U.S. has long made the free flow of data across borders a linchpin of its trade policy.

Fourth, TikTok could not have transferred all its data to the Chinese authorities without violating U.S. law. The Stored Communications Act bars companies from transferring the contents of communications to foreign authorities except under very narrow circumstances.

Fifth, there are many other ways to gather data about U.S. residents. Even weather apps can collect location data and sell it to data brokers who resell it to governments. Intelligence services certainly operate overseas. Supply-chain attacks like the SolarWinds hack, which was likely Russian in origin, suggest a particularly clever technique to exfiltrate data or compromise systems in bulk.

Sixth, TikTok was an odd target. It is not principally a private messaging platform, but rather an app that allows you to follow your interests or to share

^e U.S. Department of Justice. *Promoting Public Safety, Privacy, and the Rule of Law Around the World: The Purpose and Impact of the CLOUD Act*, n.3 (Apr. 2019).

Is the TikTok ban merely turnabout or fair play? Or does it herald a dangerous turn—when the champion of a global Internet declares it is too dangerous to tolerate?

them with the world. Users posting videos typically expect those videos to be shared publicly. Where Grindr focuses on private dating, TikTok is better known for public dancing. As the comedian Jimmy Fallon joked, “Apparently this is a very real national security threat—China’s government knowing which Americans can and can’t dance.”

Finally, subsequent history suggests the Trump Administration exaggerated the threat. Even when federal courts saw the government’s secret evidence against TikTok, they still sided with TikTok. Judge Carl Nichols, a Trump appointee, halted the TikTok ban.^f A second judge declared the government’s concerns “hypothetical.”^g And thus far, the Biden Administration has declined to pursue the Trump ban or divestiture orders further, implicitly suggesting the security threat is not as severe as presented by the prior administration. In fact, Secretary of Transportation Pete Buttigieg appeared on TikTok in April 2021. In June 2021, the Biden Administration withdrew the Trump IEEPA executive orders against TikTok and WeChat, instituting instead a broad review of applications subject to the jurisdiction of a foreign adversary. It said such a review would be based on “rigorous, evidence-based analysis and should address any unacceptable or undue risks consistent with overall national security, foreign policy, and economic objectives,

^f United States District Court, District of Columbia. *TikTok Inc. v. Trump*. *Federal Supplement, Third Series*, 490, (2020), 77.

^g United States District Court, Eastern District of Pennsylvania. *Marland v. Trump*. *Federal Supplement, Third Series*, 498, (2020), 642.

including the preservation and demonstration of America’s core values and fundamental freedoms.”^h Coupling the rescission of the prior order with this statement suggests the earlier executive orders failed to meet those standards.

Standing Up for the Global Internet

The national security rationales were overblown from the start, used to justify actions that just happened to target platforms that had proved a thorn in the side of political leaders. Trump borrowed even more of the Chinese Internet strategy than might be obvious—like the Chinese government, he sought to silence his critics.

Thankfully, independent courts proved a bulwark against such digital authoritarianism. Technologists, too, should press governments to demonstrate the actual risks, and not be content with vague hints of sinister activity too dark to reveal. After all, foreign companies can be targeted because they might carry political reports that are too controversial for domestic news media,ⁱ or because they compete with favored local corporations.

When major Internet platforms suspended Trump in the wake of the January 6, 2021 insurrection, Trump Administration Secretary of State Mike Pompeo tweeted, “Silencing speech is dangerous. It’s un-American.” He continued, “We cannot let [the Left] silence 75 [Million] Americans. This isn’t the CCP.” But Secretary Pompeo had it backward. One cannot imagine any Chinese tech platform suspending the Chinese president. Only democratic nations provide the freedom to refuse to promote or carry the views of those in power.

The U.S. should not cede its advocacy for the global Internet, one that connects people across the world.^j ■

^h See <https://bit.ly/3yLCmfx>

ⁱ Anupam Chander, *Googling Freedom*, 99 *Calif. L. Rev.* 1 (2011).

^j For a vision of national regulation that protects consumers while embracing a global Internet, see Anupam Chander, *The Electronic Silk Road*. Yale University Press, New Haven, CT, 2013.

Anupam Chander (ac1931@georgetown.edu) is a professor of Law at Georgetown University, Washington, D.C., USA.

The author thanks *Communications* column editor James Grimmelman as well as Kaitlyn Tsai and Lois Zhang for their research assistance.

Copyright held by author.

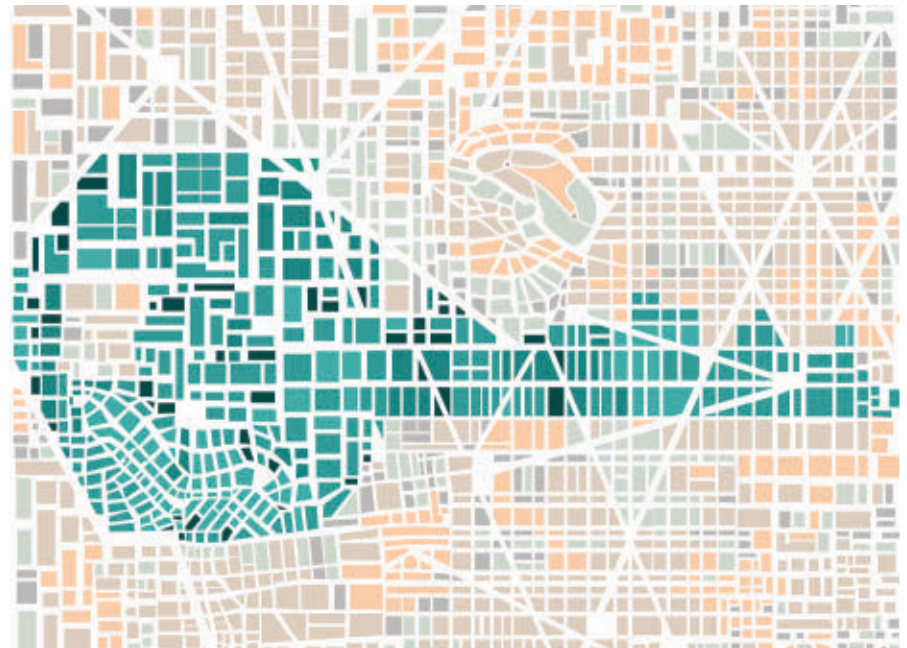
► Terry Benzel, Column Editor

Security

Security Done Right Can Make Smart Cities Wise

Seeking security improvements for smart cities.

IN 1995, AS the Internet became commercialized, visionary architect Bill Mitchell published *City of Bits*,¹ an exploration of how digital technology could profoundly change the structure and function of cities while cyberspace evolves to complement physical spaces. Information and communication technology (ICT) has embodied his title in even more ways than he might have guessed, and it promises to continue to do so. The 21st-century evolution of so-called smart cities partly realizes Mitchell's vision. Across smart cities worldwide, data is the common denominator, thanks to the various ICT applications that collect and share data, often through devices associated with the Internet of Things (IoT). The centrality of data drives many of the security concerns, as well as privacy concerns, for smart cities. Indeed, when the President's Council of Advisors on Science and Technology looked in 2016 at the range of technologies that can enhance cities, they moved from discounting smart-city hype to concluding that urban technology progress hinges on data—data collection, data analysis, and data integration.³ Notwithstanding enormous innovation and proliferating pilot projects around the world, smart cities remain in a phase of experimentation and development. Among the lessons being learned is how important security is to smart cities—to achieving the benefits of different applications and to avoid-



ing the kinds of problems observed increasingly when the confidentiality, integrity, and/or availability of data systems for infrastructure and services are compromised. It is time to ensure that security for smart cities is addressed early and often, including by engaging city residents in the process.

Why Smart Cities?

Smart cities promise genuine benefits to city governments and residents in terms of sustainability (through improved energy and water management), efficiency (through improved resource utilization and service delivery), public health (through air and water

quality monitoring and public health hot-spotting), and equity (through improved distribution of urban activity and access to services).⁵ The quest for such benefits led many of the first smart-city efforts to address challenges involving congestion and mobility. For example, multimodal transportation coordination is growing, often with a goal of diminishing the use of personal vehicles. Mobility as a service (MaaS) facilitates access to and integration of information on public transportation, micromobility (for example, bike- and scooter-sharing services), and other options. Both private and public actors are advancing MaaS. Many city govern-



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

ments seek to leverage if not directly offer MaaS services, and rideshare and transportation network companies are moving to offer information about how their service can connect to public transportation. These data-fueled mobility enhancements and the coordination benefits from evolving MaaS present the benign face of smart cities.

The realization of smart cities is more complex than just delivery of benefits. New concerns arise from the collection of the data undergirding those benefits. Some of those concerns reflect *who* is collecting the data—much of the innovation for smart cities involves companies that produce or package sensors and services that depend on the data collected. Transportation network companies tussled early with city governments about access to the data they collect, data that affects use of public infrastructure and affects demand for other mobility services. Companies supplying smart cities technology and services leverage data for competitive advantage, collecting and analyzing it in ways that are opaque to customers and governments. The March 2020 cancellation of the SidewalkLabs Toronto Quayside project appears to be a cautionary tale, associated with the publicized concerns of residents about anticipated data collection and use in that particular urban area. Who accesses and controls what data and what algorithms are themes being raised in challenging all kinds of ICT, as illustrated by controversy surrounding social media and big tech, but smart cities literally bring those concerns home. Although many contemporary concerns about widespread collection and use of data are associated with privacy, protection and stewardship of data begin with security.

More Technology Means More Security Risk

When it comes to security, smart cities connect concerns associated with specific data-collecting devices (for example, sensors of different kinds in different locations), local data storage, intermediate processing systems (expected to proliferate with the spread of 5G systems and architectures that will aggregate data), wireless communica-

tion among components, and the cloud systems that integrate data and host services. Smart cities, in short, are complex and multifaceted systems of systems. Layered communication systems, beginning with low-power wide-area networks or campus and community networks and extending to the cloud distribute the processing load for a growing volume of data generated by these systems. Each layer, of course, presents its own cyber vulnerabilities, a situation compounded with differing ownership and operation of different layers.

Concerns begin with IoT devices in homes—increasingly likely to capture images and voices even if their purposes relate to functions like temperature control—and extend to devices throughout the urban environment. In neighborhoods, technology can facilitate functions like traffic management—or if tampered with, totally confound it. The growth in ransomware attacks on city systems indicates malicious actors are tracking the growing use of ICT and data collection associated with city operations. At least as important, they are capitalizing on lagging attention to security in the acquisition and use of ICT in delivering the services on which lives depend. Security that is not sought explicitly might not be offered or supported in the configurations provided upon installation.

Surveillance has emerged as a kind of dual-use application in the context of smart cities.² Cameras are everywhere, whether or not they are visible, which increasingly is not the case. Governmental use of cameras in urban environments is not new—the U.K. is well known for its introduction of CCTV systems in the middle of the last century. What arouses contemporary concern is the combination of proliferating camera systems deployed by both public and private actors that use increasingly sophisticated software to recognize individuals from faces, gait, and other features, systems that can work even with the kinds of masks used during the COVID-19 pandemic and that might be able to detect the affect of the person observed. Cameras generate security concerns in a smart city that extend beyond their implications for privacy. Increasingly, they are combined

with other kinds of sensors dotting more and more urban surfaces—manhole covers, trash and recycling bins, streetlights, signposts, pavement, and so on. Cameras are combined with audio systems in large-event venues (for example, sports arenas) or certain urban areas where gunshots are unfortunately not uncommon. China appears to be the leader in facial recognition technology, blending a national focus on developing artificial intelligence (AI) capability with broad urban use of cameras, national citizen identifier numbers, and government commitment at all levels to smart-city development in China, and it also is active in selling such technologies to city governments in other countries.⁴

Less visible and obvious are the systems that connect physical activities of different kinds to payment and other financial systems. Transportation network companies built payment into their offerings from the outset, a feature that added to the appeal of ridehailing. MaaS more broadly features connections to payment, with examples involving links to parking (for example, SpotHero) or prepaid transit (for example, Whim) without the need for a farecard. Credit-card and other financial service providers have begun to partner with such city-focused programs. Although financial systems have high reliability and security requirements, their involvement implies that more data about individuals are collected and used than a transportation system, say, might otherwise need. Financial enterprises might be particularly attentive to cybersecurity, but where they provide third-party services to city governments questions arise about the disposition of data collected about people using public infrastructure and services.

The different trajectories of smart cities around the world make clear that local choices reflect local culture as well as capacity—what works in one place will not necessarily work in another. The extent of integration and use of surveillance seems most obvious in China. As in other countries, Chinese smart-city projects are specific to the physical city in which they unfold. They not only combine cameras and other kinds of sensors, they are advanced in a policy context that promotes smart cit-

Surveillance has emerged as a kind of dual-use application in the context of smart cities.

ies broadly, connects “smartness” with safety (“safe cities”), and integrates information about physical movement and other activity with social media and payment/finance systems. That integration yields a kind of governance labeled a social credit system, designed to facilitate or restrict activity depending on a person’s past activity and to promote trust among people as they interact and transact. The spotlight on safety implied with “safe cities” projects does not, however, guarantee security of the associated systems.

Security Should Be Baked into Smart City Governance

Globally, smart cities arise from the bottom up, with technologies deployed by retail outlets, transportation network companies, and residence owners, among others, and from the top down, with city governments procuring systems focused on single or multiple functions. These trends motivate many questions, beginning with who has access to what information associated with a given system, and how porous are both public and private systems? City governments and nonprofits (for example, FIWARE Foundation) have promoted open data for city applications, but many of the new services see competitive advantage in the data they collect and use, and various vendor-provided systems are closed. With urban systems evolving from both the bottom up and top down, cybersecurity may well fall through the cracks or at least be protected unevenly. Now that there have been a few years of smart-cities pilot projects, combined with the steady progress in numerous component technologies, it is time to take stock

and think through the large range of governance issues. Prominent among those is security.

More systematic dialogue about how technologies are designed, deployed, and used in cities is needed, perhaps catalyzed by a public awareness campaign. The last several years have seen powerful illustrations of citizens as sensors—people capturing phenomena on their cellphones and sharing, people agreeing to use those phones for pandemic contact tracing. Yet sousveillance cannot substitute for intentional and coordinated steps to promote security (and privacy) by design. Such steps can then enable clarity when urban officials and their vendors communicate with the public about the risks as well as the benefits of the systems within which we increasingly live, work, and play.

As Bill Mitchell observed, the spatial aspects of cities are “elaborate structures for organizing and controlling access.” The invisible cyberspace counterparts the Mitchell observed and anticipated might, absent explicit planning and attention to the security aspects, organize and control access and activities to a far greater degree than has been the case with physical infrastructure. If our cities are to become truly smart, then not only must technology developers and implementers explain how the AI behind some of the data processing works, they must also explain and assure the security aspects of complex urban ICT and associated data collection. Cybersecurity must be part of the new city planning process if we are to make the most of the potential of smart cities. ■

References

1. Mitchell, W.J. *City of Bits: Space, Place, and the Infobahn*, MIT Press, Cambridge, MA, 1995.
2. Muggah, R. and Walton, G. ‘Smart’ cities are surveilled cities. *Foreign Policy*. (Apr. 17, 2021).
3. President’s Council of Advisors on Science and Technology. *Technology and the Future of Cities*. Executive Office of the President, Washington, D.C., 2016.
4. Sutherland, M.D. *China’s Corporate Social Credit System*. IF11342, Congressional Research Service, U.S. Congress, Washington, D.C., 2020.
5. Winter, S.J. Who benefits? Considering the case of smart cities. *Commun. ACM* 62, 7 (July 2019).

Marjory S. Blumenthal (marjory.blumenthal@ceip.org) is Senior Fellow and Director of the Technology and International Affairs Program at the Carnegie Endowment for International Peace, Washington, D.C., USA.

The views expressed in this column are the author’s own.

Copyright held by author.

Historical Reflections

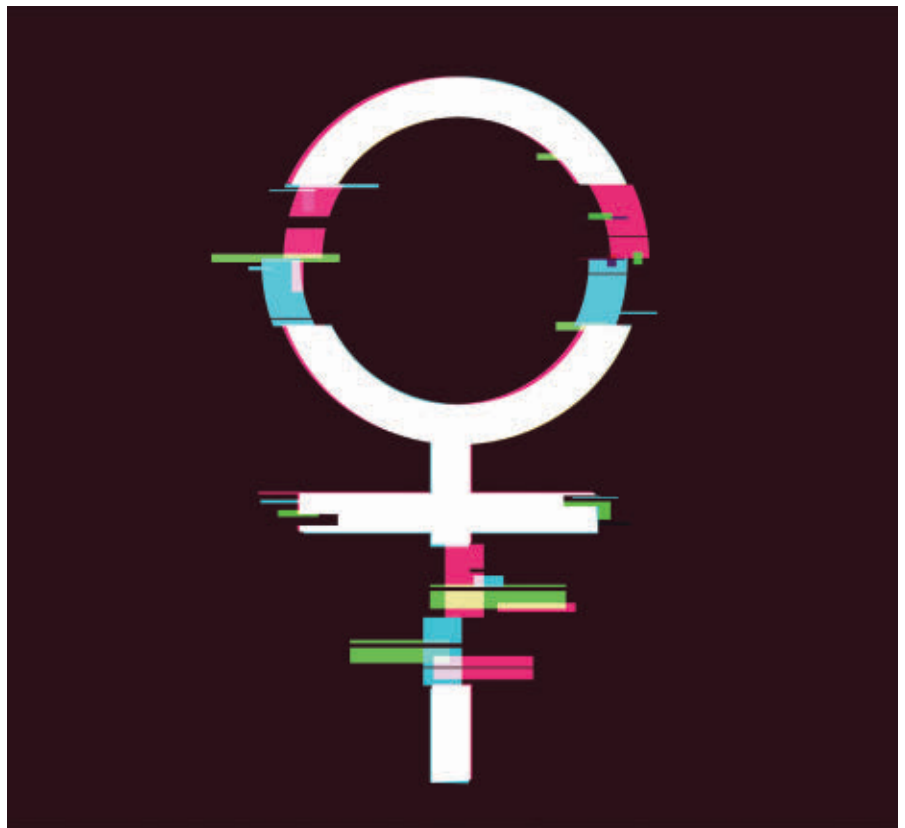
Women's Lives in Code

Exploring Ellen Ullman's 'Close to the Machine' and AMC's 'Halt and Catch Fire.'

IN THIS COLUMN, I look at two vivid depictions of programming work: Ellen Ullman's *Close to the Machine*, a memoir from 1997, and the television show "Halt and Catch Fire," which ran for four seasons starting in 2014. Both have central characters whose technology careers began in the 1970s and are followed through the mid-1990s—from the glory days of minicomputers and the first personal computers to the dawn of our current online existence. Both center on women who built their identities around computer programming, sometimes to the detriment of their personal relationships.

Getting Close to the Machine

When Ullman's book first appeared the computing world it described seemed quite different from the green screen eras described by Steven Levy in *Hackers* and Tracy Kidder in *The Soul of a New Machine* (both explored in previous "Historical Reflections" columns this year: January and April). Microsoft Windows had replaced the text interfaces of CP/M and timesharing systems. Most workplaces had already computerized and powerful personal computers were increasingly common in the home. The explosive growth of the World Wide Web was transforming the Internet from an academic enclave into a bustling shopping mall. Experienced programmers, like Ullman, were in great demand as the tech world thrilled with the excitement of unfolding possibilities.



The bigger shift, though, was literary: from the external perspective of *The Soul of a New Machine*, itself a classic of literary non-fiction, to a startlingly frank first-person voice. Most discussion of women's careers in IT focuses on sexism, hostile work and study environments, and ways to overcome barriers standing in the way of more equal participation. Ullman has surprisingly little to say about these issues, but she

is acutely aware that as a secular middle-aged Jew, bisexual woman, former communist, and Ivy League English graduate, she falls outside the typical demographic parameters of a software developer. This perhaps challenged her to think more deeply about her life and choices, and certainly equipped her to tie together the personal and professional with exceptional verve. Yet she is more concerned with telling

us what it feels like to be a programmer, specifically a programmer who tries to make sense of her own part in the evolution of capitalism, than in documenting the special challenges faced by women in IT.

This, she shows us, is what it feels like to stay up all night trying to configure a DBMS. This is how you square your career as a contract developer working for large corporations with your past as a communist agitator. And over there, Ullman confides as she continues our backstage tour of her own head, just past a prized stack of old Unix manuals and rubbing up against the fear of aging, you will see some disturbingly algorithmic sex with a callow cypherpunk named Brian who “looks exactly the way today’s computing genius is supposed to look: boyish, brilliant, and scary.”

Exploring the Work of Ordinary Developers

In some ways, though, Ullman’s experience is far more typical than that of the celebrated hackers Levy wrote about, or the billionaire entrepreneurs who receive most attention from technology writers. Most programmers, particularly back in the late 1970s when Ullman started out, did not have computer science degrees. In the 1990s, computer systems were generally much more important to people’s work lives than to their personal lives, given the investment made by most organizations to computerize their administrative processes.

Most developers produced custom database-driven application systems for the kinds of user organizations she describes, like banks, small businesses, and non-profits. Yet writers who have looked at software development focus on commercial packages and operating systems. Ullman drops hints of her past as an early employee of Sybase (the original developer of SQL Server) and mentions receiving windfalls from options at two startups. In those jobs she must have sat alongside people who went on to buy vineyards or become famous venture capitalists. But the work she chose to relate in detail is the analysis, design, and implementation of a custom application to handle the needs of local AIDS patients.

Most programmers, particularly back in the late 1970s when Ullman started out, did not have computer science degrees.

If most writing about software mimics Ayn Rand’s narrative in *The Fountainhead* of the visionary architect determined to create a monumental structure, Ullman’s programming work is more like the typical experience of a commercial architect, taking pride in designs for supermarkets or low-rise apartment buildings. Some of the book’s most interesting passages depict Ullman’s interactions with the “end users” themselves and the managers and supervisors whose desires shaped the systems she was programming. The “fleshy existence” of these users complicates the abstract versions of their needs and behaviors she has built into the system.

The book was published by City Lights books, an imprint of the legendary San Francisco bookstore. One legacy of Ullman’s immersion in radical politics, queer culture, and feminism, three things the city used to be known for, may be her unspoken conviction that her career and life deserve to be unflinchingly documented and publicly exhibited even though she did not start a famous company or invent a technology. Her determination to capture the subjective interior feelings of a character going about her ordinary business and her sense of herself as an outlier in her profession both put Ullman in a distinctively female literary tradition exemplified by pioneering modernist Virginia Woolf.

The View from Inside

In one passage Ullman contrasts the poised banalities expressed by a vice president of what appears to be Visa Inc., to whom programming seems trivial and mundane, with her own anx-

ious fascination with the interaction of humans and machines. The manager voices faith that systematic development methodologies and careful systems analysis work will assure the success of every development effort. Months later, Ullman catches the same manager in a confession that her firm’s core transaction system is an ageing mass of mainframe assembly code understood by only three programmers. “The slip-space opened before us,” writes Ullman. “The world and its transactions sat on one side. The programmers, the weird strange unherdable cats, roamed freely on the other. The vice president had peered into the abyss. Then she stepped back. ‘Don’t tell anyone I said that,’ she said.”

Ullman identifies the central challenge of software development as incompatibility between the clean, formal world of computer logic and the fuzzy humanity of the workplace: “The project begins in the programmer’s mind with the beauty of a crystal. I remember the feel of a system at the early stages of programming, when the knowledge I am to represent in code seems lovely in its structuredness. For a time, the world is a calm, mathematical place. Human and machine seem attuned to a cut-diamond-like state of grace.... Then something happens. As the months of coding go on, the irregularities of human thinking start to emerge. You write some code, and suddenly there are dark, unspecified areas.... Details and exceptions accumulate. Soon the beautiful crystal must be recut. This lovely edge and that one are gone. The whole graceful structure loses coherence. What began in a state of grace soon reveals itself to be a jumble. The human mind, as it turns out, is messy.”

Building on this insight she brilliantly explains the apparent paradox, visible also in Levy’s *Hackers*, that computer nerds can be conspicuously sloppy and disorganized in their personal life but aggressively precise in arguments. According to Ullman the “stereotype of the programmer, sitting in a dim room, growling from behind Coke cans” is rooted in their need to “talk all day to a machine that demands declarations.” “The disorder of the desk, the floor; the yellow Post-It notes everywhere; the whiteboards

covered with scrawl: all this is the outward manifestation of the messiness of human thought. The messiness cannot go into the program; it piles up around the programmer. Soon the programmer has no choice but to retreat into some private interior space, closer to the machine, where things can be accomplished.”

Fear of Obsolescence

The tech industry prizes boyishness and novelty, making Ullman’s concern with history and the passage of time a refreshing expression of humanity. She equates human aging with technological obsolescence, linking her own fears as a woman soon to enter her 50s (“a depressed, uninteresting region”) with her emotional connection to the material detritus left by relentless technological change. Early in her career, Ullman turned down an invitation to apprentice to an older programmer who “had made his peace with his own obsolescence.” The man was maintaining code on a platform developed in the 1950s. He offered her a chance to take his place, to one day become “the last human on earth who knows how to program in 1401 Auto-coder.” Ullman turned down this invitation but came to share his belief that there was “perverse dignity in knowing obsolete arcana.” Judging from the hundreds of thousands of views given to retrocomputing videos on YouTube that belief is now more widely shared.

I hope for both their sakes that Brian, her youthful “anarcho-capitalist” lover, is a heavily disguised composite or an outright invention. He personifies an emerging Silicon Valley culture that intrigues and horrifies Ullman. Brian is uncultured, almost feral, and committed to a theoretical polyamory that is in practice mostly celibate. He dreams of setting up porn servers in data havens and developing cryptographic banking systems so secure that users have to balance their own accounts. If real, Brian might have been a housemate of Peter Thiel or Elon Musk, who even then were dreaming up parts of what became PayPal.

Brian’s obvious unworthiness of our heroine is crystallized in a fireside chat early in their relationship. She shows him her prized, spiral-bound Bell Labs manual for Unix Release 3.0, issued in

June 1980. “It came,” she remembered, “from the days when I stopped being a mere programmer and was first called a ‘software engineer.’” He calls it “trash” and tells her to throw it away. Her collection of obsolete manuals must, she reflects, nevertheless contain “some threads, some concepts, some themes that transcended the details, something in computing that made it worth being alive for more than 35 years.”

Ullman gives a wonderful description of the sheer flood of paper and disks that engulfed the technologically committed in the 1990s: gigantic catalogs, updates to the Microsoft Professional Developer Network library, specialist magazines and journals, bulky manuals, new tools. Staying current means constantly mastering new technologies. Ullman relates with pride that she has taught herself “six higher-level programming languages, three assemblers, two data-retrieval languages, eight job-processing languages, seventeen scripting languages, ten types of macros, two object-definition languages, sixty-eight programming-library interfaces, five varieties of networks, and eight operating environments.” Beginning to weary of this, she wonders if perhaps the process “is simply unnatural for someone over thirty-eight,” particularly as career success tends to see developers move away from the machine and into new roles managing the people who know how to do the actual work. Yet

For me at least, the joy of discovering Ullman’s book was reading for the first time a faithful description of my own experience building client-server and Web systems.

she still feels joy when she is able to help a subcontractor find the mistake in his program. “For one more day at least,” she writes, “I would still be thought of as ‘technical.’” That meant a lot to her.

Ullman Intertwines Her Life and Work

Ullman’s book grips because she puts us right inside the mind of a 1990s software developer. In a foreword to a re-release of *Close to the Machine*, Jaron Lanier recalled his amazement on first reading it, to discover “a computer nerd who could write.” It formed “a bridge between reality at large and the empire of nerds, which seemed non-reactive and immune to subjectivity, beauty, love, or the acknowledgment of fundamental frailty.”^a More than 20 years on, I have encountered no comparably compelling memoirs written by other programmers.

Tracy Kidder and Steven Levy both serve as viewpoint characters for their readers, apparently normal people who closely observe obsessive hardware and software developers on our behalf. The importance of Silicon Valley and coding has been hard to ignore recently, but those with the skills and inclination to write about their experiences have usually been non-technical youngsters who stumble into the field. Consider, for example, the gulf between Ullman’s perspective and the 20-something memoirist and former New York publishing assistant Anna Wiener, whose recent *Uncanny Valley* critiqued startup culture from the viewpoint of a customer service worker. Plenty of novelists have tried their hand at depicting programming and other IT work but most tend to get hung up on surface detail (I mention a few exceptions, including Ullman herself, in the “Further Reading” section at the end of this column).

Ullman’s book appeared just as memoir writing was beginning to boom, with books like Frank McCourt’s *Angela’s Ashes* dominating best-seller lists and winning major awards. Most memoirs traded on the dramatic (and sometimes disputed) life experiences claimed by their creators: growing up in abject poverty, suffering tragic loss, re-

^a Lanier’s introduction appeared in the 2012 Picador edition.

covering from drug addiction, working in the sex industry, being diagnosed with terrifying diseases, fighting in wars, or building schools in Afghanistan. They offer vicarious experiences that most of us are happier to avoid encountering in person.

That is probably because coding does not provide the obvious narrative hooks of drug addiction or war. For me at least, the joy of discovering Ullman's book was reading for the first time a faithful description of my own experience building client-server and Web systems with MS Access, Oracle, SQL Server, and ColdFusion to underwrite an unusually comfortable graduate school lifestyle. When I stumbled across her book in the Center City Philadelphia branch of Borders, not far from the shelves of fat Que and O'Reilly programming manuals, I was hooked. Nobody had described so captivatingly the subjective experience of programming or the life of a freelance application developer. I gave a copy to my father and another to my dissertation advisor (an expert on late 19th-century labor), hopeful that it might communicate about new modes of work that I had not quite known how to explain myself.

The Changing Nature of Capitalism

Ullman contrasts the relentless impermanence of her own career, with its project-based alliances of convenience, virtual companies, and failed startups, with the determination of earlier forms of capitalism to present at least a façade of solidity. This is symbolized by the marble lobby of a grand bank she remembers her mother dressing up to visit, a physical space Brian hopes to replace with cryptographic algorithms.

She also remembered her father's accounting practice, built on long-term human connections, and the gamble he took to put together funding to secure her legacy: ownership of a small office building close to Wall Street. The cycle completes when she narrates a visit with her sister to meet with its struggling tenants. One complains of being unable to pay rent because of "the modems"—senior financiers are now telecommuting from the suburbs. Her career is helping to depopulate her own building. I have read a lot of books fulminating about the evils of "late capitalism," Silicon Val-

ley, and neo-liberalism but I have found nothing in them as honest and human as Ullman's efforts to reconcile the paradoxical strands of her own life. It is the opposite of the "determined solipsism" Brian shares with Levy's hackers.

Near the end of the book, Ullman returned to product development at the request of a venture capitalist, to attempt to salvage some value from a startup that is already failing. Her description echoes Kidder's experience with Data General engineers more than 15 years earlier: "A merry kind of hysteria took over the programmers. The situation was impossible, the deadline was ridiculous, they should have been completely demoralized. But, somehow, the absurdity of it all simply released them from the reality that was so depressing the rest of the company. They played silly jokes on one another. They stayed up late to see who could finish their code first. The very impossibility of success seemed to make the process of building software only that much sweeter."

They are all of them, Ullman realizes, deeply fortunate to be engineers "who built things and took our satisfaction from humming machines and running programs." Whether the company was liquidated or not, they had succeeded in transforming mere "scratchings on a white board" into something that worked. And then she breaks up with Brian and the book ends.

"Halt and Catch Fire"

The great strength of "Halt and Catch Fire," which covers the evolution of personal computing and networking from 1983 to 1995, is its ability to capture such moments of creative flow. Even at its worst, in its sputtering first season, the show has a more deeply felt connection to the work of programmers and engineers than anything else on television. As "Halt and Catch Fire" progresses it comes to share something else with Ullman's memoir and Kidder's classic book: it affirms the value of careers that do not necessarily lead to fame, power, and great wealth. The show matures into a moving examination of the creative joys and personal sacrifices its characters find in lives built around technological creativity.

Yet early on the show almost collapsed under the weight of the narrative template that has come to dominate the stories we tell about the computer industry: egotistical men becoming billionaires by bending reality to their will. "Halt and Catch Fire" only began to work after it recentered on its female characters and redefined success. It was marketed as AMC's follow-up to its hit series, and first original drama, "Mad Men." In its first season the show attempted to do the same things as "Mad Men," but in the 1980s Texan computer industry and with bad clothes. Its title, best ignored, was explained as an "early computer command" that forced "all instructions to compete for superiority at once." (HCF was actually a jokey unofficial mnemonic for an undocumented instruction that caused early Motorola processors to cycle relentlessly, probably for diagnostic purposes).^b

"Halt and Catch Fire"'s inexperienced creators, Chris Cantwell and Christopher C. Rogers, were likewise drawn to the idea of a computer industry drama patterned after "Mad Men." That show's protagonist—charismatic 1960s advertising executive Don Draper—had a tortured personal life and a traumatic backstory. He arrived amid a wave of shows centered on charismatic, brilliant, and psychologically complex antiheroes, a trend kicked off by "The Sopranos" and adopted by other acclaimed dramas such as AMC's own "Breaking Bad." "Mad Men" was the rare workplace drama that took work seriously. Its most resonant moments centered on obsolete technology and defunct brands.^c The emotional manipulation of Draper's advertising pitch for the Kodak Carousel slide projector as a personal time machine fueled by nostalgia, widely viewed as the show's finest moment, was compounded by our knowledge that the users who carefully ordered their slides to tell family stories are themselves mostly memories at this point. It is also difficult to forget a mordantly humorous incident that capped a season of asides about technological

^b Gerry Wheeler. Undocumented M6800 Instructions. *Byte* 2, 12 (Dec. 1977), 46–47; <https://bit.ly/3rbNQWX>

^c See <https://nyti.ms/3r5ztnf>

unreliability—a modern-day Jaguar marketing executive watching the show had “never been happier to see our car not start.”^d

Jobs and Woz at Compaq

Cantwell and Rogers stitched together bits of Don Draper and Steve Jobs to create Joe MacMillan, a brilliant computer marketer who left IBM under a cloud. It must have seemed like a good fit, given the reputation as an all-time great pitchman Jobs earned while introducing products such as the Macintosh and iPhone. After two movies and a blockbuster biography, Jobs is unquestionably the most famous computing innovator. The early Jobs was a real-life antihero, whose ability to convince others of his own genius created what colleagues called a “reality distortion field.” His conviction that he alone knew how things should be done led to disasters as well as triumphs. The fictional MacMillan is likewise prone to inspiring speeches and grand pronouncements but insecure, self-absorbed, and (like Don Draper) haunted by a mysterious past.

Apple Computer’s early story provides two archetypal Steves: the slick visionary who promised to “put a dent in the universe” and the hands-on hardware engineer who lived for technical challenges. By the law of narrative templates, where there is a Jobs there must be a Woz. In this case the Wozniak role is filled by unworldly engineer Gordon Clark, suckered by Joe into leading his hardware team. There are not any comparable clichés for female computer engineers, but because an all-male lead cast was out of the question the show’s creators adapted the archetype of the manic punk hacker girl, exemplified by Lisbeth Salander from *The Girl with the Dragon Tattoo*, to create the angry and damaged Cameron Howe.

Despite inheriting Jobs’ urge to create something insanely great, MacMillan’s secret plan turns out to be borrowed not from Apple but from Compaq: create a slightly faster, slightly cheaper IBM PC clone with a built-in screen and carrying handle. The show thus set out to answer a question that nobody has ever asked: What if Steve Jobs and Steve Wozniak started a PC-

clone company and hired Lisbeth Salander to write the BIOS firmware code needed to avoid infringing on IBM’s copyright? They are clearly the wrong people for that particular job. Creating a successful PC clone meant copying an effective but uninspired design while resisting the urge to make improvements that would compromise compatibility. If Jobs and Woz had founded Compaq rather than Apple then neither they nor it would be remembered today.

Perhaps the showrunners knew this all along and set out to play a long con on viewers who assumed Joe would triumph. It seems more likely, though, that they painted themselves into a corner in the pilot and spent almost an entire season trapped in what the *AV Club* called a “run of alternately humdrum and ludicrous episodes”^e before realizing they had to blow up their own show to escape. That escape is dramatic but infuriating. Cameron quits after the natural language interface she built for the computer is rejected. The others visit the Comdex trade show to unveil their computer, only to stumble onto a closed-door preview of the Apple Macintosh. Joe, abruptly realizing that his PC clone is not so special after all, then sets fire to the delivery truck holding the first batch in an overly literal interpretation of the show’s title. Many reviewers rolled their eyes when Ayn Rand’s architect hero destroyed his building because its brilliant design was tampered with. Seventy years later, the narrative gambit had not become any fresher or less juvenile.

The modest satisfactions of the first season come not from the characters but from the computer industry history worked into the background. It takes place not in California but Texas, which in the 1980s was home not just to Compaq but also to Texas Instruments, Tandy, and (a little later) Dell. The characters work at a stable mid-sized company that takes a fateful step into the personal computer market, just as firms like Texas Instruments did in real life. We get to see the cloning of a BIOS, the design of a case, and efforts to procure a display table at Comdex. Even the push for a natural

language text interface fits with the mid-1980s effort to build “conversational interfaces” for business systems.

Unexpected Greatness

After its unexpected renewal for a second season “Halt and Catch Fire” improved greatly by pushing the Jobs and Woz archetypes to the sidelines to focus on its female characters. Joe MacMillan is demoted from antihero to manipulative villain. The writers take a less indulgent view of his pathologies as they show him dating an oil heiress to sneak his way back into the computer industry. Gordon Clark spends two seasons as a bumbling supporting character and homemaker.

The show’s new central partnership is between Cameron Howe and Donna Clark, wife of Gordon, who spent most of the first season languishing in the Betty Draper template of bitterly neglected spouse with thwarted ambitions. Cameron’s brash bleached hair and swaggering aggression are gradually replaced with a much more plausible blend of defensive body language and demure clothing. Together, they found an online games company called Mutiny, modeled on the real-life Commodore 64 service Quantum Link. It begins with the ideal of leaderless hacker collective, which does not prove the most effective management structure.

Like Ullman’s memoir, “Halt and Catch Fire” rebuts the technology field’s chronic sexism by allowing its female characters to be as interesting, talented, and flawed as any male antihero. That is far more satisfying than simply setting up villainous male foils for them to overcome. The women are not perfect. Cameron, in particular, remains awkward and often selfish even as her character deepens. Her conflicts with the more pragmatic Donna are grounded in a recognizable clash between the narrow idealism of hacker culture and the compromises needed to run a business.

The show continued to improve in its universally acclaimed—yet little watched—third and fourth seasons. Critic Sean O’Neal judged this turnaround an “all-time great creative resurgence.”^f The plot moves fast, apparently because the show runners always expected their small audience to

^d See <https://bit.ly/3i3jamJ>

^e See <https://bit.ly/3hC84WT>

^f See <https://bit.ly/2ULRFpz>

lead to cancellation. I will not spoil its twists and turns with a detailed summary. Halfway through the characters relocate from Dallas to Silicon Valley, though as the entire run was filmed in Georgia there is not a particularly strong feeling of place attached to either setting. The last season jumps forward to the mid 1990s to focus on the early days of the Web.

The four main characters combine and recombine into different combinations of allies and enemies, but all have plausible motives for their mistakes or betrayals and there are no permanent villains. Keeping the same characters through these transitions requires them to demonstrate an implausible range of skills. Cameron, for example, jumps from BIOS coding to natural language processing, before settling down as a creator of video games and online services. Gordon is a microcomputer designer and electronic engineer who can reconfigure an IBM mainframe from batch operation to timesharing during a single evening or hack away to produce a pathbreaking antivirus engine.

As O'Neal noted, by its end "Halt and Catch Fire" had more in common with the funeral home family drama "Six Feet Under" than with "Mad Men." Like "Six Feet Under" it followed "perpetually unsatisfied people" whose hunt for "life-changing leaps" causes them to "shut out and repeatedly hurt the ones they love." Both shows overcame their gimmicky origins "to find their greatest resonance in small, quietly devastating, human interactions."^g Even Joe MacMillan eventually deepened into a recognizable and sympathetic human being with an awareness of his own flaws.

"Halt and Catch Fire"'s engagement with videogaming as a community-building activity is particularly effective. Toward the end of season three, Cameron explains the premise of her work in progress: a woman travels alone through space on a motorbike in search of five power-ups: a sense of proportion; a sense of humor ("to fend off most forms of attack"); a sense of self ("to appear and disappear"); decency; and "common sense, which lets her see everything more clearly." Coming after a devastating con-

g See <https://bit.ly/3xCUCeRJ>

Like Ullman's memoir, "Halt and Catch Fire" rebuts the technology field's chronic sexism by allowing its female characters to be as interesting, talented, and flawed as any male antihero.

flict in which Cameron's lack of these qualities hurt her and others, this presentation of game creation as a possible venue for psychological growth seemed like a rejoinder to Sherry Turkle's claim (discussed in my April "Historical Reflections" column) the journey of hackers to psychological maturity had been halted by a wrong turn into subculture that prized technological mastery over human connection.

"Halt and Catch Fire" builds a rich bench of supporting characters, including Bos, a folksy Texan financial executive nearing retirement age, and Diane, a successful Silicon Valley venture capitalist. Donna becomes her protégé, giving an almost unique fictional portrayal of venture capitalists as sympathetic characters. Even the soundtrack, consistently well-chosen and usually spot-on for the time periods in question, relaxes. The first season has a punk backing, from the likes of Hüsker Dü and Bad Brains. This mellows noticeably as the show progresses, to the extent of wrapping the final episode with Peter Gabriel's conspicuously unaggressive "Solsbury Hill."

Complicating Success and Failure

Writing this series of *Communications* "Historical Perspectives" columns has made me realize how much we lost when the outsized economic importance of tech from the late 1980s onward has focused most narratives of technological creativity on the pursuit and, usually, accomplishment of enor-

mous wealth and power. In contrast, "Halt and Catch Fire"'s depiction of technological careers driven by personal insecurity, a desire to build something new, and an attraction to constant change resonates with Ullman's memoir, Levy's depiction of hacker culture, and above all with Kidder's depiction of the Eagle team. Winking to this, in one of the show's final shots the camera pans slowly past a first edition of *The Soul of a New Machine*.

When the series finished, James Poniewozik wrote in the *New York Times* that its central concern turned out to be "failure as a condition of human growth."^h It is worth remembering, though, that these careers only look like failures when set against outsized definitions of success. Between them, "Halt and Catch Fire"'s four core characters found companies based on concepts an awful lot like Apple, Compaq, Dell, McAfee, AOL (as Quantum Link rebranded itself), eBay, PSINet, Netscape, Yahoo, and Google. The implausible breadth of good ideas they failed to fully exploit underlines the fact that for every famous company there were a dozen others with the same idea that are now forgotten. Statistically speaking, you were more likely to be an early employee of Excite, Lycos, or AltaVista than of Google; more likely to found Eagle Computer, Sirius Systems Technologies, or Leading Edge than Compaq. The main characters all achieve some respectable paydays. Some of them live in big houses and drive flashy cars. It is just that their lives look more like Ellen Ullman's than Bill Gates's.

I suspect this dominant narrative of successful computing careers as pathways to world domination may itself have played a part in the field's failure to attract and retain women. I wish we had more great books and shows focused on the processes, challenges, and satisfactions of more attainable technological careers. That might help more people of all genders to imagine themselves working successfully in the field. "Halt and Catch Fire" ends with a moment of potential as Donna shares a new idea with Cameron, but a few minutes earlier it allowed itself a more pointed message. Donna, by now the managing partner at her firm, con-

h See <https://nyti.ms/3xFc5PV>

venes a swanky backyard networking event for women working in Silicon Valley. Her reflection on the personal price she has paid for her success in becoming “a partner by trade and a mother and a sister by design” ends with a prediction that her teenage daughters, by then important characters in their own right, will have no need for such gatherings in their own careers. To us, 25 years later, that broken promise lands like a slap to the face.

Further Reading

Ellan Ullman has written no other books comparable to *Close to the Machine* but she did publish a collection of magazine essays written over several decades, many of them on related themes, in *Life In Code: A Personal History of Technology* (Farrar, Straus and Giroux, 2017). According to Ullman *The Bug* (Doubleday, 2003) began as an autobiographical account, becoming a novel when she decided to fictionalize her experience and give the central trauma, a long struggle to locate a simple bug in an early graphical user interface, to a male protagonist. *The Bug* is another notable portrayal

of the work of programming, though I personally found it less compelling than her memoir. Perhaps the external viewpoint makes her tortured surrogate, Ethan Levin, more difficult to sympathize with.

Richard Powers has earned a reputation as the contemporary novelist most inclined to take science and technology seriously. His breakthrough book *The Gold Bug Variations* (William Morrow, 1991) focuses on the cracking of the genetic code in the 1950s, but a parallel narrative set in the 1980s includes some great descriptions of IBM mainframes based on his own experience as a programmer and operator. His later *Ploughing the Dark* (Farrar, Straus and Giroux, 2000) centers on a young woman developing a virtual reality system for a Microsoft-like company, but I found it much less convincing. *Gain* (Farrar, Straus and Giroux, 1998) barely mentions computers but comes closer to the spirit of Ullman’s book by intertwining the history of a fictional Midwestern chemical conglomerate with the interior perspective of a woman fighting cancer. If you enjoy stories about the intertwining of human flaws and scientific

creativity you may also like Allegra Goodman’s *The Intuition* (Dial Press, 2006), a novel that tracks the work of a 1980s cancer research team riven by an allegation of misstating the results of a promising treatment.

Douglas Copeland is not a programmer, but he is a snappy describer of pop culture artifacts. Not long after coining the term “Generation X,” his quest for the generational zeitgeist took him to Microsoft and Silicon Valley in the mid-1990s for his novel *Microserfs* (HarperCollins, 1995). The opening, describing a communal house full of Microsoft workers, is wonderfully zesty (and can be read on the *Wired* website), though the book starts to bog down once they decamp to California to build something that, in retrospect, seems a lot like Minecraft. ■

Thomas Haigh (thomas.haigh@gmail.com) is a professor of history at the University of Wisconsin—Milwaukee and a Comenius visiting professor at Siegen University.

This work was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation)—Project-ID 262513311—SFB 1187 Media of Cooperation.

Copyright held by author.

ICCCQ

The Second International Conference
on Code Quality (23 Apr, online)

Static/Dynamic Analysis, Program Verification,
Bug Detection, and Software Maintenance

www.iccq.ru

CfP closes on 18 Dec

In cooperation with
ACM SIGPLAN and SIGSOFT
IEEE Computer Society





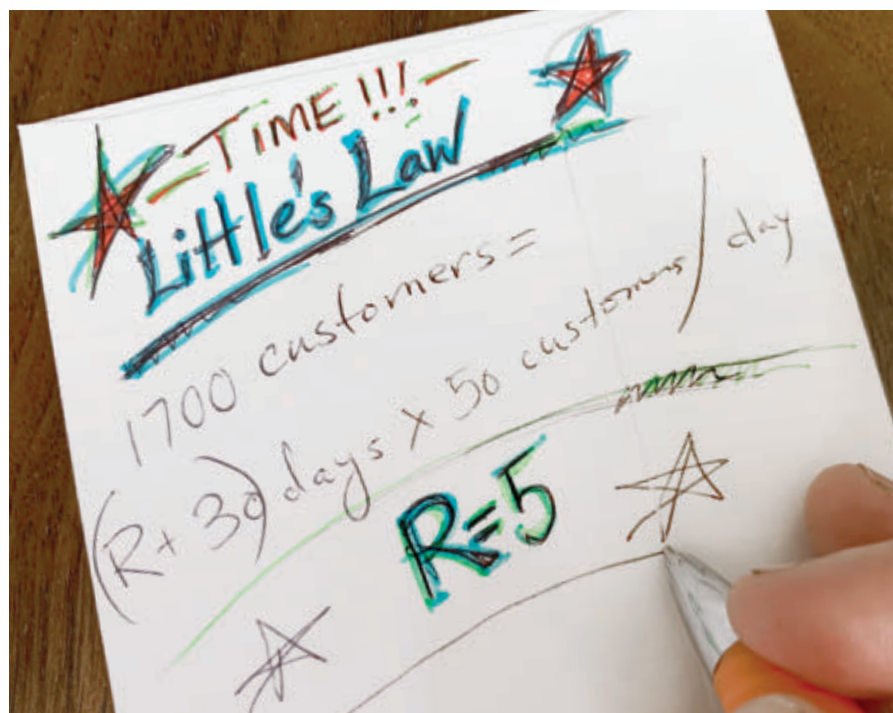
The Profession of IT Back of the Envelope

Back-of-the-envelope calculations are a powerful professional practice.

IN OUR PROFESSIONAL practice, we are often called to perform rapid, approximate calculations without a calculator. Any available scrap of paper such as an envelope will do to scribble on. These calculations are more than guesses but less than accurate mathematical proofs. Such scribbles can become the stuff of legend. How many times have we heard stories about successful and influential startup companies being born on the backs of napkins in the pub?

One consummate performer of such approximate calculations was the physicist Enrico Fermi, who famously estimated the TNT energy release equivalent of the first atomic bomb test in the New Mexico desert in July 1945 by dropping scraps of paper and measuring how far they moved as the shock wave passed by. Fermi had a quick approximation formula and the movement of the paper was its parameter. While his estimate had to be corroborated by more rigorous methods, the use of these so-called “back-of-the-envelope” calculations can often lead to startlingly useful results.

The famous Princeton statistician John Tukey was said by his colleagues to be constantly engrossed in solving problems. One day, a story goes, someone interrupted him to ask for the answer to a calculation. Without looking up, Tukey grumbled, “Would 10 digits of accuracy be sufficient?” The interrupter said, “Of course, that is more than enough!” Tukey reached the bottom drawer of his desk and pulled out a sheaf of



computer printout. He said, “This is a printout of all 10-digit numbers. Your answer is in here. Look it up.”

Since I knew and admired John Tukey, I believed the part about his irascible reaction to interruptions. But I did not believe the part about the printout in his bottom drawer. So I did a back of envelope calculation to figure out how much paper would be needed to print all the 10-digit numbers on the one-sided printouts of the day. I estimated that they could be printed four columns to a page, 60 lines to a page, for approximately 250 numbers per page. Then I took the number of 10-digit numbers, 10^{10} , and divided by 250, to conclude the printout would

contain 4×10^7 pages. Two reams of paper are $1000 = 10^3$ pages, so this is 4×10^4 reams. Clearly, 40,000 reams was much more paper than could possibly fit into the bottom drawer of his desk, much less his office, or even his statistics building. The story is a delightful yarn about Tukey’s personality rather than the truth.

This example illustrates a common practice: we perform quick calculations to convert a number representing one thing to another. For instance, how many seconds are in a year? You would perform the following calculation (with the help of your envelope):

$$60 \frac{\text{sec}}{\text{min}} \times 60 \frac{\text{min}}{\text{hour}} \times 24 \frac{\text{hour}}{\text{day}} \times 365 \frac{\text{day}}{\text{year}} = 3.15 \times 10^7 \frac{\text{sec}}{\text{year}}$$



Association for
Computing Machinery

ACM Transactions on Evolutionary Learning and Optimization (TELO)

ACM Transactions on Evolutionary Learning and Optimization (TELO) publishes high-quality, original papers in all areas of evolutionary computation and related areas such as population-based methods, Bayesian optimization, or swarm intelligence. We welcome papers that make solid contributions to theory, method and applications. Relevant domains include continuous, combinatorial or multi-objective optimization.



For further information
and to submit your
manuscript,
visit telo.acm.org

Notice a few things about these calculations:

- ▶ They are a chain.
- ▶ Each step converts one measurable quantity into another.
- ▶ The units of the ratio of conversation are shown at each step.
- ▶ The successive steps cancel the previous units of measurement, yielding the final unit at the end.

Keeping track of the units is sometimes called “dimensional analysis” and gives us trust in the results of the calculation. In other words, the artful back-of-the-envelope calculation is also an exercise in dimensional analysis.

With the aid of additional examples, I would like to show you how Little’s Law, a famous formula from statistics, can arise from dimensional analysis. You do not need to know any statistics to follow what is going on and see how incredibly useful it is.

Little’s Law

A certain Monsieur Tillet, retired algebra teacher, opens a gourmet restaurant. He maintains a wine cellar of Bordeaux red wines. He wants to serve his wines at the ideal age of 10 years. He is open 365 days per year and on average sells 50 bottles of wine each day. How big must his cellar be? The answer seems pretty obvious: over the 10-year period an average of $50 \times 365 \times 10 = 18,250$ bottles are extracted from his cellar and are all replaced. Therefore, the cellar must hold 18,250 bottles. (At 12 bottles per case, that is 1,521 cases, a fairly large cellar.)

For future reference he represents his calculation with the formula:

$$N = HX$$

Where N = average number of bottles, H = holding time, and X output rate. He verifies his formula by checking the dimensions. X is bottles/year, H is years, and therefore N has dimension “bottles.” This formula is a law—an invariant relationship between the three measured quantities.

As you can see, Monsieur Tillet found this law by a dimensional analysis of his wine cellar. He just multiplied the averages he could measure and canceled out their units so that the final result had the proper units. Of course, this formula does not account

for variations. On some days, Tillet sells more than 50 bottles, other days less. There is no guarantee that he will have enough 10-year-old wine available on any given day.

Tillet’s formula is the famous Little’s Law.⁴ In 1961, operations researcher J.D.C. Little demonstrated that, over a long period of time, the average number of items in a queue is the product of the average waiting time and the throughput. Although the proof was difficult, the formula became popular and central in queueing theory. Tillet’s Law is not strictly the same as Little’s. Despite having the same form, they are different because they rest on different assumptions. Tillet’s Law deals with directly measured quantities, Little’s Law with steady-state statistical quantities.

Tillet’s form is called “operational Little’s Law” because it is so simple that it is almost obvious when you are working with real data.^{2,3} No complicated math is needed to understand. A back-of-the-envelope calculation does the job.

Response Time Law

Further developing the business plan for his restaurant, Tillet was dismayed to discover that loyal customers were complaining about long delays to get a reservation. He had a mailing list of 1,700 customers who returned to make a reservation an average of 30 days after their last meal. He wondered if he could calculate the time it takes for his restaurant to serve the next loyal customer after they called in for a reservation.

He realized that he could use his wine cellar formula to answer the question. He assumed that each customer consumed one bottle of wine, so bottles per day output became customers

**The artful
back-of-the-envelope
calculation is also
an exercise in
dimensional analysis.**

per day served. The holding time became the average time for a customer to cycle between average time waiting for a service, R , and average return time 30 days. Filling in Little's Law,

$$1700 \text{ customers} = (R + 30) \text{ days} \times 50 \text{ customers/day}$$

Solving gives $R = 5$ days. Having confirmed there was a problem, Tillet initiated changes in his restaurant to keep these customers coming back.

Like Little's Law, this formula appears in queueing theory for the average response time of a service system that has an average of N users, average away time Z seconds, and throughput X users/sec:

$$R = \frac{N}{X} - Z$$

The away time is the time a customer spends outside the server part of the system before returning. As you can see, this is a simple rearrangement of Little's Law $N=(R+Z)X$.

Traffic Analysis

Tillet heard complaints from customers that there was a chronic traffic jam on the highway leading to his restaurant. The highway had a merge point where the road narrowed from two lanes to one. Customers approaching at 60 mph on the two-lane section suddenly had to slow and join a creeping jam 0.2 mi long before the merge, and then after the merge for another 0.1 mi before the speed picked up again to 60 mph—a total of 0.5 mi of jammed cars. Tillet wanted to calculate how long it takes to cross the jam.

Once again he decided to use Little's Law, where now H is the average time to cross the jam, X is the throughput, and N is the average number of cars inside the jam. But first, he needed to calculate X and N .

From the drivers manual and his own driving experience Tillet knew that most people adjust their speed to maintain a three-second distance from the car ahead. This means an observer on the side of the road would see the next front bumper every $3+L$ seconds, where L is the time for the car to move one car length. At speed r mi/min and standard car length 15 ft.,

$$L = \frac{1 \text{ min}}{r \text{ mi}} \times 60 \frac{\text{sec}}{\text{min}} \times \frac{1 \text{ mi}}{5280 \text{ ft}} \times 15 \frac{\text{ft}}{\text{car}} = \frac{0.17 \text{ sec}}{r \text{ car}}$$

No complicated math is needed to understand. A back-of-the-envelope calculation does the job.

At $r=1$ (60 mph), the time between cars is 3.17 sec and throughput is

$$X = \frac{1 \text{ car}}{3.17 \text{ sec}} \times 60 \frac{\text{sec}}{\text{min}} = 18.9 \frac{\text{car}}{\text{min}}$$

Because the speed is r before and after the jam, all Tillet needs now is N , the average number of cars caught inside the jam. This is easy. Inside the jam, each car occupies 20 ft. of space—15 ft. for its car length and 5 ft. as separation buffer. The total number of cars jammed into a mile would be $5,280/20 = 264$ and thus, because the actual jam occupies a total of 0.5 mi, the total in the jam is 132.

Now Tillet has what he needs for the calculation. Little's law says $H = N/X = 132/18.9 = 7$ mins. Although he was not happy with the jam, he told his customers that a seven-minute slowdown was well worth the fine food and wine they would get when they arrived.

Conclusion

Many back-of-the-envelope calculations involve dimensional analysis—build your own formula by multiplying or dividing known quantities such that the units of the final answer are correct. Dimensional analysis provides the scaffolding for detecting errors in the conversion formula. You need not remember the exact formula because you can easily construct it from dimensional analysis. There is, however, a caveat. Dimensional analysis is a quick check for the reasonableness of a calculation, but need not always yield the correct formula. For example, in physics the formula for distance traveled by an object in time t starting at rest under a constant acceleration a is $d = \frac{1}{2} at^2$. Dimensional analysis will confirm the proper units of d but won't generate the factor $\frac{1}{2}$.

Nonetheless, dimensional analysis will often permit you to create a formula that relates the quantities you know or can easily measure to the result you want. This column has demonstrated that Little's Law is really quite intuitive. If you know any two of its elements N , H , and X , you can easily calculate the third. I also showed a simple law for calculating the response time of a system that has a fixed number of recycling customers. The response time law is really Little's Law rearranged.

I went a little farther afield by applying Little's Law to simple traffic jams. At reasonable (non-jam) speeds, the throughput depends mostly on the time-separation of cars, but not their speed. Most drivers try to maintain a constant time separation from the driver ahead, except of course when in a traffic jam. It is easy to calculate the number of cars in the jam because they are packed closed together and each car occupies slightly more than a car length of space. The average number in the jam divided by the throughput is the average time a car is held in the jam.

If this method of analysis intrigues you, you can learn more about it. It is called operational analysis.¹⁻³ Besides back-of-the-envelope calculations, it enables sophisticated calculations of throughputs and response times in networks of computers.

Clearly, there is more to back-of-the-envelope calculations than simply scribbling numbers on a scrap of paper. It is a useful professional skill and, as Fermi and Tukey showed, even an art form. ■

References

1. Buzen, J.P. and Denning, P.J. Rethinking randomness: An interview with Jeff Buzen. *ACM Ubiquity* (Aug. 2016), Article No. 1; doi/10.1145/2986329
2. Denning, P.J. and Martell, C. *Great Principles of Computing*. MIT Press, 2015.
3. Denning, P.J. and Buzen, J.P. Operational analysis of queueing network models. *ACM Computing Surveys* 10, 3 (Sept. 1978), 225–261.
4. Little, J.D.C. A proof for the queueing formula: $L = \lambda W$. *Operations Research* 9, 3 (1961), 383–387.

Peter J. Denning (pjd@nps.edu) is Distinguished Professor of Computer Science and Director of the Cebrowski Institute for information innovation at the Naval Postgraduate School in Monterey, CA, is Editor of *ACM Ubiquity*, and is a past president of ACM. The author's views expressed here are not necessarily those of his employer or the U.S. federal government.

My thanks to Jeff Buzen, Britta Hale, and Phil Yaffe for their comments on the drafts.

Copyright held by author.

Viewpoint

Testing Educational Digital Games

Diversifying usability studies utilizing rapid application development.

THE DIGITAL GAME industry is a multibillion-dollar global enterprise.¹⁴ Fostered by the development of sophisticated software and hardware as well as an interest in gaming among individuals around the world, the financial impact of digital games is continuously evolving. Accordingly, new streaming platforms are launching, and existing online game systems are expanding to meet the demand.¹⁴ Traditionally referred to as video games, the term digital games is a unifying term encompassing interactive games played on consoles, smartphones, tablets, personal computers, and other devices.² The Entertainment Software Association indicates 214.4 million Americans play digital games.³ National data also suggests 75% of American households have one person that plays digital games.³

In light of the exponential increase in the use of digital games, they have become commonplace in academic settings such as elementary schools, middle schools, high schools, and postsecondary institutions. According to an article published by the American Psychological Association, educational digital games complement a myriad of instructional contexts and have the potential to enhance traditional classroom environments.⁹ Despite needing more research to validate the educational benefits of digital games, the use of educational digital games has increased in recent years.



In a study comprising 488 teachers, more than 50% of the participants used digital games in the classroom.⁵

Benefits of Educational Digital Games

Increasing amounts of scholarship have focused on the extent to which digital games promote educational outcomes. For example, Vanbecelaere et al. exploring the impact of educational digital games on academic achievement, demonstrated that educational digital games could augment students' learning outcomes.¹²

Virvou, Katsionis, and Manos, comparing aspects of an educational digital game, indicated that virtual reality games could promote student achievement.¹³ Another study, which examined qualitative data from students, reinforced the importance of assessing the impact of digital learning experiences to refine gaming software.⁴ In the study, researchers analyzed students' experiences and interactions with an educational digital game designed to teach coding skills. The research suggests educational digital games should incorporate

authentic activities that are consistent with students' expectations about the learning context.

Almeida found that undergraduate students who played an educational digital game, in addition to reading information about the topic, scored higher on a learning assessment than students who did not play the educational digital game.¹ A summary of research about the effects of educational digital games indicates that when games are used to help students learn information, the gameplay should match classroom-based instruction and incorporate assessments of student learning.² Thus, it may be advantageous to align the content, structure, and goals of educational digital games with traditional pedagogical practices.² Moreover, future research investigating educational digital games should explore the extent to which students learn information aligned with the game's instructional model. Studies should also analyze game's impact on students' motivations to learn information related to the game's subject area.

Usability Study Methods

Usability studies incorporate data and theoretical concepts to support the development and testing of software and hardware. Moreover, software testers utilize analytical frameworks to evaluate software applications. To promote a seamless analysis of educational digital games, usability researchers examine aspects of games that influence the playability of games. Therefore, usability studies that focus on educational digital games should combine model-based assessments and human-centered evaluation techniques.

Designing Educational Digital Games

Engineering pedagogical software that students can use to enhance learning outcomes has been an ongoing challenge among researchers and software engineers. The inherent difficulty in designing software to facilitate educational outcomes is that learning styles are divergent among students. Also, to further confound educational software development issues, research suggests that aligning

learning styles to instructional content may not enhance achievement outcomes.¹⁰ Considering the complexity of designing effective educational technologies, it is possible that while some educational digital games may address an element of the course content, they may not provide comprehensive coverage that matches or enhances the course material. In contrast, a student may utilize educational digital games to obtain an overview of a subject, while another student may need educational software that delineates specific concepts about an aspect of the subject.

To design effective educational digital games that consider student diversity, we may need to design inclusive game usability studies. In this regard, an article by Jakob Nielsen gives credence to the notion that game usability studies may be more impactful when the findings from a diverse array of users comprise the player-centered data.⁸ To achieve this goal, extrapolating insights from previously mentioned ideas, usability studies should incorporate research participants from underrepresented populations. Furthermore, integrating this data and diverse sampling procedures in game evaluation processes may also enhance the game's economic impact.

Diversity in Usability Studies

The diversity within the digital game industry continues to be problematic, with approximately 65% of the respondents from the International Game Developers Association's 2019 Developer Satisfaction Survey reporting that equality is an issue in the video game industry.⁶ Moreover, based on 2020 data from the Bureau of Labor Statistics, African Americans constitute approximately 6.2% of software developers, while Hispanics or Latinos comprise 5.9% of software developers.¹¹ Additionally, national data indicates African Americans comprise 12% of software quality assurance analysts and testers, while Hispanics or Latinos account for 9.2% of software quality assurance analysts and testers.¹¹ Given these statistics, it is important to consider the demographic composition associated with usability studies, such as

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



Distinguished Speakers Program

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

Students and faculty can 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 covers the cost of transportation for the speaker to travel to your event.

speakers.acm.org



Association for
Computing Machinery

The inherent difficulty in designing software to facilitate educational outcomes is that learning styles are divergent among students.

the diversity among the individuals who design, test, and evaluate educational digital games.

Rapid Application Development

An effective software process facilitates the creation, testing, and optimization of software via a strategic, measurable, and transparent integration of engineering principles. A fundamental perspective in rapid application development (RAD) is that software development will utilize iterative development phases to minimize production time while pursuing a user-centered design philosophy.⁷ RAD typically incorporates processes that involve defining functional requirements, developing a prototype, evaluating the prototype, implementing improvement processes, modifying requirements, refining the prototype, adjusting software specifications to align with user-centered expectations, and producing the software.⁷ Given its flexibility, a RAD approach may promote equity in usability studies and enable more underrepresented groups to participate in the usability testing phase for educational digital games.

Utilizing a software development process, such as RAD, which promotes adaptations and modifications throughout the educational game development life cycle, may enhance students' learning outcomes. In addition to recruiting more individuals from underrepresented racial groups and women to enter the software engineering and usability testing work-

force, it may also be advantageous to ensure that more educational digital game usability studies incorporate participants from underrepresented groups. Implementing this approach would also involve establishing performance measures to monitor how educational digital games affect learning outcomes among a diverse population of students. This strategy could also integrate the development and utilization of qualitative indicators and quantitative metrics that monitor and manifest diverse students' perspectives in digital game usability studies. Finally, while this column has focused on RAD, it should be noted that additional software process models may also advance diversity within software testing environments. □

References

1. Almeida, L.C. The effect of an educational computer game for the achievement of factual and simple conceptual knowledge acquisition. *Education Research International* (2012), 1–5.
2. Clark, D.B., Tanner-Smith, E.E., and Killingsworth, S.S. Digital games, design, and learning: A systematic review and meta-analysis. *Review of Educational Research* 86 (2016), 79–122.
3. Entertainment Software Association. 2020 essential facts about the video game industry. (2020); <https://bit.ly/3kbSRgR>
4. Esper, S. et al. CodeSpells: How to design quests to teach Java concepts. *Journal of Computing Sciences in Colleges* 29, 4 (2014), 114–122.
5. Fishman, B. et al. Empowering educators: Supporting student progress in the classroom with digital games. University of Michigan, Ann Arbor (2014); <https://bit.ly/36vww1E>
6. International Game Developers Association. Developer Satisfaction Survey 2019: Summary report; <https://bit.ly/2UFdpnc>
7. Martin, J. *Rapid Application Development*. Macmillan, New York, 1991.
8. Nielsen, J. Games user research: What's different? (Mar. 20, 2016); <https://bit.ly/2UbwmOK>
9. Novotney, A. (2015, April). Gaming to learn. *Monitor on Psychology*, 46(4). <https://bit.ly/3wBwYU0>
10. Rogowsky, B.A., Calhoun, B.M., and Tallal, P. Matching learning style to instructional method: Effects on comprehension. *Journal of Educational Psychology* 107 (2015), 64–78.
11. U.S. Bureau of Labor Statistics. Employed persons by detailed occupation, sex, race, and Hispanic or Latino ethnicity. Washington, D.C.; <https://bit.ly/3xFmzyK>
12. Vanbecelaere, S. et al. The effects of two digital educational games on cognitive and non-cognitive math and reading outcomes. *Computers and Education* 143 (2020), 1–15.
13. Virvou, M., Katsionis, G., and Manos, K. Combining software games with education: Evaluation of its educational effectiveness. *Educational Technology & Society* 8, 2 (2005), 54–65.
14. Witkowski, W. Videogames are a bigger industry than movies and North American sports combined, thanks to the pandemic. *MarketWatch* (Dec. 22, 2020); <https://on.mktw.net/3hBsVJG>

Lamont A. Flowers (lflower@clemson.edu) is the Distinguished Professor of Educational Leadership in the Department of Educational and Organizational Leadership Development in the College of Education and the Executive Director of the Charles H. Houston Center for the Study of the Black Experience in Education in the Division of Inclusion and Equity at Clemson University, Clemson, SC, USA.

Copyright held by author.

Viewpoint

Whose Smartphone Is It?

Should two private companies have complete control over the world's cellphones?

ON MAY 27, 2020, in the French National Assembly, Cédric O, the French Secretary of State for Digital Economy, forcibly expressed his government's frustration with Apple and Google in terms more appropriate to a cold war confrontation between superpowers. He noted that France and the U.K. were the two European countries building COVID-19 contact-tracing apps without these tech giants' assistance. These countries were also the only two European countries with nuclear weapons, the "acme of national sovereignty."^a

The frustration of a modern state, unable to respond to the most severe public health crisis in a century because of two private companies' decisions, should give us all pause. Apple and Google have complete and unquestionable control over the computer in your pocket and are not shy about exercising it. It is time to do something about it.

Background

In the midst of the COVID-19 epidemic, Apple and Google jointly introduced the Exposure Notification framework^{4,6} to facilitate the construction of interoperable COVID-19 contact tracing applications for iOS and Android smartphones. This framework uses Bluetooth Low Energy (BLE) advertising beacons



to discover nearby smartphones running the contact-tracing app, determine the distance between the phones, and estimate potential COVID-19 exposure between phones' users.

The DP3T group at the Swiss Federal Technical Universities EPFL and ETH developed the privacy-preserving protocol used in this framework¹ and built one of the first apps. Along with everyone else, we needed Apple's cooperation to make these apps run satisfactorily on Apple iPhones, which intentionally do not expose the functionality necessary to send and receive BLE beacons from apps running in the background. Some countries, such as Singapore and the U.K., tried to work around this limitation. The resulting apps required a phone to remain unlocked and rapidly drained its battery. Not surprisingly, user acceptance was low.

Under public and private pressure, Apple and Google jointly proposed an Exposure Notification protocol enabling the construction of COVID-19 tracing apps. The companies selected a decen-

tralized, privacy-preserving protocol, similar to the one DP3T had developed and published.³ The DP3T team worked closely with the two companies on the implementation and on the SwissCOVID app, which was the first COVID-19 tracing app in widespread testing.

Other countries, such as Germany, France, and the U.K., and U.S. states, such as North Dakota and Wyoming, wanted to build their COVID-19 tracing apps along different lines. Some preferred a centralized approach, in which a server processes all exposures. Others wanted to collect additional information, for example, when and where a contact occurred.

Apple and Google refused to allow any variance in the design of a contact-tracing app. Their expose-notification API did not reveal the received exposure keys to an app, which would be necessary to implement a centralized solution. Also, the API's license terms prevented apps from collecting physical location information. Moreover, the companies decreed that each country or state would be allowed only one COVID-19 tracing app and that the national or state health authority must produce it.

These decisions can be justified as measures to protect user privacy. Still, in the end, the technical stranglehold of these two companies, rather than the merits of the arguments, carried the day. In the U.S. and most other countries, COVID-19 tracing apps—but not France's—use the Apple and Google framework. Appeals, pressure, and threats from sovereign governments did not carry sufficient weight to change Apple and Google's decision.

^a "To date, 22 countries have chosen to develop a contact protection solution based on the interface developed by Apple and Google—the 22 countries do not include France or the U.K., which, is it a coincidence, are also the only two European countries to have their own nuclear deterrent, which is ultimately the acme of national sovereignty"; <https://bit.ly/3B0gBKY>

Discussion

The key issue here is the unprecedented degree of control that both Apple and Google exercise over the software that runs on mobile phones, today's dominant computing platform. Half of the world's people own a smartphone, with a far higher percentage in developed countries such as the U.S., China, and Europe. For many, their phones are the primary computer they use to access information, play games, or communicate with other people. In 2019, in the U.S., an average person spent 51 minutes connected to the Internet from a desktop computer but over four times as much on a smartphone.⁷ Control of smartphones is control of people's interactions with the world.

Since the early days of Apple's iPhone, and subsequently, Google's Android phones, these two companies have exercised near-total control over the functionality of software written for and distributed on "their" smartphones. Turing's work in the 1930s showed the computers are universal computing devices, capable of executing any computable function. Apple and Google are using their control of the smartphone platforms to subvert this fundamental principle, with a foreseeable cost in innovation and competitiveness.

Apple and Google exercise control at two levels. Smartphones, from the beginning, never permitted apps to access the underlying physical devices or coprocessors in a phone but instead provided application programming interfaces (APIs) that tightly constrain how a phone can be used. Beyond this, Apple limits apps' distribution to its App Store, which imposed stringent rules and a strict gatekeeping process to control which apps are acceptable.^b Google allows alternative app stores, but its dominant Play Store follows a model similar to Apple. Not only are some apps difficult or impossible to build with the APIs, but even if a creative software developer finds a way to work around the limitations, they may find it difficult or impossible to

^b Brad Smith, President of Microsoft, commented that the App Store presents a higher barrier to competition than what Microsoft was accused in its antitrust prosecution 20 years ago. See "Microsoft Says Antitrust Bodies Need to Review Apple App Store" Bloomberg (June 18, 2020); <https://bloom.bg/3iaEPcu>

Apple and Google refused to allow any variance in the design of a contact-tracing app.

distribute their app to consumers.

Both companies argue that their practices and restrictions benefit smartphone users. The companies claim to have improved software security by taking on the challenging task of scrutinizing apps in their stores for malware. Moreover, the stores offered convenient, well-known places to find any app.

At the same time, control of both the computing platform and the distribution of applications give Apple and Google unprecedented control over what software can and will be written, and hence what you can do with your smartphone.

Their control became clear in the context of the COVID-19 proximity tracing apps developed last spring. It is, however, hardly the only such incident. At the same time, Apple engaged in a public battle with Basecamp about their HEY mail reading app to force them to route payments through Apple's App Purchase, where it could take a 30% commission.² Similarly, Apple rejected game apps from tech giants Microsoft and Facebook that violated its rule against "arcade" games in its App Store.⁵

The COVID-19 incident, however, should worry us all. COVID-19 apps were not the subject of a commercial dispute. Many experts agreed that these apps could help reduce the spread of an epidemic. Despite this, Apple and Google told all of the world's governments and public health agencies: we know more than you about how to control a pandemic, and we will not allow you to bring your expertise to bear, to collect different information, or even to experiment with alternative approaches.

Although I would be happy to argue that Apple and Google made a wise choice in implementing DP3T's privacy-preserving protocol, their monopolistic

and arbitrary control over which software can run on smartphones will deaden innovation. Consumers already have an impoverished selection of apps. Moreover, the tech giant's arbitrary power furthers the competitive advantage of China's vibrant smartphone ecosystem, which is flourishing and innovating beyond these American companies' control—however, under the heavy thumb of the Chinese government.

Technical innovation alone will not resolve this problem—though improvements in security and privacy engineering might help iOS and Android achieve the goal (providing a safe smartphone user experience) that is the rationale for Apple and Google's close control. At the same time, Apple and Google employ advanced security techniques such as cryptography and secure enclaves to control which software will run on their phones. To them, malware is any software that has not gained their stamp of approval. In the end, however, it is a political and legal question whether two companies, no matter what their intent, should have the power to decide which software runs on a person's smartphone.

Fortunately, antitrust regulators in Europe and the U.S. are starting to consider the consequences of allowing two private companies complete and unquestioned control over the world's smartphones. Hopefully, these inquiries will lead to the realization that concentrated control over the computing platform in the 21st century is as dangerous to innovation and commerce as were the railroad and oil monopolies of the 19th and 20th centuries. ■

References

1. Apple and Google update joint coronavirus tracing tech to improve user privacy and developer flexibility. TechCrunch (Apr. 24, 2020); <https://tcrn.ch/3r60AhK>
2. Apple vs. HEY; <https://hey.com/apple/>
3. Decentralized Privacy-Preserving Proximity Tracing. (May 25, 2020); <https://bit.ly/3r68WGI>
4. Exposure Notification API launches to support public health agencies. (May 20, 2020); <https://bit.ly/3BOUT9t>
5. Facebook. Microsoft gripes with Apple's App Store on EU's antitrust radar. Reuters (Aug. 10, 2020); <https://reut.rs/3ra85E1>
6. Privacy-Preserving Contact Tracing. (May 20, 2020); <https://www.apple.com/covid19/contacttracing>
7. Tech companies tried to help us spend less time on our phones. It didn't work. Vox Recode, (Jan. 6, 2020); <https://bit.ly/3wG92FL>

James R. Larus (larus@larusstone.org) is Professor and Dean of the School of Computer and Communication Sciences at EPFL, the Swiss Federal Institute of Technology, in Lausanne, Switzerland.

Copyright held by author.

Viewpoint

AI Ethics: A Call to Faculty

Integrating ethics into artificial intelligence education and development.

THIS PAST YEAR has seen a significant blossoming of discussions on the ethics of AI. In working groups and meetings spanning IEEE, ACM, U.N. and the World Economic Forum as well as a handful of governmental advisory committees, more intimate breakout sessions afford an opportunity to observe how we, as robotics and AI researchers, communicate our own relationship to ethics within a field teeming with possibilities of both benefit and harm. Unfortunately, many of these opportunities fail to realize authentic forward progress during discussions that repeat similar memes. Three common myths pervade such discussions, frequently stifling any synthesis: education is not needed; external regulation is undesirable; and technological optimism provides justifiable hope.

Education

The underlying good news is that discourse and curricular experimentation are now occurring at scales that were unmatched in the recent past. World Economic Forum working groups, under the leadership of Kay Firth-Butterfield, have convened a series of expert-driven policy productions are topics including, for instance, the ethical use of chatbots in the medical field and in the financial sector. The IEEE Global Initiative on the Ethics of Autonomous and Intelligent Systems, led by John Havens, continues to make progress on in-



ternational standards regarding the ethical application of robotics and AI. These are just two of dozens of ongoing international efforts. Curricular experiments have also garnered successful publication, from single-course pilots² to whole-curricular interventions across required course sequences.³

Yet despite international policy discourse and published curricular successes, the vast majority of faculty in robotics and AI report, in private discussion, that they do not feel empowered or prepared to integrate ethics into their course materials. The substance of this hesitation rests on the notion that ‘teaching AI ethics’ is like teaching ethics itself—lecturing on utilitarian and Kantian frameworks,

for instance, which is best taught by ethics scholars. But AI ethics is not the science of ethics, but rather shorthand for the notion of applying ethical considerations to issues surfaced by AI technologies: surveillance, information ownership, privacy, emotional manipulation, agency, autonomous military operations, and so forth. As for integrating such reflection into an AI class, every case I am aware of does so, not with *sage on a stage* lecturing by the faculty member regarding Kant, but with case studies and small-group discussions on complex issues, lifting the students’ eyes up from the technology to considering its possible social ramifications. No teacher can set the stage for such discussions better than an AI



Association for
Computing Machinery

2018 JOURNAL IMPACT
FACTOR: 6.131

ACM Computing Surveys (CSUR)

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



For further information
and to submit your
manuscript,
visit csur.acm.org

expert, who can speak concretely about face recognition errors, and how such mistakes can be inequitably distributed across marginalized populations.

In a five-year experiment, I have collaborated with a professor in the College of Social Sciences at Carnegie Mellon to design and deploy AI and Humanity as a freshman course that encourages technologies and humanities students alike to develop a grammar for considering and communicating about the interplay between AI and robotics technologies and power relationships in society. We build a new grammar on the backbone of keywords, thanks to McCabe and Yanacek's outstanding analysis of critical themes, including *surveillance*, *network*, *equality*, *humanity*, *technology*.^{6,7} College freshman have shown an apt ability to conduct critical inquiry, evaluate the ethical ramifications of technology and even construct futuring visions that interrogate our possible trajectories as a society (see <http://aiandhumanity.org>).

Equally important is the question of how ethics can be integrated into extant, technical coursework throughout a department. This year, in collaboration with Victoria Dean at Carnegie Mellon, we revisit the question of curricular integration by deploying a graduate class with the capstone experience of students engaging with faculty in the Robotics Institute, studying each course's syllabus, and designing a complete ethics module for integration into each class. We believe this direct-intervention model, with case studies, futuring exercises and keywords at its heart, has the potential to affirmatively engage numerous courses and professors across our department with a low-barrier pathway to in-class ethics conversations. As Barbara Grosz and others have said, the ethics conversation should not be a one-time course, nor a one-time seminar. Thinking on societal consequences should happen regularly, so it becomes an enduring aspect of the design thinking around new AI and robotics technology research and development.

Regulation

Another common argument stems from a strong *anti-regulation* stance that embraces corporations as agents with the very best intentions. AI researchers in my workshops frequently point to

ethics review programs implemented by top corporations to show that, being global hubs of innovation, these companies have already invented the best ways to self-regulate, eliminating the need for oversight. But the existence of a few corporate ethics programs does not upend the most basic observation of all: corporate technologic titans have incentives and reward structures that are *not* directly aligned with public good, equity, and justice. The misalignment of public-private values is a perpetual temptation for corporations to veer off the ethical course to privilege private interests over public concerns.

Examples abound. Google created an ethics board, and included the president of the Heritage Foundation. When employees noted the inclusion of an individual dedicated to denying climate change and fighting LGBTQ rights, Google dissolved the entire ethics board after one week.¹⁰ In 2019, news organizations also reported that Amazon's Alexa and Google's Assistant both record audio, unbeknownst to home occupants, and that employees listen to home interactions that any reasonable user would presume to be private.⁸ When the story first took hold, Apple touted its stronger privacy positioning, boasting that, in contrast, no Apple employees ever listen to Siri.

That was the end of the news cycle, until former Apple contractor Thomas le Bonniec became a whistleblower and described a vast program in which 200 contractors in County Cork, Ireland, were listening to Siri recordings that were very private.⁴ Apple was strictly right, *employees* were not listening in, contractors were. The malintent of this fib is clear; but the larger lesson is key: corporations are beholden to their shareholders and to their own set of values and motives. We cannot expect their self-regulation to serve any purpose beyond their own value hierarchy.

We live in a world replete with examples of misaligned values that facilitate unjust outcomes; regulation of corporate technology innovation by corporations constructs a value misalignment between corporate mission and public good. As AI researchers, we derive legitimacy through our reasoned opinions regarding the arc of future technology innovation, including the use of guard rails that protect the public good. We

can best serve both corporations and the public, not by arguing that regulation stifles innovation—we are all keenly aware that poorly designed regulation does that. Rather we can innovate by helping facilitate the creation of well-designed regulation, together with policymakers and industry, that encourages the *most just* AI futures and memorializes corporate transparency for the public.

Technological Optimism

It is one of the greatest ironies of these AI workshops when researchers argue that they do not feel equipped to opine on the ethics of AI in their classes at university, yet in the same breath announcing that their AI systems will be ethical because they will design autonomous technologies to have built-in ethical governors. This disconnect arises out of a natural bias we have as innovators: we have spent entire careers practicing how to be technology-optimistic—how to imagine a future with inspiring, new inventions that we can create. This is the attitude we need as salespeople, to convince funders to make bets on our future work; and yet this optimism does a disservice when we use it *within* our institution to imagine that shortcomings in present-day AI systems will be resolved simply through innovation.

In the 1990s, the AI field was far removed from social impact because it was as impractical as theoretical mathematics. Exciting progress, at the very best, resulted in publication. That world is ancient history now. To say that AI, today, is a technical discipline is entirely naïve: it is a social, world-wide experiment. Our tools have teeth that cut into the everyday lives of all, and this leaves a collection of engineers and scientists in the awkward position of having far more impact on the future than is their due.

In earlier times, our computational peers forged Computer Professionals for Social Responsibility (CPSR), largely in response to the threat of thermo-nuclear destruction and other existential threats arising from the Strategic Defense Initiative. Because nuclear destruction was palpable, the arc from technology to personal responsibility was short and well-founded. But today our AI technology is not as obviously threatening. When misused, AI's reinforcement of bias and power configu-

The AI research community cannot sit this out.

rations in society can be insidious and sub-lethal, like petrochemical industry toxins that hurt entire communities, not as quickly as bullets, but across vastly greater scope and timescales. And unintended side effects are not limited in potential scope; when AI-led political micro-marketing directs the outcome of an election, ensuring undemocratic policy decisions *can* have existential impact on the population.

Yet publicly consumed literature ranges dramatically on the issue of technology optimism and technology realism. The singularity, espoused by Kurzweil, suggests a postmodern evolutionary pathway for a new humanity⁴ or a pathway to greater equity through low-cost robotic production.⁹ At the same time, counter-narratives explain the role of ritual surveillance in the very creation of the Internet⁵ as well as the ethical ramifications of war-fighting robots.¹ We, as public outreach specialists need to reference the existing literature on *both* sides and add to the body of counter-narratives, creating depth and sharp focus along each critical issue where society and AI technology meet, from surveillance and information ownership to authenticity and democracy.

If you are not concerned about the effects of fielded AI systems on democracy, on stakeholder capitalism, on power and bias in society, then you are operating on an unfounded level of optimism that goes against your own scientific nature.

Conclusion

The AI research community cannot sit this out. We are a critical expert group with sufficient know-how to separate authentic issues from hyperbole, to distinguish plans of action that can actually make a difference from hot air. If we do not become part of the solution, we will lose our legitimacy as well-intentioned visionaries.

Education for all stakeholders is imperative for awareness. AI is the very definition of a boundary technology that is sufficiently alien that *everyone* needs scaffolding to make informed decisions; and we cannot pass off the duty of care to create broad educational interventions to anyone else. Rule-making and regulation is equally essential. Nothing about historical corporate and governmental behavior can rationalize a *laissez-faire* approach when the consequences of inaction are so clearly inequitable. Finally, the hyperbole of techno-optimism needs to end. The public invests our opinions with significant credence, and when we state that our algorithms will be ethical innately, they actually imagine autonomous systems with human meta-cognition. There is no room for us to promulgate such a gap between computational reality and blue-sky wishes, particularly when AI is already so consequential to our lived experience. Let's embrace strong education, clear-headed regulation, and let's tone down the hyperbole of technological optimism. □

References

1. Chamayou, G. *A Theory of the Drone*. New Press, 2015.
2. Furey, H. and Martin, F. Introducing ethical thinking about autonomous vehicles into an AI course. *Thirty-Second AAAI Conference on Artificial Intelligence*. 2018.
3. Grosz, B.J. et al. Embedded EthICS: Integrating ethics across CS education. *Commun. ACM* 62, 8 (Aug. 2019), 54–61.
4. Hern, A. Apple Whistleblower goes Public over 'lack of action'. *The Guardian*, (May 20, 2020).
5. Kurzweil, R. *The Singularity Is Near: When Humans Transcend Biology*. Penguin, 2005.
6. Levine, Y. *Surveillance Valley: The Secret Military History of the Internet*. Public Affairs: Illustrated edition (Feb. 6, 2018).
7. MacCabe, C. and Yanacek, H. Eds. *Keywords for Today: A 21st Century Vocabulary*. Oxford University Press, 2018.
8. Nourbakhsh, I.R. and Keating, J. *AI and Humanity*. MIT Press, 2020.
9. O'Flaherty, K. Amazon staff are listening to Alexa conversations—Here's what to do. *Forbes*, (Apr. 12, 2019).
10. Rifkin, J. *The Zero Marginal Cost Society: The Internet of Things, the Collaborative Commons, and the Eclipse of Capitalism*. St. Martin's Press, 2014.
11. Wakefield, J. Google's ethics board shut down. *BBC News* (Apr. 5, 2019).

Illah Reza Nourbakhsh (illah@cs.cmu.edu) is Executive Director, Center for Shared Prosperity Director, CREATE Lab K&L Gates Professor of Ethics and Computational Technologies Carnegie Mellon University, Pittsburgh, PA, USA.

Copyright held by author.



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

Article development led by [acmqueue](https://queue.acm.org)
queue.acm.org

Is your organization prepared?

BY ATEFEH MASHATAN AND DOUGLAS HEINTZMAN

The Complex Path to Quantum Resistance

THERE IS A new technology on the horizon that will forever change the information security and privacy industry landscape. Quantum computing, together with quantum communication, will have many beneficial applications but will also be capable of breaking many of today's most popular cryptographic techniques that help ensure data protection—in particular, confidentiality and integrity of sensitive information. These techniques are ubiquitously embedded in today's digital fabric and implemented by many industries such as finance, health care, utilities, and the broader information communication technology (ICT) community. It is therefore imperative for ICT executives to prepare for the transition from quantum-vulnerable to quantum-resistant technologies.

This transition will be particularly complex, time-consuming, and expensive for larger organizations

with vendor dependencies and/or legacy infrastructure. Hence, it is critical that ICT leaders spend adequate time—now, while they have the luxury to do so—on planning the transition and determining their next steps. Otherwise, they may find their organizations in a chaotic state, scrambling to meet a compliance deadline or to prevent an actual loss of confidentiality or integrity of their, their customer's, or their partner's sensitive information. The absence of a well-thought-out plan could result in further delays and security vulnerabilities. Ultimately, it could have drastic implications for their core businesses and bottom lines.

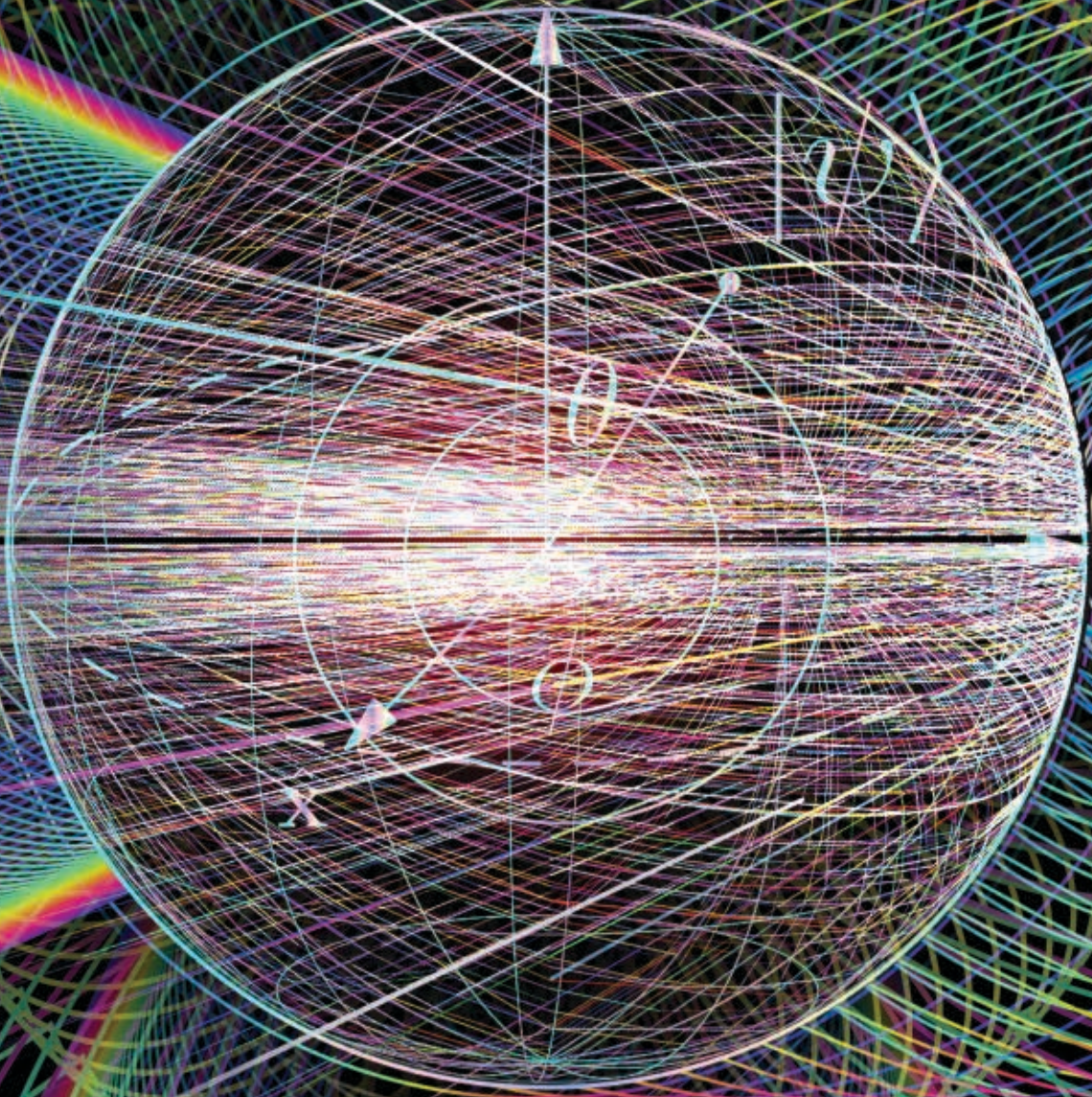
The good news is that security systems not susceptible to quantum attacks (that is, those that are *quantum-resistant*), can be implemented using today's classical computers. Organizations will not need quantum computers to resist attacks by another party's quantum computer. Several algorithms that are mathematically shown to be quantum-resistant already exist.

Standardization bodies such as the National Institute of Standards and Technology (NIST) in the U.S. and the European Telecommunications Standards Institute (ETSI) in Europe have been working on the standardization of quantum-resistant primitives since 2015,^{12,23} promising a final set of alternatives by 2025. In many cases, however, these algorithms may not be compatible with current hardware or software. For example, when existing algorithms are hardcoded in a piece of hardware, replacing the quantum-vulnerable algorithm with a quantum-resistant alternative involves swapping out the hardware.

Given the serious nature of the threat, the question organizations should be asking is how can the process of transitioning to quantum-resistant systems be accomplished in a timely and cost-effective manner, even as the solutions have yet to be standardized?

The industry's challenge is in migrating to compatible hardware platforms and ensuring the software run-

$$\hat{z} = |0\rangle$$



$$-\hat{z} = |1\rangle$$


ning on those platforms is upgraded to use quantum-resistant protocols. Depending on the needs of an organization and its approach to cryptography management, modifications to digital information security systems can range from relatively straightforward, quick, and inexpensive; to massively complex, drawn-out, and costly. The transition to a quantum-resistant state is no exception.

Competing quantum-resistant proposals are currently going through academic due diligence and scrutiny by industry leaders. Until the newly minted quantum-resistant standards are finalized, ICT leaders should do their best to plan for a smooth transition. This article provides a series of recommendations for these decision-makers, including what they need to know and do today. It will help them in devising an effective quantum transition plan with a holistic lens that considers the affected assets in *people*, *process*, and *technology*. To do so, the decision-makers first must comprehend the nature of quantum computing in order to grasp the impact of the impending quantum threat and appreciate its magnitude.


Quantum Computing

A quantum computer uses qubits (quantum bits), as opposed to classical bits, to process information. Qubits are two-state quantum systems. While a classical bit can be either a one or a zero, a qubit can be in any *quantum superposition* of zero and one. Measuring the state of a qubit causes it to collapse into one of the two states. With classical computing, four bits can have 2^4 (16) possible states but can be in only one state at a time. Superposition allows quantum computing to process all 2^4 possible states with four qubits at the same time.

Indeed, superposition paves the way for massive parallelization when searching for an answer to an equation. As the number of qubits scale, so do the number of states, but with an exponential rate. For example, with 30 qubits, you can represent and process more than a billion values at once. The use of qubits in a superposition state allows quantum computers to solve some problems significantly faster than classical computers. This speed-



As quantum computers are developed at a rapid pace, and with early models already on the market, the overall perception of the ICT community toward a quantum reality is slowly changing.



up is the reason why quantum computing is a threat to information systems' security and privacy.

Threat to Cybersecurity

Quantum computing's main potential threat to information security is in cryptography. Cryptography runs behind the scenes, out of the user's view, to keep information and communications secure. Two broad types of cryptography exist: symmetric/secret key and asymmetric/public key. Understanding the difference between the two is critical, as quantum computing impacts each differently.

Let's look at an example that illustrates the impact of a scalable quantum computer on the security of sensitive data. Many of the world's data-security practices rely on the RSA (Rivest-Shamir-Adleman) cryptosystem, which uses a product of two large prime numbers and assumes it is difficult for an adversary to factor the resulting product to find the initial prime numbers. This is known as an integer factorization problem (IFP), the intractability of which is a cornerstone in ensuring online security.

There is a good reason for this: If you choose the numbers carefully, factoring the resulting large product is indeed very difficult. A classical 2.2GHz Opteron CPU (a standard benchmark) would take about 10^{145} years to factor a 1,024-bit number, or about 7.25×10^{135} times the age of the universe. If a quantum computer is developed that can execute 100 million instructions per second (not unheard of in a desktop computer today), it could factor that number in a matter of seconds.²⁶ The main alternative cryptographic system—elliptic curve cryptography (ECC)—is not spared either. Although it is based on a different mathematical problem—namely, the discrete logarithm problem (DLP)—it too can be efficiently broken by a scalable quantum computer. At this point, the question isn't about whether quantum computers break today's encryption standards but *when* will they reach the performance level to do so.

This presents a significant challenge because the standardized RSA and ECC cryptosystems serve as the foundation of many of the cybersecurity tools and techniques that protect the world's

economy. The advent of the age of quantum computers will be massively disruptive. It is safe to assume that in the next decade or two, malicious actors (individuals or organizations) will be able to circumvent today's commonly used and trusted means of securing confidential information.

Organizations should be asking themselves now to what extent this will impact them and what they need to do to mitigate the risk.

It is not unreasonable for organizations to wonder why they should worry about the quantum threat if it is many years away. This is partially warranted. It is true that attackers cannot truly attack today's cybersecurity ecosystem until they have a scalable quantum computer, but their awareness of the inevitable availability of scalable quantum computers in the future may incentivize data harvesting breach attacks in the present. Furthermore, even if the threat is not imminent, major security changes, especially those that involve asymmetric cryptography, will take time to implement.⁵ Careful analysis, planning, and action need to be performed immediately.

When planning a response strategy, security professionals need to be concerned about two forms of attacks: *real time* and *harvest-then-decrypt*.²⁹

Real-time attacks occur when a quantum computer is in the hands of an adversary. As mentioned, asymmetric cryptography relying on IFP or DLP is catastrophically vulnerable to a scalable quantum computer attack. Symmetric cryptography is vulnerable, at least with its current key sizes, for slightly different reasons. Organizations or individuals whose communications, transactions, and authentications are still using current asymmetric or symmetric algorithms could be attacked when a scalable quantum computer is realized.¹² These real-time attacks are not currently possible, because a scaled-quantum computer is not yet available.

The harvest-then-decrypt attack happens when an adversary captures and stores encrypted data and sits on it until a quantum computer becomes available to provide a means for decryption.²⁹ Depending on the sensitivity or the shelf life of the data, this type of attack can be a serious current

threat. Malicious actors could harvest encrypted data today, put it aside for a few years, and wait for the availability of an affordable quantum computer so they can decrypt that data. Considering the many well-publicized large-scale security breaches of companies such as Yahoo in 2013 and 2014, Marriott Starwood Hotels in 2018, and Capital One in 2019, this threat is very real. (Note that in 2019 alone, four billion records were breached.²⁷) In the meantime, a constant game of cat and mouse is being played out between the attackers who seek to cause harm and the security professionals who are tasked with stopping them.

Impact on Symmetric Cryptography

Symmetric-key cryptography uses a secret key shared between two users. Party A can encrypt the data using the secret key and send the result to Party B, who uses the same key to decrypt and read the data. The secure exchange of the secret key between users, also known as key management, forms the security basis for symmetric cryptography. The vulnerability of this system is twofold: the need to transmit the key introduces the possibility that the key can be intercepted in transmission, and that quantum computers can use Grover's algorithm¹⁶ to improve the efficiency of a brute-force attack.

A secret key is generated using a source of randomness and is of a predetermined size. Thus, because of the creation process, and since there is only one key, there is no mathematical relationship to crack. The only two ways to attack symmetric algorithms are cryptanalysis and brute force. A brute-force attack involves trying every possible key to decrypt the ciphertext and obtain the plaintext. On average, an attacker would have to try half of all possible keys to obtain the correct one. Therefore, a secret key with enough entropy and length can sufficiently protect encrypted information. Grover's algorithm, however, can make use of superposition of qubits to speed up the brute-force attack by roughly a quadratic factor (that is, proportional to the square of the speed in which classical computing can make a brute-force attack on the keyspace^{6,22}). This reduces the strength of symmetric algorithms by approximately 50%. For example,

256-bit Advanced Encryption Standard (AES 256) would be able to provide only 128-bit security.²⁹

Fortunately, doubling the symmetric key sizes, when the algorithm specification can accommodate it, allows this form of cryptography to remain safe.¹² Doubling the key size is not a trivial task, however. It is reasonably straightforward when the cryptography is implemented in software because an update may allow for an efficient key-size change. But in situations where the cryptography is implemented in hardware, changing the size is more challenging and expensive. For example, some types of routers and all hardware security modules (HSMs) will need to be replaced with hardware capable of accommodating larger key sizes. Depending on the size of the organization and the extent of its symmetric cryptography use, this could be an extremely time-consuming and costly undertaking. Regardless, it will be a massive industry-wide transition, especially from a change management perspective.

Impact on Asymmetric Cryptography

Asymmetric cryptography uses two keys: public (anyone can see it) and private (only authorized people can see it). The two keys are mathematically bound, which forms the basis of asymmetric cryptography's security. One of several computationally difficult mathematical problems can be used to bind the two keys and act as the basis for security. IFP and DLP are two of the more commonly used problems. Integer factorization derives its security from the difficulty of factoring the product of two large prime numbers. Discrete logarithms involve finding an unknown integer K , from $g=b^K$, where g and b are known elements with certain mathematical attributes.

Other mathematical problems have been proposed as ways to bind the public and private pair, resulting in a variety of different asymmetric cryptosystems. Among all such cryptosystems, RSA (based on IFP) and ECC (based on DLP) have been standardized and are widely used. Indeed, they strike a nice balance between simplicity, efficiency, and security—that is, until now. When the underlying computationally difficult problems can be efficiently solved by a

Planning for Quantum Readiness

Determine transition path. Large and complex organizations tend to rely on established legislation and industry standards to inform their decisions in adopting, maintaining, and sunseting technology. These decisions are typically part of a 5- to 10-year planning and capital expense budget cycle.¹² At the same time, there is a good chance that commercial-scale quantum technology will become part of mainstream computing during the next 5–10 years.³⁴ As such, some of these organizations are already late in starting to plan for the transition to “quantum readiness,” and will be left scrambling to protect what data they can before time runs out.²² Having allocated appropriate resources, performed a risk assessment, and determined which systems may be at risk of quantum attacks in the future, organizations reach a critical decision point as follows:

- ▶ **Wait for standardization.** In some cases, organizations may choose not to act until standardization bodies announce formal recommendations for quantum-resistant security, estimated to be available from NIST by 2024. This could be an appropriate business strategy decision for ICT managers when their organization’s confidential information is perceived to be of very low value to malicious attackers, has a very short shelf life, or is transient in nature. It may also be reasonable when they rely primarily or exclusively on external vendors for security and are confident those vendors will transition to quantum-resistant security quickly.

- ▶ **Invest in crypto-agility.** Crypto-agility is the ease with which an organization is able to implement cryptographic changes. Where significant risk is present, and strategic mandate permits, organizations should invest in crypto-agility. This step is not unique to the threat of quantum attacks, but it is a prerequisite to reacting effectively once standards have been finalized.

- ▶ **Establish and maintain a quantum-resistance roadmap.** Before the certification of quantum-resistance standards, the projected timelines for an attack-capable quantum computer will be refined, as will the options for mitigating the risk. Organizations should establish a roadmap that tracks these developments as they pertain to their own specific context, and they should maintain it for the coming years until quantum-resistant security has been standardized, adopted, and implemented.

- ▶ **Implement hybrid cryptography.** For organizations with high-risk, sufficient resources, and end-to-end control over their cryptographic ecosystems, overlaying a quantum-resistant security layer on top of existing pre-quantum security can be advantageous. It retains the standardized and mandated level of mitigation against attacks by today’s classical computers, while mitigating the risk of harvest-then-decrypt attacks. Despite the additional cost, some organizations will have sufficient incentive and resources to pursue a hybrid strategy.

Remediation projects. Regardless of which alternative is chosen, once the standards are published, organizations should move quickly to implement them in accordance with the ensuing compliance mandates. Organizations that adopt the wait-and-see approach (Scenario A) will need to research their path to quantum resistance based on their respective position at that time. Organizations that maintain a roadmap but do not implement a hybrid solution (Scenario B) will execute their roadmap. Those that implement a hybrid solution (Scenario C) will simply need to make (relatively minor) adjustments to what they already have in place. All organizations will also need to determine the deprecation path for their pre-quantum cryptographic implementations (for example, RSA and ECC), as will likely be mandated in the standardization body recommendations.

scalable quantum computer, an attacker can go back in time and decrypt already harvested encrypted data through the application of Shor’s algorithm.³⁰ This algorithm takes advantage of the fact that with enough qubits in superposition, a quantum computer can simultaneously examine countless combinations of zeros and ones in parallel and, as research has demonstrated, solve the computationally difficult mathematical problems that form the basis of the public-key algorithm’s security.

The number of combinations that can be explored simultaneously is dependent on the number of qubits

available to a quantum computer. With enough qubits, a quantum computer can quickly reverse calculate the computationally difficult problem and obtain the private key.¹² In other words, the private key can be recovered from the public key, and the information being secured can be decrypted.

Other algorithms that are similarly vulnerable to quantum-enabled attacks include the Digital Signature Algorithm, Diffie-Hellman, Elliptic-Curve Diffie-Hellman, and Elliptic-Curve Digital Signature Algorithm.¹² Together with RSA and ECC, these algorithms are staples of security for

the Internet, email, virtual private networks, and the Internet of things (IoT), making the quantum threat very serious and potentially broadly impactful. If the vulnerabilities of asymmetric cryptography were all to occur at the same time, they could lead to the deterioration of the security fabric that protects the digital society.

No matter how much the size of the initial parameters is increased, resulting in a larger key, the mathematical problems that asymmetric cryptosystems rely on can be solved in polynomial time on a scalable quantum computer. Hence, standardized and widely used asymmetric cryptographic systems will be severely impacted by a sufficiently capable quantum computer.

Understanding Terminology

Quantum information computing and cybersecurity mitigation against it are being researched heavily in academia and industry. With so much research and development, varying approaches, and the inherent complexity of the topic, different terminologies are used to describe different aspects of these related fields. As the language of the vendors is added on top of this, the various terminologies inevitably get muddled. So before proceeding, it is useful to clarify some of the often-used terminology.

Quantum cryptography leverages the properties of quantum mechanics, as opposed to mathematics, to carry out cryptographic objectives such as key distribution. Quantum key distribution (QKD) is a method of transmitting a secret key over distance. It allows two parties to produce a shared random secret key known only to them, which can then be used to encrypt and decrypt messages and cannot be intercepted without the parties noticing, nor can it be reproduced by a third party. It is an *information-theoretically secure* solution to the key-management problem, which means its security is not based on any computational hardness assumption. It is the use of a quantum technology, and the need for physical infrastructure capable of transmitting quantum states, that gives rise to the label of quantum cryptography (which is not the focus of this article).

Post-quantum cryptography, also known as quantum-resistant or quantum-safe cryptography, is a subset of

classical cryptography which can be deployed on existing classical devices and are currently believed to be safe against the threat of a scalable quantum computer.

Quantum computers are not good at efficiently solving every kind of mathematical problem. In fact, there are some encryption schemes that can be run on classical devices and is based on mathematical techniques that are *quantum-resistant*. The benefit of quantum-resistant cryptography is that it does not require a new physical infrastructure to deploy. The disadvantage is that it still relies on computational security (a hard-to-solve mathematical problem).

Cryptographers have been proposing such cryptosystems since as early as 1978 when the McEliece cryptosystem was developed at NASA's Jet Propulsion Laboratory.²⁰ The reason these quantum-resistant cryptosystems have not as yet been adopted is that standardized non-quantum-resistant cryptosystems were simpler, less resource intensive (and thus less expensive to implement), and "good enough." So there seemed to be no need for the quantum-resistant systems which, after all, only protected against a theoretical future threat. These systems were viewed more as theoretical contributions within the cryptographic community.

Moreover, such cryptosystems were not as closely examined as current widely used systems and they were not standardized—*yet*. As part of any standardization process, the algorithms typically go through years of due diligence and scrutiny conducted by academics and standardization bodies such as NIST and ETSI. Anticipating that quantum development was accelerating, NIST announced a competition in 2015 for proposals for standards candidates for its quantum-resistant algorithm transition. It plans to evaluate the proposals through 2021 or 2022 and then formalize those selected into draft standards by 2024.²³

Quantum Development

Many organizations are working toward developing scalable "universal gate" quantum computers (or simply universal quantum computers), including the Institute for Quantum Computing (IQC) at the University of Waterloo,

QuTech at the University of Delft, the Yale Quantum Institute at Yale University, the Centre for Quantum Technologies in Singapore, and the Joint Quantum Institute in Maryland. Well-known companies such as Microsoft, Intel, IBM, and Google are in the race for quantum supremacy as well.⁸ There are some companies, such as D-Wave,¹¹ which build systems with thousands of qubits but use quantum annealing as opposed to universal gate architectures. These systems are very good at finding "good enough" or "local minima" solutions but not specific solutions. Quantum annealing computers cannot efficiently run Shor's algorithm and thus do not represent the same threat to cryptography that the universal quantum gate computers do. Predictions about when one of these organizations will produce a quantum computer capable of breaking much of the world's encryption vary greatly, but most estimates range between 6 and 11 years with a non-negligible success probability.²²

An important measure of quantum computing development is its ability to execute Shor's algorithm or perform a Grover's search. This depends on the number of qubits, among other factors. The current estimates of the number of qubits in a universal quantum computer needed to break RSA 2048 (asymmetric cryptography) and AES 256 (symmetric cryptography) are 4,098 and 6,681, respectively.^{15,28} This number depends on several factors, such as the efficiency of fault-tolerant error-correcting codes, physical error models, error degrees of the physical quantum computer, optimizations in quantum factoring, and the efficiency of factoring algorithms into quantum gates.²² The number of qubits that pre-

dominant companies have been able to use is summarized in Figure 1.

At first glance, it may seem that a scalable quantum computer is still a long way off and, if history is a guide, current estimates of when a quantum computer at scale will be available may be overly optimistic (or pessimistic from a cryptography point of view). In 2009, Cisco predicted that the first commercial quantum computer would be available by mid-2020.¹³ In 2016, other experts predicted a quantum computer within a decade.³ Over the years, much has been learned and significant improvements have been made, which may make more recent estimations likely more realistic. Mosca estimates a 20% chance of a universal quantum computer, with the ability to break RSA 2048, within 10 years.²² Mosca also states that "the likelihood of a scaled quantum computer in the next 20 years is an order of magnitude higher than it was 10 years ago." With the increased global interest in quantum computing and increased resources dedicated to its development, these estimates may need to be revisited.

Strategic Implications of Quantum-Resistant Security

The nature and capabilities of quantum computing and the complexities and trade-offs of quantum-resistant primitives are well-documented in the physics, computer science, and engineering literature. This body of knowledge is kept current by several established research labs as quantum computing and quantum-resistant technologies advance. A large gap exists, however, between the theoretical and technical understanding of these technologies on the one hand and practical implications for businesses

Figure 1. Qubit count by company.

Company	# of Qubits	Notes
Google	53–72	Sycamore (Universal 54 qubit non-linear superconducting resonator); Bristlecone (Universal 72 qubit superconducting)
IBM	53	IBM Q-system53 (Universal 53 qubit superconducting). Announced plans for 127 qubits in 2021, 433 qubits in 2022, and 1000 qubits in 2023.
Intel	49	TangleLake (Universal 49 qubit superconducting universal test chip)
Rigetti	28	Aspen-7 (Universal 28 qubit superconducting)
Xanadu	24	X24 (Universal 24 qubit photonics)
D-Wave	5000	D-Wave Advantage (Annealing 5000 qubit superconducting). Designed for optimization problems. Not a threat to cryptography.

on the other hand. These businesses will need to spend time and money to study the problem and adopt new technology to protect themselves.

As quantum computers are developed at a rapid pace, and with early models already on the market, the overall perception of the ICT community toward a quantum reality is slowly changing. Awareness of the issue has dramatically improved thanks in large part to the efforts of NIST¹⁰ and ETSI,¹ organizations providing quantum and quantum-resistant solutions, not-for-profit groups,^{14,17,32} and most importantly, recent academic work that bridges the technical and managerial aspects of the quantum threat and its implications.¹⁹

There are some keystone works, such as the books by Bernstein et al.⁴ and Yan,³⁵ a white paper by Accenture Labs,²⁴ as well as the article by Mosca that examines the issue comprehensively.²² These works succinctly explain which types of cryptographic algorithms are susceptible to quantum attacks and why, and they provide possible solutions that existed at the time of publication. While academic and standardization bodies such as NIST and ETSI have been considering the steps that organizations should take against the threat of quantum computers

to cryptography for many years, the issue did not gain popular media, and presumably public, attention until the National Security Agency began issuing warnings in mid-2015.^{10,12} This was also around the time when news of the first functioning quantum computers, produced by D-Wave Systems and vetted by Google and NASA, first showed up in popular media.³¹

Around the same time, the literature—both academic and popular—began to suggest business implications for organizations needing to implement quantum-resistant security measures. Some of the literature, and related research and experimentation, was measured and pragmatic, while some was more sensational.^{9,18,32} Google announced in mid-2016 it was experimenting with quantum-resistant security in its Chrome browser using the algorithm code-named “New Hope.”⁷ The company stated that this was not intended to be a new standard; rather, it was explicitly an experiment. Around the same time, Microsoft made the “LatticeCrypto” library, which developers can use to experiment with quantum-resistant key exchange.²¹ The NIST Report on Post-Quantum Cryptography in April 2016 identified the impact of quantum computing on cryptographic

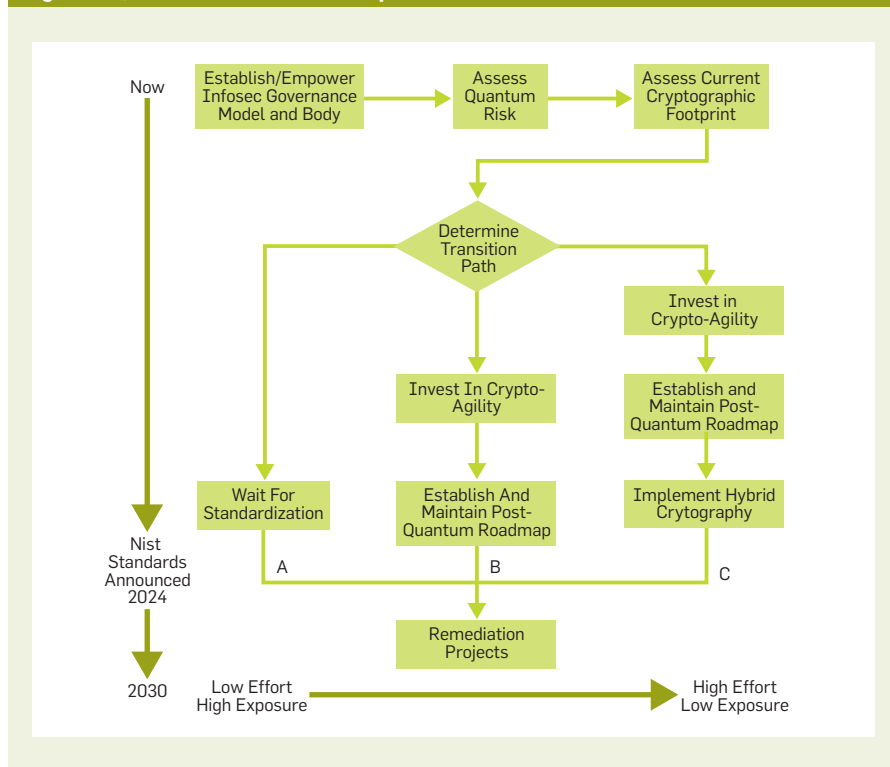
algorithms and the families of quantum-resistant primitives.¹⁰

More recently, targeted vendor articles have been written for trade websites.^{22,33} These cover the issue in more practical terms but do not consider variances among industries in terms of their information security requirements or cryptography management structures. Notably, only one example could be found involving a survey of ICT managers, and this was in a non-academic article in an online cybersecurity magazine.²⁵

While there is more awareness of the potential threat of quantum computing, this progress may not be enough. As a quantum reality gets closer and closer, the available time in which currently standardized cryptographic primitives can be relied upon is getting narrower and narrower. Organizations must get to work and move on to the next step. There needs to be more traction in response to calls for devising a plan ahead of the massive industry-wide adoption from quantum-vulnerable to quantum-resistant technologies.

ICT leaders must devise concrete plans and dedicate adequate resources in their 5- to 10-year budget allocations. To do this, they need to fully understand the threat to valuable information assets and the related business implications.

Figure 2. Quantum readiness roadmap.



Recommendations

Organizations in large, regulated sectors tend to be more aware of the threat of quantum attacks and their implications than their counterparts in smaller and unregulated sectors. Even among those aware of the threat, few are planning for mitigation steps in advance of formal recommendations from the standardization bodies. Most organizations do not have the in-house expertise needed to know what to do or when to do it. This can pose a significant threat to those industries handling sensitive data with a long expiry date, such as Social Security numbers and financial or health records.

The recommendations provided here, illustrated in Figure 2, focus on the *process* by which organizations can effectively assess and mitigate their exposure to quantum attacks. These recommendations are based on the existing literature, combined with empiri-

cal information gathered by an investigative study in which 23 ICT managers were asked about their companies' level of quantum threat awareness and plans for the transition to quantum resistance. The range of business scenarios is far broader than the current literature indicates, and the flexible recommendations presented here are intended to be more relevant to business managers than those previously proposed.

The steps detailed here are intended to aid technology executives and be applicable to organizations of various sizes and regulatory structures, and with diverse information security needs ranging from minimal to highly complex and integrated. By following these recommendations, business and technology managers can effectively size the risk of this threat to their respective organization, and then establish and execute the appropriate mitigation strategies (Scenarios A-C, described in the sidebar "Planning for Quantum Readiness").

In all cases, establishing/empowering a governance model and body is a prerequisite to following these recommendations. The next step is to conduct a thorough quantum risk assessment that determines the scope of the risk to the organization, its potential magnitude, and the likelihood of a system being compromised. The output of this assessment is specific to each organization and largely depends on the sensitivity of its assets and its current approach to safeguarding them.

The next common step is to assess the organization's current cryptographic footprint to establish which cryptographic methods are being used and exactly where (software, hardware, or network) they are being employed. Where quantum vulnerabilities are found, quantum-resistant alternatives should be selected and implemented at the appropriate time (Scenarios A-C), based on technical and organizational requirements and feasibility. By taking the appropriate recommended steps, technology managers can ensure that they are effectively minimizing the threat of quantum attacks on their respective organization.

Acknowledgments

This research was funded, in part, through a generous contribution from The Burnie Group, a Toronto-based

management consulting firm with extensive practical experience rolling out large-scale technology transformations, and the Natural Sciences and Engineering Research Council of Canada (NSERC) under Engage Grant (EGP 543598 – 19, PI: Atefeh Mashatan). The authors would like to acknowledge the efforts of Robert Fullerton and Ryan Kennedy for their assistance in conducting the preliminary stages of this research. **C**

References

- Alléaume, R., et al. Implementation security of quantum cryptography: introduction, challenges, solutions. ETSI White Paper No. 27. European Telecommunications Standards Institute, 2018; https://www.etsi.org/images/files/ETSIWhitePapers/etsi_wp27_qkd_imp_sec_FINAL.pdf.
- Barker, W., Polk, W., Souppaya, M. Getting ready for post-quantum cryptography: explore challenges associated with adoption and use of post-quantum cryptographic algorithms. Cybersecurity White Paper. U.S. Department of Commerce, National Institute of Standards and Technology, 2020; <https://nvlpubs.nist.gov/nistpubs/CSWP/NIST.CSWP.05262020-draft.pdf>.
- Bauer, B., Wecker, D., Millis, A., Hastings, B., Troyer, M. Hybrid quantum-classical approach to correlated materials. *Physical Review X* 6, 3 (2016); <https://journals.aps.org/prx/pdf/10.1103/PhysRevX.6.031045>.
- Bernstein, D.J. Introduction to post-quantum cryptography. In *Proceedings of the 2009 Post-Quantum Cryptography*. D.J. Bernstein, J. Buchmann, and E. Dahmen, Eds. Springer 1–14.
- Bindel, N., Herath, U., McKague, M., Stebila, D. Transitioning to a quantum-resistant public-key infrastructure. In *Proceedings of 2017 Post-Quantum Cryptography*. T. Lange and T. Takagi, Eds. Springer, 2017, 384–405; https://link.springer.com/chapter/10.1007/978-3-319-59879-6_22.
- Boyer, M., Brassard, G., Hoyer, P., Tapp, A. Tight bounds on quantum searching. *Fortschritte der Physik* 46, 4–5 (1998), 493–505; <https://arxiv.org/abs/quant-ph/9605034>.
- Braithwaite, M. 2016. Experimenting with post-quantum cryptography. Google Security Blog; <https://security.googleblog.com/2016/07/experimenting-with-post-quantum.html>.
- Buchmann, J., Lauter, K., Mosca, M. Postquantum cryptography, part 2. *IEEE Security & Privacy* 16, 5 (2018), 12–13; <https://ieeexplore.ieee.org/document/8490197>.
- Campagna, M. et al. Quantum-safe cryptography and security. ETSI White Paper No. 8, 2018; <https://www.etsi.org/images/files/ETSIWhitePapers/QuantumSafeWhitepaper.pdf>.
- Chen, L., Jordan, S., Liu, Y.K., Moody, D., Peralta, R., Perlner, R., Smith-Tone, D. Report on post-quantum cryptography. U.S. Department of Commerce, National Institute of Standards and Technology, 2016; <https://nvlpubs.nist.gov/nistpubs/ir/2016/NIST.IR.8105.pdf>.
- D-Wave. Processing with D-Wave; <https://www.dwavesys.com/>.
- European Telecommunications Standards Institute. Quantum-safe cryptography and security, 2015; <https://www.etsi.org/images/files/ETSIWhitePapers/QuantumSafeWhitepaper.pdf>.
- Evans, D. Top 25 technology predictions. CISCO Internet Business Solutions Group, 2009; https://www.cisco.com/c/dam/en_us/about/ac79/docs/Top_25_Predictions_121409rev.pdf.
- evolutionQ; <https://evolutionq.com/news.html>.
- Grassl, M., Langenberg, B., Roetteler, M., Steinwand, R. Applying Grover's algorithm to AES: quantum resource estimates, 2015; <https://arxiv.org/abs/1512.04965>.
- Grover, L. A fast quantum mechanical algorithm for database search. In *Proceedings of the 28th Annual ACM Symp. Theory of Computing*, 1996, 212–219; <https://dl.acm.org/doi/10.1145/237814.237866>.
- Isara. Isara's quantum-safe readiness program for enterprise; <https://www.isara.com/services/quantum-readiness-enterprise.html>.
- Majot, A., Yampolskiy, R. Global catastrophic risk and security implications of quantum computers. *Futures* 72, (2016), 17–26; <https://www.sciencedirect.com/science/article/abs/pii/S0016328715000294?via%3Dihub>.
- Mashatan, A., Turetken, O. Preparing for the information security threat from quantum computers. *MIS Q. Executive* 19, 2 (2020); <https://aisel.aisnet.org/misqe/vol19/iss2/7>.
- McEliece, R. A public-key cryptosystem based on algebraic coding theory. *DSN (Deep Space Network) Progress Report*, 1978, 42–44; https://tmo.jpl.nasa.gov/progress_report/242-44/44N.PDF.
- Microsoft. Lattice cryptography library. Microsoft, 2016; <https://www.microsoft.com/en-us/research/project/lattice-cryptography-library/>.
- Mosca, M. Cybersecurity in an era with quantum computers: will we be ready? *IEEE Security & Privacy* 16, 5 (2018), 38–41; <https://www.computer.org/csdl/magazine/sp/2018/05/msp2018050038/17D45W9KVFV>.
- National Institute of Standards and Technology. Post-quantum cryptography round 3 finalists: public-key encryption and key-establishment algorithms. NIST, 2020; <https://csrc.nist.gov/Projects/post-quantum-cryptography/round-3-submissions>.
- O'Connor, L., Dukatz, C., DiValentin, L., Farhady, N. Cryptography in a post-quantum world: Preparing intelligent enterprises now for a secure future. Accenture Labs, 2018; https://www.accenture.com/_acnmedia/pdf-87/accenture-809668-quantum-cryptography-whitepaper-v05.pdf.
- Olenick, D. Quantum leap? *SC Magazine* 26, 12 (2015), 16–17; <https://www.scmagazine.com/home/security-news/quantum-leap-the-impact-of-quantum-computing-on-encryption/>.
- Phillips, T. The mathematics behind quantum computing. <https://www.maths.stonybrook.edu/~tony/whatsnew/may07/quantumI.html>.
- Rafter, D. 2019 data breaches: 4 billion records breached so far. Norton; <https://us.norton.com/internetsecurity-emerging-threats-2019-data-breaches.html>.
- Roetteler, M., Naehrig, M., Svore, K. M., Lauter, K. Quantum resource estimates for computing elliptic curve discrete logarithms, 2017; <https://arxiv.org/abs/1706.06752>.
- Schanck, J., Whyte, W., Zhang, Z. Criteria for selection of public-key cryptographic algorithms for quantum-safe hybrid cryptography. Internet draft. IETF, 2016; <https://tools.ietf.org/html/draft-whyte-select-pkc-qsh-02>.
- Shor, P. Polynomial-time algorithms for prime factorization and discrete logarithms on a quantum computer. *SIAM J. Computing* 26, 5 (1997), 1484–1509; <https://pubs.siam.org/doi/10.1137/S0097539795293172>.
- Simonite, T. Google says it has proved its controversial quantum computer really works. MIT Technology Rev., 2015; <https://www.technologyreview.com/s/544276/google-says-it-has-proved-its-controversial-quantum-computer-really-works/>.
- Soukharev, V. InfoSec Global's roadmap to migrate to NIST's new standards. InfoSec Global, 2020; <https://www.infosecglobal.com/post/infosec-global-roadmap-to-migrate-to-nists-new-standards>.
- Totzke, S. Top five questions about using quantum-safe security in financial transactions. FinTech Futures, 2017; <https://www.fintechfutures.com/2017/07/top-five-questions-about-using-quantum-safe-security-in-financial-transactions>.
- Wallden, P., Kashefi, E. Cyber security in the quantum era. *Commun. ACM* 62, 4 (Apr. 2019), 120; <https://cacm.acm.org/magazines/2019/4/235578-cyber-security-in-the-quantum-era/fulltext>.
- Yan, S.Y. Quantum-resistant cryptosystems. In *Quantum Attacks on Public-Key Cryptosystems*, 189–203. Springer, Boston, 2013; https://doi.org/10.1007/978-1-4419-7722-9_5.

Atefeh Mashatan is an associate professor at the Ted Rogers School of Information Technology Management and the founder and director of the Cybersecurity Research Lab at Ryerson University. Mashatan's expertise at the frontlines of the global cybersecurity field was recognized by *SC Magazine* in 2019, when she was named one of the top five Women of Influence in Security.

Doug Heintzman is a technology strategist with 30 years of experience in enterprise software. He consults with companies around the world on innovation and technology disruption.

Copyright held by authors/owners.

Article development led by [acmqueue](https://queue.acm.org)
queue.acm.org

**A discussion with Michael Gardiner,
Alexander Truskovsky, George Neville-Neil,
and Atefeh Mashatan.**

Quantum-Safe Trust for Vehicles: The Race Is Already On

THE THEORY OF quantum computing has been with us for nearly three decades, courtesy of a quantum mechanical model of the Turing machine proposed by physicist Paul Benioff in the early 1980s. For most of that time, the notion has seemed more a far-off vision than an impending reality. That changed abruptly with a 2019 claim by Google AI, in conjunction with NASA, that it had managed to perform a quantum computation infeasible on a conventional computer.

While many have eagerly anticipated the new vistas that could open with the arrival of quantum computing, cryptographers and security experts have not generally shared that enthusiasm since one of the most anticipated quantum advantages comes in

integer factorization, which is critical to RSA (Rivest-Shamir-Adleman)-based security. Also, as far back as 1994, MIT mathematician Peter Shor developed a quantum algorithm capable of solving the discrete logarithm problem central to Diffie-Hellman key exchange and elliptic curve cryptography.

Now that it seems quantum-computing capabilities could become commercially available within the next decade or two—likely in the form of cloud-based services—security professionals have turned with an intensified sense of urgency to the challenge of how to respond to the threat of quantum-powered attacks.

One domain where this is particularly true is in the automotive industry, where cars now coming off assembly lines are sometimes referred to as “rolling datacenters” in acknowledgment of all the entertainment and





communications capabilities they contain. The fact that autonomous driving systems are also well along in development does nothing to allay these concerns. Indeed, it would seem the stakes of automobile cybersecurity are about to become immeasurably higher just as some of the underpinnings of contemporary cybersecurity are rendered moot.

To explore the implications of this in the discussion that follows, acmqueue brought together some of the people who are already working to build a new trust environment for the automotive industry: **Alexander Truskovsky**, director of technical strategy at ISARA Corporation, where efforts are being made to develop quantum-safe cryptographic roots of trust; **Mike Gardiner**, a solutions architect at Thales who has been central to efforts to tailor quantum-safe protections for the automotive industry;

Atefeh Mashatan, director of the Cybersecurity Research Lab at Ryerson University; and **George Neville-Neil**, director of Engineering Operational Security at JUUL Labs, who is better known to many as Kode Vicious.

ATEFEH MASHATAN: What do you see as your greatest concerns when it comes to quantum vulnerability in the automobile industry?

MICHAEL GARDINER: One of the big concerns has to do with over-the-air software updates for smart cars—like a Tesla, for example—where somebody with a quantum computer could potentially issue malicious firmware while creating the illusion it comes from the manufacturer. There's also risk associated with the telemetry data the car sends back to the manufacturer, which could be intercepted or tampered with to make it appear the vehicle went somewhere it didn't actually go.

MASHATAN: Why should we even live with the exposure associated with over-the-air updates?

GARDINER: Now that our cars are becoming smart, they have essentially turned into datacenters on wheels. Which is to say they are now increasingly composed of software components, all of which contain bugs just by their very nature. Auto manufacturers can use software updates not only to deliver new features that keep the car's general entertainment system up to date, but also to correct defects as they surface in other systems.

ALEXANDER TRUSKOVSKY: Just to provide some sense of scale, vehicles such as Ford's F-150 come with more than 100 million lines of code. You can easily imagine a fair number of bugs to deal with there. It's not really a question of *whether* software updates will be necessary, but rather how many and



MICHAEL GARDINER

In the automotive industry, a lot of the components were coded a long time ago and haven't necessarily been looked at recently or vetted by third parties. So, the entertainment systems you find in cars now generally are able to talk to the same CAN bus used by the safety-critical systems.



how often. Typically, you would rather not burden the owner of the car with the expense and inconvenience of coming into the dealership each and every time those updates need to be administered. It's also anticipated that by 2022 each vehicle sold will have some degree of autonomy built into it. This, of course, makes it all the more critical that there be some mechanism in place for updating that software in a prompt and efficient manner.

GEORGE NEVILLE-NEIL: It's one thing to say a car has 100 million lines of code, but most people who build systems containing both safety-critical and nonsafety-critical components are smart enough to know they need to separate those things from each other. How confident are we that over-the-air updates for the safety-critical components won't end up getting bundled along with those for the entertainment system? I ask since that entertainment system is just one big hideous Linux box full of every open-source library some clown wanted to include so people would be able to play music, videos, and games in the car.

The brake system, on the other hand, is something that was presumably written by adults and ideally has been firewalled off from everything else—and not just by a digital firewall either, but also by an air gap. I trust there will be software updates for those safety-critical systems that are sent out separately from those delivered for the car's general entertainment system.

GARDINER: In the automotive industry, a lot of these components were coded a long time ago and haven't necessarily been looked at recently or vetted by third parties. So, the entertainment systems you find in cars now generally are able to talk on the same CAN [Controller Area Network] bus that the safety-critical systems use.

NEVILLE-NEIL: That's a little disturbing.

TRUSKOVSKY: I agree, but at this stage, over-the-air updates are mostly used just for the entertainment systems. Some manufacturers such as Tesla can enable some other functionality via software update, but, for the most part, only the entertainment systems are being updated in this way. Also, with the shift to greater computerization, many vehicles are now being switched to dif-

ferent networking systems. Which is to say you're going to see more Ethernet-based communication between components since a CAN bus simply cannot handle the load that comes along with the current autonomous-driving-system requirements.

So today, vehicles are being designed to be updatable as well as to accommodate some more advanced computer systems. But bear in mind that vehicle design cycles are lengthy—generally five to eight years—meaning that vehicles set to debut five years from now have already been designed.

MASHATAN: What's being done to secure software updates for cars at this point?

TRUSKOVSKY: If handled at a dealership, a mechanic can use a USB key to download the software update, generally without signatures—even though that's not an advisable practice. Over-the-air software updates, in contrast, absolutely require code-signing, and that calls for a public infrastructure with the trust anchor being a root certificate embedded in the vehicle—where the private key belongs to the original equipment manufacturer. With that in place, updates might be delivered to cars in much the same way they're currently sent to mobile phones or laptops in the sense that they would be digitally signed and the vehicle then would take additional steps to verify the authenticity of the software before applying it.

The problem is the embedded trust anchors at the heart of this system are based on classic public-key cryptography, which will be easily broken once attackers are able to use quantum computers. Changing out those dated trust anchors for new ones that have been hardened against quantum-based attacks will require vehicles to be brought in for servicing. In some cases, that might be accomplished easily just by updating the public key, but more often, the upgrade will require some sort of hardware replacement.

This takes us to another problem related to the emergence of autonomous vehicles that include sensors that talk to ECUs [engine control units], which in turn talk to the brakes and the steering system so the vehicle knows where to steer and when to brake. In that scenario, there will be messages that are

relayed between different components and need to be authenticated so the ECU knows they are indeed coming from the vehicle's actual brake sensor or collision sensor—and *not* being impersonated by some hacker trying to take control of the vehicle.

All of which is to say that this is a zero-trust infrastructure where every single message needs to be authenticated and autonomous driving decisions must be fully authenticated. The cryptography being evaluated for use in this environment has yet to be standardized by NIST [National Institute of Standards and Technology]. Still, while some of the parameters of the core quantum-safe algorithms can be modified, the fundamentals of those algorithms—that is, the key sizes, the speeds, and the ways in which things are executed—are not going to change. This means these algorithms can start being tested on vehicle components so that auto manufacturers will be able to start releasing new models that include hardware capable of supporting post-quantum cryptography as soon as possible.

Also, in parallel, work can begin on embedding quantum-safe trust anchors in vehicles since the math used for code signing is essentially ready to roll today. Then, a few years from now, once the final standards become available, that quantum-safe software update channel we've been talking about can be used to supplement the trust anchors with any additional quantum-safe functionality developed in the interim, most of which is expected to relate to requirements for autonomous driving.

MASHATAN: What happens if the NIST standard proves to be not entirely compatible with the quantum-resistant algorithm you're currently working with?

TRUSKOVSKY: You have to hedge your bets, meaning you need to provide for every type of crypto algorithm—lattices, multivariate, code-based, hash-based... you name it. If the lattice-based approaches prove to be broken, then you need to be ready to employ hash-based and multivariate. So, you really need to be able to port all of them.

MASHATAN: Beyond over-the-air software updates, what should auto manufacturers be particularly concerned

about once quantum computing becomes commercially available?

TRUSKOVSKY: Actually, there's another matter related to software updates we should talk about first. In the case of autonomous vehicles, there are occasions when the manufacturer needs to send various authenticated commands to the vehicle. Providing for the security of those commands is pretty similar to what it takes to protect software updates—which is to say, both need to be authenticated in the same way.

The sorts of commands I'm talking about are those that might be sent to an autonomous vehicle following an accident. In that event, an authenticated command could be sent to the vehicle to direct it to a particular service facility or to get the car to move itself out of traffic, over to the shoulder of the road. Clearly, these commands need to be quantum safe as well. I'm talking largely in terms of authentication here, but encryption also plays a big part since we need to provide privacy protection for the user.

GARDINER: There's another aspect of this: Because users will have connectivity to smart vehicles from their mobile devices, any commands they send and any information the car sends back to their mobile devices will also need to be protected for privacy. There's a lot of potential here for hackers to obtain sensitive private information.

MASHATAN: Are these communications between users and their vehicles currently protected by some form of encryption?

NEVILLE-NEIL: My impression is that communications within a car are not encrypted as yet, nor are they likely to be in the near future. They really ought to be, given how many people have tapped into the CAN bus and now will start tapping into the Ethernet. But I haven't seen a standard that says encryption for this is going to be required.

GARDINER: Yes, that's my understanding as well.

NEVILLE-NEIL: Let's also not forget that a huge amount of location data goes to and from these newer vehicles, and there definitely are a lot of issues with that.

MASHATAN: Yes, and that goes beyond privacy concerns since access to that location data could even conceivably be used to enable abductions or other

violent crimes. Is that something people working on post-quantum security are thinking about?

GARDINER: Quantum-enabled attacks would be able to negate the privacy of that telemetry information completely since it's protected only by asymmetric cryptography. In any event, most security efforts are focused on the integrity of software updates at this point.

TRUSKOVSKY: Yes, that and vehicle safety. In the case of a car with an autonomous driving system, information from other vehicles and other sources about, say, a collision just up the road could alter the vehicle's driving instructions. But that data is forgeable, which clearly represents a threat to the safety of the occupants of that vehicle.

NEVILLE-NEIL: In light of this and the other security risks we've been discussing, what do you see as an optimal timeline for getting quantum-resistant cryptography and PKI [public-key infrastructure] deployed throughout the automotive world? And how does that compare with what actually seems feasible?

GARDINER: Even if we were to start deploying right now, we would be at the mercy of supply-chain issues in the automotive industry. It would take roughly five years at the going rate to get these sorts of design changes into a car that's in production.

TRUSKOVSKY: The average on-the-road lifespan of a vehicle is about 11½ years. Then we need to work our way backwards from that to account for the five years or so it takes to design a vehicle. That means the design decisions being made today ought to still make sense 16½ years from now. The recommendation from NIST and other organizations is that mitigations be in place for quantum-computing threats by 2030. Clearly, there's already a fair amount of urgency when it comes to the question of when we should start quantum-proofing vehicles.

For all the pressure to move quickly, the automotive industry can probably be counted upon to proceed stepwise toward the production of vehicles provisioned for quantum-safe trust. Almost without a doubt, the first of those steps will focus on ensuring that the embedded compute devices installed

in cars are actually up to the challenge.

That alone will represent quite a departure for automakers that historically have relied on the least expensive off-the-shelf microcontrollers available. But that simply won't suffice when it comes to providing for an array of complex quantum-safe encryption algorithms or the throughput demands that autonomous driving systems are sure to place on controllers required to communicate continuously with a variety of sensors.

NEVILLE-NEIL: Let's talk about some of the challenges the automotive industry will face once it comes to implementing quantum-resistant PKI. It won't be exactly like deploying PKI within a datacenter or for something that's always online. What are some of the key differences you've been working through?

GARDINER: When you're thinking about something like the trust roots for conducting financial transactions by way of a web browser, all of the trusted CAs [certificate authorities] you're going to encounter are ones that have previously been agreed upon by the browser makers and the CA/Browser Forum [a consortium of CAs and vendors of browser software, operating systems, and other PKI-enabled applications]. The rules around how you can use those certificates and for which keys and for how long have already been established.

In the automotive space, there's no equivalent to that. An automotive manufacturer ought to be able to roll a quantum-resistant PKI into its vehicles early on without first needing to obtain broad industry acceptance.

NEVILLE-NEIL: Does that actually make things easier? Or does it just end up looking like the same problem that surfaced back when all the browser people had to find some way to agree?

GARDINER: I don't think so—at least not until the industry starts talking about vehicle infrastructure or vehicle-to-vehicle communications. That will require wider industry agreements on what should and shouldn't be allowed. But, for now, if we're talking just about firmware updates for some particular car model or what it takes to secure te-

lemetry information between the manufacturer and the car or between a user's mobile and the car, that can be handled on a manufacturer-by-manufacturer basis.

NEVILLE-NEIL: Given that, what does the roadmap look like for rolling out these protections? You say we're talking about defending objects that have an expected 16½-year design span. So, if today is day one, what does the next year or two need to look like for automakers in terms of implementing something along these lines?

TRUSKOVSKY: First, they should be able to do at least a couple of things in parallel. One is that, as they're working on a new vehicle design, they can migrate their CAN bus to Ethernet while also updating all the embedded compute devices that serve as their controllers. These are things they can either design themselves or shop for off the shelf. Either way, it will be necessary to evaluate these devices to make sure they're capable of supporting all the available encryption algorithm families. Some of those algorithms might not be used for years, but the automakers should at least be confident that the hardware they're installing today will be capable of handling them.

At the same time, they can also ensure that their current software/firmware update capabilities will be able to take advantage of the most secure algorithms now available for that purpose—specifically, stateful hash-based signatures.

With both of those goals accomplished, an automaker will have assurance that it has hardware capable of supporting quantum-safe cryptography over the long term, along with a quantum-safe channel through which to push additional quantum-safe functionality over the years to come.

MASHATAN: Given that 16½-year design span for cars, will the compute devices that auto manufacturers are currently embedding be up to all that?

TRUSKOVSKY: That's a good question. A lot of the algorithms we'll see in the near future will be pretty complicated. And yet, auto manufacturers typically won't spend any more for these compute devices than is absolutely necessary—meaning the devices they buy are usually quite limited. This could prove interesting since the

quantum-safe algorithms are definitely going to be far more demanding than the encryption algorithms they're currently running.

MASHATAN: Can you quantify that?

TRUSKOVSKY: It all depends. Some algorithms, like the lattices, run pretty fast, but they also have large keys and signatures. And then you have algorithms like supersingular isogenies that have much smaller keys but run much slower.

Since one of the considerations here has to do with the operation of autonomous vehicles, some thought will also have to be given to throughput. That is, you need to be able to handle perhaps 100 messages per second since there's always going to be some number of sensors talking to some number of controllers—and all of that has to happen in real time. Also, some of those messages will likely be encrypted or require signatures, which is going to add considerably to processing time. But the manufacturers still have to make sure they can meet those real-time requirements. That could prove to be quite a challenge.

MASHATAN: What about that other potential challenge—over-the-air updates? How many automakers are likely to start moving in that direction?

TRUSKOVSKY: I've read that, over the next couple of years, the automotive industry is expected to save \$35 billion by doing software updates over the air rather than handling them in person at dealerships.

MASHATAN: Has anyone done a risk assessment of the auto industry's potential exposure to quantum-enabled attacks?

GARDINER: We have rated the potential as low in the short term, moving up to medium to high over time. But because we see the impact of any attack as being critical, we're treating this as a medium risk at minimum, even over the short term.

NEVILLE-NEIL: What is the implementation timeline for getting quantum-resistant PKIs out there? What needs to happen first? And when do you think that's going to happen?

GARDINER: I think the first thing automotive manufacturers are looking to attain is increased cryptographic agility out of the resources in their cars. Which is to say that right at the top of

their list is gaining the ability to handle quantum-safe algorithms, firmware updates, and telemetry communications. In terms of what happens inside the vehicles, that's probably less of a concern for the time being simply because that requires physical access.

Anyway, just the sourcing of components capable of handling the increased load will, in itself, represent a huge change since these organizations are accustomed to looking at just small microcontrollers that offer the essential built-in functions but little beyond that. Now they're going to have to think more about the future without knowing exactly what that future is going to look like or how much extra capacity that's going to require. I imagine it's going to take them four or five years to go from planning to getting something on the road.

NEVILLE-NEIL: Do you think the path is at least somewhat going to resemble what happened back when the concern was embedding secure compute elements in desktop systems? If you were looking to do SSL [Secure Sockets Layer] 10 or 15 years ago, you had to add specialized cryptographic components to your server. Do you see the first push here being made with the same microcontrollers that automakers were using before, along with some added cryptographic components? Or do you think they're going to need to ditch those microcontrollers and move up to full-on modern processors that include built-in cryptographic instructions? If so, since all those cryptographic instructions now are asymmetric rather than quantum-resistant, how's that going to work?

GARDINER: At first the automakers are either going to have to build in more general-purpose compute devices so they can achieve the required flexibility, or they are going to need to look at FPGA [field-programmable gate array] technology in order to solve that requirement. That's because all the ASICs and other hardware out there right now may not be able to handle these new quantum-safe algorithms. There are some other things they could try, but those might not fit with whatever the standard for this proves to be by 2024.

NEVILLE-NEIL: FPGAs—or anything

else along those lines—are going to represent quite a cost bump. Do you think the automakers will be willing to take that on?

GARDINER: I'm not sure. But I suppose they could consider relying on a symmetric key scheme that's internal to the vehicle and then try to handle integrity and encryption that way. With that approach, they might be able to get away with just one centralized FPGA that's responsible for all the translation between the internal car world and the external world. That still probably wouldn't line up with whatever the standard becomes within the next few years.

Since quantum crypto standards have not yet crystallized, we'll likely see adjustments on many fronts for years to come. But one thing is certain—quantum computing is coming. And it's no longer comfortably far off in the distant future.

Reassuringly, though, organizations are coming to realize that they can't afford to be caught flat-footed once that day comes. From experience, they already know it takes considerable time and effort just to move from one encryption algorithm to another. The shift to quantum-safe algorithms will involve far more than that, and the stakes when it comes to getting everything right will also be much higher.

MASHATAN: In terms of anticipating challenges ahead, are there any lessons to be learned from looking at some of the cryptographic changes made in the past?

GARDINER: Probably so. SHA-1 (Secure Hash Algorithm 1), for example, has probably been broken for a few years now, and yet there still are things out there in the wild that continue to use it. Unless we start preparing for the post-quantum challenge now, we're going to find ourselves in that same position, where the industry continues to rely upon cryptography that's no longer effective well after the arrival of quantum computing. As a general rule, the cybersecurity industry tends to be quite risk-averse when it comes to tackling new things. But this is one



ALEXANDER TRUSKOVSKY

Over the past 16 years, neither the U.S. government nor the Canadian government has managed to complete its migration to Suite B Cryptography. This gives you a pretty good idea about just how long it takes really large organizations to migrate from one algorithm to another.





GEORGE NEVILLE-NEIL

My impression is communications within a car are not encrypted as yet, nor are they likely to be in the near future. They really ought to be, given how many people have tapped into the CAN bus and now will start tapping into the Ethernet.



of those cases where people need to consider that they're likely to put themselves at a much higher risk if they don't start preparing now, since there's no doubt that quantum computing is coming.

TRUSKOVSKY: Another good example from the past would be that the NSA [National Security Agency] announced Suite B Cryptography in 2005, and it was mandated that all government agencies should implement it. But then, 10 years later, came another announcement saying that anyone who hadn't already completed the migration to Suite B ought to pause and wait for the new quantum-safe standards to emerge. Which is to say that, over the past 16 years, neither the U.S. government nor the Canadian government has managed to complete its migration to Suite B Cryptography. This gives you a pretty good idea about just how long it takes really large organizations to migrate from one algorithm to another.

Some of these quantum-safe algorithms behave quite differently from what security experts have become accustomed to. That is, a typical signature algorithm has one private key and one public key. The private key signs while the public key verifies, and you can do that as many times as you want. With the quantum-safe algorithms, you also have one public key that verifies the signatures, but the private key is very different. Basically, it's a collection of one-time keys that have been organized in a binary tree, where each key is a leaf and the root of the tree is your public key. During a signing verification operation, you sign with one of those private keys, but then that key has to be discarded.

What you effectively have is a large number of exhaustible private keys that you need to manage and maintain the state for. This just hasn't been done before. So now, PKI organizations that use these schemes to create root certificates and sign entity certificates need to do a good deal more planning than before. That's because, in the case of a root certificate, the height of the tree determines the number of potential signatures, and there's a trade-off between that number as it grows and the amount of efficiency that can be achieved. This

means you need to plan so you can determine the maximum number of signatures you want to accept from a particular key. You also need to provide for high availability and plan for disaster recovery.

With a large tree of private keys, it's also important to be careful about state. If you back up in multiple locations, you need to share that state across all those locations, which is very impractical, of course. The point is that you end up changing your whole operational plan just because you're now dealing with an exhaustible key. At some point, even though the public key that signs your certificates is still valid, you may run out of private keys. It's easy to see how people who aren't accustomed to these sorts of issues could become pretty frustrated with the quantum-safe algorithms.

GARDINER: The other big cryptographic change people need to get used to is the lack of a general-purpose, jack-of-all-trades key such as RSA. Today RSA is used for encrypting and for signing and for exchanging keys. But these newer algorithms don't really allow multiple operations with a single key, or even with a pair of keys.

NEVILLE-NEIL: Just exactly how large are these trees?

TRUSKOVSKY: There actually are both single-tree and multi-tree variants. The single-tree variants range from tree height 5, where you've got 32 possible signing verification key pairs, to tree height 25, where you've got around 32 million keys. You can then nest these sets in a number of multi-tree formats that allow for an essentially infinite number of keys. But that naturally can lead to significant complications. State management, as already noted, can also quickly become *very* complicated.

These schemes are not recommended for general-purpose use, but they can work really well in those instances where you sign something once and then verify it many times. This applies to root certificates and even intermediate certificates since these are things you create once while also signing a bunch of certs, and then those certs are used over and over again. It also applies to code signing, where you sign once and then many vehicles just need to verify the signa-

ture, which actually makes the job a lot easier.

NEVILLE-NEIL: OK, so it's going to be PKI that handles these trees when it comes to handing the keys out. What's the thinking in terms of how that stash of keys is to be maintained in the car?

TRUSKOVSKY: That's a very good question, but it doesn't actually apply to the car. Instead, it applies where the software updates are actually signed, which is in a hardware security module [HSM] located in the auto manufacturer's datacenter. The corresponding part of the system you'll find on the vehicle side is the public key that serves as the root node of the binary tree, which is only 60 octets [with each octet consisting of eight bits] in length for stateful, hash-based signature schemes—which is to say it's very small.

The vehicle is actually responsible only for doing the easy part here. It just needs to verify the signature, and that's relatively easy since the scheme relies on hashing, which is something all the current automotive hardware nodes are capable of.

With that said, it still amounts to somewhat more hashing than the auto-makers are accustomed to. But then that's really all there is to it. The whole signature verification process involves doing only a couple hundred hashes. This is why that scheme is so suitable for deployment today. In fact, it could be used for software updates right now since even the computer hardware currently found in vehicles is capable of supporting it. Meanwhile, the real difficulties that come along with the new private key can be relegated to a datacenter where a couple of HSMs can be used to handle all the signing and backup requirements.

NEVILLE-NEIL: Taking all this into account, what would you say are the quantum-related issues people should be most concerned about right now? And why?

TRUSKOVSKY: There's certainly no need to worry about everything at the same time. It really gets back to that matter of product life spans and design cycles. For example, the financial services industry has its own quantum concerns to address, but there's no need for those folks to drop everything and start rethinking credit-card

security just yet. That's because credit-card transactions are short-lived, and the cards themselves are replaced every few years. Even in the worst case, a new credit card could be issued within just a few days. In any event, any credit card you have in your wallet right now is likely to have been replaced a number of times before universal quantum-compute capabilities are made available to potential adversaries.

Auto manufacturers, on the other hand, need to account for much longer product life cycles. By the same token, they don't have to contend with the truly daunting product life-span concerns faced by the aerospace industry, where jet engines often are in service for several decades and, of course, cannot be readily replaced. That's a field already well into a timeframe where they need to be deeply concerned about looming quantum security threats—meaning they'll soon need to have answers for all aspects of that problem.

Auto manufacturers can just turn their focus initially to ensuring that whatever engine hardware is designed and built today is capable of handling the cryptography that will become essential once attackers are able to take advantage of quantum-compute capabilities. Once the auto industry has a handle on that, it can turn its attention to making sure it also has the ability to deliver software and firmware updates in a quantum-safe manner.

NEVILLE-NEIL: Can you think of any industries or organizations that already seem to be approaching this correctly?

GARDINER: From an industry perspective, you have ETSI [European Telecommunications Standards Institute], which has already started to figure out how its standards are going to evolve with the addition of both quantum-resistant cryptography and quantum-key distribution in parallel with the work that's being done at NIST. There also are efforts going on in the CA/Browser Forum on standardizing post-quantum certificates, with DigiCert being a particularly loud voice in that space.

At a more organizational level, there are some great examples coming to light now at Microsoft, Google, and

Cloudflare. They've all been doing some very public experiments on integrating quantum-resistant cryptography into TLS [Transport Security Layer], SSH [Secure Shell], and VPN [virtual private network] connections. A lot of the code they have built so far can be found in open-source repositories, so others can take advantage of it. I'd say these organizations also have a great focus on the practicalities of the upcoming transition and what needs to be done to ensure that real-world systems will be ready for whatever is ultimately standardized. One of the promising takeaways from these experiments is that they've shown overall performance in these schemes is still dominated by network transmission, meaning there's probably no need for concern that user experience is going to suffer unduly.

Among those who are currently tackling this challenge, the common thread seems to be a focus on cryptographic agility rather than on attempting to anticipate which of the proposed schemes is going to end up being certified. This suggests that, at an organizational level at least, there's an understanding that the efforts to modify these standards can continue to move forward in parallel so long as they're all built to be agile. Organizations also seem to be taking a pragmatic approach with their own efforts by assuming that a hybrid of quantum and classical schemes might prove necessary in order to meet compliance targets.

This would be my advice for the automotive industry as well: Keep cryptographic agility as a primary focus and don't overoptimize for any specific implementation. This should help with the quantum transition while also allowing for adaptability to changing regional requirements. For companies with limited resources that need to focus their efforts, I'd say integrity and identity issues are the ones to concentrate on. Roots of trust tend to be the most difficult to swap out since people believe they can be trusted over a long period of time, and yet, in the end, all issues of integrity in a distributed system rely on shared roots of trust.

CONTRIBUTORS

KHALED AMMAR
Borialis AI

RENZO ANGLES
University of Talca

WALID AREF
Purdue University

MARCELO ARENAS
PUC & IMFD

MACIEJ BESTA
ETH Zürich

PETER A. BONCZ
CWI

KHUZAIMA DAUDJEE
University of Waterloo

EMANUELE DELLA VALLE
Polytechnic University of Milan

STEFANIA DUMBRAVA
ENSIE

OLAF HARTIG
Linköping University

BERNHARD HASLHOFER
Austrian Institute of Technology

TIM HEGEMAN
VU University Amsterdam

JAN HIDDERS
Birkbeck, University of London

KATJA HOSE
Aalborg University

ADRIANA IAMNITCHI
University of South Florida

VASILIKI KALAVRI
Boston University

HUGO KAPP
Oracle Labs Switzerland

WIM MARTENS
Universität Bayreuth

M. TAMER ÖZSU
University of Waterloo

ERIC PEUKERT
Universität Leipzig

STEFAN PLANTIKOW
Neo4j

MOHAMED RAGAB
University of Tartu

MATEI R. RIPEANU
University of British Columbia

SEMIH SALIHOGLU
University of Waterloo

CHRISTIAN SCHULZ
Heidelberg University and Universität Wien

PETRA SELMER
Neo4j

JUAN F. SEQUEDA
data.world

JOSHUA SHINAVIER
Uber Engineering

GÁBOR SZÁRNYAS
Budapest Univ. of Technology and Economics

RICCARDO TOMMASINI
University of Tartu

ANTONINO TUMEO
Pacific Northwest National Lab

ALEXANDRU UTA
VU University Amsterdam

ANA LUCIA VARBANESCU
University of Amsterdam

HSIANG-YUN WU
TU Wien

NIKOLAY YAKOVETS
TU Eindhoven

DA YAN
The University of Alabama

EIKO YONEKI
University of Cambridge

Ensuring the success of big graph processing for the next decade and beyond.

BY SHERIF SAKR, ANGELA BONIFATI,
HANNES VOIGT, AND ALEXANDRU IOSUP

The Future Is Big Graphs: A Community View on Graph Processing Systems

GRAPHS ARE, BY nature, ‘unifying abstractions’ that can leverage interconnectedness to represent, explore, predict, and explain real- and digital-world phenomena. Although real users and consumers of graph instances and graph workloads understand these abstractions, future problems will require new abstractions and systems. What needs to happen in the next decade for big graph processing to continue to succeed?



A Joint Effort by the Computer Systems and Data Management Communities

The authors of this article met in Dec. 2019 in Dagstuhl for Seminar 19491 on Big Graph Processing Systems.^a The seminar gathered a diverse group of 41 high-quality researchers from the data management and large-scale-systems communities. It was an excellent opportunity to start the discussion about next-decade opportunities and challenges for graph processing.

This is a community publication. The first four authors co-organized the community event leading to this article and coordinated the creation of this manuscript. All other authors contributed equally to this research. Unfortunately, Sherif Sakr passed away during the period following the event and the completion of this article. This article is published in memoriam.

a <https://www.dagstuhl.de/en/program/calendar/semhp/?semnr=19491>

We are witnessing an unprecedented growth of interconnected data, which underscores the vital role of graph processing in our society. Instead of a single, exemplary (“killer”) application, we see big graph processing systems underpinning many emerging but already complex and diverse data management ecosystems, in many areas of societal interest.^a

To name only a few recent, remarkable examples, the importance of this field for practitioners is evidenced by the large number (more than 60,000) of people registered^b to download the Neo4j book *Graph Algorithms*^c in just over one-and-a-half years, and by the enormous interest in the use of graph processing in the artificial intelligence (AI) and machine learning (ML) fields.^d

a As indicated by a user survey¹² and by a systematic literature survey of 18 application domains, including biology, security, logistics and planning, social sciences, chemistry, and finance. See <http://arxiv.org/abs/1807.00382>

b See <https://app.databox.com/datawall/551f309602080e2b2522f7446a20adb705cabbde8>

c See <https://www.oreilly.com/library/view/graph-algorithms/9781492047674/>

d Many highly cited articles support this statement, including “Inductive Representation Learning on Large Graphs” by W. Hamilton et al. (2017) and “DeepWalk: Online Learning of Social Representations” by B. Perozzi et al. (2014); <https://arxiv.org/pdf/1403.6652.pdf>

» key insights

- Graphs are ubiquitous abstractions enabling reusable computing tools for graph processing with applications in every domain.
- Diverse workloads, standard models and languages, algebraic frameworks, and suitable and reproducible performance metrics will be at the core of graph processing ecosystems in the next decade.

Furthermore, the timely Graphs 4 COVID-19 initiative^e is evidence of the importance of big graph analytics in alleviating the pandemic.

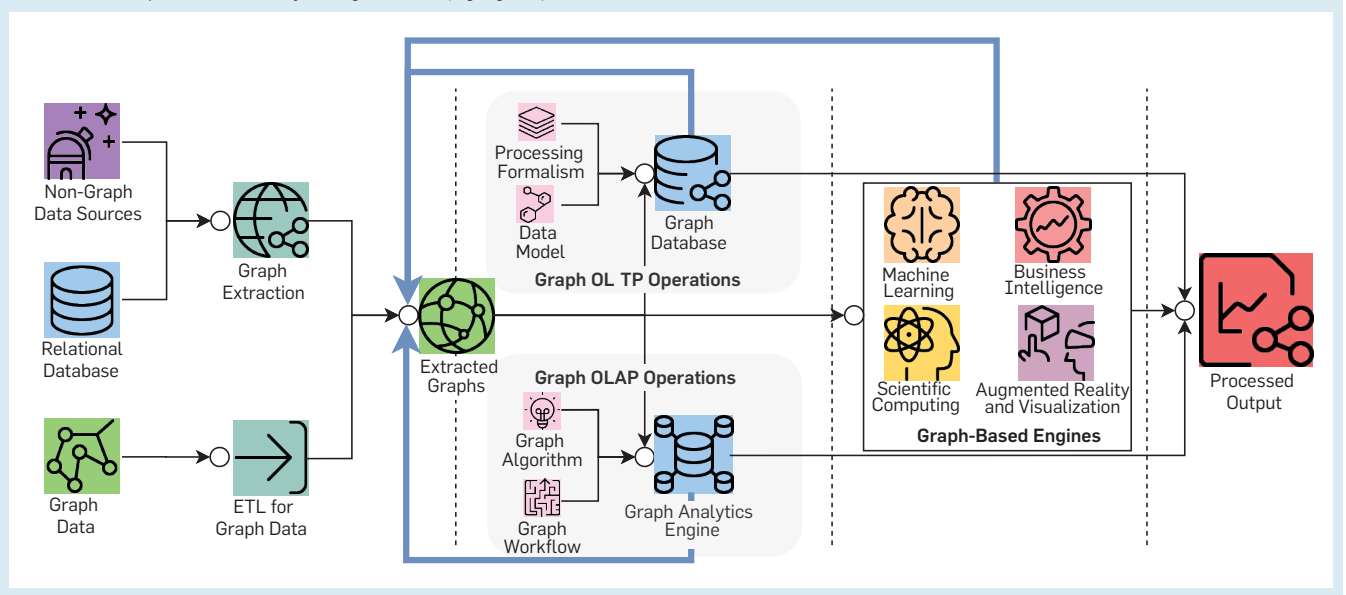
Academics, start-ups, and even big tech companies such as Google, Facebook, and Microsoft have introduced various systems for managing and processing the growing presence of big graphs. Google’s PageRank (late 1990s) showcased the power of Web-scale graph processing and motivated the development of the MapReduce programming model, which was originally used to simplify the construction of the data structures used to handle searches, but has since been used extensively outside of Google to implement algorithms for large-scale graph processing.

Motivated by scalability, the 2010 Google Pregel “think-like-a-vertex” model enabled distributed PageRank

e See <https://neo4j.com/graphs4good/covid-19/>

Figure 1. Illustration of a complex data pipeline for graph processing.

Data flows left to right, from data source to output, via a series of functionally different processing steps. Feedback and loopbacks flow mainly through the blue (highlighted) arrows.



computation, while Facebook, Apache Giraph, and ecosystem extensions support more elaborate computational models (such as task-based and not always distributed) and data models (such as diverse, possibly streamed, possibly wide-area data sources) useful for social network data. At the same time, an increasing number of use cases revealed RDBMS performance problems in managing highly connected data, motivating various startups and innovative products, such as Neo4j, Sparksee, and the current Amazon Neptune. Microsoft Trinity and later Azure SQL DB provided an early distributed database-oriented approach to big graph management.

The diversity of models and systems led initially to the fragmentation of the market and a lack of clear direction for the community. Opposing this trend, we see promising efforts to bring together the programming languages, ecosystem structure, and performance benchmarks. As we have argued, there is no killer application that can help to unify the community.

Co-authored by a representative sample of the community (see the sidebar, “A Joint Effort by the Computer Systems and Data Management Communities”), this article addresses the questions: What do the next-decade big-graph processing systems look like from the perspectives of the data management and the large-scale-systems communities?^f What can we say today about the guiding design principles of these systems in the next 10 years?

Figure 1 outlines the complex pipeline of future big graph processing systems. Data flows in from diverse sources (already graph-modeled as well as non-graph-modeled) and is persisted, managed, and manipulated with online transactional processing (OLTP) operations, such as insertion, deletion, updating, filtering, projection, joining, uniting, and intersecting. The data is then analyzed, enriched, and condensed with online analytical processing (OLAP) operations, such as grouping, aggregating, slicing, dicing, and rollup. Finally, it is disseminated and consumed by a variety of applications, including machine learning, such as



What needs to happen in the next decade for big graph processing to continue to succeed?



ML libraries and processing frameworks; business intelligence (BI), such as report generating and planning tools; scientific computing; visualization; and augmented reality (for inspection and interaction by the user). Note that this is not typically a purely linear process and hybrid OLTP/OLAP processes can emerge. Considerable complexity stems from (intermediate) results being fed back into early-process steps, as indicated by the blue arrows.

As an example, to study coronaviruses and their impact on human and animal populations (for example, the COVID-19 disease), the pipeline depicted in Figure 1 could be purposed for two major kinds of analysis: network-based ‘omics’ and drug-related search, and network-based epidemiology and spread-prevention. For the former, the pipeline could have the following steps:

1. Initial genome sequencing leads to identifying similar diseases.
2. Text (non-graph data) and structured (database) searches help identify genes related to the disease.
3. A network treatment coupled with various kinds of simulations could reveal various drug targets and valid inhibitors, and might lead to effective prioritization of usable drugs and treatments.

For the latter, social media and location data, and data from other privacy-sensitive sources, could be combined into social interaction graphs, which could be traversed to establish super-spreaders and super-spreading events related to them, which could result in the establishment of prevention policies and containment actions. However, the current generation of graph processing technology cannot support such a complex pipeline.

For instance, on the COVID-19 knowledge graph,^g useful queries can be posed against individual graphs^h inspecting the papers, patents, genes, and most influential COVID-19 authors. However, inspecting several data sources in a full-fledged graph processing pipeline across multiple graph datasets, as illustrated in Figure 1, raises many challenges for current graph da-

^g See <https://covidgraph.org/>

^h See <https://github.com/covidgraph/documentation/blob/master/helpful-queries.md>

^f The summary of the Dagstuhl seminar. See <https://www.dagstuhl.de/19491>

tabase technology. In this article, we formulate these challenges and build our vision for next-generation, big-graph processing systems by focusing on three major aspects: *abstractions*, *ecosystems*, and *performance*. We present expected data models and query languages, and inherent relationships among them in lattice of abstractions and discuss these abstractions and the flexibility of lattice structures to accommodate future graph data models and query languages. This will solidify the understanding of the fundamental

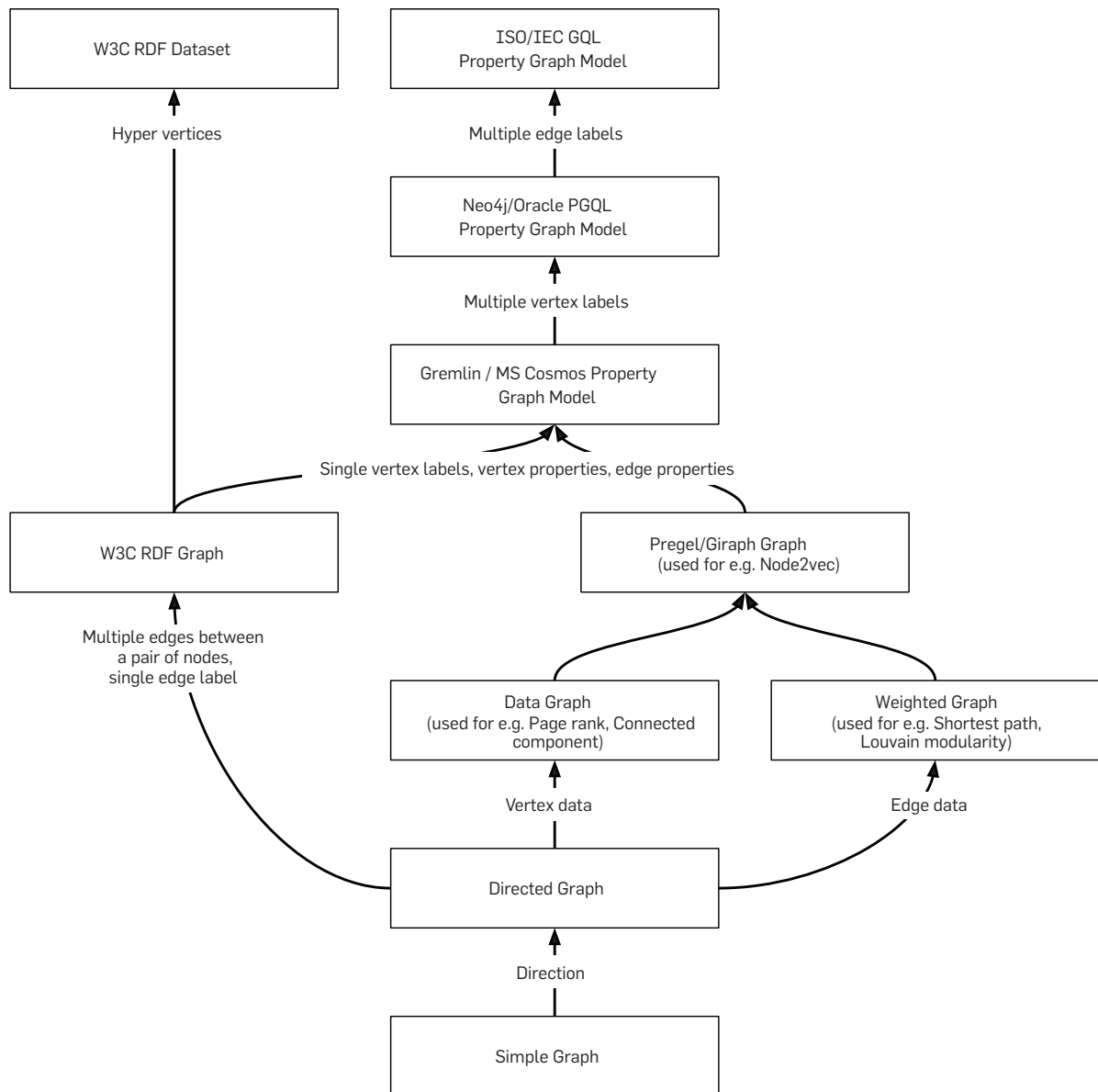
principles of graph data extraction, exchange, processing, and analysis, as illustrated in Figure 1.

A second important element, as we will discuss, is the vision of an ecosystem governing big graph processing systems and enabling the tuning of various components, such as OLAP/OLTP operations, workloads, standards, and performance needs. These aspects make the big processing systems more complicated than what was seen in the last decade. Figure 1 provides a high-level

perception of this complexity in terms of inputs, outputs, processing needs, and final consumption of graph data.

A third element is how to understand and control performance in these future ecosystems. We have important performance challenges to overcome, from methodological aspects about performing meaningful, tractable, and reproducible experiments to practical aspects regarding the trade-off of scalability with portability and interoperability.

Figure 2. Example lattice shows graph data model variants with their model characteristics.⁸



Abstractions

Abstractions are widely used in programming languages, computational systems, and database systems, among others, to conceal technical aspects in favor of more user-friendly, domain-oriented logical views. Currently, users have to choose from a large spectrum of graph data models that are similar, but differ in terms of expressiveness, cost, and intended use for querying and analytics. This ‘abstraction soup’ poses significant challenges to be solved in the future.


Understanding data models. Today, graph data management confronts many data models (directed graphs, RDF, variants of property graphs, and so on) with key challenges: deciding which data model to choose per use case and mastering interoperability of data models where data from different models is combined (as in the left-hand side of Figure 1).

Both challenges require a deeper understanding of data models regarding:

1. How do humans conceptualize data and data operations? How do data models and their respective operators support or hinder the human thought process? Can we measure how “natural” or “intuitive” data models and their operators are?

2. How can we quantify, compare, and (partially) order the (modeling and operational) *expressive power* of data models? Concretely, Figure 2 illustrates a lattice for a selection of graph data models. Read bottom-up, this lattice shows which characteristic has to be added to a graph data model to obtain a model of richer expressiveness. The figure also underlines the diversity of data models used in theory, algorithms, standards, and relevant industry systems. How do we extend this comparative understanding across multiple data model families, such as graph, relational, or document? What are the costs and benefits of choosing one model over another?

3. Interoperability between different data models can be achieved through mappings (semantic asser-



We are witnessing an unprecedented growth of interconnected data, which underscores the vital role of graph processing in our society.



tions across concepts in different data models) or with direct translations (for instance, W3C’s R2RML). Are there general ways or building blocks for expressing such mappings (category theory, for example)?

Studying (1) requires foremost investigators working with data and data models, which is uncommon in the data management field and should be conducted collaboratively with other fields, such as human-computer interaction (HCI). Work on HCI and graphs exists, for example, in HILDA workshops at Sigmod. However, these are not exploring the search space of graph data models.

Studying (2) and (3) can build on existing work in database theory, but can also leverage findings from neighboring computer science communities on comparison, featurization, graph summarization, visualization, and model transformation. As an example, graph summarization²² has been widely exploited to provide succinct representations of graph properties in graph mining¹ but they have seldom been used by graph processing systems to make processing more efficient, more effective, and more user centered. For instance, approximate query processing for property graphs cannot rely on sampling as done by its relational counterpart and might need to use quotient summaries for query answering.

Logic-based and declarative formalisms. Logic provides a unifying formalism for expressing queries, optimizations, integrity constraints, and integration rules. Starting from Codd’s seminal insight relating logical formulae to relational queries,¹² many first order (FO) logic fragments have been used to formally define query languages with desirable properties such as decidable evaluation. Graph query languages are essentially a syntactic variant of FO augmented with *recursive* capabilities.

Logic provides a yardstick for *reasoning* about graph queries and graph constraints. Indeed, a promising line of research is the application of formal tools, such as model checking, theorem proving,¹⁵ and testing to establish the *functional correctness* of complex graph processing systems, in general, and of graph database systems, in particular.

The influence of logic is pivotal not

ⁱ The figure does not aim to provide a complete list of Graph DBMS products. Please consult, for example, <https://db-engines.com/en/ranking/graph+dbms> and other market surveys for comprehensive overviews.

Known Properties of Graph Processing Workloads

Graph workloads may exhibit several properties:

1. Graph workloads are useful for many, vastly diverse domains.^{24,25,26}

Notable features include edge orientation, such as properties/timestamps for edges and nodes; graph methods (neighborhood statistics, pathfinding and traversal, and subgraph mining); programming models (think-like-a-vertex, think-like-an-edge, and think-like-a-subgraph); diverse graph sizes, including trillion-edge graphs;²⁶ and query and process selectivities.⁹

2. Graph workloads can be highly irregular, mixing (short-term) data-intensive and compute-intensive phases.²⁶ The source of irregularity, such as different datasets, algorithms, and computing platforms, greatly affects performance. Their interdependency forms the Hardware-Platform-Algorithm-Dataset (HPAD) Law.²⁹

3. Graph processing uses a complex pipeline, combining a variety of tasks other than querying and algorithms.^{1,24} From traditional data management, workloads include: transactional (OLTP) workloads in multi-user environments, with many short, discrete, likely atomic transactions; and analytical (OLAP) workloads with fewer users but complex and resource-intensive queries or processing jobs, with longer runtime (minutes). Popular tasks also include extract, transform, load (ETL); visualization; cleaning; mining; and debugging and testing, including synthetic graph generation.

4. Scalability, interactivity, and usability affect how graph users construct their workloads.²⁴

only to database languages, but also as a foundation for combining logical reasoning with statistical learning in AI. Logical reasoning derives categorical notions about a piece of data by logical deduction. Statistical learning derives categorical notions by learning statistical models on known data and applying it to new data. Both leverage the topological structure of graphs (ontologies and knowledge graphs^j or graph em-

beddings such as Node2vec^d to produce better insights than on non-connected data). However, both happen to be isolated. Combining both techniques can lead to crucial advancements.

As an example, deep learning (unsupervised feature learning) applied to graphs allows us to infer structural regularities and obtain meaningful representations for graphs that can be further leveraged by indexing and querying mechanisms in graph databases and exploited for logical reasoning. As another example, probabilistic models and causal relationships can be naturally encoded in property graphs and are the basis of advanced-graph neural networks.^k Property graphs allow us to synthesize more accurate models for ML pipelines, thanks to their inherent expressivity and embedded domain knowledge.

These considerations unveil important open questions as follows: How can statistical learning, graph processing, and reasoning be combined and integrated? Which underlying formalisms make this possible? How can we weigh between the two mechanisms?

Algebraic operators for graph processing. Currently, there is no standard graph algebra. The outcome of the Graph Query Language (GQL) Standardization Project could influence the design of a graph algebra alongside existing and emerging use cases.²⁵ However, next-generation graph processing systems should address questions about their algebraic components.

What are the fundamental operators of this algebra compared to other algebras (relation, group, quiver or path, incidence, or monadic algebra comprehensions)? What core graph algebra should graph processing systems support? Are there graph analytical operators to include in this algebra? Can this graph algebra be combined and integrated with an algebra of types to make type-systems more expressive and to facilitate type checking?

A “relational-like” graph algebra able to express all the first-order queries¹¹ and enhanced with a graph pat-

tern-matching operator¹⁶ seems like a good starting point. However, the most interesting graph-oriented queries are *navigational*, such as reachability queries, and cannot be expressed with limited recursion of relational algebra.^{3,8} Furthermore, relational algebra is a closed algebra; that is, input(s) and output of each operator is a relation, which makes relational algebra operators composable. Should we aim for a closed-graph algebra that encompasses both relations and graphs?

Current graph query engines combine algebra operators and ad hoc graph algorithms into complex workloads, which complicates implementation and affects performance. An implementation based on a single algebra also seems utopic. A query language with general Turing Machine capabilities (like a programming language), however, entails tractability and feasibility problems.² Algebraic operators that work in both centralized and distributed environments, and that can be exploited by both graph algorithms and ML models such as GNNs, graphlets, and graph embeddings, could be highly desirable for the future.

Ecosystems

Ecosystems behave differently from mere systems of systems; they couple many systems developed for different purposes and with different processes. Figure 1 exemplifies the complexity of a graph processing ecosystem through high-performance OLAP and OLTP pipelines working together. What are the ecosystem-related challenges?

Workloads in graph processing ecosystems. Workloads affect both the functional requirements (what a graph processing ecosystem will be able to do) and the non-functional (how well). Survey data²⁵ points to pipelines, as in Figure 1: complex workflows, combining heterogeneous queries and algorithms, managing and processing diverse datasets, with characteristics summarized in the sidebar “Known Properties of Graph Processing Workloads.”

In Figure 1, graph processing links to general processing, including ML, as well as to domain-specific processing ecosystems, such as simulation and numerical methods in science

j A recent practical example is the COVID-19 Knowledge Graph: <https://covidgraph.org/>

k “A Comprehensive Survey on Graph Neural Networks” by Z. Wu et al, 2019; abs/1901.00596.

and engineering, aggregation and modeling in business analytics, and ranking and recommendation in social media.

Standards for data models and query languages. Graph processing ecosystem standards can provide a common technical foundation, thereby increasing the mobility of applications, tooling, developers, users, and stakeholders. Standards for both OLTP and OLAP workloads should standardize the data model, the data manipulation and data definition language, and the exchange formats. They should be easily adoptable by existing implementations and also enable new implementations in the SQL-based technological landscape.

It is important that standards reflect existing industry practices by following widely used graph query languages. To this end, ISO/IEC started the GQL Standardization Project in 2019 to define GQL as a new graph query language. GQL is backed by 10 national standards bodies with representatives from major industry vendors and support from the property graph community as

represented by the Linked Data Benchmarks Council (LDBC).¹

With an initial focus on transactional workloads, GQL will support composable graph querying over multiple, possibly overlapping, graphs using enhanced regular path queries (RPQs),³ graph transformation (views), and graph updating capabilities. GQL enhances RPQs with pattern quantification, ranking, and path-aggregation. Syntactically, GQL combines SQL style with visual graph patterns pioneered by Cypher.¹⁴

Long-term, it would also be worthwhile to standardize building blocks of graph algorithms, analytical APIs and workflow definitions, graph embedding techniques, and benchmarks.²⁸ However, broad adoption for these aspects requires maturation.

Reference architecture. We identify the challenge of defining a reference architecture for big graph processing. The early definition of a reference architecture has greatly benefited the discussion around the design, develop-

ment, and deployment of cloud and grid computing solutions.¹³

For big graph processing, our main insight is that many graph processing ecosystems match the common reference architecture of datacenters,¹⁸ from which Figure 3 derives. The Spark ecosystem depicted here is one among thousands of possible instantiations. The challenge is to capture the evolving graph processing field.

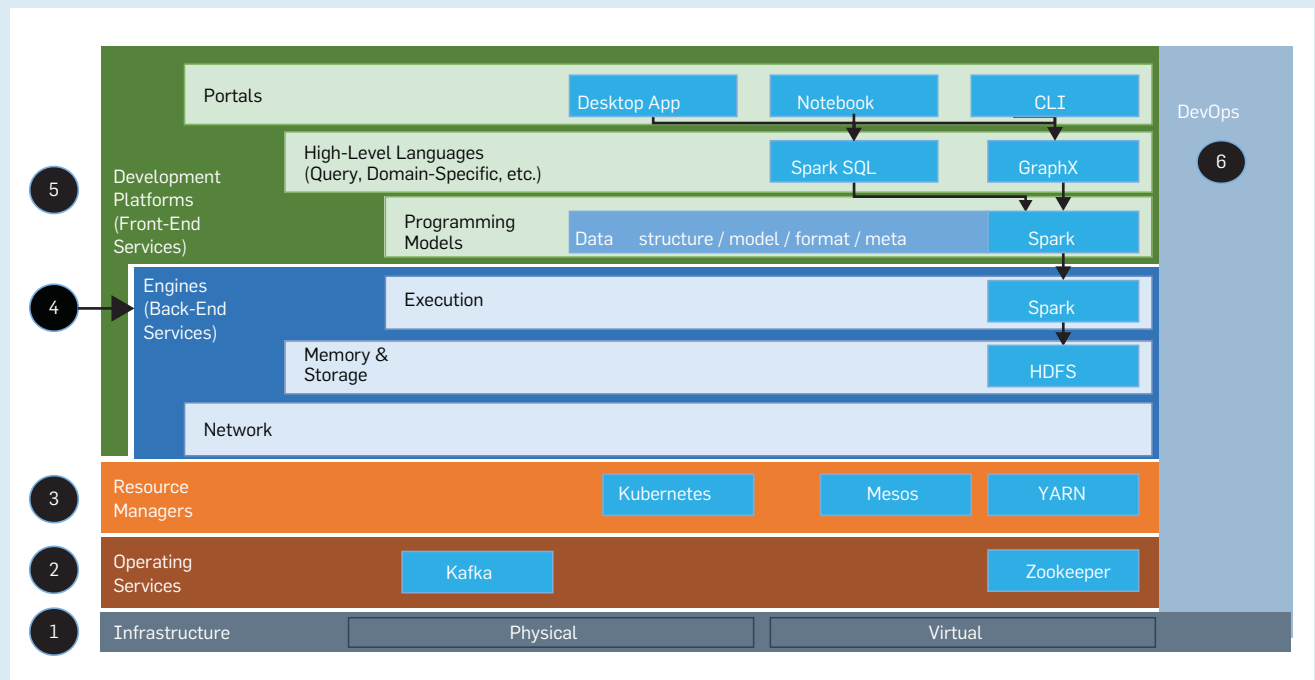
Beyond scale-up vs. scale-out. Many graph platforms focus either on scale-up or scale-out. Each has relative advantages.²⁷ Beyond merely reconciling scale-up and scale-out, we envision a *scalability continuum*: given a diverse workload, the ecosystem would automatically decide how to run it, and on what kind of heterogeneous infrastructure, meeting service-level agreements (SLAs).

Numerous mechanisms and techniques exist to enforce scale-up and scale-out decisions, such as data and work partitioning, migration, offloading, replication, and elastic scaling. All decisions can be taken statically or dynamically, using various optimization and learning techniques.

1 See <http://ldbncouncil.org/>

Figure 3. A reference architecture for graph processing ecosystems.

Layer 1, the infrastructure layer, provides physical and virtual resources. Layer 2, the operating services layer, provides services across resources, including data streaming and synchronization. Resource managers, in layer 3, provide static and dynamic resource management and scheduling across resources. Back-end and front-end layers (layers 4 and 5, respectively) represent specialization efforts. Conversely, layers 2 and 3 may generalize techniques initially developed in layers 4 and 5.



Dynamic and streaming aspects.


Future graph processing ecosystems should cope with dynamic and streaming graph data. A *dynamic graph* extends the standard notion of a graph to account for updates (insertions, changes, deletions) such that the current and previous states can be seamlessly queried. Streaming graphs can grow indefinitely as new data arrives. They are typically unbounded, thus the underlying systems are unable to keep the entire graph state. The sliding window semantics⁶ allow the two notions to be unified, with insertions and deletions being considered as arrivals and removals from the window.

Since current streaming processing technologies are fairly simple, for instance aggregations and projections as in industrial graph processing libraries (such as Gelly on Apache Flink), the need for “complex graph data streams” is evident, along with more advanced graph analytics and ML ad hoc operators. Another research challenge is to identify the graph-query processing operators that can be evaluated on dynamic and streaming graphs while taking into account recursive operators^{7,23} and path-oriented semantics, as needed for standard query languages such as GQL and G-Core.⁴


Graph processing platforms are also dynamic; discovering, understanding, and controlling the *dynamic phenomena* that occur in complex graph processing ecosystems is an open challenge. As graph processing ecosystems become more mainstream and are embedded in larger data-processing pipelines, we expect to increasingly observe known systems phenomena, such as performance variability, the presence of cascading failures, and autoscaling resources. What new phenomena will emerge? What programming abstractions²⁰ and systems techniques can respond to them?

Performance

Graph processing raises unique performance challenges, from the lack of a widely used performance metric other than response time to the methodological problem of comparing graph processing systems across architectures and tuning processes to performance portability and reproducibility. Such



Instead of a single, exemplary (“killer”) application, we see big graph processing systems underpinning many emerging but already complex and diverse data management ecosystems.



challenges become even more daunting for graph processing ecosystems.

Benchmarks, performance measurement, and methodological aspects. Graph processing suffers from methodological issues similar to other computing disciplines.^{5,24} Running comprehensive graph processing experiments, especially at scale, lacks tractability⁹—that is, the ability to implement, deploy, and experiment within a reasonable amount of time and cost. As in other computing disciplines,^{5,24} we need new, reproducible, experimental methodologies.

Graph processing also raises unique challenges in performance measurement and benchmarking related to complex workloads and data pipelines (Figure 1). Even seemingly minute HPAD variations, for example the graph’s degree distribution, can have significant performance implications.^{17,26} The lack of interoperability hinders fair comparisons and benchmarking. Indexing and sampling techniques might prove useful to improve and predict the runtime and performance of graph queries,^{8,21,30} challenging the communities of large-scale systems, data management, data mining, and ML.

Graph processing systems rely on complex runtimes that combine software and hardware platforms. It can be a daunting task to capture system-under-test performance—including parallelism, distribution, streaming vs. batch operation—and test the operation of possibly hundreds of libraries, services, and runtime systems present in real-world deployments.

We envision a combination of approaches. As in other computing disciplines,^{5,24} we need new, reproducible experimental methodologies. Concrete questions arise: How do we facilitate quick yet meaningful performance testing? How do we define more faithful metrics for executing a graph algorithm, query, program, or workflow? How can we generate workloads with combined operations, covering temporal, spatial, and streaming aspects? How do we benchmark pipelines, including ML and simulation? We also need organizations such as the LDBC to curate benchmark sharing and to audit benchmark usage in practice.

Specialization vs. portability and

interoperability. There is considerable tension between specializing graph processing stacks for performance reasons and enabling productivity for the domain scientist, through portability and interoperability.

Specialization, through custom software and especially hardware acceleration, leads to significant performance improvements. Specialization to graph workloads, as noted in the sidebar, focuses on diversity and irregularity^m in graph processing: sheer dataset-scale (addressed by Pregel and later by the open source project, Giraph), the (truncated) power-law-like distributions for vertex degrees (PowerGraph), localized and community-oriented updates (GraphChi), diverse vertex-degree distributions across datasets (PGX.D, PowerLya), irregular or non-local vertex access (Mosaic), affinity to specialized hardware (the BGL family, HAGGLE, rapids.ai), and more.

The high-performance computing domain proposed specialized abstractions and C++ libraries for them, and high-performance and efficient runtimes across heterogeneous hardware. Examples include BGL,²⁸ CombBLAS, and GraphBLAS. Data management approaches, including Neo4j, GEMS,¹⁰ and Cray’s Urika, focus on convenient query languages such as SPARQL and Cypher to ensure portability. Ongoing work also focuses on (custom) accelerators.

Portability through reusable components seems promising, but no standard graph library or query language currently exists. More than 100 big graph processing systems exist, but they do not support portability: graph systems will soon need to support constantly evolving processes.

Lastly, interoperability means integrating graph processing into broader workflows with multi-domain tools. Integration with ML and data mining processes, and with simulation and decision-making instruments, seems vital but is not supported by existing frameworks.

A memex for big graph processing systems. Inspired by Vannevar Bush’s

^m Irregularity could be seen as the opposite of the locality principle commonly leveraged in computing.

1940s concept of personal memex, and by a 2010s specialization into a Distributed Systems Memex,¹⁹ we posit that it would be both interesting and useful to create a Big Graph Memex for collecting, archiving, and retrieving meaningful operational information about such systems. This could be beneficial for learning about and eradicating performance and related issues, to enable more creative designs and extend automation, and for meaningful and reproducible testing, such as feedback building-block in smart graph processing.

Conclusion

Graphs are a mainstay abstraction in today’s data-processing pipelines. How can future big graph processing and database systems provide highly scalable, efficient, and diversified querying and analytical capabilities, as demanded by real-world requirements?

To tackle this question, we have undertaken a community approach. We started through a Dagstuhl Seminar and, shortly after, shaped the structured connections presented here. We have focused in this article on three interrelated elements: abstractions, ecosystems, and performance. For each of these elements, and across them, we have provided a view into what’s next.

Only time can tell if our predictions provide worthwhile directions to the community. In the meantime, join us in solving the problems of big graph processing. The future is big graphs. **□**

References

1. Aggarwal, C.C. and Wang, H. Managing and mining graph data. *Advances in Database Systems 40*. Springer, (2010).
2. Aho, A.V. and Ullman, J.D. Universality of data retrieval languages. In *Proceedings of the 6th ACM SIGACT-SIGPLAN Symposium on Principles of Programming Languages* (1979) 110–119.
3. Angles, R. et al. Foundations of modern query languages for graph databases. *ACM Computing Surveys 50*, 5 (2017), 68:1–68:40.
4. Angles, R. et al. G-CORE: A core for future graph query languages. *SIGMOD Conf.* (2018), 1421–1432.
5. Angriman, E. et al. Guidelines for experimental algorithmics: A case study in network analysis. *Algorithms 12*, 7 (2019), 127.
6. Babcock, B., Babu S., Datar, M., Motwani, R., and Widom, J. Models and issues in data stream systems. *PODS* (2002), 1–16.
7. Bonifati, A., Dumbrava, S., and Gallego Arias, E.J. Certified graph view maintenance with regular datalog. *Theory Pract. Log. Program.* 18, 3–4 (2018), 372–389.
8. Bonifati, A., Fletcher, G.H.L., Voigt, H., and Yakovets, N. Querying graphs. *Synthesis Lectures on Data Management*. Morgan & Claypool Publishers (2018).
9. Bonifati, A., Holubová, I., Prat-Pérez, A., and Sakr, S. Graph generators: State of the art and open


- challenges. *ACM Comput. Surv.* 53, 2 (2020), 36:1–36:30.
10. Castellana, V.G. et al. In-memory graph databases for web-scale data. *IEEE Computer* 48, 3 (2015), 24–35.
11. Chandra, A.K. Theory of database queries. *PODS* (1988), 1–9.
12. Codd, E.F. A relational model of data for large shared data banks. *Commun. ACM* 13, 6 (June 1970), 377–387.
13. Foster, I. and Kesselman, C. *The Grid 2: Blueprint for a New Computing Infrastructure*. Elsevier (2003).
14. Francis, N. et al. Cypher: An evolving query language for property graphs. *SIGMOD Conference* (2018), 1433–1445.
15. Gonthier, G. et al. A machine-checked proof of the odd order theorem. *Intern. Conf. Interactive Theorem Proving* (2013), 163–179.
16. He, H. and Singh, A.K. Graphs-at-a-time: Query language and access methods for graph databases. *SIGMOD Conference* (2008), 405–418.
17. Iosup, A. et al. LDBC Graphalytics: A benchmark for large-scale graph analysis on parallel and distributed platforms. In *Proc. VLDB Endow.* 9, 13 (2016), 1317–1328.
18. Iosup, A. et al. Massivizing computer systems: A vision to understand, design, and engineer computer ecosystems through and beyond modern distributed systems. *ICDCS* (2018), 1224–1237.
19. Iosup, A. et al. The AtLarge vision on the design of distributed systems and ecosystems. *ICDCS* (2019), 1765–1776.
20. Kalavri, V., Vlassov, V., and Haridi, S. High-level programming abstractions for distributed graph processing. *IEEE Trans. Knowl. Data Eng.* 30, 2 (2018), 305–324.
21. Leskovec, J. and Faloutsos, C. Sampling from large graphs. *KDD* (2006), 631–636.
22. Liu, Y., Safavi, T., Dighe, A., and Koutra, D. Graph summarization methods and applications: A survey. *ACM Comput. Surv.* 51, 3 (2018) 62:1–62:34.
23. Pacaci, A., Bonifati, A., and Özsu, M.T. Regular path query evaluation on streaming graphs. *SIGMOD Conf.* (2020), 1415–1430.
24. Papadopoulos, A.V. et al. Methodological principles for reproducible performance evaluation in cloud computing. *IEEE Trans. Software Engineering* (2020), 93–94.
25. Sahu, S. et al. The ubiquity of large graphs and surprising challenges of graph processing: Extended survey. *Proc. VLDB Endow. J.* 29, 2 (2020), 595–618.
26. Saleem, M. et al. How representative is a SPARQL benchmark? An analysis of RDF triplestore benchmarks. *WWW Conf.* (2019), 1623–1633.
27. Salihoğlu, S. and Özsu, M.T. Response to “Scale up or scale out for graph processing.” *IEEE Internet Computing* 22, 5 (2018), 18–24.
28. Siek, J.G., Lee, L.Q., and Lumsdaine, A. The boost graph library: User guide and reference manual. Addison-Wesley (2002).
29. Uta, A., Varbanescu, A.L., Musaafir, A., Lemaire, C., and Iosup, A. Exploring HPC and big data convergence: A graph processing study on Intel Knights Landing. *CLUSTER* (2018), 66–77.
30. Zhao, P. and Han, J. On graph query optimization in large networks. In *Proc. VLDB Endow.* 3, 1 (2010), 340–351.

Sherif Sakr was a professor at the Institute of Computer Science at University of Tartu, Estonia. He passed away on March 25, 2020 at the age of 40.

Angela Bonifati (angela.bonifati@univ-lyon1.fr) is a professor at Lyon 1 University and Liris CNRS in Villeurbanne, France.

Hannes Voigt is a software engineer at Neo4j, Germany.

Alexandru Iosup is a professor at Vrije Universiteit Amsterdam and a visiting professor at Delft University of Technology, The Netherlands.

 This work is licensed under a <https://creativecommons.org/licenses/by-nc-sa/4.0/>



Watch the authors discuss this work in the exclusive *Communications* video. <https://caacm.acm.org/videos/the-future-is-big-graphs>

DOI:10.1145/3434641

Organizational distrust, not compensation, is more likely to send IT pros packing.

BY KELLY IDELL, DAVID GEFEN, AND ARIK RAGOWSKY

Managing IT Professional Turnover

IT EMPLOYEE TURNOVER is a major concern of CIOs and senior IT managers.³² It has been for many years. The most recent annual Society for Information Management (SIM) survey of IT managers confirms that concern and attempts to get to the bottom of the issue. Since 2014, IT employee turnover has been on the rise—9% in 2014, 8.6% in 2015, 8% in 2016, 7.3% in 2017, and 8.2% in 2018, with 69.9% of those being voluntary. As if that were not troublesome enough, 6.9% of the IT workforce is projected to retire in the next five years.³²

These trends have a direct impact on the bottom line of organizations employing IT professionals. Some managers believe the ‘revolving door’ of IT adds an estimated 20% to their expected costs. Overall, compensation accounts for 35%–37% of the entire IT budget.³² When an employee leaves a company, it bears the burden of selecting, recruiting, and training a replacement,⁵ which amounts to up to 150% of the employee’s annual salary, considering the time spent to search for, recruit, and interview a replacement.²⁹

Maybe partially reflecting these costs, the percent of IT budget spent on training has accordingly been rising in recent years, standing at a projected 5.9% in 2019, up from 5.1% in 2018 and 2.9% in 2017.³²

When trying to explain why IT professionals are leaving, IT managers surveyed by SIM blame a strong job market. Indeed, IT talent retention has consistently been ranked in the survey as the second or third “Most Important/Worrisome IT Management Issues” since 2013. Only the combined category of security/cybersecurity/privacy outranks it.³²

This survey, which also looks at the issue from a socio-psychological angle, finds IT employee trust—in direct managers, teams, and organizations—as the top motivating factor. This perspective compares and contrasts IT employee distrust in the broad organization—which is often beyond the control of managers but toxic to work environments nonetheless¹⁷—with the constructive actions that IT managers can take to offset that distrust, namely by fostering trust in managers as well as the teams on which IT employees work. This comparison is made within the context of assessing those trust and distrust beliefs against things such as satisfaction with pay, having perceived alternatives, and a sense of obsolescence in one’s current position.

Results show that when all the correlations are analyzed together, turnover intentions correlate significantly only to *distrust in the organization*. That is not to say that having perceived

» key insights

- IT professional turnover is a continuing major concern in the industry.
- Turnover is related to increased distrust in the organization, a distrust that is increased when IT professionals feel their present position increases their obsolescence, possibly because it suggests their managers do not care to keep them up to date.
- IT managers can reduce that distrust also by increasing trust within the IT team.



alternatives, feeling obsolete, or lacking trust in managers or teams does not significantly correlate to turnover intentions. Rather, they paled in comparison to such an extent that they became insignificant when analyzed together with distrust in the organization. As important as they may be, they are actually mediated through reduced distrust. Other demographics, including age, gender, education level, organization size, and organizational tenure mentioned in other studies, for instance Zaza et al.,³⁵ were mostly insignificant.

Trust and Distrust

So why is distrust in the organization, namely, in the broader organizational environment, so important? To understand that one must first look at what trust is about. Trust is “the willingness

of a party to be vulnerable to the actions of another party based on the expectation the other will perform a particular action important to the trustor, irrespective of the ability to monitor or control that other party.”²² Such trust is crucial in many contexts involving intangible social exchanges.

To understand what those intangible social exchanges are and why trust is so crucial in them, one must first look at what a tangible exchange is. A tangible exchange is like going to the store to buy a pint of milk; you know exactly what you are getting and exactly what you are paying. There is little to no emotional or trusting element involved in a tangible exchange because the rules are set in advance and there is no dependency on the goodwill of the other party.

In contrast, in a social exchange the rules are not explicit and are in-

tangible, even if the costs and benefits can still be substantial. An intangible exchange may be telling an IT employee how important she is and showing it by respecting her opinion. The value of that respect is clearly higher than the equivalent minor dollar cost of the time it takes the manager to do so, but it is invaluable. Likewise, showing that the manager is trustworthy and can be counted on to treat employees with honesty, benevolence, and capability²² is something that cannot be easily measured. Nor can one contract that trustworthiness, because there are no explicit rules of reciprocity in this case.


You cannot tell your IT employees that you will be honest and care about them in exchange for their extra commitment and loyalty. It is expected, but not explicit. Doing so builds trust by

confirming one's trustworthiness; not doing so breeds distrust by creating suspicion.¹⁷ Moreover, if one were to make that exchange explicit, as in, "I will care for you but in exchange you must work a bit harder," the ensuing relationship would be anything but one based on trust.


Compensation is a tangible, mostly economic, exchange, but relying on one's manager, knowing that the team cares about you, and reciprocating accordingly is what makes it a social exchange. And this is the catch: it is precisely because there are no explicit and enforced rules in a social exchange that it is so dependent on creating trust through reciprocity.^{2,11} And, by an equivalent measure, lack of reciprocity can create distrust. Distrust is one of the reasons why organizations fail,¹⁷ as well as why countries do not develop economically.⁹ Combined, trust enables and determines how employees perceive the fairness of their organization,⁶ while distrust ruins it.^{9,17}

Practically, this is not to say that typical IT managers should refrain from creating distrust. Sometimes it is just inevitable that the organization as a whole might unintentionally create some level of distrust—because it is so remote and its actions are not always broadcasted or due to the intangible aspects of its social exchanges with the IT employees. For example, an organization inadvertently creates distrust by not making it clear how much the endeavors and overtime invested by IT employees are appreciated. Fortunately, an IT manager can, to some extent, counter that negativity by creating trust in the more immediate environment of the manager and team.

Trust and distrust are not necessarily opposites. When we trust, we make assumptions about how others will behave when they cannot be enforced.¹⁰ This is often based on how trustworthy those others were in the past.^{22,26} Making assumptions is necessary in many cases because people are, in essence, free agents and not always even rational ones at that. Without assuming that others will behave in an acceptable manner—in other words, trusting them—the social world would often be cognitively overwhelming. Trust allows one to assume away many possible behaviors by others, and in doing so, re-



When all the correlations are analyzed together, turnover intentions correlate significantly only to distrust in the organization.



duce the otherwise overwhelming complexity of the social world.^{10,20}

Distrust, in contrast, is an emotional aversion.²³ More on distrust can be found in Bobko et al.,³ on the neuroscience aspect and how trust and distrust are related in that line of research in Krueger et al.,¹⁹ on the societal effect on trust in Zhang et al.,³⁶ on the interplay between institutional and personal trust in Vries et al.,³⁴ and the interplay between cognitive and affective trust in Fan et al.⁸ Research has indeed shown that, accordingly, trust is associated with neural correlates of brain regions known to be involved in rational decision-making, while distrust is associated with those dealing with fear and aversion.⁷

Collaboration is the result of trusting those one interacts with. Seclusion and withdrawal are the result of distrust in the broad social context.^{9,17} Distrust can make employees assume malicious intentions even when they were not intended¹⁸ and may result in a loss of organizational control as employees are drawn into passivity, secrecy, isolation, avoidance, blame,¹⁸ cynicism, and lack of motivation.²⁸

The Study

This study, which includes 793 organizations representing 23.3% of the U.S. GDP, looks at turnover intention through those lenses, comparing the importance of the trust that IT employees have in those they work closely with—their manager and team—with their overall distrust in the organization. The items in each section of the questionnaire were rated on a seven-point Likert scale dealing with *Turnover Intention*,⁴ *Satisfaction with Compensation*,¹³ *Trust in Direct Manager*,⁶ *Trust in Team*,^{1,27} *Distrust in Organization*,^{14,30} *Perceived Threat of Professional Obsolescence*,¹⁶ and *Perceived Alternatives*.³³ Demographics known to also affect turnover intentions³⁵ were added as controls. In this conceptualization, it is assumed that trust-building acts by IT management might offset inadvertent distrust-creating actions by the organization, such as creating an impression of unfairness.

Data was collected through Qualtrics, an online survey distribution company. Each potential participant re-

ceived a recruitment letter via email with a link to a survey, where participants were informed that participation was anonymous and optional. Participants were paid to participate in the survey, which took 5–12 minutes. The survey, restricted to individuals located in the U.S. who self-identified as an “IT professional,” was open for two weeks, during which 258 completed responses were collected. Of the 258 respondents, 67% were married, most aged 25–34 (36%) and 35–44 (43%). Not surprising for IT professionals, most respondents were male (71%) and had a four-year (43%) or professional degree (28%). Organizational tenure was mostly 5–10 years (40%), 10–15 years (24%), and 0–5 years (22%).

Data was analyzed using MPlus,²⁴ a structural equation modeling package that enables analyzing models in which there are many layers of dependent (predicted) and independent (predictor) variables, where those variables may include both explicit measures (such as age) and latent constructs (such as trust) that cannot be measured directly but can be measured as they are reflected through several items in a questionnaire. The latent constructs in Table 1 are shown in bold italics, with the items reflecting each construct below the construct’s name. MPlus runs a simultaneous maximum likelihood estimation of the entire model, including both the measurement model (as a confirmatory factor analysis of how measurement items load on their assigned latent constructs) and the structural model (how the constructs relate to each other). The questionnaire items and their standardized loadings in the model appear in Table 1.

All the loadings are significant at the .001 level. Table 2 shows the descriptive statistics of the resulting constructs while Table 3 shows the correlations among the constructs. In both Table 3 and Figure 1, a single asterisk means the path is significant at the .05 level, a double at .01, and a triple at .001. Overall model fit was $\chi^2_{341}=495.20$, RMSEA=.059, CFI=.94, TLI=.93. These values show overall good fit.¹² The degree of explained variance (R^2) was .73 for *Turnover Intentions* and .62 for *Dis-trust in Organization*.

The analysis included all those

Table 1. Items by scale and their standardized loadings.

Construct/Items Wording	Loading (std.)
<i>Turnover Intentions</i>	
How often have you considered leaving your job?	.86 (.04)
How often are you frustrated when not given the opportunity at work to achieve your personal work-related goals?	.83 (.04)
How often do you dream about getting another job that will better suit your personal needs?	.82 (.04)
<i>Satisfaction with Compensation</i> (Scale was ‘Extremely Satisfied’ to ‘Extremely Unsatisfied’)	
My take-home pay	.93 (.02)
My current salary	.94 (.01)
My overall level of pay	.96 (.01)
<i>Perceived Threat of Professional Obsolescence</i>	
I fear technical obsolescence	.87 (.03)
I feel intimidated	.88 (.03)
I feel the threat of obsolescence	.95 (.02)
<i>Trust in Direct Manager</i>	
I would be comfortable giving my direct manager a task or problem that was critical to me, even if I could not monitor his/her actions.	.74 (.06)
If someone questioned my direct manager’s motives, I would give my direct manager the benefit of the doubt.	.78 (.05)
I would be willing to let my direct manager have complete control over my future in this company.	.76 (.05)
<i>Trust in Team</i>	
Members of my team show a great deal of integrity.	.88 (.02)
I can rely on those with whom I work in this team.	.90 (.02)
Overall, the people in my team are very trustworthy.	.89 (.02)
We are usually considerate of one another’s feelings in this team.	.89 (.02)
The people in my team are friendly.	.83 (.03)
We have confidence in one another in this team.	.84 (.03)
<i>Dis-trust in Organization</i>	
I am not sure I fully trust my organization.	.83 (.04)
My organization is not always honest and truthful.	.81 (.04)
I don’t think my organization treats me fairly.	.80 (.04)
<i>Perceived Alternatives</i>	
I have many alternative job opportunities including some that are different from what I do now.	.82 (.05)
There are many jobs available similar to mine.	.80 (.05)
I can find another job doing exactly what I am doing now.	.67 (.06)

paths, but in the interest of readability, Figure 1 shows only the significant paths. This means that *Dis-trust in Organi-zation* fully mediated the effects of the other independent variables and the controls. Paths that do not appear in Figure 1, such as between *Satisfaction with Compensation* and between *Turn-over Intentions*, are insignificant. (Add-ing correlations between any of the con-

trols and any of the independent variables into the model in Figure 1 shows that in addition to what is shown in Figure 1, there were significant correlations between *Gender and Satisfaction with Compensation*, *Gender and Profes-sional Obsolescence*, and *Satisfaction with Compensation and Professional Ob-solescence*.)

The model in Figure 1 indicates

that distrust mitigates the effects of the other reasons leading to intended turnover. Practically, those results may be interpreted as suggesting that distrust of the organization causes employees to want to leave, while trusting one's more immediate team and other positive experiences and actions can counter that distrust and indirectly alleviate that intent. On a practical level,

distrust also highly correlates with a sense of being professionally obsolete or with employees feeling they have alternatives, possibly indicating telltale signs that managers may wish to pay attention to.

Identifying such employees, and taking steps to keep them from suddenly leaving, resonates with input we received from senior IT managers who

complain their employees sometimes do not bother with their contractual requirement of giving a 14-day leave notice.

The Takeaway

Throwing money at a problem is an easy and tempting approach. Or, in this case, arguing that IT employees leave because they are not satisfied with their compensation. That may be true, but satisfaction with compensation is not the reason in this survey. In fact, the data suggests that in the context of IT employee turnover, it might be primarily a matter of the organization's overall policies that result in distrust and push employees away, rather than about convincing them to stay through issues that are at the direct discretion of their managers, such as building trust in the team they manage.

Such a conclusion is perhaps reminiscent of other research on employee turnover, which argued that organizational commitment, arguably missing when the employee distrusts the organization, reduces turnover intentions.²¹ Nonetheless, taking steps to build trust within the IT team, something CIOs and IT managers can take on, can reduce distrust and, hence, turnover intentions. Moreover, as transparency increases trust,³¹ managers may wish to consider that line of action too when it comes to reducing distrust. As a caveat to that conclusion, though, it should be added that the direction of the implied causality among the perception of having alternatives or professional obsolescence and distrust might be more complex

Table 2. Construct descriptive statistics table.

Variable	Mean	Std.	Min	Max
Turnover Intentions	3.61	1.67	1	7
Satisfaction with Compensation	2.47	1.41	1	7
Perceived Threat of Professional Obsolescence	4.16	1.89	1	7
Trust in Direct Manager	2.53	1.23	1	7
Trust in Team	2.13	1.09	1	6.83
Distrust in Organization	4.27	1.80	1	7
Perceived Alternatives	2.94	1.37	1	7

Figure 1. Structural model.

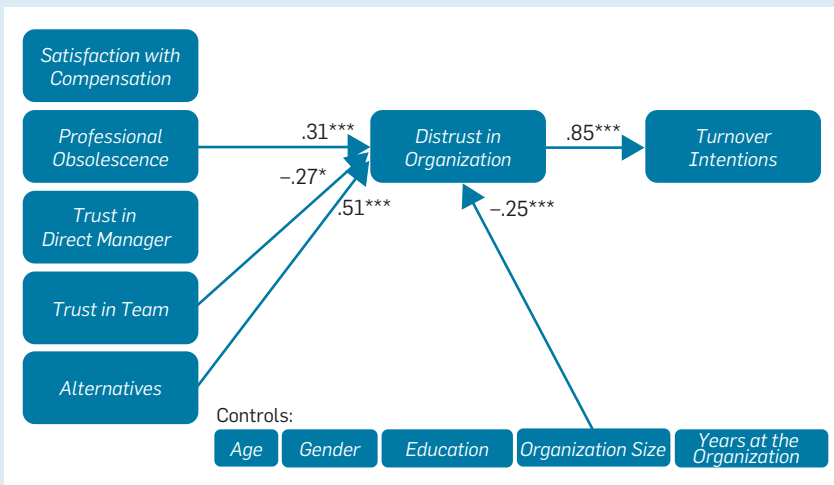


Table 3. Correlation among the constructs.

	Turnover intentions	Compensation Satisfaction	Threat of Obsolescence	Trust in Manger	Trust in Team	Distrust in Org.	Perceived Alternatives
Turnover Intentions	1						
Satisfaction with Compensation	-.04	1					
Perceived Threat of Professional Obsolescence	.49***	.24***	1				
Trust in Direct Manager	-.13*	.56***	.15*	1			
Trust in Team	-.18**	.50***	.11	.62***	1		
Distrust in Organization	.66***	-.12	.44***	-.16**	-.30***	1	
Perceived Alternatives	.43***	.23***	.34***	.21***	.18**	.36***	1
Age	.18**	-.02	.16*	-.05	-.06	.16**	.09
Org. Size	-.09	-.07	.04	.06	.02	-.15*	.01
Org. Tenure	.11	-.12	-.02	-.12	-.04	.06	.03
Gender	.15*	.16*	.10	.06	.04	.04	.10
Education	-.08	-.14*	-.18**	-.11	-.08	-.09	-.15*

than discussed. It should also be added as a note of caution that common method variance is unavoidable in questionnaire data.

Also noteworthy is that *Satisfaction with Compensation* is not only an insignificant predictor of turnover intentions when other factors are included (see Figure 1), but it is even uncorrelated with turnover intentions (see Table 3). This may suggest a need to reconsider, at least in this context, previous suggestions that pay leads to job satisfaction which leads to less turnover.^{15,20} Again, it is distrust in the organization that pushes IT employees to consider leaving, rather than pay that convinces them to stay. That it is such a broad, organization-wide issue which pushes IT employees away may also be reflected in the demographics.

As might be expected, employees in larger organizations felt more distrust in their organization (see Table 3). Such a tendency may reflect the consequences of more interpersonal trust-building relationships in smaller organizations compared to the perhaps inevitable greater social distance and disconnect with decision-making in large organizations. The high correlations between *Trust in Direct Manager* and both *Satisfaction with Compensation* and *Trust in Team* merit future research.

The fact that trust in the team reduced overall distrust in the organization may have broader implications for IT managers about the need to better manage the interpersonal relationships within these teams. That trust in the team is highly correlated with trust in the manager may not be that surprising because a key function of any successful manager is team building.

But that trust in the team is highly correlated with overall pay satisfaction is suggestive. A team that is trustworthy allows people to grow and improve their skills, and that is important to people. Satisfaction with compensation, then, might not be just about money; rather, it may also be related to being part of a team one can trust. Being part of such a team also correlates with the perception of having alternatives, adding to the possible risk that an employee might leave, but the strength of the correlations in Table 3

suggests that being part of a trustworthy team carries more weight in affecting satisfaction with compensation than with increasing perceived alternatives. What constitutes a trustworthy team is reflected in the questions that measured it. Trustworthy IT teams are those where their members feel that there is integrity, dependability, and friendliness.

As Ian Fleming once put it: “Surround yourself with human beings, my dear James. They are easier to fight for than principles.” That truism might apply to managing IT employee retention, too. Management principles such as compensation,^{15,25} high-level managerial policies, and constructive behavior⁶ are clearly important. But, at their core, IT employees are human. Fighting for them to retain them is a matter of creating an organization they do not distrust enough to want to leave. **C**

References

1. Allen, N.J. and Meyer, J.P. The measurement and antecedents of affective, continuance and normative commitment to the organization. *J. Occupational Psychology* 63, 1 (1990), 1-18.
2. Blau, P.M. *Exchange and Power in Social Life*. J. Wiley, (1964).
3. Bobka, P., Barelka, A.J., and Hirshfield, L.M. The construct of state-level suspicion: A model and research agenda for automated and information technology (IT) contexts. *Human Factors* 56, 3 (2013), 489-508.
4. Bothma, C.F. and Roodt, G. The validation of the turnover intention scale. *SA J. Human Resource Management* 11, 1 (2013), 1-12.
5. Boushey, H. and Glynn, S.J. There are significant business costs to replacing employees. *Center for American Progress*, (2012), 16.
6. Colquitt, J.A. and Rodell, J.B. Justice, trust, and trustworthiness: A longitudinal analysis integrating three theoretical perspectives. *Academy of Management J.* 54, 6 (2011), 1183-1206.
7. Dimoka, A. What does the brain tell us about trust and distrust? Evidence from a functional neuroimaging study. *MIS Quarterly* 34, 2 (2010), 373-377.
8. Fan, H. and Lederman, R. Online health communities: How do community members build the trust required to adopt information and form close relationships? *European J. of Information Systems* 27, 1 (2017), 62-89.
9. Fukuyama, F. *Trust: The Social Virtues and the Creation of Prosperity*. The Free Press, (1995).
10. Gefen, D., Karahanna, E., and Straub, D.W. Trust and TAM in online shopping: An integrated model. *MIS Quarterly* 27, 1 (2003), 51-90.
11. Gefen, D. and Pavlou, P.A. The boundaries of trust and risk: The quadratic moderating role of institutional structures. *Information Systems Research* 23, 3 (2012), 940-959.
12. Gefen, D., Rigdon, E., and Straub, D.W. An update and extension to SEM guidelines for administrative and social science research. *MIS Quarterly* 35, 2 (2011), III-IV.
13. Heneman III, H.G. and Schwab, D.P. Pay satisfaction: Its multidimensional nature and measurement. *Intern. J. of Psychology* 20, 1 (1985), 129-141.
14. Igbaria, M. and Siegel, S.R. The reasons for turnover of information systems personnel. *Information & Management* 23, 6 (1992), 321-330.
15. Joseph, D., Ng, K.-Y., Koh, C., and Ang, S. Turnover of information technology professionals: A narrative review, meta-analytic structural equation modeling, and model development. *MIS Quarterly* 31, 3 (2007), 547-577.
16. Joseph, D., Tan, M.L., and Ang, S. Is updating play or work?: The mediating role of updating orientation

- in linking threat of professional obsolescence to turnover/turnaway intentions. *Intern. J. of Social and Organizational Dynamics in IT* 1, 4 (2011), 37-47.
17. Kanter, R.M. Leadership and the psychology of turnarounds. *Harvard Business Rev.* 81, 6 (2003), 58-67.
18. Kramer, R.M. The sinister attribution error: Paranoid cognition and collective distrust in organizations. *Motivation and Emotion* 18, (1994), 199-230.
19. Krueger, F. and Meyer-Lindenberg, A. Toward a model of interpersonal trust drawn from neuroscience, psychology, and economics. *Trends in Neurosciences* 49, 2 (2019), 92-101.
20. Luhmann, N. *Trust and Power*. John Wiley and Sons, (1979).
21. Mathieu, C., Fabi, B., Lacoursière, R., and Raymond, L. The role of supervisory behavior, job satisfaction and organizational commitment on employee turnover. *J. Management & Organization* 22, 1 (2016), 113-129.
22. Mayer, R.C., Davis, J.H., and Schoorman, F.D. An integrative model of organizational trust. *Academy of Management Rev.* 20, 3 (1995), 709-734.
23. McKnight, H.D. and Choudhury, V. Distrust and trust in B2C e-commerce: Do they differ? In *Proc. of the Intern. Conf. Electronic Commerce*, (2006).
24. Muthén, L.K. and Muthén, B.O. *Mplus 6 User's Guide*. Muthén and Muthén, (2010).
25. Paré, G. and Tremblay, M. The influence of high-involvement human resources practices, procedural justice, organizational commitment, and citizenship behaviors on information technology professionals' turnover intentions. *Group & Organization Management* 32, 3 (2007), 326-357.
26. Pavlou, P.A. and Gefen, D. Psychological contract violation in online marketplaces: Antecedents, consequences, and moderating role. *Information Systems Research* 16, 4 (2005), 372-399.
27. Pearce, J., Sommer, S., Morris, A., and Friderger, M. A configurational approach to interpersonal relations: Profiles of workplace social relations and task interdependence. *University of California, Irvine Working Paper No. GSM# 0892015*, (1992).
28. Pugh, S.D., Skarlicki, D.P., and Passell, B.S. After the fall: Layoff victims' trust and cynicism in re-employment. *J. Occupational and Organizational Psychology* 76, 2 (2003), 201-212.
29. Righoni, B. and Nelson, B. Many millennials are job-hoppers—but not all. *Gallup*, 2016.
30. Robinson, S.L. and Rousseau, D.M. Violating the psychological contract: Not the exception but the norm. *J. of Organizational Behavior* 15, 3 (1994), 245-259.
31. Schnackenberg, A.K. and Tomlinson, E.C. Organizational transparency: A new perspective on managing trust in organization-stakeholder relationships. *J. Management* 42, 7 (2020), 1784-1810.
32. The SIM IT trends study: 2018 comprehensive report: Issues, investments, concerns, & practices of organizations and their IT executives. *SIM*, (2019).
33. Thatcher, J.B., Stepina, L.P., and Boyle, R.J. Turnover of information technology workers: Examining empirically the influence of attitudes, job characteristics, and external markets. *J. Management Information Systems* 19, 3 (2002), 231-261.
34. Vries, J.R.D., Zee, E.v.d., Beunen, R., Kat, R., and Feindt, P.H. Trusting the people and the system. The interrelation between interpersonal and institutional trust in collective action for agri-environmental management. *Sustainability* 11, 24 (2019), 1-18.
35. Zaza, I. and Armstrong, D.J. Information technology professionals' turnover intentions: A meta-analysis of perceived organizational factors, (2017).
36. Zhang, Y. and Xin, Z. Rule comes first: The influences of market attributes on interpersonal trust in the marketization process. *Special Issue: The Social Psychology of Neoliberalism* 75, 1 (2019), 286-313.

Kelly Idell recently received a Doctorate of Business Administration from Drexel University, Philadelphia, PA, USA.

David Gefen (gdfend@drexel.edu) is academic director of the Doctorate in Business Administration (DBA) Program and provost distinguished research professor at Drexel University, Philadelphia, PA, USA.

Arik Ragowsky is an associate professor of Information Systems Management in the Mike Ilitch School of Business at Wayne University, Detroit, MI, USA.

Copyright held by authors/owners.

DOI:10.1145/3433637

As robots begin to interact closely with humans, we need to build systems worthy of trust regarding the safety and quality of the interaction.

BY HADAS KRESS-GAZIT, KERSTIN EDER, GUY HOFFMAN, HENNY ADMONI, BRENNAN ARGALL, RÜDIGER EHLERS, CHRISTOFFER HECKMAN, NILS JANSEN, ROSS KNEPPER, JAN KŘETÍNSKÝ, SHELLY LEVY-TZEDEK, JAMY LI, TODD MURPHEY, LAUREL RIEK, AND DORSA SADIGH

Formalizing and Guaranteeing Human-Robot Interaction

ROBOT CAPABILITIES ARE maturing across domains, from self-driving cars to bipeds to drones. As a result, robots will soon no longer be confined to safety-controlled industrial settings; instead, they will directly interact with the general public. The growing field of human-robot interaction (HRI) studies various aspects of this scenario, from social norms to collaborative manipulation to human-robot teaming, and more.

Researchers in HRI have made great strides in developing models, methods, and algorithms for robots acting with and around humans,²⁹

but these “computational HRI” models and algorithms do not generally come with formal guarantees and constraints on their operation. To enable human-interactive robots to move from the lab to real-world deployments, we must address this gap.

Demonstrating trustworthiness in various forms of automation through formal guarantees has been the focus of validation, verification, and synthesis efforts for years. For instance, aircraft control systems must meet guarantees—such as correctly handling transitions between discrete modes of operation (take-off, cruise, landing)—while simultaneously providing a guarantee on *safety* (for example, not being in both take-off and landing modes at the same time) and *liveness*, the ability to eventually achieve a desirable state (for instance, reaching cruise altitude).

Formal methods, such as model checking and theorem proving, play a central role in helping us understand when we can rely on automation to do what we have asked of it. Formal methods can be used to create correct-by-construction systems, provide proofs that properties hold, or find counterexamples that show when automation might fail.

Formalizing HRI can enable the creation of trustworthy systems and, just as importantly, support explicit reasoning about the context of guarantees. First, writing formal models of aspects of HRI would enable verification, validation, and synthesis, thus

» key insights

- **To obtain formal guarantees, specifications that capture the requirements and constraints of “good” HRI are essential.**
- **Our analysis of different HRI domains highlights the importance of understanding human behavior for successful HRI.**
- **HRI brings unique challenges and opportunities for formal methods, from defining specifications to dealing with adaptation and variability in humans.**

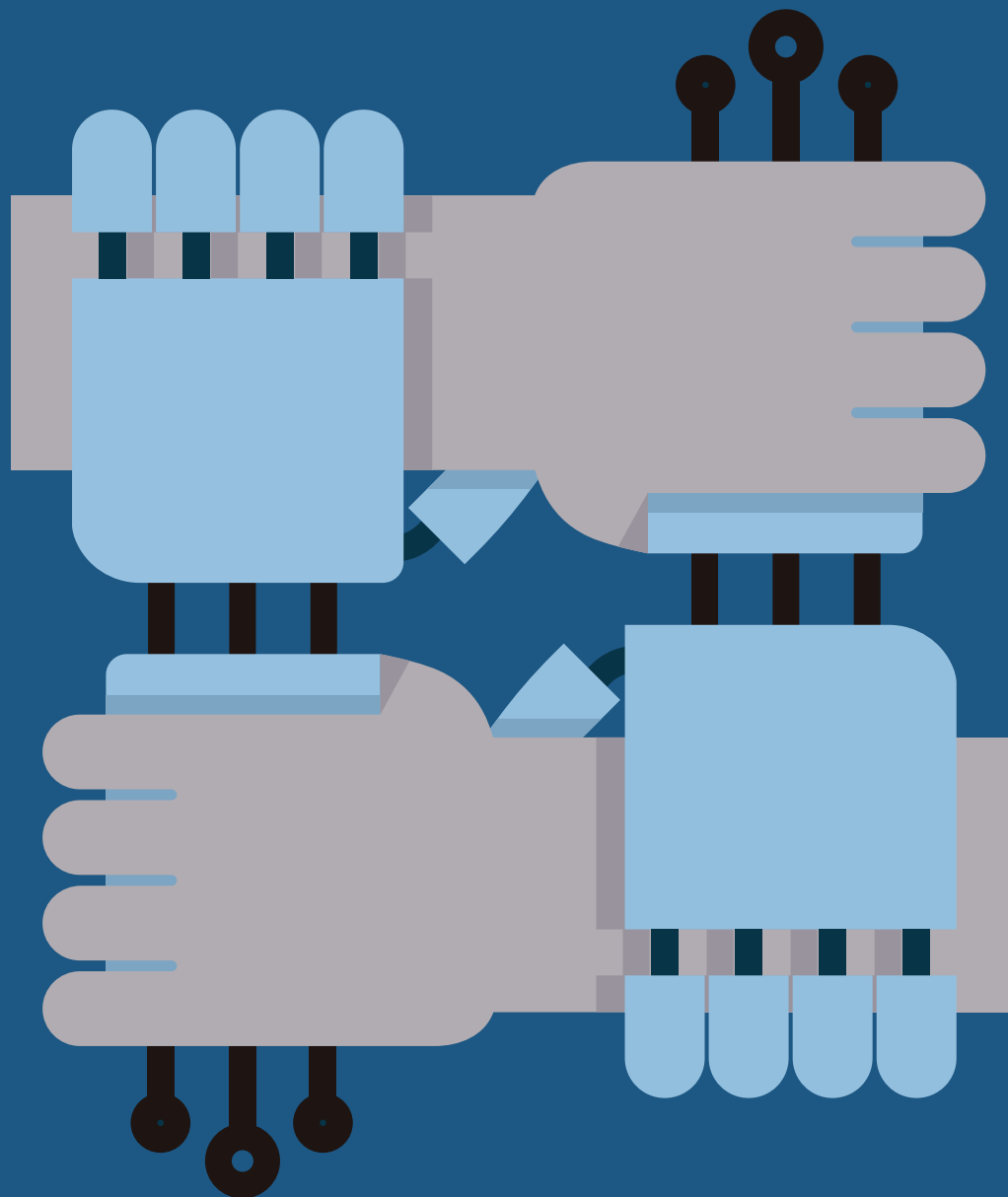


IMAGE BY PEDRO NEVES DESIGN

providing some guarantees on the interaction. Second, it is unrealistic to verify a complete human-robot system due to the inherent uncertainty in physical systems, the unique characteristics and behaviors of people, and the interaction between systems and people. Thus, a formal model requires us to articulate explicit assumptions regarding the system, including the human, the robot, and the

environments in which they are operating. Doing so exposes the limits of the provided guarantees and helps in designing systems that degrade gracefully when assumptions are violated.

In this article, we discuss approaches for creating trustworthy systems and identify their potential uses in a variety of HRI domains. We conclude with a set of research challenges for the community.

Techniques for Demonstrably Trustworthy Systems

We divide the techniques for gaining confidence in the correctness of a system into four approaches: synthesis of correct-by-construction systems, formal verification at design time, runtime verification or monitoring, and test-based methods. Common to all of these approaches is the need to articulate specifications—descriptions of


what the system should and should not do. Specifications typically include both safety and liveness properties and are defined in a formal language, for example temporal logics over discrete and/or continuous states, or satisfiability modulo theory (SMT) formulas (for example in Clarke et al.⁸).

The four approaches outlined below are listed in decreasing order of exhaustiveness and, as a result, computational complexity. Less exhaustive approaches can typically handle more complex systems at a greater level of realism. Synthesis is the most computationally expensive approach and requires the coarsest abstraction, but it can automatically create a system with guarantees. Test-based methods, however, can handle the most complex systems but do not provide formal guarantees regarding the satisfaction of the specifications. In practice, a combination of techniques is required, as no single technique can be relied upon on its own.³¹


Synthesis is the process of automatically generating a system from the specifications. In the context of robotics, there are different techniques for doing so,¹⁴ including offline generation of state machines or policies satisfying discrete and probabilistic temporal logic specifications, online receding horizon optimization for continuous temporal logics, and online optimization with SMT solvers.

Formal verification techniques span methods that exhaustively explore the system (model checking, reachability analysis⁸) to those that reason about the system using axioms and proof systems (theorem proving¹⁰). Techniques vary from deterministic, worst-case analysis to probabilistic reasoning and guarantees, and from discrete spaces to continuous ones. Such methods are typically applied at design time and either determine that the specification is met in every possible system behavior or provide a counterexample—a system execution that violates the specification—which may then be used to further refine the design or the specification.

Runtime monitoring is the process of continuously checking system accuracy during execution using *monitors* that check specifications, either



When the interaction involves shared human-robot control, equally important to the idea of humans trusting the robot is the notion of whether and to what extent the robot can trust the human.



created manually or automatically through synthesis.²⁰ This type of verification is, in a sense, the most lightweight way to integrate formal methods into a design. It does not alter the design; it enables the detection of failures or deviations from expected/formalized behavior, allowing the robot to be shut down or switched into a safe mode. An additional benefit of runtime-checkable specifications is that they allow us to “probe” the system at design time using methods such as statistical model checking.¹⁵

Test-based methods complement formal methods during verification and validation. In particular, simulation-based testing² can expose the system under test to more realistic stimuli than the often highly abstract scenarios that can be verified formally. From a performance point of view, simulation-based testing can reach verification goals faster and with less effort than conventional testing in the real world. *Coverage* is a measurement of verification progress, allowing engineers to track the variety of tests used and determine their effectiveness in achieving verification goals. Assertion monitors act as test oracles, much like the monitors used for runtime verification. Model-based testing is a specific technique that asserts the conformance of a system under test to a given formal model of that system.³⁰ This is particularly important when guarantees or code generation rely on the correctness of a model.

Validation, verification, and synthesis techniques are always related to given specifications. These specifications can never cover the full behavior of a physical system in the world; rather, they include assumptions and abstractions to make the problem tractable. Therefore, guarantees are provided with respect to the specification, enabling us to gain confidence in the system’s overall correctness, narrow down the sources of problems, and understand the constraints that limit deployment.

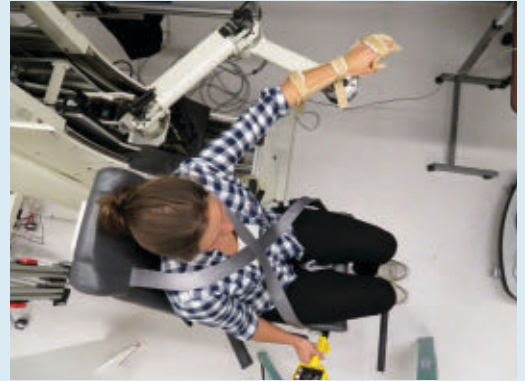
HRI Domains and Their Unique Challenges

Many HRI domains could benefit from formal methods, and each domain brings about unique challenges:

Domains of HRI that could benefit from formal methods.

Clockwise from top left, physical HRI (construction), physical HRI in healthcare (rehabilitation), autonomous driving, social HRI, and cognitive healthcare.

Images courtesy of (clockwise from top left): Michael Suguitan, Kathleen Fitzsimons, Wendy Ju, Yaxin Hu, and The Levy-Tzedek Lab.



Physical HRI involves systems in which the physical states of an automation interact with the physical states of a human;⁴ for example, a robotic wheelchair carrying a person or a construction-assistant robot and a person carrying a heavy load together. In addition to physical states interacting, their *internal* states interact, since both the robot and human often have a model of the task they are working on to achieve as well as a model of each other.

For example, in a setting where rehabilitation robots assist an individual with motion, the robot may be responsible for physical safety (keeping someone upright) while simultaneously maximizing therapy benefit, requiring it to stay out of the way as much as possible. Thus, the system is tasked to assist, but not to over-assist. This fundamental tension between the two purposes of the automation with respect to the human leads to challenging questions in terms of specification (for example, how does one articulate the notion of safety while avoiding over-assisting?) and verification (for instance, how does one prove that the control methods satisfy the specification?).

Healthcare Robotics: There are a variety of robots being developed to assist people with activities of daily living, including physical mobility, manipulation, medicine reminders, and cognitive support. Robots are also developed to support clinicians, caregivers, and other stakeholders in healthcare contexts.²⁷

For example, physically assistive robots, such as exoskeletons and robotic prostheses, can help individuals perform movements, such as walking and reaching. Socially assistive robots can help individuals engage in positive health behaviors, such as exercise and wellness.²³ People have different abilities and disabilities that may change over short- and long-term horizons. Therefore, modeling a person's ability and personalizing the system is crucial for creating successful HRI systems in the healthcare domain.

Autonomous Driving: Recent years have seen significant advances in autonomous driving. As these fully or semi-autonomous vehicles appear on the road, challenges arise due to interactions with humans. Humans, in the context of autonomous driving, fall into three main categories: drivers or riders in the autonomous vehicle,

drivers of other vehicles around the autonomous car, and pedestrians or bicyclists interacting with autonomous vehicles on roads.

An obvious specification in this domain is safety—no collisions. However, that specification is not enough. When merging onto a highway, the safest course of action is to wait until no other vehicles are nearby. On busy roads this is not a reasonable course of action. Therefore, the specification needs to go beyond addressing the challenges of driving a single vehicle by formalizing desirable behavior when cars interact with other vehicles and road users.²⁸ The challenges of this domain are to model and validate acceptable road behavior; reason about the expected and unexpected behavior of people in the above categories; and provide certification, diagnosis, and repair techniques that will enable autonomous vehicles to drive on our roads.

Social Collaboration: In addition to the contexts listed above, there are many instances in which humans and robots will engage in predominantly social, rather than physical, interactions.⁵ For example, information-kiosk robots at an airport might engage

people in conversations to get them to their destinations.

Social collaborations across many domains can be characterized by the use of social actions, such as verbal and nonverbal communication, to achieve a shared goal. Social collaboration typically requires the agents involved to maintain a *theory of mind* about their partners, identifying what each agent believes, desires, and aims to achieve. In social collaboration, it is important that the robot follows social norms and effective collaboration practices, for example not interrupting the speaker and providing only true and relevant information.⁹ If a robot fails to follow such conventions, it risks failing at the collaboration due to lack of trust or other social effects. One major challenge in formalizing social collaborations is how to encode social norms and other behavior limitations as formal constraints. Researchers interested in verifying or synthesizing social collaborations will have to identify which social behaviors and which elements of the task are important for the collaboration to succeed.

Work in Formalizing HRI

Researchers in computational HRI²⁹ have developed models for human behavior, for human-robot collaboration and interaction, and algorithms that have been demonstrated in various HRI domains. Whereas these approaches are evaluated qualitatively and quantitatively, the HRI research community has not often formalized what constitutes correct behavior. Generally speaking, there are very few examples of formal specifications, or algorithms that can verify or synthesize such specifications.

In the past few years, collaborations between HRI researchers and researchers studying formal methods, verification, and validation have begun to address the challenge of formalizing specifications and creating demonstrably trustworthy HRI systems. Some efforts have explored linear temporal logic as a formalism to capture and verify norms in an interaction²⁵ and to synthesize human-in-the-loop cyber-physical systems.²¹ Other examples include using satisfiability modulo theories for encoding social navigation specifications,⁷ signal temporal logic for handover behavior,¹⁶ and automata-based

assertion monitors for robot-to-human handover tasks.³

Other researchers have focused on socio-cyber physical systems, for instance by including human factors—ranging from specific roles of humans, their intentions, legal issues, and levels of expertise—into cyber-physical systems.⁶ Other work models an assisted-living scenario as a Markov decision process,²² making use of the probabilistic model-checker PRISM.¹⁷

Challenges for the Research Community

Work described above suggests the promise of introducing formal methods and techniques into HRI domains. That said, creating and reasoning about trustworthy HRI requires addressing HRI's unique aspects and rethinking current approaches to system verification, validation, and synthesis. In this section, we distill three unique aspects of HRI research posing a challenge for formal methods: designing useful HRI specifications, dealing with expected human adaptation to the automated system, and handling the inherent variability of humans. For each challenge domain, we identify high-priority research directions that could drive progress toward creating trustworthy HRI systems.

Designing formal HRI specifications: Whenever verifying, testing, or synthesizing a system, one needs to formalize the system by defining the state space of the model and the specification of interest. For example, in the context of autonomous cars obeying the law and social conventions, the state space may include the position and velocity of the car and any other cars in the environment. The specification may represent a requirement of the form, ‘the car never exceeds the speed limit and always maintains a safe distance from all other cars.’ In the context of HRI, designing useful specifications raises several research questions:

► **What should be the space of specifications?** In HRI, simply modeling the physical state of the robot and the human is usually not enough. The physical state does not capture requirements such as avoiding over-assisting a person or maintaining social and cultural norms. We need to create richer spaces that enable the writing of such

specifications while balancing the complexity of the algorithms that will be used for verification and synthesis in these spaces.

► **How do we write specifications that capture trust?** A human will only trust a robot to react in a safe way if it obviously and demonstrably does so. Hence, the robot needs to not only be safe but also be *perceived* as safe, which may require a considerable safety margin. On the other hand, when the interaction involves shared human-robot control, equally important to the idea of humans trusting the robot is the notion of whether and to what extent the robot can trust the human. This plays a role in determining under what circumstances the robot should step in and in what manner. Particularly in safety-critical scenarios, and when the robot is filling a gap in the human's own capabilities, reasoning about trust in the human is key. Critical factors are to measurably assess the human's ability to actually perform the task, and the human's current state, for instance accounting for levels of fatigue. These notions of trust go beyond typical safety and liveness specifications and require specification formalisms that can capture them.

► **What should be the definition of failure?** Beyond failure with respect to physical safety, which is well studied in the literature, interaction failures may have varying impacts. A small social gaffe, such as intruding on personal space, may not be an issue, but a large mistake, such as dropping a jointly manipulated object, might have a long-term effect on interaction. We need to be able to define specifications that capture the notion of social failure and develop metrics or partial orders on such failures, so that the systems can fail gracefully.

► **How can we formalize the human's behavior during an interaction?** A common technique in verification is assume-guarantee reasoning, where a system's behavior is verified only under the assumption that its input satisfies a well-defined specification. If the input violates the assumption, the system behavior is no longer guaranteed. Given our understanding and observations of human-human and human-robot interactions, a challenge for synthesizing and verifying HRI is to


formalize assumptions on the behavior of the human—who provides the input of the HRI system—in a way that supports verification, is computationally tractable, and captures the unique characteristics of humans.

Adapting to human adaptation: During interaction, humans and robots will engage in mutual adaptation processes.¹² For example, people become less cautious operators of machines (cutting corners, giving a narrower berth to obstacles) as they become more familiar with them. Therefore, any models used to represent the interaction and reason about it must capture this adaptation. To complicate matters, the temporal adaptation may occur at different time scales: short time scales, such as morning vs. evening fatigue, and longer time scales, such as functional ability improvement or deterioration over months.^{11,12} Changing models in and of themselves makes formalizing HRI more complicated, but it is the diversity of the ways humans adapt to a task and a teammate that makes their accurate modeling even more challenging. This property brings up the following research challenges:


► **Which mathematical structures can capture non-stationary models?** Mutual adaptations are common in human-human interaction. For example, humans build conventions when communicating with each other through repeated interactions using language or sketches.³² Studying these interactions and formalizing them can form the basis for new HRI models. When developing such models, an important consideration is how to capture the different time scales of adaptation.

► **How can the robot detect and reason about the human's adaptation?** As humans adapt to the interaction, their behavior (and thus the input to the interaction) may change. For example, people may become less emotionally expressive as the novelty of the interaction wears off, or they may give less control input as they trust the autonomy of the system more. This creates a challenge at runtime when a robot is attempting to ascertain how the human adapted. We need to develop runtime verification algorithms that can detect such adaptation and influence the interaction.

► **How to model feedback loops?** As the robot and the human adapt to each



Modeling a person's ability and personalizing the system is crucial for creating successful HRI systems in the healthcare domain.



other, it is important to reason about the positive and negative feedback loops that emerge and their effect on the resulting interaction. These feedback loops can take the human-robot systems to desirable or undesirable equilibria. For example, the difference between driving cultures around the world may be explained by repeated interactions between drivers causing behavioral feedback loops, leading to emergent locally distinct conventions. We need to study the long-term behavior of repeated interactions and adaptations and verify the safety of the resulting emergent behaviors.

Variability among human interactants: While we can reasonably assume that the model of a particular type of robot is the same for all robots of that type, there does not exist a model of a “typical” human—one size does not fit all. Even identifying the proper parameters or family of parameters that encapsulate the types of variability in people is a seemingly impossible task. People differ across backgrounds, ages, and abilities, which raises the important question of how much to personalize the model and specification to a specific individual or population:

► **Can we identify general specifications for which one, simple human model is enough?** Is it possible to create a basic, human-centric and application-agnostic model of human behavior that indicates a basic specification, such as loss of engagement of a human in the interaction? Such a generic model can detect behavior outside the expected, for example distraction or lack of attention, and could be used to trigger safety measures irrespective of the specific application area. A current example for such a model is used in driver assist systems; they measure where the driver is looking, suggesting the driver take a break if they detect staring or lack of eye movement—universal signs of sleepiness.

► **What levels of personalization are needed?** Refining the research question above, it is important to study not only the formalisms that allow models and specifications to be personalized but also to what extent personalization is required for smooth interaction, what are the trade-offs between the complexity of the model and improved interaction, and what metrics enable

reasoning about the trade-offs. For this purpose, models of mental representations (for example, levels of cognitive control for error-free decision-making²⁶) could be useful.


► **How can we model the human’s ability level?** The interaction should be appropriate for the ability level of the person. When humans are better off completing a task on their own, too much assistance is not desirable; for example, in therapeutic and educational settings. In other cases, too little assistance can be frustrating and lead to disengagement. It is important to model both the ability and the modes of interactions that are most appropriate for each task.

► **How do we formalize experiential considerations?** People from different backgrounds may have different assumptions²⁴ and expectations¹⁹ from robots and may perceive the interaction with the robot differently. Since meeting user expectations is important for fostering trust between the human and the robot,^{13,18} the personalization of the interaction should consider the experiential background of the user, who may expect the robot to be, for example, more assertive and active, or more reserved and passive.

Conclusion

As robots begin to interact closely with humans, we need to build systems worthy of trust regarding both the safety and the quality of the interaction. To do so, we have to be able to formalize what a “good” interaction is, and we need algorithms that can check that a given system produces good interactions or can even synthesize such systems.

To make progress, we must first acknowledge that a human is not another dynamic physical element in the environment, but has beliefs, goals, social norms, desires, and preferences. To address these complexities, we must develop models, specifications, and algorithms that use our knowledge about human behavior to create demonstrably trustworthy systems. In this article, we identified a number of promising research directions and we encourage the HRI and formal methods communities to create strong collaborations to tackle these and other questions toward the goal of trustworthy HRI.

Acknowledgment. This article is a result of fruitful discussions at the Dagstuhl seminar on Verification and Synthesis of Human-Robot Interaction.¹ The authors thank all fellow participants and the Schloss Dagstuhl—Leibniz Center for Informatics, for their support. 

References

1. Alami, R. et al. Verification and synthesis of human-robot interaction (Dagstuhl Seminar 19081). *Dagstuhl Reports* 9, 2 (2019), 91–110. <https://doi.org/10.4230/DagRep.9.2.91>.
2. Araiza-Illan, D. et al. Coverage-driven verification—An approach to verify code for robots that directly interact with humans. *Hardware and Software: Verification and Testing* (2015), 69–84.
3. Araiza-Illan, D. et al. Systematic and realistic testing in simulation of control code for robots in collaborative human-robot interactions. *Towards Autonomous Robotic Systems Conference* (2016), 20–32.
4. Argall, B.D. and Billard, A.G. A survey of tacitile human-robot interactions. *Robotics and Autonomous Systems* 58, 10 (Oct. 2010), 1159–1176. <https://doi.org/10.1016/j.robot.2010.07.002>.
5. Breazeal, C. et al. Social robotics. *Springer Handbook of Robotics*, Springer, (2016), 1935–1972.
6. Calinescu, R. et al. Socio-cyber-physical systems: Models, opportunities, open challenges. *2019 IEEE/ACM 5th Intern. Wkshp. Soft. Eng. for Smart Cyber-Physical Systems* (2019), 2–6.
7. Campos, T. et al. SMT-based control and feedback for social navigation. *Intern. Conf. Robotics and Automation, 2019, Montreal, QC, Canada*, 5005–5011.
8. Clarke, E.M. et al. eds. *Handbook of Model Checking*, Springer (2018).
9. Grice, H.P. Logic and conversation. *Syntax and Semantics, Vol. 3: Speech Acts*. P. Cole and J.L. Morgan, eds. Academic Press. (1975) 41–58.
10. Hoare, C.A.R. An axiomatic basis for computer programming. *Commun. ACM* 12, 10 (Oct. 1969), 576–580. <https://doi.org/10.1145/363235.363259>.
11. Hoffman, G. Evaluating fluency in human–robot collaboration. *IEEE Transactions on Human-Machine Systems* 49, 3 (2019), 209–218.
12. Iqbal, T. and Riek, L.D. Human-robot teaming: Approaches from joint action and dynamical systems. *Humanoid Robotics: A Reference* (2019), 2293–2312.
13. Kellmeyer, P. et al. Social robots in rehabilitation: A question of trust. *Science Robotics* 3, 21 (2018). <https://doi.org/10.1126/scirobotics.aat1587>.
14. Kress-Gazit, H. et al. Synthesis for robots: Guarantees and feedback for robot behavior. *Ann. Review of Control, Robotics, and Auton. Systems* 1, 1 (2018). <https://doi.org/10.1146/annurev-control-060117-104838>.
15. Kretínský, J. Survey of statistical verification of linear unbounded properties: Model checking and distances. *Leveraging Applications of Formal Methods, Verification and Validation: Foundational Techniques* (2016), 27–45.
16. Kshirsagar, A. et al. Specifying and synthesizing human-robot handovers. In *Proc. of the IEEE/RSJ Intern. Conf. on Intelligent Robots and Systems* (2019).
17. Kwiatkowska, M. et al. PRISM 4.0: Verification of probabilistic real-time systems. In *Proc. of the 23rd Intern. Conf. on Computer-Aided Verification* (Berlin, Heidelberg, 2011), 585–591.
18. Langer, A. et al. Trust in socially assistive robots: Considerations for use in rehabilitation. *Neuroscience & Biobehavioral Reviews* 104 (2019), 231–239. <https://doi.org/https://doi.org/10.1016/j.neubiorev.2019.07.014>.
19. Lee, H.R. et al. Cultural design of domestic robots: A study of user expectations in Korea and the United States. *2012 IEEE RO-MAN: The 21st IEEE Intern. Symp. on Robot and Human Interactive Communication* (2012), 803–808.
20. Leucker, M. and Schallhart, C. A brief account of runtime verification. *The J. Logic and Algebraic Programming* 78, 5 (2009), 293–303. <https://doi.org/http://dx.doi.org/10.1016/j.jlap.2008.08.004>.
21. Li, W. et al. Synthesis for human-in-the-loop control systems. *Tools and Algorithms for the Construction and Analysis of Systems—20th Intern. Conf.* (2014), 470–484.
22. Mason, G. et al. Assurance in reinforcement learning using quantitative verification. *Advances in*

Hybridization of Intelligent Methods: Models, Systems and Applications. I. Hatzilygeroudis and V. Palade, eds. Springer International Publishing, 71–96.

23. Mataric, M.J. and Scassellati, B. Socially assistive robotics. *Springer Handbook of Robotics*. Springer (2016), 1973–1994.
24. Nomura, T. Cultural differences in social acceptance of robots. *26th IEEE Intern. Symp. on Robot and Human Interactive Communication (RO-MAN)* (Aug. 2017), 534–538.
25. Porfiro, D. et al. Authoring and verifying human-robot interactions. In *Proc. of the 31st Ann. ACM Symp. on User Interface Software and Tech.* (2018), 75–86.
26. Rasmussen, J. Mental models and the control of action in complex environments. *Selected Papers of the 6th Interdisciplinary Wkshp. on Informatics and Psychology: Mental Models and Human-Computer Interaction 1* (NLD, 1987), 41–69.
27. Riek, L.D. Healthcare robotics. *Commun. ACM* 60, 11 (2017), 68–78. <https://doi.org/10.1145/3127874>.
28. Sadigh, D. et al. Planning for cars that coordinate with people: Leveraging effects on human actions for planning and active information gathering over human internal state. *Autonomous Robots (AURO)* 42, 7 (Oct. 2018), 1405–1426.
29. Thomaz, A. et al. Computational human-robot interaction. *Found. Trends Robotics* 4, 2–3 (Dec. 2016), 105–223. <https://doi.org/10.1561/23000000049>.
30. Tretmans, G.J. Test generation with inputs, outputs and repetitive quiescence. *Centre for Telematics and Information Technology (CTIT)*.
31. Webster, M. et al. A corroborative approach to verification and validation of human–robot teams. *The Intern. J. Robotics Research* 39, 1 (2020), 73–99. <https://doi.org/10.1177/0278364919883338>.
32. Wilkes-Gibbs, D. and Clark, H.H. Coordinating beliefs in conversation. *J. Memory and Language* 31, 2 (1992), 183–194.

Hadas Kress-Gazit is the Geoffrey S.M. Hedrick Sr. Professor in the Sibley School of Mechanical and Aerospace Engineering at Cornell University, Ithaca, NY, USA.

Kerstin Eder is a professor of Computer Science at the University of Bristol, UK.

Guy Hoffman is an associate professor and the Mills Family Faculty Fellow in the Sibley School of Mechanical and Aerospace Engineering at Cornell University, Ithaca, NY, USA.

Henny Admoni is the A. Nico Habermann Assistant Professor of Robotics at Carnegie Mellon University, Pittsburgh, PA, USA.

Brenna Argall is an associate professor of Mechanical Engineering, Computer Science, and Physical Medicine & Rehabilitation at Northwestern University, Evanston, IL, USA.

Rüdiger Ehlers is a professor for Embedded Systems at Clausthal University of Technology in Clausthal-Zellerfeld, Germany.

Christoffer Heckman is an assistant professor and the Pankove Faculty Fellow of Computer Science at the University of Colorado, Boulder, CO, USA.

Nils Jansen is an assistant professor with the Institute for Computing and Information Sciences at Radboud University Nijmegen, The Netherlands.

Ross Knepper is a roboticist and computer scientist specializing in the algorithmic aspects of autonomous robots and human-robot systems.

Jan Křetínský is a professor of Computer Science at the Technical University of Munich, Germany.

Shelly Levy-Tzedek is an associate professor at Ben-Gurion University, Beersheba, Israel.

Jamy Li is an assistant professor in Industrial Engineering at Ryerson University in Toronto, Canada.

Todd Murphey is a professor of Mechanical Engineering at Northwestern University in Evanston, IL, USA.

Laurel Riek is associate professor of Computer Science and Engineering at the University of California San Diego, USA.

Dorsa Sadigh is an assistant professor of Computer Science at Stanford University, Stanford, CA, USA.

Publish Your Work Open Access With ACM!

ACM offers a variety of Open Access publishing options to ensure that your work is disseminated to the widest possible readership of computer scientists around the world.



Please visit ACM's website to learn more about ACM's innovative approach to Open Access at:
<https://www.acm.org/openaccess>

A blueprint for leveraging the tremendous opportunities the IoT has to offer.

BY ATHMAN BOUGUETTAYA, QUAN Z. SHENG, BOUALEM BENATALLAH, AZADEH GHARI NEIAT, SAJIB MISTRY, ADITYA GHOSE, SURYA NEPAL, AND LINA YAO

An Internet of Things Service Roadmap

THE INTERNET OF THINGS (IOT) is taking the world by storm, thanks to the proliferation of sensors and actuators embedded in everyday things, coupled with the wide availability of high-speed Internet⁵⁰ and evolution of the 5th-generation (5G) networks.³⁴ IoT devices are increasingly supplying information about the physical environment (for example, infrastructure, assets, homes, and cars). The advent of IoT is enabling not only the connection and integration of devices that monitor physical world phenomena (for example, temperature, pollution, energy consumption, human activities, and movement), but also data-driven and AI-augmented *intelligence*. At all levels, synergies from advances in IoT, data analytics, and artificial intelligence (AI) are firmly recognized as strategic priorities for digital transformation.^{10,41,50}

IoT poses two key challenges:³⁶ *Communication* with things and *management* of things.⁴¹ The service paradigm is a key mechanism to overcome these challenges by transforming IoT devices into IoT services, where they will be treated as first-class objects through the prism of services.⁹ In a nutshell, services are at a higher level of abstraction than data. Services descriptions consist of two parts: functional and non-functional, such as, Quality of Service (QoS) attributes.²⁷ Services often transform data into an *actionable knowledge* or achieve physical state changes in the operating context.⁹ As a result, the service paradigm is the perfect basis for understanding the transformation of data into actionable knowledge, that is, making it useful. Despite the increasing uptake of IoT services, most organizations have not yet mastered the requisite knowledge, skills, or understanding to craft a successful IoT strategy. As a result, we do not have an adequate understanding of the ways by which we might leverage IoT opportunities.

From a service engineering perspective, IoT services may present difficult challenges, with many unsolved theoretical and technical questions.⁹ Such challenges stem from the scale of the systems contemplated, changes in service environments, quality of generated data and

» key insights

- Serendipity of IoT services will lead to highly innovative applications, including the crowdsharing of a wide array of services such as wireless energy services and other digital services.
- The service paradigm lends itself very nicely to the modeling of, and delivering on IoT. Each “thing” is modeled as a service with a set of purposes, that is, functionalities, delivered with a set of quality of services, that is, non-functional properties. The QoS can then be used as a discriminant to select the best IoT service.
- Augmenting IoT with services promises to deliver the same exciting outcomes as what the Web has achieved when it augmented the Internet. This led to a fundamental positive change in how humanity engages in all aspects of life.



IMAGE BY ANDRÉJ BORYS ASSOCIATES, USING SHUTTERSTOCK

enabled services, the inherent heterogeneity and uncertainty of ubiquitous environments including connectivity, and growing concerns about the unintended consequences of the digital age—security and privacy breaches.⁴⁹ For example, IoT devices can crowdsource a wide range of service types such as computing services, wireless energy sharing services, and environmental sensing services to other IoT devices in close proximity. In energy sharing services,²² IoT devices can wirelessly send energy to other nearby devices. However, because IoT services are crowdsourced, they are highly susceptible to improper and malicious usage. Stealing credit card information and sensitive medical histories, cyber-attacks, denial of service attacks, and privacy violations are examples of improper and malicious usages of IoT services.⁸

We see the evolution of the work in IoT services as mirroring in a way, at least conceptually, the work done in the World Wide Web (WWW). These efforts over the last 30 years led to generic abstractions and computation techniques that enabled a holistic computing environment in which users, information, and applications establish on-demand interactions, to realize useful experiences and to obtain services. The benefit of such an environment originates from the added value generated by the possible interactions. We believe IoT services will require similar building blocks in terms of useful models and techniques to build the added value promised by the ubiquity as well as the *serendipity* of IoT services. We also believe that providing enhanced simplicity, agility, efficiency, and robustness in engineer-

ing and provisioning of IoT services will unlock the IoT service paradigm at a global scale. The realization of this vision, however, poses formidable computing challenges to bring IoT services to the masses. While initial research outcomes exist which could be leveraged, significant progress is needed to make IoT services a tangible reality.

In this article, we identify key criteria of IoT services via an analogy analysis of the Internet and the Web and discuss the emerging technologies for the IoT environment. We also describe the major challenges in IoT services and present a research roadmap for the identified challenges.

An Analogy Analysis

We argue that for IoT to reach its full potential (from “technology” to “services”), there is a need to analyze

similar trajectories of other recent technological trends. We propose there is such an analogy with the Internet and the Web (see the accompanying figure). While the Internet was created as a “technology” for worldwide digital communication, the Web has transformed the Internet into *meaningful* services.⁷ We identify three key impacts of Web over the Internet:

► **Democratization:** The term “democratization” has its roots in political science and refers to the process of transitioning to a democratic form of government. More generally, it can be thought of as the process of removing the barriers of privilege and of offering equal rights, access, and authority. The WWW achieved the democratization of access to information. Before the advent of the WWW, information often resided in repositories with privileged access or in places where the barriers to access were onerous. The Web “democratized” access to information by removing (for the most part) barriers to access. In addition to democratizing access to information, the Web has also democratized the ability to publish information. Almost anyone can create a website and post information on it—those simple steps making that information accessible to everyone. In the early Internet, the information flow was in only one direction, which was static, with no way for users to add to or interact with the information. However, the Web emphasizes the importance of people’s interactions with the Internet. Everyone has an opportunity to contribute to the Web. And, by paying attention to what users are looking

for and doing online, better services (for example, recommender systems) are designed over the years.⁴⁰

► **Commoditization:** The Web *re-defines* the way businesses were performed. The Web enables e-commerce platform technology using the Internet as a backbone and gives birth to today’s platform economy. The platform technology has profoundly affected everyday life and how business and governments operate. Commercial transactions are conducted in electronic marketplaces that are supported by the platform technology. Transaction-oriented marketplaces include large e-malls, consumer-to-consumer auction platforms, multichannel retailers, and many millions of e-retailers. Examples of popular such transaction platform include Amazon, Airbnb, Uber, and Baidu. Massive business-to-business marketplaces have been created on the Web.²⁶ Moreover, the platform economy enables more efficient use of resources. Almost *instantaneous access* to services is made available by on-demand platforms using the Web. Such service orientation is not possible using only the concept of the Internet. Consumers have access to services and products from anywhere, and the price becomes a core factor in decision making. The bookstores are closing; big retailers are complaining about the penetration of online services such as Amazon; the emergence of Uber is disturbing the taxi industry. Established brick-and-mortar companies are competing to get a share of online customers. For example, big hotel chains are competing with Airbnb that does not own a single

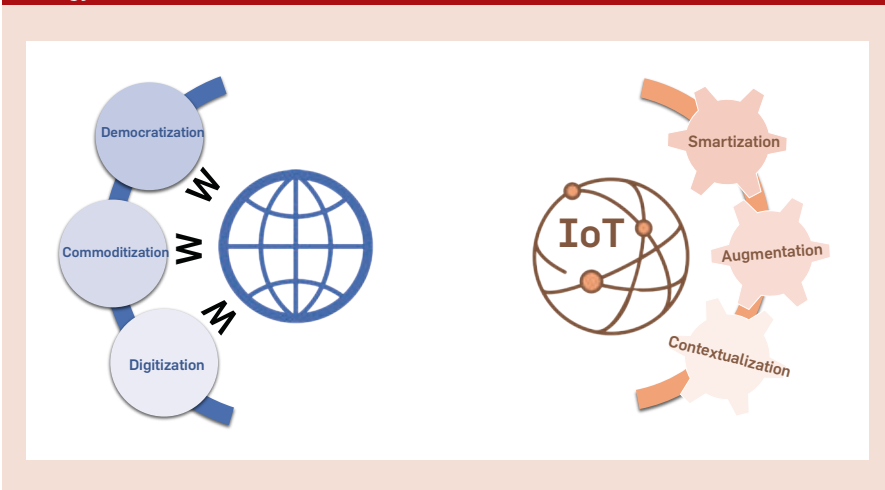
house. Every sector from IT to retails are affected by this phenomenon.

► **Digitization** is the process of converting information into digital formats. The Web is the single most important enabler for digitizing the information.⁷ The Web played a key role in moving the world from analog to digital form due to the increasing access to digital information. The advantages offered by digitization are the increasing access and *preservation* of information. Moreover, digitization enables enhanced services. Innovative services can be created using existing digital information in response to user demands. Such services have a direct impact on several industries. Libraries have been closing; traditional print media companies are struggling to survive in the face of social media, and online blogs and forums; more people consume news from Twitter and Facebook than newspapers and television. Policies are debated and elections are fought on social media. A huge amount of information is being created and consumed in digital form directly. State libraries are collecting social media data for information preservation. Our daily life activities are stored and shared in real time. Every individual, organization, industry, and government is impacted by this transformation.

There is a striking analogy of the IoT/services with the Internet/Web as services are the technology that transforms the IoT into a meaningful and useful framework. We outline three key criteria for defining novel IoT services, as shown in the right part of the figure.

► **Smartization** refers to the *process of introducing intelligence* to traditional systems to achieve sustainable, efficient, and convenient services. IoT services are instrumental in achieving this goal by working collaboratively to enable smart services for the people.⁸ Examples of such systems are smart cities, smart homes, and smart health. Consider an example of a smart city. The rapid industrialization has built global mega cities. These cities are now going through the post-industrialization development phase, where efficiency, sustainability, and livability become important factors for economic growth. These factors are largely addressed through smart city initiatives.

Analogy between the Internet vs. the Web and IoT vs. IoT service.



For example, Virtual Singapore application enables city planners in Singapore to simulate various scenarios including emergency evacuation. Similarly, smart transportation systems will automate our roadways, railways, and airways, transform passenger experiences, and reshape the way cargo and merchandise are tracked and delivered.

We believe the IoT service is the *foundation* to build next-generation intelligent systems and to transform all aspects of our life. In the process of “smartization,” IoT services bring together data, analytics, and decision-making services within a single platform and ensure they can work seamlessly and ubiquitously to provide an enhanced user experience. Hence, the smartization plays a significant role in making intelligent cyber-physical systems.

► **Augmentation** refers to the process of creating new services on demand by analyzing interactions among the devices and human to enhance the human experience. The self-driving vehicles, robots in aged care, and home automation are some examples of augmentation using IoT services. Augmentation has become a reality due to the availability of AI-assisted intelligent assistance (IA).^{14,42} We believe that future IA learns human behaviors, attitudes, and emotions, and creates new services on-demand to meet individual’s needs.

One emerging application area that will be enhanced by IoT services is augmented reality (AR). IoT services enabled AR can be used to visualize and interact with data from thousands of sensors simultaneously in real time. With such services, for example, a farmer can walk through his fields and get all information about soil, crops, water, moisture, temperature, and pest control in real time with precise location. The farmer can interact with such services and get even better insights on demand in real time.

We believe the IoT service is the key enabler to build intelligent systems on-demand.¹⁸ IoT services move the process of digitization from the artefacts in the Internet era to “everything” in the IoT world—our cities, hospitals, transport systems, as well as human beings. The IoT is a key enabling pillar of digitizing “everything” since “things” in the systems have *embedded*

capability to collect data, which captures the holistic view of the systems. The digitization process goes beyond the physical systems. For example, the “brain wearables” help to digitize people’s thoughts, feelings, and emotion and understand the neuroplasticity of our brain. These IoT services are very helpful to develop advanced augmented tools for people suffering from many physical and mental sicknesses (for example, anxiety, depression, paralysis). Hence, augmentation plays a significant role in enhancing interactions between human and physical systems.

► *Contextualization* refers to IoT services being aware of the situation and quickly adapting to the environment. The adaptation is not only limited to transforming or filtering sensor data in a meaningful and useful way to fit for the purpose, but also instantiating appropriate actuators. Giving contexts to IoT services through sensor data and actuator actions becomes an important criterion to build personalized services.^{8,41,42}

Data is the present while *context* is the future. Context will be the key to all industries. For example, a service will be created on demand to help a customer to buy milk to fit with his dietary requirement and health conditions. Similarly, a personalized service will help people to pack their luggage while going on holidays, that is, one’s suitcase, clothes, and weather forecast can interact with each other as IoT services and create a new advisory service on demand. A personalized temperature service is created on-demand for an individual to maintain the room temperature at home depending on individual’s preferences, meaning that an air-conditioning unit in the wall can interact with wearable sensors and create a personalized service.

Many start-up companies have emerged in recent times that were driven by contextualization. This indicates that contextualization will grow exponentially in the coming years and most of our current products and services will be personalized. Hence, we believe that contextualization plays a significant role in enhancing the user experience by creating personalized services on demand in real time.^{18,41,42}

The service computing research community has been continuing to de-

sign and develop IoT services for the last decade. Although there are incremental advancements, we argue that service computing has not fully explored to its potential in designing IoT services. This roadmap aims at outlining the vision and the underlying IoT service research challenges.

Emerging Technologies and IoT Services

The advancements of existing computing paradigms such as data science, deep learning, and cloud computing, and emerging technologies such as Edge computing, 5G networks, and blockchain are creating opportunities for innovative IoT services. These different paradigms or technologies have been explored in the context of IoT applications and platforms and are equally important for IoT services. They must be coordinated to develop a distributed and dynamic IoT service.³⁸ However, coordinating different computing paradigm with IoT services poses several research challenges. For example, the integration of data science with IoT needs to solve several research issues including the heterogeneity of IoT data formats, the real-time analytics, the data provenance, the dynamic data management and the IoT application orchestration.³⁸ Several research directions are proposed to address these challenges from the data science perspective.³⁸ Enhanced software abstraction of the IoT computation units such as MicroElement (MEL) and standard data integration protocols have the potential to resolve the IoT data heterogeneity issue.⁴⁷ A MEL consists of microelements such as data, computing, and actuators to deploy integration and computing solutions.

The *IoT service development platform* should have the ability to efficiently store and analyze real-time large IoT data streams from different types of physical and social sensors. The Edge computing is the potential research platform where IoT data are stored and analyzed at the edge of the IoT network instead of cloud services.⁴³ However, integrating Edge computing in the IoT service has several key research issues: programmability; naming; network and resource constraints management; QoS reliability; and security.¹²

Programmability refers to the development of service on heterogeneous edge nodes. Several novel approaches such as the development of computing streams are proposed to address the programmability in Edge computing.¹⁷ Naming refers to the standard way to discover and to communicate with a large amount of IoT services. Traditional naming approaches such as Domain Name Search (DNS) or Uniform Resource Identifier (URI) are not capable to serve the dynamic and large number of IoT services. Hence, novel naming approaches are required for dynamic IoT services.

Another key challenge in integrating IoT service with Edge computing is to enabling large computing task with the resource-constrained edge nodes. The computation tasks are not preferred to migrate to the cloud as it may increase the network latency and hinder real-time decision makings. The research community is addressing this research issue by proposing different edge architectures and distributed task scheduling models.⁴⁸ The application programmers have difficulties in ensuring the QoS of the IoT services due to diverse Edge infrastructures and fault events. Hence, the application QoS requirements and the underlying edge and fog infrastructures should be considered in building a QoS-aware IoT services.⁴⁸

The *IoT service infrastructure* produces a large amount of sensor data that must be analyzed to bring smartization in different applications such as smart homes and smart cities. Deep learning is a powerful analytic tool to extract new features and to bring intelligence in real-world applications.²³ However, integrating deep learning into IoT services has several key research issues, particularly, learning from noisy sensor data, and enabling resource-constrained edge computing for deep learning algorithms. Data preprocessing is an important step for deep learning approaches. As IoT data is heterogeneous and generated from different sources, the accurate preprocessing or data curation is complex for real-time services. To the best our knowledge, existing approaches propose to use layered-based learning frameworks where intermediate features are



We believe the IoT service is the foundation to build next-generation intelligent systems and to transform all aspects of our life.



learned in edge servers and the final output layer is processed in the cloud.²⁹ New learning acceleration engines are proposed for edge servers.¹⁹ These approaches are yet to adopt the full potential of deep learning for IoT services.

Security, privacy, and data trust is another key research issue for integrating the data science into the IoT service. It is proposed to adapt the blockchain technology to bring data provenance in the IoT service.³⁸ The Blockchain technology can create a trusted, decentralized, and autonomous system. Several blockchain-based IoT application framework is proposed in the existing literature.³³ However, integrating blockchain in an IoT service has several challenges such as resource limitations, interoperability of security protocols, and the dynamic trust management.²¹

Existing research roadmaps on IoT services mainly focus on utilizing emerging technologies from the data science perspective.^{29,33,38,43} We focus on integrating emerging technologies from the service computing perspective.

Challenges in IoT Services Research: A Roadmap

Actuation. The IoT will achieve the *democratization of actuation*, that is, invoking Internet-addressable things to take state-altering actions. Actuation has not received much attention in the current discourse on IoT but is likely to become a major focus of attention soon. Accessible actuation entails that the ability to use IoT devices to take action can, in principle, be made available to all. The ability to operate IoT-enabled home devices remotely is already a well-recognized use case.

Actuation over the Internet (we might refer to this as *open actuation*) will have far-reaching and game-changing consequences that we have not yet started to fathom.^{42,50} We have seen a simpler version of this phenomenon in *tele-operation*, but the impact of open actuation will be orders of magnitude greater. The tele-operation is typically point-to-point, that is, an operator invokes operations on a single device. The open actuation can be point-to-multipoint, where a single operator invokes multiple actuators over the IoT. The tele-operation is typically preconfigured, that is, a tele-operation link is

set up between an operator and a device by prior design (and often with investments in the physical infrastructure to make the teleoperation possible). The open actuation can be *emergent*. An operator might identify devices on the fly whose operation would help to achieve the operator's goals. A bespoke infrastructure for tele-operation is not necessary.

A combination of sensing and actuation gives us the ability to monitor and manage physical systems. Remote management of physical systems over the IoT can lead to the crowdsourced models of managing physical infrastructures. For instance, citizen groups might volunteer to manage specific civic spaces, such as a park or a community hall. For a park, they might be able to monitor turf health through sensors, while using remotely operated actuators such as sprinklers to water the turf when required. Citizen groups could manage neighborhood safety through similar means.

We have witnessed an exponential growth of autonomous systems in the last decade leading to the industrial revolution to realize the vision of Industry 4.0.³⁹ These autonomous systems are equipped with sensors and actuators and support its self-operation. The self-driving car is a good example of such system. Autonomous vehicles are also in operations in many research and commercial activities.¹⁵ Examples include autonomous vehicles in mining, robots in health-care, an underwater vehicle in climate study, among others. These autonomous systems are expected to interact with each other as well as their physical environments, building an autonomous Cyber Physical System (CPS).³⁷

The fine-grained IoT-enabled device-level levers for sensing and actuation will make automation far more ubiquitous. The democratization of the management of physical infrastructures will also enable greater delegation and autonomy. The services of physical devices could be globally shared.

Servitization. Servitization involves the wrapping of an existing product or system in a service-oriented model. The IoT service will lead to a greater, and potentially ubiquitous servitization. The IoT service can transform existing devices into ones that offer

value-added services. In this regard, IoT devices can harness service-oriented notions of *publication, discovery, and composition*.^{9,18,50} For example, servitization can enable IoT devices to publish their functionalities and QoS guarantees in device registries which can be searched to discover new devices and their associated services. In the case of composition, servitized IoT devices can be composed using new service composition techniques to obtain desired functionalities that meet the QoS constraints. The servitization can also lead to new conceptions of markets which regulate the usage of devices. For instance, servitized IoT devices may form a market for carbon credits that incentivizes the use of more carbon-friendly devices in more eco-friendly ways.

Governments around the world are struggling to deal with legal and social policies arising from the tremendous growth in the use of IoT devices in citizens' daily life. Though many of the policies from Internet governance could be applicable to IoT devices, it requires a new thinking due to the complexity, scale, and heterogeneity they bring. Servitization of IoT devices plays an important role to fill the gap and build policy, regulation, and governance for them.⁶

IoT service discovery. Future IoT is expected to be 50- to 100-times bigger than the current Internet, and the environments interacted by dynamic IoT services also evolve constantly.^{32,46} We identify a new set of challenges for IoT service discovery to enable the querying of billions of IoT resources to find the right service at the right time and location. We identify two different techniques that an IoT service discovery approach can adopt. The first technique is *semantic annotations* for IoT service descriptions and their associated sensory data. For instance, the OpenIoT project^a exploits a semantic sensor network (SSN) ontology from W3C for the sensor discovery and dynamic integration. The Hydra project^b adopts OWL (that is, an ontology language for Semantic Web) and SAWSDL (a semantic annotation of WSDL) to semantically annotate IoT services. A

a www.openiot.eu

b www.hydramiddleware.eu

number of ontologies have been proposed to represent IoT resources and services including Ontology Web Language for Things (OWL-T),²⁴ IoT-Lite Ontology,^c Comprehensive Ontology for IoT (COIoT),⁴⁵ and IoT-Stream.¹³ The Sensor Modeling Language SensorML which is a part of the OGC sensor Web enablement suite of standards, supports semantic descriptions of IoT services based on standardized XML tags. However, given the diversity and rapid IoT technological advances, it is challenging to reach an agreement on a single ontological standard for describing IoT services, and to maintain it.

Regarding IoT semantic reasoning, similar approaches to those described in Maarala et al.²⁵ and Chen et al.¹¹ may be used. The second technique uses the textual descriptions associated with IoT devices to locate the IoT services. Examples of IoT service discovery approaches based on the textual description are MAX,⁵² and Microsearch.⁴⁴ A research challenge is the *natural order ranking* of IoT contents.

The natural order ranking sorts contents by their intrinsic characteristics, rather than their relevance to a given query. In large data collections where a massive number of entities may be relevant to a query, natural order ranking mechanisms become crucial to deliver the most relevant results. PageRank is a well-known natural order ranking mechanism, which orders Web pages based on their importance via link analysis. Due to the size of IoT, another promising direction is to develop new natural order ranking mechanisms for the IoT contents to provide an effective and efficient IoT service discovery.⁴⁶ It is important to define the natural order that is applicable across heterogeneous IoT contents and has scalability. One potential solution could rely on QoS metrics of IoT services. Another possible solution is to construct a network of hidden links between IoT services and apply link analysis algorithms which are similar to PageRank. Discovering implicit relationships among IoT devices has been reported in recent research work.⁵¹ Considering the aforementioned techniques, the further work is to develop scalable approaches for the IoT service discovery.

c <https://www.w3.org/Submission/iot-lite/>

Security, privacy, and trust. IoT services become key pillars of automation and augmentation. Building trust in IoT services is the key to their success. Building trusted ecosystems among IoT services needs appropriate security, privacy, and trust measures between IoT services which are enabled by sensors and actuators, and their interactions with human being.^{8,49} Like all other Internet-based services in the past, IoT-based services are also being developed and deployed without security consideration. IoT devices are inherently vulnerable to malicious cyber threats because of the following reasons: they do not have well-defined perimeters; they are highly dynamic and heterogeneous; they are continuously changing because of mobility; and, they cannot be given the same protection that is received by enterprise services. In addition, due to *billions* of such IoT services, traditional human interaction-driven security solutions do not scale for security analysts or IoT service end-users to carry out security activities. Those activities may include approving the granting of permissions to IoT devices and setting up access control policies and configurations. The IoT-enabled augmented and automated decision-making systems will also “encourage” malicious cyber threats due to the high value of such systems. Hence, coordinated efforts are required from the research community to address resulting concerns.⁸

Is there such a thing as privacy in IoT services? With the prevalence of smartphones, social media, and people who tend to share so much information directly or indirectly, some researchers are starting to assert that there is no such thing as true privacy in the digital world. The impact of a data breach in an individual’s life and regular targeted e-commerce activities by the corporates have strengthened the view that privacy is more important than ever in the presence of IoT. In the beginning, the privacy concerns were limited to data, that is, personal records, images, video, and so on. With the adoption of smartphones, the privacy concern is moved from data to physical location as the location-based services are collecting an individual’s location in real time. With the emergence of “brain wearable” technology, one would be

able to read people’s mind and capture thoughts, feelings, hence raising the concern of mental privacy.

The technology trends in security are moving in two conflicting directions in terms of IoT services. On the one hand, the advancement of quantum computing makes the current security technologies obsolete, as they can be broken within seconds. Consequently, we must develop *quantum resistance schemes*. On the other hand, current security technologies cannot be applied to many IoT systems, as they cannot operate on power constraints environment. As a result, the IoT services demand *lightweight* quantum resistance security schemes.

Crowdsourcing IoT services. IoT devices are typically set up in fixed facilities or carried by humans. IoT users may crowdsource the functions of nearby IoT devices to suit their needs, such as WiFi hotspot sharing and wireless charging.^{1,22} The service paradigm can be applied as a key mechanism to abstract IoT devices and their functions along with their non-functional attributes (QoS) as crowdsourced IoT services from IoT users’ perspectives. These services will run as proxies of IoT devices. Crowdsourcing IoT services is a new and promising direction for the IoT service platform.⁹ Since crowdsourcing is more likely to be used if there are financial rewards and other incentives, an appropriate incentive model is required to motivate IoT service providers to form various types of crowdsourced IoT service.³⁰

The interactions among crowdsourced IoT services have a greater complexity than traditional service-oriented applications due to a large number of the expected IoT applications in the crowdsourced environment. This induces some unique challenges on trust management for crowdsourced IoT services. Firstly, the expected large number of newly deployed IoT services will likely have historical records to show any initial trustworthiness credentials. Therefore, traditional feedback-driven trust management would not be a realistic approach for crowdsourced IoT environments. In this regard, trust management of crowdsourced IoT services requires an alternative trust anchor instead of historical records. It can be IoT services’

inherent characteristics, which can generally reflect their trustworthiness. For example, IoT devices that are manufactured under a high-level security standard by a reputable manufacturer, are likely to be safely employed. The manufacturer’s reputation can be the trust anchor of IoT devices. There may exist multiple trust anchors (for example, the reputations of the manufacturers or owners of IoT devices), the aggregation of which would reflect the overall trustworthiness of IoT services.⁴ Secondly, the sheer diversity coupled with the expected large number of IoT application scenarios will redefine dynamism in service trustworthiness. The trustworthiness of an IoT service is greatly influenced by its application contexts and service users’ trust preferences and requirements.^{2,3} As a result, the trust management in IoT environments should address the dynamic and fluid nature of IoT services. Thirdly, the traditional centralized trust management would be quite costly and inefficient for crowdsourced IoT services because of the expected large number of IoT devices. The distribution of trust-related information on IoT services is expected to be decentralized. A key challenge for IoT service consumers is, therefore, the trustworthy access to reliable trust information for the IoT trust evaluation in a decentralized way.

Experiential IoT services. Experiential computing deals with digitally represented human experiences in everyday activities through every day “things” that have embedded computing capabilities. IoT services enable the realization of the vision of experiential computing by creating an experiential environment through the mediation of four dimensions of human experiences (that is, time, space, actors, and things). In this environment, users can explore and experience everyday events from multiple perspectives and revisit these events as many times as they wish to obtain the desired results.²⁰ The computation paradigm in such environments moves from the current data analytics to experience analytics, where the computation will be performed on digitally represented user experiences. This brings several new research challenges:

► Can my autonomous vehicle give

me the same experience that was felt by someone else?

- ▶ How can one generate an experience from a massive amount of data collected from IoT devices?

- ▶ Can one transfer his/her experience from one environment (for example, home) to another environment (for example, office)?

Requirements-driven IoT service design. The challenge of designing a device infrastructure, composed of both sensors and actuators, is difficult. Although the designing problem has some similarities with the problem of requirements-driven service composition, there are significant differences.¹⁸ In the service composition problem, a catalog of services is available a priori. In requirements driven IoT service design, there are challenging questions that need to be addressed as follows:

- ▶ What are the data requirements of the problem? What data would the decision modules and actuators need to be able to deliver on the required functionality? In the era of data analytics and the deployment of sophisticated AI systems, these are non-trivial problems, for example, the challenges associated with *feature engineering*. The many-to-many mapping between requirements and data items can be complex and requires equally complex reasoning to compute.

- ▶ What collections of sensors will be necessary to meet the data requirements of the problem? What hardware configurations would support the relevant nonfunctional requirements? Where should sensors be located? What hardware performance guarantees would be necessary to ensure that overall system-level performance guarantees are met?

- ▶ In a similar spirit, what actuators would the system require? What locations would be appropriate, what hardware configurations would be necessary and what hardware performance guarantees would satisfy the overall non-functional requirements?

- ▶ What kinds of coordination models would be necessary to orchestrate the behaviors of sensors and actuators to meet the stated requirements? Will existing schemes for specifying coordination models (such as process modeling notations) suffice?



Data is the present while context is the future. Context will be the key to all industries.



Computing complex compositions of sensors and actuators. As discussed previously, the problem of the IoT system design takes us into uncharted territory. The hardware dimensions of the problem, that is, finding the appropriate hardware configurations for sensors and actuators and the spatio-temporal dimensions need to be integrated and addressed.^{18,31} Furthermore, the Internet-of-Everything (IoE) aspects²⁸ add greater complexity. The autonomous human elements of the system and the AI-enabled computation components whose behavior would be emergent and not entirely predictable at design time need to be considered. In this regard, we identify the following challenges:

- ▶ **Managing resource-constrained sensing and actuation:** IoT systems often need to operate under significant resource constraints. This necessitates a significant reworking of standard approaches to system design, which leads to a novel conception of *resource-aware design*. In the spirit of earlier thinking on sensor networks, we must design sensing behaviors that consider finite energy reserves on sensor batteries. Similarly, actuator behavior would need to account for the finite capacity of actuator power sources.

- ▶ **Managing sensors and actuators at scale:** The IoT will enable us to address individually (for example, resource locators) devices at a very fine-grained level, and consequently on a very large scale.^{10,35,46} However, system design and management might not be very effective at these levels of granularity. In some cases, assigning individual addresses or resource locators at these low levels of granularity might also be challenging. We will, therefore, require novel abstractions that allow us to aggregate (and disaggregate) groups of sensors and actuators. An example is abstractions for classes of mutually interchangeable sensors and actuators. Interchangeability could be parametric. A set of sensors could be swapped for each other under a given set of functional requirements but not for a different set of functional requirements. Protocols for invoking sensor or actuator behavior will also need to exploit these abstractions. A range of similar issues also needs to be addressed for managing IoT devices at scale.

Large-scale IoT experimental facilities. While IoT-based digital strategies and innovations provide industries across the spectrum with exciting capabilities to create a competitive edge and build more value into their services, as what the Internet has done in the past 25 years, there are still significant gaps in making IoT a reality. One such gap lies on the missing of a large-scale, real-world experimental testbed for research and experimentation of new IoT service technologies.^{10,41}

The current IoT research infrastructures are largely in small scale, fragmented. There are not, therefore, suitable for IoT research and development. There is an urgent need to create such a unique research facility to stimulate advanced experimental research and realistic assessment of IoT technologies. Fueling the use of such a facility among the scientific community, end users, and service providers would increase the understanding of the technical and societal barriers in IoT adoption. The IoT-Lab^d is a recent effort in this trend. IoT-Lab test beds are located at six different sites across France, which are publicly accessible. A similar effort is also currently happening in Australia, aiming at establishing a nationwide IoT testbed across seven sites in major Australian cities. Digital Twins is a recent technology development that have attracted both industry and academia, and can be exploited to build large-scale IoT experimental facilities.^e

IoT data analytical services. The IoT analytics aims at delivering domain-specific solutions by aggregating and distilling heterogeneous IoT data to obtain information and actionable knowledge of appropriate quality and integrity. There is a need for a new paradigm of advanced IoT analytical services, which effectively and efficiently provide the underlying intelligence via harnessing the combination of physical and cyber worlds to turn IoT data into IoT intelligence. The following are some of the key identified challenges:

► **Dynamic contextual changes:** IoT data are tightly associated with multifaceted dynamic contexts, including user's internal contexts (for example,



The democratization of the management of physical infrastructures will also enable greater delegation and autonomy. The services of physical devices could be globally shared.



users' activities), external contexts (for example, location and time), and things' contexts (for example, expiration, usage status, and locations). Therefore, the effective IoT data analytical services are required to be capable of capturing both the salient changes and subtle ones of real-time contexts.

► **Tangled complex relationships:** IoT data exhibits highly heterogeneous and multi-dimensional correlations. For instance, user behaviors on things are intrinsically correlated both spatially and temporally.^{16,51} The new paradigm of IoT data analytical services needs to decode and leverage the heterogeneous nature of complex relationships.

► **Real-time distributed analytics:** IoT data is generated with high volume from scattered sources on a continuous basis, and the value of data might exponentially decay over timestamps for many IoT applications. This requires the analytical models to derive useful patterns and actionable knowledge with quality summarizations and then use these for provisioning streaming IoT analytics.

► **Reducing bias and ensuring fairness in IoT data analytics:** IoT data analytics would heavily rely on advanced machine learning techniques. The fairness in AI/ML technologies is an active research in itself and several techniques have been developed.⁵ However, we need to understand what it means to IoT services and should apply the same rigorous scrutiny to IoT data and services.

Conclusion

The IoT is widely considered as a new revolution of the Internet where billions of everyday objects are connected to empower human interactions with both virtual and physical worlds in a manner that is simply unprecedented. We believe that advancements made in the service computing over the past decades have not fully explored its potential in the designing of IoT services. We have identified three key criteria that define IoT services, namely smartization, augmentation, and contextualization. We outlined 10 main challenges in developing an IoT service. Designing and engineering scalable and robust IoT based solutions remains a deeply challenging problem. We identify critical direc-

d <https://www.iot-lab.info>

e <https://azure.microsoft.com/en-au/services/digital-twins/>

tions spanning discovery, security, privacy, and analytics. Interesting future research directions include:

► Actuation over the Internet should be further investigated to provide ubiquitous automation.

► Existing policies and regulations for Internet governance should be enhanced to enable servitization of IoT devices.

► IoT service discovery approaches should be dynamic and scalable to cater to the gigantic size and diversity of IoT and rapid IoT technological advances.

► Lightweight quantum security schemes should be explored for power-constrained IoT services.

► The trust management framework for crowdsourcing IoT services should be decentralized to manage the dynamism of service trust.

► Data analytics approaches should be translated to experience analytics to create an experiential IoT environment.

► Complex feature engineering should be investigated to address requirements-driven IoT service design.

► Computing complex compositions of sensors and actuators should be AI-driven and follow the resource-aware design. C

References

- Abusafia, A., Bouguettaya, A., and Mistry, S. Incentive-based selection and composition of IoT energy services. In *Proceedings of IEEE Intern. Conf. Services Computing*, 2020, 304–311.
- Bahutair, M., Bouguettaya, A., and Neiat, A. Adaptive trust: Usage-based trust in crowdsourced IoT services. In *Proceedings of 2019 IEEE Intern. Conf. on Web Services*. IEEE, 2019, 172–179.
- Bahutair, M., Bouguettaya, A., and Neiat, A. Just-in-time memoryless trust for crowdsourced IoT services. In *Proceedings of 2020 IEEE Intern. Conf. Web Services*. IEEE, 2020, 1–8.
- Bahutair, M., Bouguettaya, A., and Neiat, A. Multi-perspective trust management framework for crowdsourced IoT services. *IEEE Trans. Services Computing*, 2021.
- Bellamy, R., Dey, K., Hind, M., Hoffman, S., Houde, S., Kannan, K., Lohia, P., Martino, J., Mehta, S. et al. AI fairness 360: An extensible toolkit for detecting and mitigating algorithmic bias. *IBM J. Research and Development* 63, 4/5 (2019), 4–1.
- Berman, F. and Cerf, V. Social and ethical behavior in the internet of things. *Commun. ACM* 60, 2 (Feb. 2017), 6–7.
- Berners-Lee, T., Cailliau, R., Luotonen, A., Nielsen, H., and Secret, A. The world-wide web. *Commun. ACM* 37, 8 (Aug. 1994), 76–82.
- Bertino, E., Choo, K., Georgakopoulos, D., and Nepal, S. Internet of things (IoT): Smart and secure service delivery. *ACM Trans. Internet Technol.* 16, 4 (Dec. 2016), 22:1–22:7.
- Bouguettaya, A. et al. A service computing manifesto: The next 10 years. *Commun. ACM* 60, 4 (Apr. 2017), 64–72.
- Cerf, V. Things and the net. *IEEE Internet Computing* 16, 6 (2012), 96.
- Chen, G., Jiang, T., Wang, M., Tang, X., and Ji, W. Modeling and reasoning of IoT architecture in semantic ontology dimension. *Computer Communications* 153 (2020), 580–594.
- Chiang, M. and Zhang, T. Fog and IoT: An overview of research opportunities. *IEEE Internet of Things J.* 3, 6 (2016), 854–864.
- Elsaleh, T., Enshaeifar, S., Rezvani, R., Acton, S., Janeiko, V., and Bermudez-Edo, M. IoT-stream: A lightweight ontology for internet of things data streams and its use with data analytics and event detection services. *Sensors* 20, 4 (2020), 953.
- Fattah, S., Sung, N., Ahn, I., Ryu, M., and Yun, J. Building IoT services for aging in place using standard-based IoT platforms and heterogeneous IoT products. *Sensors* 17, 10 (2017), 2311.
- Gerla, M., Lee, E., Pau, G., and Lee, U. Internet of vehicles: From intelligent grid to autonomous cars and vehicular clouds. In *Proceedings of IEEE World Forum on Internet of Things*. IEEE, 2014, 241–246.
- Gharineiat, A., Bouguettaya, A., and Ba-hutair, M. A deep reinforcement learning approach for composing moving IoT services. *IEEE Trans. Services Computing*, 2021.
- Hao, Z., Novak, E., Yi, S., and Li, Q. Challenges and software architecture for fog computing. *IEEE Internet Computing* 21, 2 (2017), 44–53.
- Huang, B., Bouguettaya, A., and Neiat, A. Convenience-based periodic composition of IoT services. In *Proceedings of 16th Annual Intern. Conf. Service-Oriented Computing*. (Hangzhou, China, Nov. 12–15, 2018) 660–678.
- Huang, Z., Lin, K., Tsai, B., Yan, S., and Shih, C. Building edge intelligence for online activity recognition in service-oriented IoT systems. *Future Generation Computer Systems* 87 (2018), 557–567.
- Jain, R. Experiential computing. *Commun. ACM* 46, 7 (July 2003), 48–55.
- Khan, M. and Salah, K. Iot security: Review, blockchain solutions, and open challenges. *Future Generation Computer System* 82 (2018), 395–411.
- Lakhdari, A., Bouguettaya, A., Mistry, S., and Neiat, A. Composing energy services in a crowdsourced IoT environment. *IEEE Trans. Services Computing*, 2020, 1.
- Li, H., Ota, K., and Dong, M. Learning IoT in edge: deep learning for the internet of things with edge computing. *IEEE Network* 32, 1 (2018), 96–101.
- Maamar, Z., Fati, N., Kajan, E., Asim, M., and Qamar, A. Owl-t for a semantic description of IoT. In *Proceedings of European Conf. Advances in Databases and Information Systems*. Springer, 2020, 108–117.
- Maarala, A., Su, X., and Rieki, J. Semantic reasoning for context-aware internet of things applications. *IEEE Internet of Things J.* 4, 2 (2016), 461–473.
- Medjahed, B., Benatallah, B., Bouguettaya, A., Ngu, A., and Elmagarmid, A. Business-to-business interactions: issues and enabling technologies. *The VLDB J.* 12, 1 (2003), 59–85.
- Medjahed, B., Bouguettaya, A., and Elmagarmid, A. Composing web services on the semantic web. *The VLDB J.* 12, 4 (2003), 333–351.
- Miraz, M., Ali, M., Excell, P., and Picking, R. A review on internet of things (IoT), internet of everything (IoE) and internet of nano things (IoNT). *Internet Technologies and Applications*. IEEE, 2015, 219–224.
- Mohammadi, M., Al-Fuqaha, A., Sorour, S., and Guizani, M. Deep learning for IoT big data and streaming analytics: A survey. *IEEE Communications Surveys & Tutorials* 20, 4 (2018), 2923–2960.
- Neiat, A., Bouguettaya, A., and Mistry, S. Incentive-based crowdsourcing of hotspot services. *ACM Trans. Internet Technology* 19, 1 (2019), 5.
- Neiat, A., Bouguettaya, A., Sellis, T., and Mistry, S. Crowdsourced coverage as a service: Two-level composition of sensor cloud services. *IEEE Trans. Knowledge and Data Engineering* 29, 7 (2017), 1384–1397.
- Ngu, A., Gutierrez, M., Metsis, V., Nepal, S., and Sheng, Q. IoT middleware: A survey on issues and enabling technologies. *IEEE Internet of Things J.* 4, 1 (Feb. 2017), 1–20.
- Novo, O. Blockchain meets IoT: An architecture for scalable access management in IoT. *IEEE Internet of Things J.* 5, 2 (2018), 1184–1195.
- Palattella, M., Dohler, M., Grieco, A., Rizzo, G., Torsner, J., Engel, T., and Ladid, L. Internet of things in the 5G era: Enablers, architecture, and business models. *IEEE J. Selected Areas in Communications* 34, 3 (Mar. 2016), 510–527.
- Patel, P., Ali, M., and Sheth, A. On using the intelligent edge for IoT analytics. *IEEE Intelligent Systems* 32, 5 (2017), 64–69.
- Raggett, D. The web of things: Challenges and opportunities. *Computer* 48, 5 (2015), 26–32.
- Rajkumar, R., Lee, I., Sha, L., and Stankovic, J. Cyber-physical systems: The next computing revolution. In *Proceedings of Design Automation Conf. IEEE*, 2010, 731–736.
- Ranjan, R. et al. The next grand challenges: Integrating the internet of things and data science. *IEEE Cloud Computing* 5, 3 (2018), 12–26.
- Schlingensiepen, J., Nemtanu, F., Mehmood, R., and McCluskey, L. Autonomic transport management systems? Enabler for smart cities, personalized medicine, participation, and industry grid/industry 4.0. *Intelligent Transportation Systems—Problems and Perspectives*. Springer, 2016, 3–35.
- Shadbolt, N., Berners-Lee, T., Hendler, J., Hart, C., and Benjamins, R. The next wave of the web. In *Proceedings of the 15th Intern. Conf. World Wide Web*. ACM, 2006, 750–750.
- Sheng, M., Qin, Y., Yao, L., and Benatallah, B. *Managing the Web of Things: Linking the Real World to the Web*. Morgan Kaufmann, 2017.
- Sheth, A., Srivastava, B., and Michahelles, F. IoT-enhanced human experience. *IEEE Internet Computing* 22, 1 (2018), 4.
- W. Shi, J. Cao, Q. Zhang, Y. Li, and L. Xu. Edge computing: Vision and challenges. *IEEE Internet of Things J.* 3, 5 (2016), 637–646.
- Tan, C., Sheng, B., Wang, H., and Li, Q. Microsearch: A search engine for embedded devices used in pervasive computing. *ACM Trans. Embedded Computing Systems* 9, 4 (2010), 43, 2010.
- Tayur, V. and Suchithra, R. A comprehensive ontology for Internet of Things (coIoT). In *Proceedings of 2019 2nd Intern. Conf. Advanced Computational and Communication Paradigms*. IEEE, 2019, 1–6.
- Tran, N., Sheng, Q., Babar, M., and Yao, L. Searching the web of things: State of the art, challenges, and solutions. *ACM Computing Surveys* 50, 4 (2017), 55.
- Villari, M., Fazio, M., Dustdar, S., Rana, O., Chen, L., and Ranjan, R. Software defined membrane: Policy-driven edge and Internet of Things security. *IEEE Cloud Computing* 4, 4 (2017), 92–99.
- Villari, M., Fazio, M., Dustdar, S., Rana, O., and Ranjan, R. Osmotic computing: A new paradigm for edge/cloud integration. *IEEE Cloud Computing* 3, 6 (2016), 76–83.
- Voas, J., Kuhn, R., Kolias, C., Stavrou, A., and Kambourakis, G. Cybertrust in the IoT age. *Computer* 51, 7 (2018), 12–15.
- Want, R., Schilit, B., and Jensen, S. Enabling the Internet of Things. *Computer* 48, 1 (Jan. 2015), 28–35.
- Yao, L., Sheng, Q., Ngu, A., Li, X., and Benatallah, B. Unveiling correlations via mining human-thing interactions in the web of things. *ACM Trans. Intelligent Systems and Technology* 8, 5 (2017), 62.
- Yap, K., Srinivasan, V., and Motani, M. Max: Wide area human-centric search of the physical world. *ACM Trans. Sensor Networks* 4, 4 (2008), 26.

Athman Bouguettaya is a professor in the School of Computer Science at the University of Sydney, Australia.

Quan Z. Sheng is a professor in the Department of Computing at Macquarie University, Sydney, Australia.

Boualem Benatallah is a Scientia professor in the School of Science and Engineering at the University of New South Wales, Australia.

Azadeh Ghari Neiat is a lecturer in mobile apps computing at Deakin University, Melbourne, Australia.

Sajib Mistry is a lecturer in the School of EECMS at Curtin University, Australia.

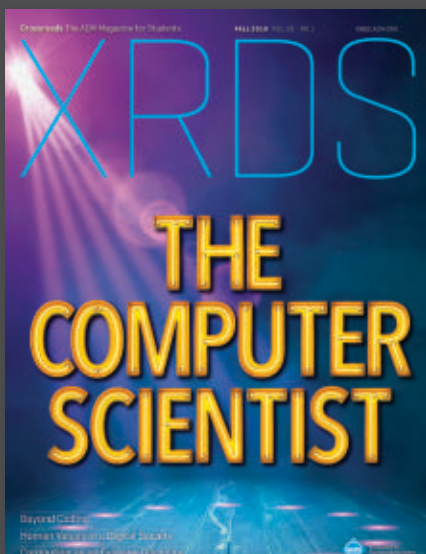
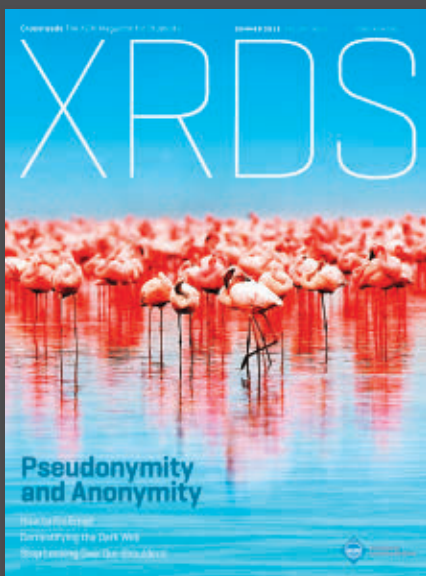
Aditya Ghose is a professor of computer science and director of the Decision Systems Lab at the University of Wollongong, Australia.

Surya Nepal is principal research scientist at CSIRO Data61, Epping, Australia.

Lina Yao is a Scientia associate professor in the School of Science and Engineering at the University of New South Wales, Australia.

Acknowledgment. This work was supported in part by Australian Research Council under Grants DP160103595 and LE180100158. The statements made here are solely the responsibility of the authors.

Copyright held by authors/owners.



XRDS

At XRDS, our mission is to empower computer science students around the world. We deliver high-quality content that makes the complexity and diversity of this ever-evolving field accessible. We are a student magazine run by students, for students, which gives us a unique opportunity to share our voices and shape the future leaders of our field.

Accessible, High-Quality, In-Depth Content We are dedicated to making cutting-edge research within the broader field of computer science accessible to students of all levels. We bring fresh perspectives on core topics, adding socially and culturally relevant dimensions to the lessons learned in the classroom.

Independently Run by Students XRDS is run as a student venture within the ACM by a diverse and inclusive team of engaged student volunteers from all over the world. We have the privilege and the responsibility of representing diverse and critical perspectives on computing technology. Our independence and willingness to take risks make us truly unique as a magazine. This serves as our guide for the topics we pursue and in the editorial positions that we take.

Supporting and Connecting Students At XRDS, our goal is to help students reach their potential by providing access to resources and connecting them to the global computer science community. Through our content, we help students deepen their understanding of the field, advance their education and careers, and become better citizens within their respective communities.

XRDS is the flagship magazine for student members of the Association for Computing Machinery [ACM].

www.xrds.acm.org



Association for
Computing Machinery

research highlights

P. 98

**Technical
Perspective**
**The Importance
of WINOGRANDE**

By Leora Morgenstern

P. 99

**WinoGrande:
An Adversarial Winograd
Schema Challenge at Scale**

By Keisuke Sakaguchi, Ronan Le Bras,
Chandra Bhagavatula, and Yejin Choi

P. 107

**Technical
Perspective**
**Does Your
Experiment Smell?**

By Stefano Baliatti

P. 108

**PlanAlyzer: Assessing
Threats to the Validity
of Online Experiments**

By Emma Tosch, Eytan Bakshy, Emery D. Berger,
David D. Jensen, and J. Eliot B. Moss

Technical Perspective

The Importance of WINOGRANDE

By Leora Morgenstern

EXCELLING AT A test often does not translate into excelling at the skills the test purports to measure. This is true not only of humans but also of AI systems, and the more so the greater the claims of the test's significance.

This became evident less than a decade after the introduction of the Winograd Schema Challenge (WSC),³ a test designed to measure an AI system's commonsense reasoning (CSR) ability by answering simple questions. An example would be, given the information: *The sculpture rolled off the shelf because it wasn't anchored*, answering: *What wasn't anchored?*

There are multiple AI systems² that achieve human performance on the WSC but are not capable of performing CSR. This would seem to be good reason to retire the WSC to the dustheap of benchmarks which have been conquered for little gain. But Yejin Choi and her colleagues at AI2 have sought to re-engineer the WSC as a more meaningful benchmark of a system's CSR ability. WINOGRANDE is one of a series of groundbreaking papers in which Choi and her team explore new methods of dataset development and adversarial filtering, expressly designed to prevent AI systems from making claims of smashing through benchmarks without making real progress.

Why try to fix the WSC? Why not simply develop a new dataset better suited to measuring CSR ability? The WSC's appeal lies partly in the test's radical simplicity and partly in what success might entail. Levesque proposed that the common task of pronoun resolution—determining which entity a pronoun referred to—could substitute as a test of CSR ability and intelligence. For example, consider the question: *Anna did better than Lucy on the test because she had studied so hard. Who studied hard?* Humans easily infer it is Anna who studied hard: We know studying hard generally leads to better grades.


But a machine without CSR ability likely cannot answer correctly.

Levesque sought to minimize bias in a sentence's structure toward a particular referent by collecting *pairs* of sentences that were nearly identical. For example, the above sentence could be rewritten as: *Anna did worse than Lucy on the test because she had studied so hard. Who studied hard?* In this case the answer changes: it is Lucy who studied hard. The reasoning is similar, but the substitution of *worse* for *better* leads to a different answer. Such pairs of sentences, named *Winograd Schemas*, were intended to eliminate the possibility of such structural bias.

Achieving near human performance on Winograd Schemas seemed beyond the capability of AI systems five years ago. But by using deep learning frameworks such as BERT,¹ which combine a transformer architecture, statistical natural language processing techniques, and a massive pre-trained language model, AI researchers rapidly developed high-performing systems—on the WSC as well as other benchmarks, for example, SuperGLUE⁶—while hardly moving the needle on more general AI measures.⁴

How to fix the WSC to prevent overestimation of machine performance? WINOGRANDE combines two closely intertwined strategies: generating a large corpus (a drawback of the original WSC was the tiny training corpus released) and filtering out *biased* examples. The WINOGRANDE corpus was generated by Mechanical Turkers (MTs), who wrote pairs of sentences using anchor words and obeying constraints. Other MTs ensured humans could easily infer pronoun referents in these sentences. Then the corpus was processed using a filtering algorithm to retain only examples that minimize *representation bias*. Removed pairs include those with dataset specific polarity basis (for example, *advanced rock climbing is more strongly associated with being strong than being weak*). The result is a corpus (~44K ex-

amples) for which the best system's accuracy in 2019 was 79.1%, considerably below human level. This effect, to prevent AI systems achieving human performance levels in the absence of genuine reasoning ability, was a desired goal.

What is the long-term impact? A year later, the Choi team's UNICORN can solve WINOGRANDE problems with an almost human-level 91.28% accuracy, as indicated by the WINOGRANDE leaderboard. AI systems will likely soon solve WINOGRANDE at human level—without necessarily having made real progress on the underlying task of CSR. Arguably, this indicates that solving either the WSC or WINOGRANDE does not indicate CSR ability. The contributions of WINOGRANDE, however, go far beyond performance on specific datasets. Importantly, the methodologies introduced in the paper are independent of the WINOGRANDE dataset. Methods used to help MTs generate large-scale corpora can be adapted to create other corpora. The filtering algorithm introduced here can be modified to filter bias and other sources of error more aggressively. These techniques will remain useful, whether AI systems prematurely achieve human-level performance on any of the multiple corpora that researchers currently target. 

References

- Devlin, J., Chang, M-W, Lee, K. and Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding, 2018; *arXiv:1810.04805*.
- Kocijan, V., Lukaszewicz, T., Davis, E., Marcus, G. and Morgenstern, L. A review of Winograd Schema Challenge datasets and approaches, 2020; *arXiv:2004.13831*.
- Levesque, H., Davis, E., and Morgenstern, L. The Winograd Schema Challenge. In *Proceedings of the 13th Intern. Conf. the Principles of Knowledge Representation and Reasoning*, 2012.
- Marcus, G. and Davis, E. *Rebooting AI: Building artificial intelligence we can trust*. Vintage, 2019.
- Nicholas, L., Le Bras, R., Bhagavatula, C. and Choi, Y. UNICORN on RAINBOW: A universal commonsense reasoning model on a new multitask benchmark. AAAI, 2021.
- Wang, A. et al. SuperGlue: A stickier benchmark for general-purpose language understanding systems, 2019; *arXiv:1905.00537*.

Leora Morgenstern is a Principal Scientist at PARC, a Xerox company, in Palo Alto, CA, USA.

Copyright held by author.

WinoGrande: An Adversarial Winograd Schema Challenge at Scale

By Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi

Abstract

Commonsense reasoning remains a major challenge in AI, and yet, recent progresses on benchmarks may seem to suggest otherwise. In particular, the recent neural language models have reported above 90% accuracy on the Winograd Schema Challenge (WSC),²² a commonsense benchmark originally designed to be unsolvable for statistical models that rely simply on word associations. This raises an important question—whether these models have truly acquired robust commonsense capabilities or they rely on spurious biases in the dataset that lead to an overestimation of the true capabilities of machine commonsense.

To investigate this question, we introduce WinoGrande, a large-scale dataset of 44k problems, inspired by the original WSC, but adjusted to improve both the scale and the hardness of the dataset. The key steps of the dataset construction consist of (1) large-scale crowdsourcing, followed by (2) systematic bias reduction using a novel AFLITE algorithm that generalizes human-detectable *word associations* to machine-detectable *embedding associations*. Our experiments demonstrate that state-of-the-art models achieve considerably lower accuracy (59.4%–79.1%) on WINOGRANDE compared to humans (94%), confirming that the high performance on the original WSC was inflated by spurious biases in the dataset.

Furthermore, we report new state-of-the-art results on five related benchmarks with emphasis on their dual implications. On the one hand, they demonstrate the effectiveness of WINOGRANDE when used as a resource for transfer learning. On the other hand, the high performance on all these benchmarks suggests the extent to which spurious biases are prevalent in all such datasets, which motivates further research on algorithmic bias reduction.

1. INTRODUCTION

Commonsense reasoning has been a long-standing open research question in AI.⁵ The Winograd Schema Challenge (WSC),²² proposed as an alternative to the Turing Test,³⁹ has been regarded as a prototypical benchmark to test commonsense capabilities in AI. WSC are designed to be pronoun resolution problems (see examples in Table 1) that are trivial for humans but hard for machines that merely rely on statistical patterns such as word associations without true commonsense understanding. One of the difficulties in commonsense reasoning comes from “reporting bias” in language¹⁵; commonsense knowledge is often too obvious for people to explicitly state in text, which can confuse the models that rely on statistical patterns in language.

However, recent advances in neural language models have saturated most major benchmarks, such as a variant of

WSC dataset where the models now achieve around 90% accuracy. This raises a curious question:

Have neural language models successfully acquired commonsense or are we overestimating the true capabilities of machine commonsense?

This question about the potential overestimation leads to another crucial question regarding potential unwanted biases that the large-scale neural language models might be exploiting, essentially solving the problems *right*, but for *wrong* reasons. Indeed, although WSC questions are carefully crafted by experts, recent studies have shown that they are nevertheless prone to incidental biases. Trichelair et al.³⁶ have reported *word-association* (13.5% of the cases, see Table 1 for examples) as well as other types of *dataset-specific* biases. Although such biases and annotation artifacts are not apparent for individual instances, they get introduced in the dataset as problems as authors subconsciously repeat similar problem-crafting strategies.

To investigate this question about the true estimation of the machine commonsense capabilities, we introduce **WinoGrande**, a new dataset with 44k problems that are inspired by the original design of WSC, but modified to improve both the scale and hardness of the problems. The key steps in WINOGRANDE construction consist of (1) a carefully designed crowdsourcing procedure, followed by (2) a novel algorithm AFLITE that generalizes human-detectable biases based on *word* occurrences to machine-detectable biases based on *embedding* occurrences. The key motivation of our approach is that it is difficult for humans to write problems without accidentally inserting unwanted biases.

Although humans find WINOGRANDE problems trivial with 94% accuracy, the best state-of-the-art results, such as those from RoBERTa,²⁵ are considerably lower (59.4%–79.1%) depending on the amount of training data provided (from 800 to 41k instances). Furthermore, we also demonstrate that WINOGRANDE provides transfer learning to other existing WSC and related benchmarks, achieving new state-of-the-art (SOTA) performances.

Although the improvements of SOTA over multiple challenging benchmarks are exciting, we cautiously note that these positive results must be taken with a grain of salt. The result might also indicate the extent to which spurious effects are prevalent in existing datasets, which runs the risk of overestimating the true capabilities of

The original version of this paper was published in the *Proceedings of the 34th AAAI Conference on Artificial Intelligence* (Feb. 2020).

Table 1. WSC problems are constructed as pairs (called *twin*) of nearly identical questions with two answer choices.

Twin sentences			Options (answer)
✓ (1)	a	The trophy doesn't fit into the brown suitcase because it's too <i>large</i> .	trophy/suitcase
	b	The trophy doesn't fit into the brown suitcase because it's too <i>small</i> .	trophy/suitcase
✓ (2)	a	Ann asked Mary what time the library closes, <i>because</i> she had forgotten.	Ann/Mary
	b	Ann asked Mary what time the library closes, <i>but</i> she had forgotten.	Ann/Mary
✗ (3)	a	The tree fell down and crashed through the roof of my house. Now, I have to get it <i>removed</i> .	tree/roof
	b	The tree fell down and crashed through the roof of my house. Now, I have to get it <i>repaired</i> .	tree/roof
✗ (4)	a	The lions ate the zebras because they are <i>predators</i> .	lions/zebras
	b	The lions ate the zebras because they are meaty.	lions/zebras

The questions include a *trigger word* that flips the correct answer choice between the questions. Examples (1)–(3) are drawn from WSC²² and (4) from DPR.³¹ Examples marked with ✗ have language-based bias that current language models can easily detect. Example (4) is undesirable because the word “predators” is more often associated with the word “lions,” compared to “zebras.”

machine intelligence on commonsense reasoning. More generally, human-crafted problems and tasks (regardless of whether they are crowd-sourced or by experts) contain annotation artifacts in many cases, and algorithmic bias reduction such as AFLITE is essential to mitigate such dataset-specific bias.

Our work suggests a new perspective for measuring progress in AI. Instead of constructing *static* benchmark datasets and asking the community to work on them for years, we propose the use of *dynamic* datasets that evolve together with the state-of-the-art models.

2. CROWDSOURCING WINOGRANDE AT SCALE

WSC problems have been considered challenging to craft by crowdsourcing due to the structural constraints of twins and the requirement of linguistic knowledge (Table 1). Nevertheless, we present an effective approach to creating a large-scale dataset (WINOGRANDE) of WSC problems while maintaining its original properties—that is, trivial for humans but hard for AI systems. Our approach consists of a carefully designed crowdsourcing task followed by a novel adversarial filtering algorithm (§3) that systematically removes biases in the data.

Enhancing crowd creativity. Creating twin sentences from scratch puts a high cognitive load on crowd workers who thereby subconsciously resort to writing pairs that are lexically and stylistically repetitive. To encourage creativity and reduce their cognitive load, we employed *creativity from constraints*³⁵—a psychological notion which suggests that appropriate constraints can help structure and drive creativity. In practice, crowd workers are primed by a randomly chosen topic as a suggestive context (details here), although they are asked to follow precise guidelines on the structure of the curated data.

Crowdsourcing task. We collect WINOGRANDE problems via crowdsourcing on Amazon Mechanical Turk (AMT). Workers are asked to write twin sentences (as shown in Table 1) that meet the requirements for WSC problems (e.g., avoiding word association, nonzero but small edit distance). To avoid repeating the same topics, workers were instructed to randomly pick an *anchor* word(s) from a randomly assigned WikiHow article and to ensure that the twin sentences contain the *anchor* word. The *anchor* word does not have to be a *trigger* word, but we ensured that it is not a function word

such as *it*, *the*, and *of*. In our pilot experiments, we found that this constraint drastically improves the worker’s creativity and diversity of topics. Additionally, workers were instructed to keep twin sentence length in between 15 and 30 words although maintaining at least 70% word overlap between a pair of twins. Following the original WSC problems, we aimed to collect twins in two different domains—(i) social commonsense: a situation involving two same gender people with contrasting attributes, emotions, social roles, etc., and (ii) physical commonsense: a context involving two physical objects with contrasting properties, usage, locations, etc. In total, we collected 77k questions (i.e., 38k twins).

Data validation. We validated each collected question through a distinct set of three crowd workers. Following a rigorous process, a question is deemed valid if (1) the majority of the three workers chooses the correct answer option, (2) they agree that the two answer options are unambiguous (one option is clearly more plausible than the other), and (3) the question cannot be answered simply by word association in which the local context around the target pronoun is given (e.g., “because **it** was going so fast.” (**race car/school bus**)). As a result, 68% of the questions (53k) were deemed valid and we discarded the invalid questions.

Although our crowdsourcing procedure addresses some amount of instance-level biases such as word association, it is still possible that the constructed dataset has *dataset-specific* biases, especially after it has been scaled up. To address this challenge, we propose a method for systematic bias reduction.

3. ALGORITHMIC DATA BIAS REDUCTION

Several recent studies^{16, 29, 38, 27, 12} have reported the presence of *annotation artifacts* in large-scale datasets. Annotation artifacts are unintentional patterns in the data that leak information about the target label in an undesired way. State-of-the-art neural models are highly effective at exploiting such artifacts to solve problems *correctly*, but for *incorrect* reasons. To tackle this persistent challenge with dataset biases, we propose AFLITE—a novel algorithm that can systematically reduce biases using the state-of-the-art contextual representation of words.

The workers met minimum qualification in AMT: 99% approval rate, 5k approvals. The reward was \$0.4 per twin sentences.

Lightweight adversarial filtering. Our approach builds upon the adversarial filtering (AF) algorithm proposed by Zellers et al.,⁴¹ but makes two key improvements: (1) AFLITE is much more broadly applicable (by not requiring over-generation of data instances) and (2) it is considerably more lightweight (not requiring retraining a model at each iteration of AF). Overgenerating machine text from a language model to use in test instances runs the risk of distributional bias where a discriminator can learn to distinguish between machine generated instances and human-generated ones. In addition, AF depends on training a model at each iteration, which comes at extremely high computation cost when being adversarial to a model such as BERT.⁷

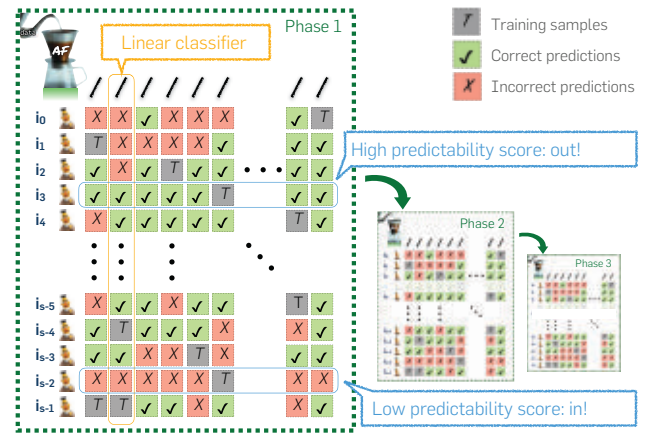
Instead of manually identified lexical features, we adopt a dense representation of instances using their *pre-computed* neural network embeddings. In this work, we use RoBERTa²⁵ fine-tuned on a small subset of the dataset. Concretely, we use 6k instances (5k for training and 1k for validation) from the dataset (containing 53k instances in total) to fine-tune RoBERTa (referred to as RoBERTa_{embed}). We use RoBERTa_{embed} to pre-compute the embeddings for the rest of the instances (47k) as the input for AFLITE. We discard the 6k instances from the final dataset.

Next, we use an ensemble of linear classifiers (logistic regressions) trained on random subsets of the data to determine whether the representation used in RoBERTa_{embed} is strongly indicative of the correct answer option. If so, we discard the corresponding instances and proceed iteratively.

Figure 1 provides an illustration of AFLITE algorithm. The algorithm takes as input the *pre-computed* embeddings and labels, along with the size n of the ensemble, the training size m for the classifiers in the ensemble, the size of the filtering cutoff, and the filtering threshold τ . At each filtering phase, we train n linear classifiers on different random partitions of the data and we collect their predictions on their corresponding validation set. For each instance, we compute its *score* as the ratio of correct predictions over the total number of predictions. We rank the instances according to their score and remove the top- k instances whose score is above threshold τ . We repeat this process until we remove fewer than k instances in a filtering phase or there are fewer than m remaining instances. When applying AFLITE to WINOGRANDE, we set $m = 10,000$, $n = 64$, $k = 500$, and $\tau = 0.75$.

This approach is also reminiscent of recent work in NLP on adversarial learning.^{3, 1, 9} Belinkov et al.¹ proposed an adversarial removal technique for NLI, which encourages models to learn representations that are free of hypothesis-only biases. When proposing a new benchmark, however, we cannot enforce that any future model will purposefully avoid learning spurious correlations in the data. In addition, although the hypothesis-only bias is an insightful bias in NLI, we make no assumption about the possible sources of

Figure 1. Illustration of the AFLITE algorithm. It takes as input the pre-computed representations of each instance (e.g., BERT embeddings). An ensemble of linear classifiers are trained on different random partitions of the data and used to compute the predictability score for each instance. The algorithm filters out the instances with the highest scores and proceeds iteratively to the next filtering phase.



bias in WINOGRANDE. Instead, we adopt a more proactive form of bias reduction by relying on the state-of-the-art (statistical) methods to uncover undesirable dataset shortcuts.

Assessment of AFLITE. We assess the impact of AFLITE relative to two baselines: random data reduction and pointwise mutual information (PMI) filtering. In random data reduction, we randomly subsample the dataset to evaluate how a decrease in dataset size affects the bias. In PMI filtering, we compute the difference (f) of PMIs for each twin (t) as follows:

$$f(t_1, t_2) = \sum_{w \in t_1} \text{PMI}(y=1; w) - \sum_{w \in t_2} \text{PMI}(y=1; w).$$

Technically, we first pre-computed PMI between a word and the label $y = 1$ for each word in the dataset, following a method proposed by Gururangan et al.¹⁶ The sum of PMI value of each token in a given sentence indicates the likelihood of the label $y = 1$ for the sentence. We only retain the twins that have a small difference in their PMI values as it corresponds to the twins that are hard to discriminate.

Figure 2 plots RoBERTa pre-computed embeddings whose dimension is reduced to 2D (*top*) and 1D (*bottom*) using Principal Component Analysis (PCA). We observe that WINO-GRANDE_{all} and the two baselines exhibit distinct components between the two correct answer options (i.e., $y \in \{1, 2\}$), whereas such distinction becomes less salient in WINO-GRANDE_{debiased}, which implies that AFLITE successfully reduces the spurious correlation in the dataset (between instances and labels). To quantify the effect, we compute the KL divergence between the samples with answer options. We find that the random data reduction does not reduce the KL divergence (2.53 \rightarrow 2.51). It is interesting to see that PMI-filtering marginally reduces the KL divergence (\rightarrow 2.42), although the principal component analysis on the PMI-filtered subset still leads to a significant separation between the labels. On the other hand, in WINOGRANDE_{debiased}, AFLITE reduces the KL divergence

The AFLITE algorithm is published with further development.²⁰

AFLITE is designed for filtering instances so that the resulting dataset is less biased, whereas the original AF algorithm⁴¹ is designed for “generating and modifying” individual instances, such as by creating better distractors. AFLITE and AF are therefore different in their goals and hence difficult to compare directly.

Figure 2. The effect of debiasing by AfLite. RoBERTa pre-computed embeddings (applied PCA for dimension reduction) are shown in two-dimensional space (top row) and histograms regarding d_1 (bottom row) with the bin size being 100. Data points are colored depending on the label (i.e., the answer y is option 1 (blue) or 2 (red)). In the histograms, we show the KL-divergence between $p(d_1, y=1)$ and $q(d_1, y=2)$.

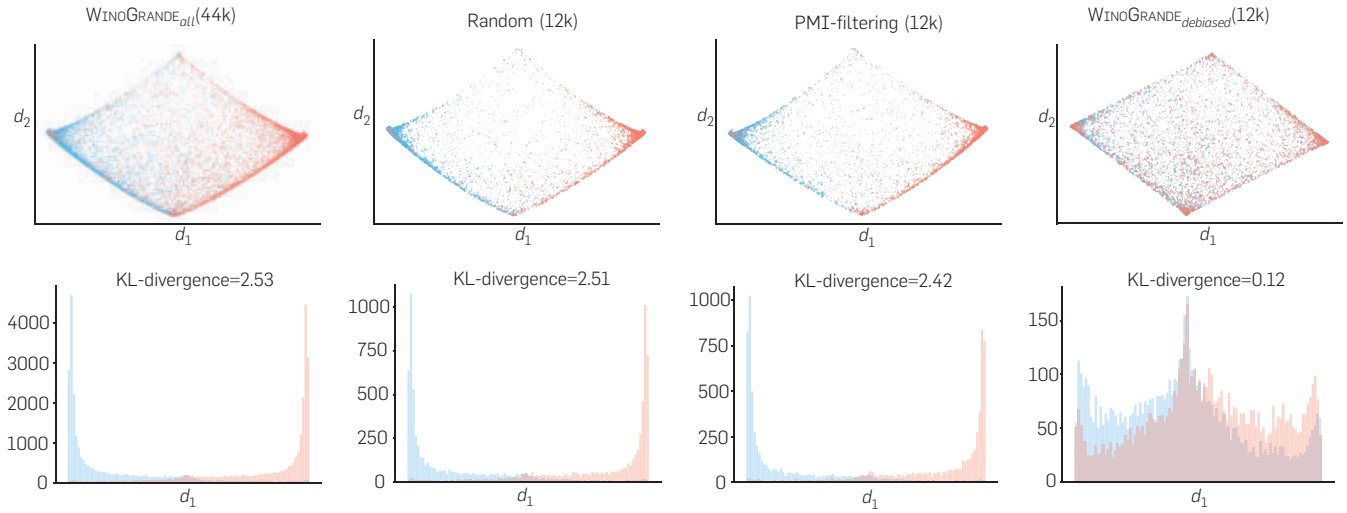


Table 2. Examples that have dataset-specific bias detected by AfLite (marked with X).

	Twin sentences	Options (answer)
X	The monkey loved to play with the balls but ignored the blocks because he found them exciting . The monkey loved to play with the balls but ignored the blocks because he found them dull .	balls /blocks balls/ blocks
X	William could only climb beginner walls while Jason climbed advanced ones because he was very weak . William could only climb beginner walls while Jason climbed advanced ones because he was very strong .	William /Jason William/ Jason
✓	Robert woke up at 9:00 am while Samuel woke up at 6:00 am, so he had less time to get ready for school. Robert woke up at 9:00 am while Samuel woke up at 6:00 am, so he had more time to get ready for school.	Robert /Samuel Robert/ Samuel
✓	The child was screaming after the baby bottle and toy fell. Since the child was hungry , it stopped his crying. The child was screaming after the baby bottle and toy fell. Since the child was full , it stopped his crying.	baby bottle /toy baby bottle/ toy

The words that include (dataset-specific) polarity bias (§3) are highlighted (positive and negative). For comparison, we show examples selected from WinoGrande_{debiased} (marked with ✓).

dramatically ($\rightarrow 0.12$), which suggests that this debiased dataset should be challenging for statistical models that solely rely on spurious correlation.

What bias has been actually detected by AfLite? Is the bias really spurious and undesirable according to the original WSC’s goal? Table 2 presents examples that AFLITE has detected as a dataset-specific bias. We see a structural pattern in the first two twins, where the sentiment between the answer option and the target pronoun is highly correlated. In other words, these problems can be easily answered by simply exploiting the pattern of the polarity (positive or negative). Importantly, this dataset-specific bias is structural rather than at the token level, contrasting with the biases that have been identified in the NLI literature,^{16, 29} and it is hard to detect these biases using heuristics such as lexical PMI-filtering. Instead of depending on such heuristics, AFLITE is able to detect the samples that potentially have such biases algorithmically.

After applying the AFLITE algorithm, we obtain a *debiased* dataset of 12,282 instances split into training (9,248), development (1,267), and test (1,767) sets. We also release 31k

problems that are filtered out by AFLITE for additional training set (§4) and resource (§5), resulting in a total number of problems in WINOGRANDE_{all} to be 43,432 (40,398 for training, 1,267 for development, and 1,767 for test).

WinoGrande versus the Original WSC. Although WINOGRANDE is inspired by the original WSC, we make a few design choices that deviate from the original design guidelines of WSC in order to scale up the dataset considerably while ensuring the hardness of the dataset.

First, WINOGRANDE is formatted as a fill-in-the-blank problem where the blank corresponds to the mention of one of the two names in the context, following the same modification made by other recent WSC variants such as Trinh and Le.³⁷ By contrast, the original WSC explicitly places a pronoun (instead of a blank). From the modeling stand point, the use of blanks instead of explicit pronouns do not make the problem any easier.

Second, although we originally collected all problems in twins, the final questions in the filtered WINOGRANDE_{debiased} are not always twins, because it is possible that AFLITE filters out only one of the twin sentences. In WINOGRANDE_{debiased}?

about 1/3 of questions are not twins. We also release WINOGRANDE_{all} (training set) that all consists of twins.

Third, unlike the original WSC problems that were composed by just a few linguistics experts, WINOGRANDE is authored by crowdworkers. Thus, the language used in WINOGRANDE reflects the more diverse and noisy language used by crowds. Importantly, laymen still find WINOGRANDE problems easy to solve, with 94% accuracy (§4).

4. EXPERIMENTAL RESULTS

4.1. Baseline models

We evaluate the WINOGRANDE_{debiased} (dev and test) on the methods/models that have been effective on the original WSC.

Wino knowledge hunting. Wino Knowledge Hunting (WKH) by Emami et al.¹⁰ is based on an information retrieval approach, where the sentence is parsed into a set of queries and then the model looks for evidence for each answer candidate from the search result snippets.

Ensemble neural LMs. Trinh and Le³⁷ is one of the first attempts to apply a neural language model, which is pre-trained on a very large corpora (such as LM-1-Billion, CommonCrawl, SQuAD, and Gutenberg Books). In this approach, the task is treated as fill-in-the-blank question with binary choice. The target pronoun in the sentence is replaced by each answer candidate, and the neural language model provides the likelihood of the two resulting sentences. This simple yet effective approach outperforms previous IR-based methods.

BERT. BERT⁷ is another pre-trained neural language model that has bidirectional paths and consecutive sentence representations in hidden layers. We finetune BERT with splitting the input sentence into context and option using the candidate answer as delimiter. The input format becomes [CLS] context [SEP] option [SEP]; for example, *The trophy doesn't fit into the brown suitcase because the _____ [SEP] is too large. [SEP]* (The blank _____ is filled with either option 1 or 2), and the [CLS] token embedding is used to classify which answer option is correct. We used grid-search for hyper-parameter tuning: learning rate $\{1e-5, 3e-5, 5e-5\}$, number of epochs $\{3, 4, 5, 8\}$, and batch-size $\{8, 16\}$ with three different random seeds.

RoBERTa. RoBERTa²⁵ is an improved variant of BERT that adds more training data with larger batch sizes and training time, as well as other refinements such as dynamic masking. RoBERTa performs consistently better than BERT across many benchmark datasets.

Word association baseline. Using BERT and RoBERTa, we also run the word association baseline (*local-context-only*) to check if the dataset can be solved by language-based bias. In this baseline, the model is trained with only local contexts ($w_{t-2:EOS}$) surrounding the blank to be filled (w_t) (e.g., because the _____ [SEP] is too large. [SEP]). This is analogous to the *hypothesis-only* baseline in NLI,²⁹ where the task (dataset) does not require the full context to achieve high performance.

Table 3. Performance of several baseline systems on WinoGrande_{debiased} (dev and test).

Methods	Dev acc. (%)	Test acc. (%)
WKH	49.4	49.6
Ensemble LMs	53.0	50.9
BERT	65.8	64.9
RoBERTa	79.3	79.1
BERT (local context)	52.5	51.9
RoBERTa (local context)	52.1	50.0
BERT-DPR*	50.2	51.0
RoBERTa-DPR*	59.4	58.9
Human Perf.	94.1	94.0

The star (★) denotes that it is zero-shot setting (e.g., BERT-DPR* is a BERT model fine-tuned with the DPR dataset and evaluated on WinoGrande_{debiased}).

Table 4. Performance of RoBERTa with different training sizes.

Training size	Dev acc. (%)	Test acc. (%)
XS (160)	51.5	50.4
S (640)	58.6	58.6
M (2,558)	66.9	67.6
L (10,234)	75.8	74.7
XL (40,398)	79.3	79.1

Fine-tuning on DPR dataset. Definite Pronoun Resolution (DPR) Dataset, collected by Rahman and Ng,³¹ consists of 1,886 WSC style problems written by 30 undergraduate students. Kocijan et al.¹⁹ have recently shown that BERT finetuned with DPR boosts the performance on WCS (72.2% accuracy). As additional baselines, we finetune BERT and RoBERTa with DPR and evaluate on WINOGRANDE. This allows us to compare the difficulty of WSC and WINOGRANDE empirically.

Human evaluation. In addition to the methods described above, we compute human performance as the majority vote of three crowd workers for each question.

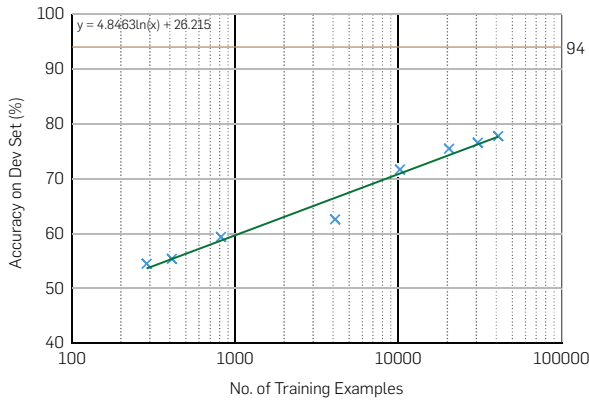
4.2. Results

Table 3 shows the results. Two baselines, WKH and Ensemble LMs, only achieve chance-level performance (50%). The best model, RoBERTa, achieves 79.1% test-set accuracy, whereas human performance achieves 94.0%, indicating that the WINOGRANDE_{debiased} is still easy for humans to answer as desired. Regarding the word association (i.e., local context) baselines, both BERT and RoBERTa achieve close to chance-level performance, illustrating that most WINOGRANDE_{debiased} problems cannot be answered by local context only. Finally, BERT and RoBERTa finetuned with DPR achieve chance-level to below 60% accuracy, which contrast with the performance boosts on WSC (72% by BERT (Kocijan et al.¹⁹) and 83% in RoBERTa) and other existing WSC-style problems (as shown in §5.3). This indicates that WINOGRANDE_{debiased} consists of more challenging problems than WSC and existing variants.

When we use the debiased training set (9248), both BERT and RoBERTa showed only chance level performance.

https://github.com/tensorflow/models/tree/master/research/lm_commonsense

Figure 3. Learning curve on the dev set of Wino-Grande. Each point on the plot is the best performance for a given number of randomly selected training examples, computed over 10 random seeds.



Learning curve. In order to see the effect of training size, Table 4 shows the performance by RoBERTa trained on different training sizes from 160k to 40k questions. Figure 3 shows the learning curve of the best model, RoBERTa, on the WINOGRANDE_{debiased} dev set. RoBERTa’s performance ranges from 59% to 79% when the size of training data is varied from 800 (2% of the training data) to 41K (100% of the training data) instances. To achieve human-level performance, the current state-of-the-art models would need over 118K training instances.

Importantly, the lower end of the available training data (~800) in the learning curve roughly matches the size of the training data made available in previous variants of WSC (see Table 5). For most of these datasets, state of the art already reaches around 90% (§5). By contrast, when we control for the training set size in WINOGRANDE, RoBERTa’s performance is considerably lower (59%), demonstrating that our dataset construction method is able to compose WSC problems that are collectively considerably harder than previous datasets.

5. TRANSFER LEARNING FROM WINOGRANDE

WINOGRANDE contains a large number of WSC style questions. In addition to serving as a benchmark dataset, we use WINOGRANDE as a resource—we apply transfer learning by first fine-tuning a model on our dataset and evaluating its performance on related datasets: WSC, PDP, SuperGLUE-WSC, DPR, KnowRef, KnowRef, and Winogender. We establish the state-of-the-art results across several of these existing benchmark datasets.

5.1. Existing WSC and related datasets

We briefly describe existing WSC variants and other related datasets. Table 5 provides their summary statistics.

Since the original publication of this paper, there have been several updates with higher performance such as Lin et al.²³ and Khashabi et al.¹⁸ that rely on similar models with even larger parameters and data sources, implying that the models detect annotation artifacts better than RoBERTa. This indicates that we need *dynamic* datasets that evolve together with the evolving state-of-the-art algorithms.

Table 5. Statistics on WSC and related datasets (§5.1).

Dataset	#Probs	Avg Len	#Vocab
WSC	273	19.1	919
PDP	80	39.5	594
SuperGLUE-WSC	804	28.4	1711
DPR	1886	15.9	4127
KnowRef	1269	19.3	5310
COPA	1000	13.3	3369
Winogender	720	15.6	523
WINOGRANDE _{debiased}	12,282	21.1	11,408
WINOGRANDE _{all}	43,432	20.6	16,469

WSC.²² This is the original Winograd Schema Challenge dataset, which consists of 273 problems. The problems are manually crafted by the authors to avoid word association bias as much as possible, although Trichelair et al.³⁶ later report that 13.5% of the questions may still have word-association bias.

PDP.²⁶ Pronoun Disambiguation Problems (PDP) dataset is closely related to the original WSC, and used in the 2016 running of the Winograd Schema Challenge. The dataset consists of 80 pronoun disambiguation problems. It is formulated as a multiple choice task, in which a pronoun must be resolved to one of up to 5 (but mostly binary) possible antecedents.

SuperGLUE-WSC.⁴⁰ SuperGLUE contains multiple datasets such as a modified version of WSC, which we will refer to as SuperGLUE-WSC. This dataset aggregates the original WSC, PDP and additional PDP-style examples, and recasts them into True/False binary problems (e.g., “Pete envies **Martin** because *he* is very successful.” Q: Does *he* refer to **Martin**? A: True). The number of problems are roughly doubled from WSC and PDP, although the size is still relatively small (804 in total). We converted WinoGrande to the True/False binary problems.

DPR.³¹ Definite Pronoun Resolution Dataset (DPR) introduces 1,886 additional WSC problems authored by 30 undergraduate students. Trichelair et al.³⁶ point out that this dataset is overall less challenging than the original WSC due to an increased level of language-based or dataset-specific biases. We split the original training set (1,332) into training (1,200) and development (122) sets, DPR does not have an official split for it.

KnowRef.¹¹ KnowRef provides over 8k WSC-style coreference resolution problems that are extracted and filtered with heuristic rules from 100 million web sentences (Reddit, Wikipedia, and OpenSubtitles). We report results on the publicly available *test* set (1.2k problems).

COPA.³² This dataset introduces 1000 problems that aim to test commonsense reasoning focusing on script knowledge, formulated as a binary choice about *causes* and *effects* of given premises. Because COPA does not provide a training set, we split the original development set (500) into training (400) and development (100) sets in the same way as SuperGLUE-COPA.⁴⁰

Winogender.³³ This dataset introduces 720 problems focusing on pronouns resolution with respect to people, with distinct goal of measuring gender bias in coreference resolution systems.

5.2. Experimental setup

Our model is based on RoBERTa finetuned with WINOGRANDE (train and dev sets). To compare different corpora used as a resource, we also finetune RoBERTa on DPR (train and test sets). For hyper parameter search, we use the same grid search strategy as in §4.

Additional human evaluation. We also report human performance for WSC, PDP, and DPR to calibrate the quality of our crowd worker pool as well as support previous findings. To our knowledge, this is the first work to report human performance on the DPR dataset.

5.3. Experimental results

Tables 6 and 7 show the results of applying transfer learning from WINOGRANDE to other WSC variants. Overall, RoBERTa fine-tuned on WINOGRANDE helps improve the accuracy on all the related tasks (Table 6), and performs consistently better than when it is fine-tuned on DPR.

Although improvements on some related datasets (particularly WSC, PDP, and DPR) might seem expected, the significant improvement on COPA is not so. The COPA task—identifying causes and effects—is very different from that in WINOGRANDE. This significant improvement on an unrelated task indicates that WINOGRANDE can serve as a resource for commonsense knowledge transfer.

Important implications. We consider that although these positive results over multiple challenging benchmarks are highly encouraging, they may need to be taken with a grain of salt. In particular, these results might also indicate the extent to which spurious dataset biases are prevalent in existing datasets, which runs the risk of overestimating the true capabilities of machine intelligence on commonsense reasoning.

Our results and analysis indicate the importance of continued research on debiasing benchmarks and the increasing need for algorithmic approaches for systematic bias reduction, which allows for the benchmarks to evolve together with evolving state of the art. We leave it as a future research question to further investigate how much of our improvements are due to dataset biases of the existing benchmarks as opposed to true strides in improving commonsense intelligence.

5.4. Diagnostics for gender bias

Winogender is designed as diagnostics for checking whether a model (and/or training corpora) suffers from gender bias. The bias is measured by the difference in accuracy between the cases where the pronoun gender matches the occupation’s majority gender (called “non-gotcha”) or not (“gotcha”). Formally, it is computed as follows:

$$\Delta F = \text{Acc}_{(\text{Female, Non-gotcha})} - \text{Acc}_{(\text{Female, Gotcha})}$$

$$\Delta M = \text{Acc}_{(\text{Male, Non-gotcha})} - \text{Acc}_{(\text{Male, Gotcha})}$$

for female and male cases, respectively.

Large values of ΔF or ΔM indicate that the model is highly gender-biased, whereas $|\Delta F| = |\Delta M| = 0$ (along with high accuracy) is the ideal scenario. In addition, if ΔF or ΔM is largely *negative*, it implies that the model is biased in the other way around.

The result of the gender-bias diagnostics is shown in Table 7. Although we find that the RoBERTa models

Table 6. Accuracy (%) on existing WSC-related tasks (test set).

WSC ²²	
Liu et al. ²⁴	52.8
WKH ¹⁰	57.1
Ensemble LMs ³⁷	63.8
GPT2 ³⁰	70.7
BERT-DPR* ¹⁹	72.2
HNN ¹⁷	75.1 [†]
RoBERTa-DPR* (This work)	83.1
RoBERTa-WinoGrande* (This work)	90.1
Humans ²	92.1
Humans (This work)	96.5
PDP ²⁶	
Liu et al. ²⁴	61.7
Trinh and Le ³⁷	70.0
RoBERTa-DPR* (This work)	86.3
RoBERTa-WinoGrande* (This work)	87.5
HNN ¹⁷	90.0[†]
Humans ⁵	90.9
Humans (This work)	92.5
SuperGLUE-WSC ⁴⁰	
Majority baseline	65.1
RoBERTa-DPR-ft (This work)	83.6
RoBERTa-WinoGrande-ft (This work)	85.6
RoBERTa-ensemble ²⁵	89.0
Humans ⁴⁰	100
DPR ³¹	
Rahman and Ng ³¹	73.0
Peng et al. ²⁸	76.4
BERT-WinoGrande* (This work)	84.9
RoBERTa-ft (This work)	91.7
RoBERTa-WinoGrande* (This work)	92.5
RoBERTa-WinoGrande-ft (This work)	93.1
Humans (This work)	95.2
KnowRef ¹¹	
Emami et al. ¹¹	65.0
RoBERTa-DPR* (This work)	84.2
RoBERTa-WinoGrande* (This work)	85.6
Humans ¹¹	92.0
COPA ³²	
Gordon et al. ¹⁴	65.4
Sasaki et al. ³⁴	76.4
RoBERTa-WinoGrande* (This work)	84.4
RoBERTa-ft (This work)	86.4 [‡]
RoBERTa-WinoGrande-ft (This work)	90.6
Humans ¹³	99.0

The star (★) denotes that it is zero-shot setting. “-ft” indicates *fine-tuning* on the targeted dataset (train and dev). RoBERTa-X-ft denotes sequential fine-tuning with dataset X followed by the targeted dataset. The dagger (†) indicates that the evaluation data is not exactly the same from ours. The double dagger (‡) denotes that we could not reproduce the same number as in SuperGLUE leaderboard.⁴⁰

fine-tuned on WINOGRANDE and DPR both demonstrate very high-accuracy, the gender gap in RoBERTa-WinoGrande is smaller than RoBERTa-DPR.

6. CONCLUSION

We introduce WINOGRANDE, a new collection of 44k WSC-inspired problems that is significantly larger than existing variants of the WSC dataset. To create a dataset that is robust

Table 7. Accuracy (%) and gender bias on Winogender dataset.

Winogender ³³					
	Gotcha	Female	Male	ΔF	ΔM
RULE	No	38.3	51.7	28.3	14.2
	Yes	10.0	37.5		
STATS	No	50.8	61.7	5.0	21.7
	Yes	45.8	40.0		
NEURAL	No	50.8	49.2	14.1	2.5
	Yes	36.7	46.7		
RoBERTa-DPR (This work)	No	98.3	96.7	1.6	0.9
	Yes	96.7	95.8		
RoBERTa-WG (This work)	No	97.5	96.7	0.8	0.8
	Yes	96.7	97.5		


“Gotcha” indicates whether the target gender pronoun (e.g., she) is minority in the correct answer option (e.g., doctor). |ΔF| and |ΔM| show the system performance gap between “Gotcha” and “non-Gotcha” for each gender (lower the better). The first three baselines are adopted from Rudinger et al.³³, Rule is Lee et al.,²¹ Stats is Durrett and Klein,⁸ and Neural is Clark and Manning.⁴

against spurious dataset-specific bias, we also present AFLITE—a novel lightweight adversarial filtering algorithm for systematic bias reduction. The resulting dataset is considerably more challenging for existing state-of-the-art models while still being trivially easy for humans. In addition, using WINOGRANDE as a resource, we demonstrate effective transfer learning and achieve state-of-the-art results on several related benchmarks.

In parallel, we also emphasize the potential risk of overestimating the performance of the state-of-the-art methods on the existing commonsense benchmarks; these models might be solving the problems *right* for the *wrong* reasons, by relying on spurious statistical patterns (annotation artifacts).

Our work suggests a new perspective for designing benchmarks for measuring progress in AI. Unlike past decades where the community constructed a *static* benchmark dataset to work on for many years to come, we now need AI algorithms to compose challenges that are hard enough for AI, which requires *dynamic* datasets that evolve together with the evolving state-of-the-art.

ACKNOWLEDGMENTS

We thank the anonymous reviewers, Dan Weld, Noah Smith, Luke Zettlemoyer, Hannaneh Hajishirzi, Oren Etzioni, Leora Morgenstern, Ernest Davis, Gary Marcus, and Yuling Gu, for their thoughtful feedback. This research was supported in part by NSF (IIS-1524371, IIS-1714566), DARPA under the CwC program through the ARO (W911NF-15-1-0543), and DARPA under the MCS program through NIWC Pacific (N66001-19-2-4031). 


References

- Belinkov, Y., Poliak, A., Shieber, S., Van Durme, B., Rush, A. On adversarial removal of hypothesis-only bias in natural language inference. *SEM* (2019), 256–262.
- Bender, D. Establishing a human baseline for the winograd schema challenge. *MAICS* (2015), 30–45.
- Chen, X., Cardie, C. Multinomial adversarial networks for multi-domain text classification. *NAACL* (2018), 1226–1240.
- Clark, K., Manning, C.D. Deep reinforcement learning for mention-ranking coreference models. *EMNLP* (2016), 2256–2262.
- Davis, E., Marcus, G. Commonsense reasoning and commonsense knowledge in artificial intelligence. *Commun. ACM* 58, 9 (Aug. 2015), 92–103.

- Davis, E., Morgenstern, L., Ortiz, C. Human tests of materials for the winograd schema challenge. Unpublished manuscript (2016). <https://cs.nyu.edu/faculty/davise/papers/WS2016SubjectTests.pdf>, 2016.
- Devlin, J., Chang, M.-W., Lee, K., Toutanova, K. BERT: Pre-training of deep bidirectional transformers for language understanding. arXiv:1810.04805 (2018).
- Durrett, G., Klein, D. Easy victories and uphill battles in coreference resolution. *EMNLP* (2013), 1971–1982.
- Elazar, Y., Goldberg, Y. Adversarial removal of demographic attributes from text data. *EMNLP* (2018), 11–21.
- Emami, A., Trischler, A., Suleman, K., Cheung, J.C.K. A generalized knowledge hunting framework for the winograd schema challenge. *NAACL: SRW* (2018), 25–31.
- Emami, A., Trichelair, P., Trischler, A., Suleman, K., Schulz, H., Cheung, J.C.K. The KnowRef coreference corpus: Removing gender and number cues for difficult pronominal anaphora resolution. *ACL* (2019), 3952–3961.
- Geva, M., Goldberg, Y., Berant, J. Are we modeling the task or the annotator? An investigation of annotator bias in natural language understanding datasets. arXiv:1908.07898 (2019).
- Gordon, A., Kozareva, Z., Roemmele, M. SemEval-2012 task 7: Choice of plausible alternatives: An evaluation of commonsense causal reasoning. *SEM* (2012), 394–398.
- Gordon, A.S., Bejan, C.A., Sagae, K. Commonsense causal reasoning using millions of personal stories. *AAAI* (2011), 1180–1185.
- Gordon, J., van Durme, B. Reporting bias and knowledge acquisition. *AKBC* (2013), 25–30.
- Gururangan, S., Swamydipta, S., Levy, O., Schwartz, R., Bowman, S., Smith, N.A. Annotation artifacts in natural language inference data. *NAACL* (2018), 107–112.
- He, P., Liu, X., Chen, W., Gao, J. A hybrid neural network model for commonsense reasoning. arXiv:1907.11983 (2019).
- Khashabi, D., Khot, T., Sabharwal, A., Tafjord, O., Clark, P., Hajishirzi, H. Unifiedqa: Crossing format boundaries with a single qa system. arXiv preprint arXiv:2005.00700, (2020).
- Kocijan, V., Cretu, A.-M., Camburu, O.-M., Yordanov, Y., Lukaszewicz, T. A surprisingly robust trick for the winograd schema challenge. *ACL* (2019), 4837–4842.
- Le Bras, R., Swamydipta, S., Bhagavatula, C., Zellers, R., Peters, M., Sabharwal, A., Choi, Y. Adversarial filters of dataset biases. *ICML* (2020).
- Lee, H., Peirsman, Y., Chang, A., Chambers, N., Surdeanu, M., Jurafsky, D. Stanford’s multi-pass sieve coreference resolution system at the CoNLL-2011 shared task. *CoNLL: Shared Task* (2011).
- Levesque, H.J., Davis, E., Morgenstern, L. The winograd schema challenge. In *AAAI Spring Symposium: Logical Formalizations of Commonsense Reasoning* (2011).
- Lin, S.-C., Yang, J.-H., Nogueira, R., Tsai, M.-F., Wang, C.-J., Lin, J. Ttttackling winogrande schemas. arXiv preprint arXiv:2003.08380 (2020).
- Liu, Q., Jiang, H., Ling, Z.-H., Zhu, X., Wei, S., Hu, Y. Commonsense knowledge enhanced embeddings for solving pronoun disambiguation problems in winograd schema challenge. arXiv:1611.04146 (2016).
- Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M.S., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L.S., Stoyanov, V. Roberta: A robustly optimized bert pretraining approach. ArXiv, abs/1907.11692 (2019).
- Morgenstern, L., Davis, E., Ortiz, C. L. Planning, executing, and evaluating the winograd schema challenge. *AI Magazine* 37, 1 (2016), 50–54.
- Niven, T., Kao, H.-Y. Probing neural network comprehension of natural language arguments. *ACL* (2019), 4658–4664.
- Peng, H., Khashabi, D., Roth, D. Solving hard coreference problems. *NAACL* (2015), 809–819.
- Poliak, A., Naradowsky, J., Haldar, A., Rudinger, R., Van Durme, B. Hypothesis only baselines in natural language inference. *SEM* (2018), 180–191.
- Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I. Language models are unsupervised multitask learners. *OpenAI Blog* (2019), 777–789.
- Rahman, A., Ng, V. Resolving complex cases of definite pronouns: The winograd schema challenge. *EMNLP-CoNLL* (2012).
- Roemmele, M., Bejan, C.A., Gordon, A.S. Choice of plausible alternatives: An evaluation of commonsense causal reasoning. In *AAAI Spring Symposium: Logical Formalizations of Commonsense Reasoning* (2011).
- Rudinger, R., Naradowsky, J., Leonard, B., Van Durme, B. Gender bias in coreference resolution. *NAACL* (2018), 15–20.
- Sasaki, S., Takase, S., Inoue, N., Okazaki, N., Inui, K. Handling multiword expressions in causality estimation. *IWCS* (2017).
- Stokes, P.D. *Creativity from Constraints: The Psychology of Breakthrough*. Springer Publishing Company, New York, NY, 2005.
- Trichelair, P., Emami, A., Cheung, J.C.K., Trischler, A., Suleman, K., Diaz, F. On the evaluation of commonsense reasoning in natural language understanding. arXiv:1811.01778 (2018).
- Trinh, T.H., Le, Q.V. A simple method for commonsense reasoning. arXiv:1806.02847 (2018).
- Tsuchiya, M. Performance impact caused by hidden bias of training data for recognizing textual entailment. *LREC* (2018), 1506–1511.
- Turing, A.M. Computing machinery and intelligence. *Mind* 59, 236(1950), 433–460.
- Wang, A., Pruksachatkun, Y., Nangia, N., Singh, A., Michael, J., Hill, F., Levy, O., Bowman, S.R. SuperGlue: A stickier benchmark for general-purpose language understanding systems. arXiv:1905.00537 (2019).
- Zellers, R., Bisk, Y., Schwartz, R., Choi, Y. Swag: A large-scale adversarial dataset for grounded commonsense inference. *EMNLP* (2018), 93–104.

Keisuke Sakaguchi, Ronan Le Bras, and Chandra Bhagavatula ([keisukes, ronanlb, chandrab]@allenai.org), Allen Institute for AI, Seattle, WA, USA.

Yejin Choi ([yejin]@cs.washington.edu), University of Washington & Allen Institute for AI, Seattle, WA, USA.

 This work is licensed under a <https://creativecommons.org/licenses/by/4.0/>

Technical Perspective

Does Your Experiment Smell?

By Stefano Ballelli

ONLINE HUMAN-BEHAVIOR EXPERIMENTATION is pervasive, manifold, and unavoidable. Leading digital companies routinely conduct over 1,000 A/B tests every month with millions of users. Online labor markets boast hundreds of thousands of workers to hire for crowdsourcing tasks, including experimentation and beta-testing. Outside industry, academic researchers utilize online labor markets to run behavioral experiments that span from cooperation games to protein folding tasks.

Hidden behind a deceiving façade of simplicity, implementing a human-behavior experiment for unbiased statistical inference is a task not to be taken lightly. It requires knowledge of computer programming, statistical inference, experimental design, and even behavioral insights. This unique mix of skills is generally honed with practice, heart-breaking mistakes, and “code smells.” Popularized by Martin Fowler’s book, *Refactoring: Improving the Design of Existing Code*, code smells indicate certain code structures that violate fundamental design principles and increase the risk of unintended software behavior. Code smells that lead to failures of randomization—the process of assigning observation units (users, devices, and so on) to treatments—are a threat to the validity of experiments. For instance, a probability incorrectly set may bar users to enter a particular treatment, or a degraded user experience in one treatment might lead to a higher attrition rate (that is, dropouts).


Detecting the source of a smell is not always trivial because experiments interact with multiple components, including external systems. The presence of several points of failure and the lack of a mathematical formalism to validate experiments in the context of their programming language have made human expert review the gold standard for assessing their correctness. But even experts are fallible. Therefore, two complementary practices are common in “smell-hunting”: simulations and pilots. Both are useful and both have

drawbacks. Simulations involve an array of bots randomly clicking their way through the experiment. They can catch internal bugs, but they cannot detect faulty interactions with external systems or failures in randomization due to idiosyncratic population characteristics or differential attrition rates. These issues are addressed by pilots, scaled-down versions of the experiment with real users. However, pilots require additional time and money and may frustrate participants if the user experience is poor. Moreover, a failed pilot may tarnish an experimenter’s reputation in crowdsourcing markets. Finally, in some cases, pilots are not possible at all, for example, in one-shot field experiments.

For all these reasons, I welcome the PlanAnalyzer software as detailed in the following paper by Tosch et al. PlanAnalyzer is a linter for PlanOut, a framework for online experiments popular in corporate settings, in particular Facebook, where it was originally developed. In addition to flagging code smells, PlanAnalyzer also reports in a human-readable fashion which hypotheses a PlanOut script can and cannot test statistically. How does PlanAnalyzer achieve this goal? It takes a PlanOut script as input and translates it into an internal representation that assigns special labels to variables in the code. It then builds a data dependence graph, based on which it establishes reliable causal paths between those specially labeled variables (that is, the contrasts). To account for missing information and interactions with external systems, manually

annotated labels may be integrated.

PlanAnalyzer was validated against a corpus of actual PlanOut scripts created and deployed at Facebook. The results are very encouraging: PlanAnalyzer replicated 82% of all contrasts manually annotated by domain experts and achieved a precision and recall of 92% each in detecting code smells in a synthetically mutated dataset. Moreover, the authors unveiled a collection of common bad coding practices, including ambiguous type comparisons, modulus operators applied to fractions, and the use of PlanOut scripts for application configuration only. Future work in this area might focus on automatically correcting errors in the code, generating statistical code to analyze the output of an experiment (another potential source of smell), or introducing reasoning about hypotheses (for example, whether non-proportional sampling of observation units is valid).

PlanAnalyzer is the first tool to statically check the validity of online experiments. It is cheaper, faster, and possibly safer than deploying bots or running a pilot. In sum, it is a major milestone. Together with recent advances in AI-driven methods for choosing optimal values of experimental parameters, adaptively ordering survey questions, and imputing missing responses, it shows how computer-assisted methods for the design, validation, and analysis of experiments are gaining a foothold. As this pattern will continue to grow in the future, we should expect two things: consolidation in the extremely fragmented landscape of tools for online experimentation, and the establishment of a validated set of coding standards. Both outcomes will boost the replicability of experimental results, paving the way for faster progress in the study of online human-behavior in industry and academia. 

PlanAnalyzer is the first tool to statistically check the validity of online experiments.

Stefano Ballelli is a fellow at the Mannheim Center for European Social Science Research and a postdoc at the Alfred-Weber Institute of Economics at Heidelberg University, Germany.

Copyright held by author.

PlanAlyzer: Assessing Threats to the Validity of Online Experiments

By Emma Tosch, Eytan Bakshy, Emery D. Berger, David D. Jensen, and J. Eliot B. Moss

Abstract

Online experiments are an integral part of the design and evaluation of software infrastructure at Internet firms. To handle the growing scale and complexity of these experiments, firms have developed software frameworks for their design and deployment. Ensuring that the results of experiments in these frameworks are trustworthy—referred to as *internal validity*—can be difficult. Currently, verifying internal validity requires manual inspection by someone with substantial expertise in experimental design.

We present the first approach for checking the internal validity of online experiments statically, that is, from code alone. We identify well-known problems that arise in experimental design and causal inference, which can take on unusual forms when expressed as computer programs: failures of randomization and treatment assignment, and causal sufficiency errors. Our analyses target PLANOUT, a popular framework that features a domain-specific language (DSL) to specify and run complex experiments. We have built PLANALYZER, a tool that checks PLANOUT programs for threats to internal validity, before automatically generating important data for the statistical analyses of a large class of experimental designs. We demonstrate PLANALYZER's utility on a corpus of PLANOUT scripts deployed in production at Facebook, and we evaluate its ability to identify threats on a mutated subset of this corpus. PLANALYZER has both precision and recall of 92% on the mutated corpus, and 82% of the contrasts it generates match hand-specified data.

1. INTRODUCTION

Many organizations conduct online experiments to assist decision-making.^{3,13,21,22} These organizations often develop software components that make designing experiments easier, or that automatically monitor experimental results. Such systems may integrate with existing infrastructure that perform such tasks as *recording* metrics of interest or *specializing* software configurations according to features of users, devices, or other experimental subjects. One popular example is Facebook's PLANOUT: a domain-specific language for experimental design.²

A script written in PLANOUT is a procedure for assigning a *treatment* (e.g., a piece of software under test) to a *unit* (e.g., users or devices whose behavior—or *outcomes*—is being assessed). Treatments could be anything from software-defined bit rates for data transmission to the layout of a Web page. Outcomes are typically metrics of interest to the firm, which may include click-through rates, time spent on a page, or the proportion of videos watched to completion. Critically, treatments and outcomes must be

recorded in order to estimate the *effect* of treatment on an outcome. By abstracting over the details of how units are assigned treatment, PLANOUT has the potential to lower the barrier to entry for those without a background in experimental design to try their hand at experimentation-driven development.

Unfortunately, the state of the art for validating *experimental designs* (i.e., the procedure for conducting an experiment, here encoded as a PLANOUT program) is a manual human review. The most common experimental design on the Web is the A/B test, which entails a fairly simple analysis to estimate the treatment effect. However, more complex experiments may require more sophisticated analyses, and in general there is a many-to-many relationship between design and analyses. Many experiments written in a domain-specific language (DSL) such as PLANOUT can be cumbersome to validate manually, and they cannot be analyzed using existing automated methods. This is because experiments expressed as programs can have errors that are unique to the intersection of experimentation and software.

We present the first tool, PLANALYZER, for statically identifying the sources of statistical bias in programmatically defined experiments. Additionally, PLANALYZER automatically generates *contrasts* and *conditioning sets* for a large class of experimental designs (i.e., between-subjects designs that can be analyzed using *average treatment effect* (ATE) or *conditional average treatment effect* (CATE)); because ATE is a special case of CATE, when the distinction between the two is not necessary, we will refer to them collectively as (C)ATE). We make the following contributions:

Software for the static analysis of experiments. PLANALYZER produces three key pieces of information: (1) a list of the variables in the environment that are actually being randomly assigned; (2) the variables that are recorded for analysis; and (3) the variables that may be legitimately compared when computing causal effects. These three pieces of information are required in order to determine whether there are any valid statistical analyses of the recorded results of an experiment, and, when possible, what those analyses are.

Characterizing errors and bad practices unique to programmatically defined experiments. Traditional errors in offline experimentation can take on unusual forms in programmatically defined experiments. Additionally, some coding practices can lead to faults during downstream

The original version of this paper was published in the *Proceedings of ACM on Programming Languages*, OOPSLA, Oct. 2019.

statistical analysis, highlighting the potential utility of defining “code smells” for bad practices in experiments.⁸ We introduce errors and code smells that arise from the intersection of experiments and software.

Empirical analysis of real experiments. We report PLANALYZER’s performance on a corpus of real-world PLANOUT scripts from Facebook. Due to the vetting process at Facebook, few errors exist naturally in the corpus. Therefore, we perform mutation analysis to approximate a real-world distribution of errors. We also consider the set of author-generated contrasts (the set of variable values that reallocated to be compared, necessary for estimating causal effects) for each script. We demonstrate PLANALYZER’s effectiveness in finding major threats to validity and in automatically generating contrasts.

2. LANGUAGE CHARACTERISTICS

As a DSL is built by domain experts, PLANOUT implements functionality only relevant to experimentation. Consequently, PLANOUT is not Turing complete: it lacks loops, recursion, and function definition. It has two control flow constructs (`if/else` and `return`) and a small core of built-in functions (e.g., `weightedChoice`, `bernoulli-Trial`, and `length`).

Although not required for an experimentation language, PLANOUT also allows for runtime binding of external function calls and variables. This allows for easy integration with existing (but more constrained) systems for experimentation, data recording, and configuration. We expect PLANOUT scripts to be run inside another execution environment, such as a Web browser, and have access to the calling context in order to bind free variables and functions.

PLANOUT abstracts over the sampling mechanism, providing an interface that randomly selects from pre-populated partitions of unit identifiers, corresponding to samples from the population of interest. The PLANOUT framework provides a mechanism for extracting the application parameters manipulated by a PLANOUT script and hashes them, along with the current experiment name, to one or more samples. The mapping avoids clashes between concurrently running experiments, which is one of the primary challenges of online experimentation.^{12,13} Readers interested in the specifics of PLANOUT’s hashing method for scaling concurrent experiments can refer to an earlier paper²; it is not relevant to PLANALYZER’s analyses.

On its surface, PLANOUT may appear to share features with probabilistic programming languages (PPLs).^{11,16} PPLs completely describe the data generating process; by contrast, PLANOUT programs specify only one part of the data generating process—how to randomly assign treatments—and this code is used to control aspects of a product or service that is the focus of experimentation.

There are two critical features of PLANOUT that differentiate it from related DSLs, such as PPLs: (1) the requirement that all random functions have an explicit unit of randomization, and (2) built-in control of data recording via the truth value of PLANOUT’s `return`. Only named variables on paths that terminate in `return true` are recorded. This is similar to the discarded executions in the

implementation of conditional probabilities in PPLs. A major semantic difference between PLANOUT and PPLs is that we expect PLANOUT to have deterministic execution for an input. Variability in PLANOUT arises from the population of inputs; variability in PPLs comes from the execution of the program itself.

3. VALIDATION OF STATISTICAL CONCLUSIONS

Statistical conclusions of a randomized experiment typically estimate the effect of a treatment T on an outcome Y for some population of units. The function that estimates the causal effect of T on Y may take many forms. Nearly all such functions can be distilled into estimating the true difference between an outcome under one treatment and its *potential outcome(s)* under another treatment.

In the case of a randomized experiment, if T is assigned completely at random, for example, according to:

```
T = uniformChoice(choices=[400, 750], unit=userid);
```

then the causal effect of T (the *average treatment effect* (ATE)) can be estimated by simply taking the difference of the average outcome for units assigned to $T = 400$ and $T = 750$: $Avg(Y | T = 400) - Avg(Y | T = 750)$. Such an experiment could be useful for learning how some outcome Y (e.g., video watch time) differs for equivalent individuals experiencing videos at the 400 or 750kbps setting.

It is not uncommon to use different probabilities of treatment for different kinds of users; we refer to the partition of users as a *subgroup* S . We can still estimate causal effects, but must instead compute the difference in means separately for different values of the variables in S . This is often referred to as *subgroup analysis*. This estimand is known as the *conditional average treatment effect* (CATE). The variables that define the subgroup are referred to as the *conditioning set* and can be thought of as a constraint on the units that can be compared for any given contrast. Average effect estimators like (C)ATE over finite sets of treatments can be expressed in terms of their valid contrasts: knowing the assignment probabilities of $T = 400$ versus $T = 750$ is sufficient to describe how to compute the treatment effect.

Typically, experts must manually verify that the estimators comport with the experimental design. There are some exceptions: some systems for automatically monitoring very simple experiments like A/B tests, where the treatment is a single variable that takes on one of the two values and the estimand is ATE.

As a DSL, PLANOUT provides a mechanism for more complex experimental designs. Control-flow operators, calls to external services, and in-language mechanisms for data recording prohibit simple automatic variable monitoring. For example, an experiment that sets variables differently on the basis of the current country of the user cannot naïvely aggregate results across all the participants in the experiment. Such an experiment would require additional adjustment during post-experiment analysis, because a user’s current country is a *confounder* (i.e., a variable that causes both the treatment and outcome). PLANALYZER

automatically produces the appropriate analyses, including the contrasts and conditioning sets.

4. VALIDATION OF EXPERIMENTAL DESIGNS

Shadish et al.¹⁹ enumerate a taxonomy of nine well-understood design errors for experimentation, referred to as *threats to internal validity*—that is, the degree to which valid causal conclusions can be drawn within the context of the study. Seven of these errors can be avoided when the researcher employs a *randomized experiment* that behaves as expected. The two remaining threats to validity that are *not* obviated by randomization are *attrition* and *testing*. Attrition may not have a meaningful definition in the context of online experiments, especially when outcomes are measured shortly after treatment exposure. Testing in experimental design refers to taking an initial measurement and then using the test instrument to conduct an experiment. Analysis may not be able to differentiate between the effect that a test was designed to measure and the effect of subjects learning the test itself. Testing is a form of *within-subjects* analysis that is not typically employed in online field experiments and whose analyses are outside the scope of this work. Therefore, *failed randomized assignment* is the primary threat to internal validity that we consider. Randomization failures in programs manifest differently from randomization failures in the physical world: for example, a program cannot disobey an experimental protocol, but data flow can break randomization if a probability is erroneously set to zero.

We characterize the ways in which syntactically valid PLANOUT programs can fail to randomize treatment assignment. Note that because there is currently no underlying formalism for the correctness of online field experiments that maps cleanly to a programming language context, we cannot define a soundness theorem for programmatically defined experiments. Some of the threats described here would be more properly considered code smells, rather than outright errors.⁸

4.1. Randomization failures

There are three ways a PLANOUT program may contain a failure of randomization: when it records data along a path that is not randomized, when the units of randomization have low cardinality, and when it encounters path-induced determinism. PLANALYZER detects all three automatically.

Recording data along nonrandomized paths occurs when there exists at least one recorded path through the program that *is* randomized and at least one recorded path through the program that *is not* randomized:

```
if (inExperiment(userid=userid)) {
  T = bernoulliTrial(p=0.5, unit=userid);
} else {
  T = true;
}
return true;
```

Such programs can typically be fixed by adding a `return false` for the appropriate path(s).

Units of randomization, such as `userid` or `deviceid`, must

have significantly higher cardinality than experimental treatments to ensure that each treatment is assigned a sufficient number of experimental units to make valid statistical inferences about the population. If the unit is an external variable unfamiliar to PLANALYZER, it will assume that the variable has low cardinality. PLANALYZER allows user-defined annotations to make its analyses more precise. Therefore, PlanOut users can correct their programs by either annotating the unit of randomization as having high cardinality, or reassessing their choice of unit.

Data-flow failures of randomization occur when inappropriate computations flow into units. PLANOUT allows units to be the result of arbitrary computations. For example, one PLANOUT script in our evaluation corpus sets the unit of randomization to be `userid * 2`. A PLANOUT user might want to do this when rerunning an experiment, to ensure that at least some users are assigned to a new treatment. However, this feature can lead to deterministic assignment when used improperly. The following is a syntactically valid PLANOUT program that triggers an error in PLANALYZER:

```
T1 = uniformChoice(choices=[400, 900], unit=userid);
T2 = bernoulliTrial(p=0.3, unit=T1);
```

When writing this code, the researcher may believe that there are four possible assignments for the pair of variables. However, because the assignment of input units to a particular value is the result of a deterministic hashing function, every user who is assigned `T1=400` is assigned the same value of `T2` because the input to the hash function for `bernoulliTrial` is always 400. Therefore, they will never record both (400, true) and (400, false) in the data, which likely contradicts the programmer's intent.

4.2. Treatment assignment failures

PLANALYZER requires that all assigned treatments along a path have the possibility of being assigned to at least one unit and that at least some treatments may be compared. There are three ways a PLANOUT program may contain a failure of treatment assignment: when some treatment has a zero probability of being assigned, when there are fewer than two treatments that may be compared along a path, and when dead code blocks contain treatment assignment.

Detecting the latter two cases are standard tasks in static program analysis. We note that for the first case, syntactically correct PLANOUT code permits authors to set probabilities or weights to zero, either directly or as the result of evaluation. Detecting this kind of value-dependent behavior is not unusual in program analysis either, but the reason why we wish to avoid it may not be obvious: to establish a causal relationship between variables, there must be at least two alternative treatments under comparison.

4.3. Causal sufficiency errors

One of the main assumptions underlying causal reasoning is *causal sufficiency*, or the assumption that there are no unmeasured confounders in the estimate of treatment

effect. Barring runtime failures, we have a complete picture of the assignment mechanism in PLANOUT programs. Unfortunately, a PLANOUT program may allow an unrecorded variable to bias treatment assignment.

Consider a program that assigns treatment on the basis of user country, accessed via a `getUserCountry` function:

```
if (getUserCountry(userid=userid) == 'US') {
  T = uniformChoice(choices=[7, 9], unit=userid);
} else {
  T = uniformChoice(choices=[4, 7, 9], unit=userid);
}
```

Treatment assignment of `T` depends on user country, so the user country is a potential confounder. Because this variable does not appear in the input program text, it cannot be recorded by the PLANOUT framework’s data recording system. Therefore, the program and resulting analyses will violate the causal sufficiency assumption.

If PLANALYZER encounters a static error or threat, it reports that the script failed to pass validation and gives a reason to the user. Some of the fixes are easy to determine from the error and could be applied automatically. We leave this to future work. Other errors require a more sophisticated understanding of the experiment the script represents and can only be determined by the script’s author.

5. PLANALYZER STATIC ANALYSIS TOOL

PLANALYZER is a command-line tool written in OCaml that performs two main tasks: it checks whether the input script represents a randomized experiment, and it generates all valid contrasts and their associated conditioning sets for scripts that can be analyzed using (C)ATE. Figure 1 provides an overview of the PLANALYZER system.

5.1. PlanOut intermediate representation (IR)

Upon parsing, PLANALYZER performs several routine program transformations, including converting variables to an identification scheme similar to SSA, performing constant propagation, and rewriting functions and relations (such as

equality) in A-normal form.^{1,4,15,18} Because it may not be possible to reason about the final values of a variable defined in a PLANOUT program due to the presence of external function calls, PLANALYZER reasons about intermediate values instead and reports results over a partially evaluated program.¹⁰

After these routine transformations, PLANALYZER splits the program into straight line code via tail duplication such that every path through the program may be evaluated in isolation of the others. Although this transformation is exponential in the number of conditional branches, in practice the branching factor of PLANOUT programs is quite small.

PLANALYZER then converts guards into assertions and uses the Z3 SMT solver to ensure variables assigned along paths are consistent with these assertions.⁵ For each assertion, PLANALYZER queries Z3 twice—first to obtain a satisfying solution, and then to test whether this solution is unique. Evaluation of the intermediate representation may contain unevaluated code, so if there is more than one solution, PLANALYZER keeps the code chunk abstract.

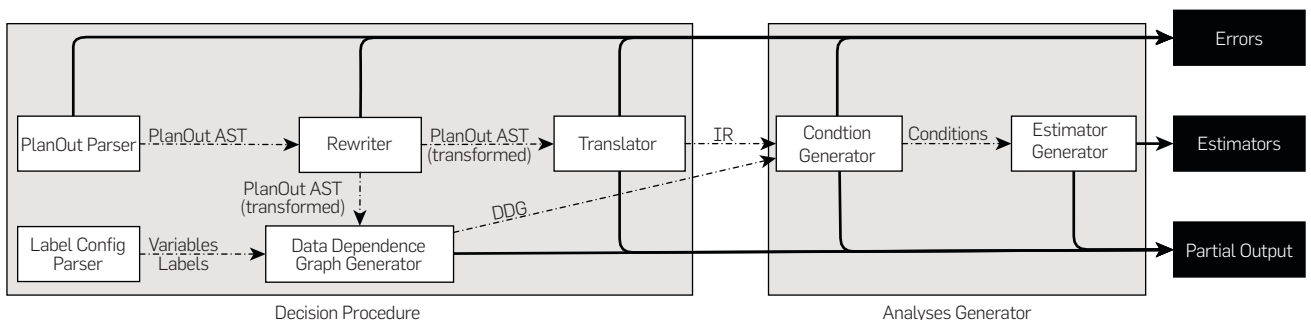
PLANALYZER uses SSA and A-normal form because they aid in contrast generation: a single execution of a PLANOUT program corresponds to the assignment of a unit to a treatment. However, additional intermediate variables can have somewhat ambiguous semantics when attempting to model a programmatically defined experiment causally; although they aid in, for example, the detection of causal sufficiency errors, they make reasoning about causal inference using methods such as causal graphical models quite difficult.

5.2. Variable labels for causal inference

The PLANOUT language contains only some of the necessary features for reasoning about the validity of experiments. Given only programs written in PLANOUT, PLANALYZER may not be able to reason about some common threats to internal validity. The interaction between random operators and control flow can cause variables to lose either their randomness or their variation. Furthermore, we need some way of guaranteeing that external operators do not introduce confounding.

To express this missing information, we introduce a

Figure 1. The PlanAnalyzer system transforms input PlanOut programs, possibly with user-provided variable labels, into a normalized form before translating the program to the intermediate representation (IR). PlanAnalyzer produces a data dependency graph in order to generate the data dependence graph (DDG) and resulting estimators. At each step in the analyses, PlanAnalyzer may produce errors. When there is insufficient information to produce estimators, but the input program has no known threats to validity, PlanAnalyzer provides as much partial output as possible.



4-tuple of variable labels (*rand*, *card*, *tv*, *corry*) that PLANALYZER attempts to infer and propagate for each PLANOUT program it encounters.^{17,6} Unsurprisingly, inference may be overly conservative for programs with many external functions or variables. To increase the scope of experiments PLANALYZER can analyze, users may supply PLANALYZER with global and local configuration files that specify labels.

Randomness (*rand*). PLANOUT may be used with existing experimentation systems; this means that there may already be sources of randomness available and familiar to users. Furthermore, as PLANOUT was designed to be extensible, users may freely add new random operators.

Cardinality (*card*). The size of variables' domains (cardinality) impacts an experiment's validity. Simple pseudorandom assignment requires high cardinality units of randomization to properly balance the assignment of units into conditions.

Time Variance (*tv*). For the duration of a particular experiment, a given variable may be constant or time-varying. Clearly, some variables are always constant or always time-varying. For example, *date-of-birth* is constant, whereas *days-since-last-login* is time-varying. However, there are many variables that cannot be globally categorized as either constant or time-varying. The *tv* label allows experimenters to specify whether they expect a variable to be constant or time-varying over the duration of a given experiment.

Because (C)ATE assumes subjects receive only one treatment value for the duration of the experiment, PLANALYZER cannot use them to estimate the causal effect of treatments or conditioning set variables having a *tv* label. A PLANOUT program may contain other valid contrasts assigned randomly, and independently from the time-varying contrasts; PLANALYZER will still identify these treatments and their conditioning sets as eligible for being analyzed via (C)ATE.

Covariates and Confounders (*corry*). Many experiments use features of the unit to assign treatment, which may introduce confounding. PLANALYZER automatically marks external variables and the direct results of nonrandom external calls as correlated with outcome (i.e., Y). This signals that, if the variable is used for treatment assignment, either their values must be recorded or sufficient downstream data must be recorded to recover their values.

5.3. Data dependence graph (DDG)

PLANALYZER builds a DDG to propagate variable label information.⁷ Because PLANOUT only has a single, global scope, its data dependence analysis is straightforward:

1. Assignment induces a directed edge from the references on the right-hand side to the variable name.
2. Sequential assignment of var_i and var_{i+1} induces no dependencies between var_i and var_{i+1} , unless the r-value of var_{i+1} includes a reference to var_i .
3. For an if-statement, PLANALYZER adds an edge from each of the references in the guard to all assignments in the branches.
4. In the case of an early return, PLANALYZER adds edges

from the variables in dependent guards to all variables defined after the return.

Random, independent assignment implies independence between potential causes, so long as the (possibly empty) conditioning set has been identified and recorded. PLANALYZER computes the DDG for the full script and uses the full DDG to determine when it is possible to marginalize over some variables.

Propagating variable labels. PLANALYZER marks variables directly assigned by built-in random functions or external random functions as random. The randomness label takes a tuple of identifiers as its argument. This tuple denotes the unit(s) of randomization, used for reasoning about causal estimators. Any node with a random ancestor is marked as random (with the exception of variables that do not vary), with units of randomization corresponding to the union of the ancestors' units.

If a random operator uses a low-cardinality unit of randomization, it will be marked as nonrandom. Note, however, that if the unit of randomization for a random function is a tuple with at least one high cardinality variable, then the resulting variable will remain random.

PLANALYZER propagates time-varying labels in the same manner as random labels. Unlike randomness, there is no interaction between the time-varying label and any other labels.

Converting DDGs to causal graphical models. Readers familiar with graphical models may wonder whether the DDG can be transformed into a directed graphical model. Programmatically defined experiments have two features that, depending on context, make such a transformation either totally inappropriate or difficult to extract: (1) deterministic dependence and (2) conditional branching. These two features can induce what is known as "context-sensitive independence," which limits the effectiveness of existing algorithms that would otherwise make graphical models an appealing target semantics. Although some work has sought to remedy branching, treatment of context-sensitive independence in graphical models more broadly is an open research problem.¹⁴ Furthermore, from a practical perspective, it is unclear how the versioned variables in the DDG ought to be unified, and some variables simply do not belong to a CGM (e.g., *userid*).

6. PLANOUT CORPORA

We analyze a corpus of PlanOut scripts from Facebook to evaluate PlanAnalyzer. We also make use of a corpus of manually specified contrasts that were used in the analysis of the deployed experimentation scripts. Scripts do not contain any user data, but may contain deidentified IDs (such as those of employees testing the scripts). Each experiment may have a temporary (but syntactically valid) representation captured by a snapshotting system, leading to multiple versions of a single experiment. Although we do not have access to the custom analyses of more complex experiments (e.g., database queries, R code, etc.), we can infer some characteristics of the intended analysis by partitioning the corpus into three subcorpora. Although we ana-

lyzed all three, we focus on just one here:

PlanOut-A. This corpus contains scripts that were analyzed using some form of ATE (i.e., $Avg(Y|T_1 = t_1^0, \dots, T_n = t_n^0) - Avg(Y|T_1 = t_1^1, \dots, T_n = t_n^1)$), where the variables T_1, \dots, T_n were manually specified and automatically recorded during the duration of the experiment. Users may manually specify that a subset of the recorded variables be continuously monitored for pairwise ATE. Neither the recording nor the data analysis tools have any knowledge of PLANOUT. This is the main corpus we will use for evaluating PLANALYZER, because the goal of PLANALYZER is to automate analyses that firms such as Facebook must now do manually.

Note that users of PlanOut at Facebook are typically either experts in the domain of the hypotheses being tested or they are analysts working directly with domain experts.

6.1. Characterizing representative PLANOUT programs

We designed PLANALYZER’s analyses on the basis of the universe of syntactically valid PLANOUT programs and our domain knowledge of experimentation. We built PLANALYZER from the perspective that (1) PLANOUT is the primary means by which experimenters design and deploy experiments, but (2) they can use other systems, if they exist. Facebook uses many experimentation systems and has a variety of human and code-review methods for the functionality that PLANALYZER provides. Therefore, we wanted to know: what are some characteristics of PLANOUT programs that people actually write and deploy?

We found that engineers and data scientists at Facebook used PLANOUT in a variety of surprising ways and had coding habits that were perhaps indicative of heterogeneity in the programming experience of authors. Through conversations with engineers at Facebook, we have come to understand that most PLANOUT authors can be described along the two axes depicted in Table 1.

Table 2 enumerates the errors raised by PLANALYZER over the three corpora. Each warning does not necessarily indicate an error during deployment or analysis, due to the fact that there are preexisting mechanisms and idiosyncratic usages of PLANOUT.

PLANOUT-A contains our highest quality data: all scripts were vetted by experts before deployment, with some component analyzed using ATE. Figure 2 provides a lightly

Table 1. Experience matrix for PlanOut authors.

		Programming	Experience
		High	Low
Experimental Design Experience	Low High	I	II
		III	IV

The horizontal axis represents programming experience or ability; the vertical axis represents experience in experimental design. We believe most authors represented in the PlanOut corpora are in quadrants I and II. PlanAlyzer’s novel analyses target experiment authors in quadrants I-III and may be especially useful for authors in III, whom we believe are underrepresented in the corpora. We conjecture, but cannot verify, that most of the errors PlanAlyzer flags in the corpora belong to authors in II.

anonymized example program that PLANALYZER identified as having a potential error. Its style and structure is a good representation of real-world PLANOUT programs.

We found the following coding practices in PLANOUT-A:

Ambiguous Semantics and Type Errors. Because PLANALYZER must initially perform type inference, it found 87 scripts in PLANOUT-A that had type errors, which suggest

Table 2. The counts of code smells, static script errors, and tool failures found when running PlanAlyzer on the PlanOut-A corpus A PlanAlyzer error does not necessarily indicate the experiment was run in error.

Output category	Scripts (566)	Exps. (240)
Not an experiment	10	10
Low cardinality unit	7	1
Ambiguous semantics	5	2
Type inconsistencies	10	4
Causal sufficiency errors	111	54
False positive	47	23
Testing code	23	8
Possible random assignment	41	23
Recorded no randomization	25	11
Missed paths (tests)	4	1
No randomization (config)	12	7
Possible random assignment	9	3
Random variable no variation	2	2
No positivity	7	3
Dead code	5	4
Feature not implemented in tool	29	8

A single experiment may have many script versions, not all of which were deployed. The numbers for PlanOut-A reflect the state of the corpus after adjustments for easily fixed type inconsistencies (initially 87), because we know those scripts ran in production, and wanted to see if PlanAlyzer could find more interesting errors or smells.

Figure 2. A representative, lightly edited and anonymized experiment written in PlanOut. This script mixes testing code with experimentation code. Lines 5–12 set values for the author of the script whose userid is AUTHOR_ID and records those values. The actual experiment is in lines 14–26. It is only conducted on the population defined by the external predicate and the user being recorded in (represented here when the userid is 0). PlanAlyzer raises an error for this script.

```

1  show_feature = true;
2  in_exp = false;
3  if (userid == AUTHOR_ID) {
4    in_exp = true;
5    if (post_has_photo == true) {
6      show_feature = false;
7    }
8    if (post_has_video == true) {
9      show_feature = false;
10   }
11   return true;
12 }
13 in_pop = extPred(ep="in_pop", userid=userid);
14 if (!in_pop || userid == 0) {
15   return false;
16 }
17 in_exp = bernoulliTrial(p=1/100, unit=post_id)
18 ;
19 if (!in_exp) {
20   return false;
21 }
22 if (post_has_photo == true) {
23   show_feature = false;
24 }

```

there might be some utility in providing our type checking facility to users of PLANOUT.

We also found three scripts from one experiment that applied the modulus operator to a fraction; because PLANOUT uses the semantics of its enclosing environment for numeric computation, this script will return different values if it is run using languages with different semantics for modulus, such as PHP versus JavaScript.

Modifying deployment settings within experimentation logic. Some of the scripts marked as not experiments begin with `return false` and had an unreachable and fully specified experiment below the `return` statement. PLANALYZER flags dead code in PLANOUT programs, because it can be the result of a randomly assigned variable causing unintended downstream control flow behavior. However, every dead code example we found had the form `condition = false; if (condition)...` These features occurred exclusively in experiments that had multiple scripts associated with them that did not raise these errors. After discussing our findings with engineers at Facebook, we believe that this might be a case of PLANOUT authors modifying the experiment although it is running to control deployment, rather than leaving dead-code in by accident, as it appears from PLANALYZER’s perspective.

Using PlanOut for Application Configuration. One of the most surprising characteristics we found in PLANOUT-A was the prevalence of using PLANOUT for application configuration, à la Akamai’s ACMS system or Facebook’s Gatekeeper.^{20, 21} When these scripts set variables, but properly turned off data recording (i.e., returned false), PLANALYZER marked them as not being experiments. When they did not turn off logging, they were marked as recording paths without randomization. Some instances of application configuration involved setting the support of a randomly assigned variable to a constant or setting a weight to zero. Because experiments require variation for comparison, PLANALYZER raises an error if the user attempts to randomly select from a set of fewer than two choices. Three scripts contained expressions of the form `uniformChoice (choices = [v], unit=userid)` for some constant value v .

As a result, engineers who aim to use PLANOUT as a configuration system have no need for PLANALYZER, but anyone writing experiments would consider these scripts buggy.

Mixing external calls to other systems. Almost 20% of the scripts (106) include calls to external experimentation systems. In a small number of cases, PLANOUT is used exclusively for managing these other systems, with no calls to its built-in random operators.

Nonread-only units. One of the other firms we spoke to that uses PLANOUT treats units of randomization as read-only, unlike other variables in PLANOUT programs. Facebook does not do this. Therefore, programs that modify the unit of randomization may be valid; for instance, the aforementioned instance where the unit was set to `userid * 2`. We also observed a case where the unit was set to be the result of an external call—without knowing the behavior of this external call, it is assumed to be low cardinality. In this case, the experiment was performing cluster random assignment, which is not covered by ATE

and out of scope for PLANALYZER.

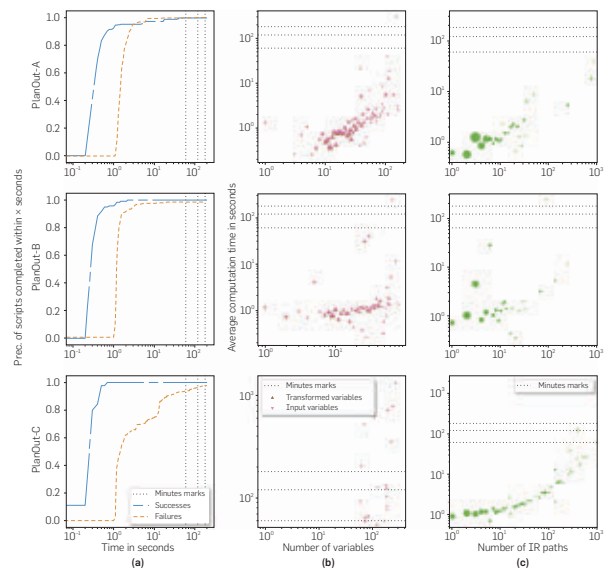
6.2. Planalyzer performance

All analyses were run on a MacBook Air with a 1.6 GHz Intel Core i5 processor having four logical cores. The longest runtime for any analysis was approximately three min; runtime scales linearly with the number of “paths” through the program, where a path is defined according to the transformed internal representation of the input PLANOUT program and is related to the number of conditioning sets. PLANALYZER uses the Z3 SMT solver⁵ to ensure conditioning sets are satisfied and to generate treatments,^{23, 9} so both the number of variables in the program and the number of paths in the internal representation could cause a blowup in runtime. We found that runtime increases linearly with the number of internal paths, but possibly exponentially with the number of variables, as depicted in Figure 3.

PLANALYZER produces meaningful contrasts that are comparable with the human-specified gold standard, automatically generating 82% of our eligible gold-standard contrasts. PLANALYZER runs in a reasonably short amount of time; likely due to PLANOUT’s generally small program sizes.

Summary. We did not expect to see any real causal sufficiency errors due to the expert nature of the authors of PLANOUT-A. Rather, we expect to see some false positives due to the fact that PLANALYZER is aggressive about flagging

Figure 3. Wall-clock timing data for the PlanOut corpus. Plots in column (a) depict the empirical CDF of all scripts on a log-scale. Plots in columns (b) and (c) show the relationship between the running time and features of the PlanOut script we might expect to affect running time, on log-scale on both axes. Plots in column (b) show both the number of variables in the input PlanOut script, and the number of variables in the transformed, intermediate representation of the PlanOut program. Plots in column (c) depict the relationship between the number of paths through PlanOut programs and their running time. The times depicted in both (b) and (c) are averages over scripts satisfying the x-axis value, and the size of the points are proportional to the number of scripts used to compute that average. We chose this representation, rather than reporting error bars, because the data is not iid.



potential causal sufficiency errors. We made this design choice because the cost of unrecorded confounders can be very high.

PLANOUT scripts in deployment at Facebook represent a range of experimental designs. We observed factorial designs, conditional assignment, within-subjects experiments, cluster random assignment, and bandits experiments in the scripts we examined.

7. EVALUATION

Real-world PLANOUT scripts unsurprisingly contained few errors, because they were primarily written and overseen by experts in experimental design. Therefore, to test how well PLANALYZER finds errors, we selected a subset of fifty scripts from PLANOUT-A and mutated them. We then validated a subset of the contrasts PLANALYZER produced against a corpus of hand-selected contrasts monitored and compared by an automated tool used at Facebook. Finally, we reported on PLANALYZER's performance, because its effectiveness requires accurately identifying meaningful contrasts within a reasonable amount of time.

7.1. Mutation methodology

We first identified scripts that were eligible for this analysis. We modified the PLANOUT-A scripts that raised errors when it was appropriate to do so. For example, we updated a number of the scripts that erroneously raised causal sufficiency errors so that they would not raise those errors anymore. We excluded scripts that, for example, contained testing code or configuration code. This allowed us to be reasonably certain that most of the input scripts were correct.

All of our mutations operate over input PLANOUT programs, rather than the intermediate representation. We believed this approach would better stress PLANALYZER. We perform one mutation per script.

We considered two approaches when deciding how to perform the mutations:

1. Randomly select a mutation type, and then randomly select from the eligible AST points for that mutation.
2. Generate all of the eligible AST points for all of the mutations, and then randomly select from this set.

Method 1 leads to an even split between the classes of mutations in the test corpus; method 2 leads to frequencies that are proportional to the frequencies of the eligible AST nodes. We chose the latter because we believed it would lead to a more accurate representation of real programming errors.

To select the subset of scripts to evaluate, we sampled 50 experiments and then selected a random script version from that experiment. We then manually inspected the mutated script and compared the output of the mutation with the original output.

Findings: fault identification over mutated scripts. When analyzing our sample of 50 mutated scripts, Planalyzer produced only one false positive and only one false negative. The precision and recall were both 92%. On the one hand, this is very surprising, given both the false positive rate in

the PLANOUT-A corpus for causal sufficiency errors (CSE) at 8%, and the proportion of CSE mutations in this sample (28%). However, we found that most of the CSE mutations caused the program to exit before random assignment, causing PLANALYZER to raise legitimate errors about recorded paths with no randomization. The rest were true causal sufficiency errors (i.e., they would cause bias in treatment). The one false negative we observed occurred in a script that redefined the treatment variable for two userids, in what appears to be testing code. The mutation wrapped the redefined treatment, so this is a case where PLANALYZER should have raised a “no randomization error” in both the input script as well as the mutated script.

7.2. Validation against human-generated contrasts

We decided whether an experiment should be in the subset according to the following three criteria: (1) all variables in the human-generated contrasts appeared in the original script; (2) PLANALYZER was able to produce at least one contrast for the experiment; and (3) PLANALYZER produced identical contrasts across all versions of the experiment. Criteria (1) and (2) ensure that analysis does not require knowledge unavailable to PLANALYZER. Criteria (3) is necessary because the tool that monitors contrasts logs them per experiment, not per version. If the possible contrasts change between versions, we cannot be sure which version corresponded to the data. Ninety-five of the unique experiments met these criteria.

Findings: contrast generation. PLANALYZER found equivalent contrasts for 78 of the 95 experiments. For 14 experiments, it produced either partial contrasts or no contrasts. In each of these cases, the desired contrast required summing over some of the variables in the program (marginalization), or more sophisticated static analysis than the tool currently supports. Because it is computationally expensive to produce every possible subset of marginalized contrasts, we consider the former to be an acceptable shortcoming of the tool. Finally, three experiments had issues with their human-generated contrasts (no contrasts, or ambiguous or unparsable data).

8. CONCLUSION

The state of the art for auditing experiments and for generating their associated statistical analyses is almost entirely a manual process. This is the first work that analyzes field experiments statically. We propose a new class of errors and threats unique to the programmatic specification of experimentation. We have implemented a tool that, for the most common class of experiments, automatically identifies threats and generates statistical analyses. We compare the output of PLANALYZER against human-generated analyses of real PLANOUT scripts and find that PLANALYZER produces comparable results.

ACKNOWLEDGMENTS

This research was in part conducted while Emma Tosch was an employee of Facebook. Although at the University of Massachusetts Amherst, Emma Tosch was supported in part by the United States Air Force under Contract No. FA8750-17-C-0120. Any opinions, findings, and conclusions or recommendations expressed in this material are those of

the author(s) and do not necessarily reflect the views of Facebook nor the United States Air Force. 

References

1. Aho, A.V., Sethi, R., Ullman, J.D. *Compilers, Principles, Techniques*. Addison Wesley, Boston, Massachusetts, 1986.
2. Bakshy, E., Eckles, D., Bernstein, M.S. Designing and deploying online field experiments. In *Proceedings of the 23rd International Conference on World Wide Web, WWW '14* (New York, NY, USA, 2014), ACM, New York, 283–292.
3. Crook, T., Frasca, B., Kohavi, R., Longbotham, R. Seven pitfalls to avoid when running controlled experiments on the web. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining* (New York, NY, USA, 2009), ACM, New York, 1105–1114.
4. Cytron, R., Ferrante, J., Rosen, B.K., Wegman, M.N., Zadeck, F.K. Efficiently computing static single assignment form and the control dependence graph. *ACM Trans. Programming Languages and Systems (TOPLAS)* 13, 4 (1991), 451–490.
5. De Moura, L., Bjørner, N. Z3: An efficient SMT solver. In *International conference on Tools and Algorithms for the Construction and Analysis of Systems* (Berlin, Germany, 2008), Springer, Berlin, Germany, 337–340.
6. Denning, D. E. A lattice model of secure information flow. *Commun. ACM* 19, 5 (1976), 236–243.
7. Ferrante, J., Ottenstein, K.J., Warren, J.D. The program dependence graph and its use in optimization. *ACM Trans. Programming Languages and Systems (TOPLAS)* 9, 3 (1987), 319–349.
8. Fowler, M. *Refactoring: Improving the Design of Existing Code*. Addison-Wesley Professional, Boston, MA, USA, 2018.
9. Fredrikson, M., Jha, S. Satisfiability modulo counting: A new approach for analyzing privacy properties. In *Proceedings of the Joint Meeting of the 23rd EACSL Annual Conference on Computer Science Logic (CSL) and the 29th Annual ACM/IEEE Symposium on Logic in Computer Science (LICS)* (New York, NY, USA, 2014), ACM, New York, 42.
10. Futamura, Y. Partial evaluation of computation process—an approach to a compiler-compiler. *Higher-Order Symbolic Comput.* 12, 4 (1999), 381–391.
11. Gordon, A.D., Henzinger, T.A., Nori, A.V., Rajamani, S.K. Probabilistic programming. In *Proceedings of the on Future of Software Engineering* (New York, NY, USA, 2014), ACM, New York, 167–181.
12. Kohavi, R., Deng, A., Frasca, B., Walker, T., Xu, Y., Pohlmann, N. Online controlled experiments at large scale. In *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining* (New York, NY, USA, 2013), ACM, 1168–1176.
13. Kohavi, R., Longbotham, R.,

- Sommerfield, D., Henne, R.M. Controlled experiments on the web: Survey and practical guide. *Data Mining and Knowledge Discovery* 18, 1 (2009), 140–181.
14. Minka, T., Winn, J. Gates. In *Advances in Neural Information Processing Systems* (Online, 2009), JMLR: W&CP, 1073–1080.
15. Muchnick, S.S. *Advanced Compiler Design Implementation*. Morgan Kaufmann, Burlington, MA, USA, 1997.
16. Pfeffer, A. *Practical Probabilistic Programming*. Manning Publications Co., Shelter Island, NY, USA, 2016.
17. Sabelfeld, A., Myers, A.C. Language-based information-flow security. *IEEE J. Selected Areas Commun.* 21, 1 (Jan 2003), 5–19.
18. Sabry, A., Felleisen, M. Reasoning about programs in continuation-passing style. *Lisp Symbolic Comput.* 6, 3-4 (1993), 289–360.
19. Shadish, W.R., Cook, T.D., Campbell, D.T. *Experimental and Quasi-Experimental Designs for Generalized Causal Inference*. Houghton Mifflin Company, Boston, MA, USA, 2002.
20. Sherman, A., Lisiecki, P.A., Berkheimer, A., Wein, J. Acms: The akamai configuration management system. In *Proceedings of the 2nd conference on Symposium on Networked Systems Design & Implementation, Volume 2* (Berkley, CA, USA, 2005), USENIX Association, USA, 245–258.
21. Tang, C., Kooburat, T., Venkatachalam, P., Chander, A., Wen, Z., Narayanan, A., Dowell, P., Karl, R. Holistic configuration management at facebook. In *Proceedings of the 25th Symposium on Operating Systems Principles* (New York, NY, USA, 2015), ACM, New York, 328–343.
22. Tang, D., Agarwal, A., O'Brien, D., Meyer, M. Overlapping experiment infrastructure: More, better, faster experimentation. In *Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (New York, NY, USA, 2010), ACM, New York, 17–26.
23. Valiant, L.G. The complexity of computing the permanent. *Theor. Comput. Sci.* 8, 2 (1979), 189–201.

Emma Tosch (etosch@uvm.edu), University of Vermont, Burlington, VT, USA.
Eytan Bakshy (ebakshy@fb.com), Facebook, Inc., Menlo Park, CA, USA.
Emery D. Berger (emery@cs.umass.edu), University of Massachusetts, Amherst, MA, USA.

David D. Jensen (jensen@cs.umass.edu), University of Massachusetts, Amherst, MA, USA.
J. Eliot B. Moss (moss@cs.umass.edu), University of Massachusetts, Amherst, MA, USA.

© 2021 ACM 0001-0782/21/9 \$15.00

ACM Student Research Competition

Attention: Undergraduate and Graduate Computing Students



The ACM Student Research Competition (SRC), sponsored by Microsoft, 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 undergraduate and graduate students to earn many tangible and intangible rewards from participating:

- **Awards:** cash prizes, medals, and ACM student memberships
- **Prestige:** Grand Finalists and their advisors are invited to the Annual ACM Awards Banquet, where they are recognized for their accomplishments
- **Visibility:** opportunities to meet with researchers in their field of interest and make important connections
- **Experience:** opportunities to sharpen communication, visual, organizational, and presentation skills in preparation for the SRC experience

Learn more about ACM Student Research Competitions: <https://src.acm.org>

Daegu Gyeongbuk Institute of Science & Technology (DGIST)

DGIST Second half of 2021 Tenure-Track Faculty Public Invitation

It is an honor to have a professor with excellent ability to realize the vision of a Convergence University that changes the world with innovation through convergence education and leading high-tech research along with respect.

POSITIONS

Graduate School

- ▶ Emerging Materials Science
 - Analytical Chemistry
 - Polymer Chemistry
 - Materials Physics
- ▶ Information and Communication Engineering
 - All areas in electrical engineering and computer science including but not limited to the following
 - Machine Learning Theory, Reinforcement Learning, NLP, Computer Vision, and other related areas in AI and ML
 - Computer Security, Database/Data Mining
 - Areas related to Autonomous Vehicles, 6G Communications/Networks, Bio Sensors and Systems, Radar/LiDAR Sensors and Systems, Power/RF Technologies
 - High Speed ADC and RF Circuit, Highly Energy-Efficient Wireless Power Transfer Systems, Emerging Communication Systems Design, Digital Circuits for AI/Security
 - Neuromorphic Device & Emerging Electronic Device

- ▶ Robotics Engineering
 - AI theories and applications for robotics: AI algorithm, deep learning, machine learning, motion planning, intelligent control and other related topics
 - Autonomous vehicle technology: computer vision, SLAM, vehicle control, intelligent transportation system and other related research topics
 - General robotics: cooperative robot, industrial robot, humanoid, surgery / rehabilitation robot, exoskeleton, mobile robot and other related research topics
 - All areas in mechanical engineering and electrical engineering related to robotics for exceptional candidates

- ▶ Energy Science and Engineering
 - All areas in chemistry, physics, materials science, electrical engineering, and chemical engineering related to energy conversion, storage, and saving as well as other semiconductor-related topics
 - Design, synthesis, fabrication, and characterization of materials and devices
 - Characterization of structure, properties, dynamics, and functions in energy materials
 - Transmission electron microscope (TEM), high-resolution STEM, EELS, tomography
 - Spectroscopy (pump-probe, time-resolved instantaneous PL & absorption, ultrafast multidimensional spectroscopy)
 - Device physics, fabrication, and characterization related to electronic and optoelectronic devices

- ▶ Brain and Cognitive Science
 - All areas in brain and cognitive sciences
 - Other areas in biological sciences that are applicable to neuroscience, such as advanced biological techniques (e.g. Omics, Imaging, Bioinformatics) and classical approaches (e.g. Biochemistry, Cell Biology)

- ▶ New biology
 - Bioinformatics/Microbial metagenomics
 - Chemical Biology
 - Immunology
 - Plant Development and Biochemistry
- ▶ Interdisciplinary Engineering
 - All areas in interdisciplinary engineering including, but not limited to, the following

<ICT area>

- Innovative electronic devices
- Advanced driver assistance systems(ADAS) for automated driving platform
- Machine learning
- Radar system design & super-resolution radar
- Human-robot interaction

<Materials area>

- Electromagnetic processing of materials
- Polymer chemistry
- Nano materials
- Thin film solar cells
- Photonic/optical materials

<Biotechnology area>

- Disease Diagnosis
- Prevention and therapeutic technology
- Candidates with the following cooperative experiences preferred:
 - Teaching experience in relevant fields
 - Practical experience in relevant research fields
 - Experience in venture start-up or technology transfer

College of Transdisciplinary Studies

School of Undergraduate Studies

The School of Undergraduate Studies at DGIST is seeking to fill a tenure-track faculty position in the area of Technology Management, Marketing Management, Business Strategy, or Management Information Systems. The appointment may be made at the Assistant Professor or Associate Professor levels, commensurate with qualifications and experience.

The successful candidate will be expected to teach and mentor at both the undergraduate and graduate levels, if necessary, in English. Applicants must have a doctoral degree. Applicants with experience with venture start-up/incubation and technology management are preferred. The appointed candidate will be committed to teaching an Entrepreneurship and Social Responsibility course and developing other entrepreneurship-related courses for undergraduate students. The appointed candidate will also be developing and managing the Technical Venture Academy and MA entrepreneurship program.

Date of Appointment

- ▶ March 1st, 2022 (Appointment date can be adjusted in consultation with department)

Qualification

- ▶ Encourage support for female scientists
- ▶ With no reasons for disqualification based on related Korean Law (STATE PUBLIC OFFICIALS ACT Article 33)
- ▶ Ph.D Holder with ability to teach in English required
- ▶ Without distinction of nationality

How to Apply

- ▶ Apply after accessing <https://faculty.dgist.ac.kr/>
- ▶ Application Period - July 15th, 2021 (Thur) ~ August 3th, 2021 (Tue) 18:00 (GMT+09:00)



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



ACM BOOKS

Collection II

This book introduces the concept of Event Mining for building explanatory models from analyses of correlated data. Such a model may be used as the basis for predictions and corrective actions. The idea is to create, via an iterative process, a model that explains causal relationships in the form of structural and temporal patterns in the data. The first phase is the data-driven process of hypothesis formation, requiring the analysis of large amounts of data to find strong candidate hypotheses. The second phase is hypothesis testing, wherein a domain expert's knowledge and judgment is used to test and modify the candidate hypotheses.

The book is intended as a primer on Event Mining for data-enthusiasts and information professionals interested in employing these event-based data analysis techniques in diverse applications. The reader is introduced to frameworks for temporal knowledge representation and reasoning, as well as temporal data mining and pattern discovery. Also discussed are the design principles of event mining systems. The approach is reified by the presentation of an event mining system called EventMiner, a computational framework for building explanatory models. The book contains case studies of using EventMiner in asthma risk management and an architecture for the objective self. The text can be used by researchers interested in harnessing the value of heterogeneous big data for designing explanatory event-based models in diverse application areas such as healthcare, biological data analytics, predictive maintenance of systems, computer networks, and business intelligence.

<http://books.acm.org>

<http://store.morganclaypool.com/acm>



Event Mining *for explanatory modeling*

Laleh Jalali
Ramesh Jain

ISBN: 978-1-4503-8483-4

DOI: 10.1145/3462257

[CONTINUED FROM P. 120] game of Go through the trial-and-error learning that we see in reinforcement learning.

Eventually, you built a system that learned to play Go on a smaller, nine-by-nine sized board.

We had some successes in the early days on the small-sized boards. Our system did learn, through these very principled trial-and-error reinforcement learning techniques, to associate different patterns with whether they would lead to winning or losing the game. Then I started collaborating Sylvain Gelly at the University of Paris on a project called MoGo, which became the first nine-by-nine Go championship program.

Later, you reconnected with your former Cambridge University classmate and DeepMind co-founder, Demis Hassabis, to continue that work with AlphaGo—which became the first computer program to beat a professional player on a full-sized 19x19 Go board.

I was very keen to have another look at computer Go when I arrived at DeepMind, because it felt like deep learning represented a very promising new possibility. So we started with a research question, namely whether deep learning could address the position evaluation problem. If you look at a pattern of stones on the board, can you predict who's going to win? Can you identify a good move? As we started working on that research question, it quickly became apparent that the answer was "yes." My feeling was that if we could build a system that could achieve the level of amateur *dan* through neural networks that simply examined a position and picked a move, with no precepts whatsoever—and none of the expertise that game engines have always had—it was time to hit the accelerate button.

In 2016, AlphaGo beat world champion Lee Sedol, who has said that the experience has made him a better player. What do you make of that?

There are multiple books now written on how human players should use AlphaGo strategies in their own playing. It has challenged people to think more holistically about the game, rath-

“AlphaZero will often give up a lot of material in a way that can be quite shocking to chess players, to gain a long-term edge over its opponent.”

er than in terms of local contributions to the score.

The same thing has happened in chess with AlphaGo's successor, AlphaZero, a program that has achieved superhuman performance despite starting without any human data or prior knowledge except the game's rules.

In contrast to the way that previous computer programs played the game, AlphaZero has encouraged people to be more flexible; to move away from material evaluation and understand that there are positions that can be enormously valuable in the long run. AlphaZero will often give up a lot of material in a way that can be quite shocking to chess players, to gain a long-term edge over its opponent.

AlphaGo suffered from what you called ‘delusions’, that is, persistent holes in its evaluation of a play that led it to make mistakes. How did you address these delusions in AlphaZero?

We tried many different things, but ultimately, it came down to being more principled. The more you trust your trial-and-error learning to correct its own errors, the fewer delusions the system will suffer from. We started off with a dataset that contained 100 different delusional positions. By the time we trained up AlphaZero, it got every single one of those delusional positions correct in its understanding. The more iterations of training it went through, the more those delusions it could correct.

So there was no piece of specific additional training that was required?

The fundamental process of reinforcement learning is one of recogniz-

ing the holes in your own knowledge and getting the opportunity to correct them. That correction process leads to better results, and we really need to trust it. We would rerun the same algorithm again from new random weights and see it track the same progress, fixing the same delusions in roughly the same order, as if it were peeling its own onion layer by layer.

AlphaZero has mastered a number of different games, from Shogi to Space Invaders. Others have found even broader applications.

The beautiful thing about creating a general-purpose algorithm is that you end up being surprised by the ways in which it is used, and I think that's been true here as well. One group used AlphaZero to do retro-chemical synthesis and found that it outperformed all previous baselines. Another group used it to solve one of the outstanding problems in quantum computation, namely to optimize the quantum dynamics. A startup in North Africa used AlphaZero to solve logistical problems. It is quite nice when other people take your algorithm to achieve good results.

Where is that work taking you next?

I try to ask what seems like the deepest science question. In this case, it felt to me that rather than trying another game, we should address what happens in applications where you don't know the rules—where you're interacting with people or with the real world or where you're dealing with complicated, messy dynamics that no one tells you about. We built a version of this approach that we call MuZero. MuZero is able to learn a model of the rules or dynamics and uses that to plan and solve problems. It is kind of amazing; we plugged it back into ChessGo and Shogi, and found that it could reach superhuman performance just as quickly, even without telling it the rules of the game. It was also able to beat baseline results in some of the more traditional reinforcement learning benchmarks, like Atari, where we'd previously been limited to model-free techniques without any lookahead planning.

Leah Hoffmann is a technology writer based in Piermont, NY, USA.

© 2021 ACM 0001-0782/21/9 \$15.00

Q&A

Playing With, and Against, Computers

2019 ACM Computing Prize recipient David Silver on developing the AlphaGo algorithm, his fascination with Go, and on teaching computers to play.

GAMES HAVE LONG been a fertile testing ground for the artificial intelligence community, and not just because of their accessibility to the popular imagination. Games also enable researchers to simulate different models of human intelligence, and to quantify performance. No surprise, then, that the 2016 victory of DeepMind's AlphaGo algorithm—developed by 2019 ACM Computing Prize recipient David Silver, who leads the company's Reinforcement Learning Research Group—over world Go champion Lee Sedol generated excitement both within and outside of the computing community. As it turned out, that victory was only the beginning; subsequent iterations of the algorithm have been able to learn without any human data or prior knowledge except the rules of the game and, eventually, without even knowing the rules. Here, Silver talks about how the work evolved and what it means for the future of general-purpose AI.

You grew up playing games like chess and Scrabble. What drew you to Go?

I learned the game of Go when I was a young kid, but I never really pursued it. Then later on, when I first moved to London, I started playing in a club in Hampstead in a crypt at the bottom of a church. It is a fascinating and beautiful game. Every time you think you know something about Go, you discover—like peeling an onion—there is another level of complexity to it.



“Every time you think you know something about Go, you discover—like peeling an onion—there is another level of complexity to it.”

When did you start thinking about teaching computers to play?

I think it was always in my mind. One of the things that drew me to Go as a human player was the understanding that it was a challenging game for computers to play. Humans possess an intuition in Go that appears far beyond the scope of brute force computation, and this—along with the rich history of the game—lends the game a certain mystique. That subsequently led to my work understanding it as a computer scientist.

After a few years working in the games industry, you went to the University of Alberta to get your Ph.D. and see if reinforcement learning techniques could help computers crack Go.

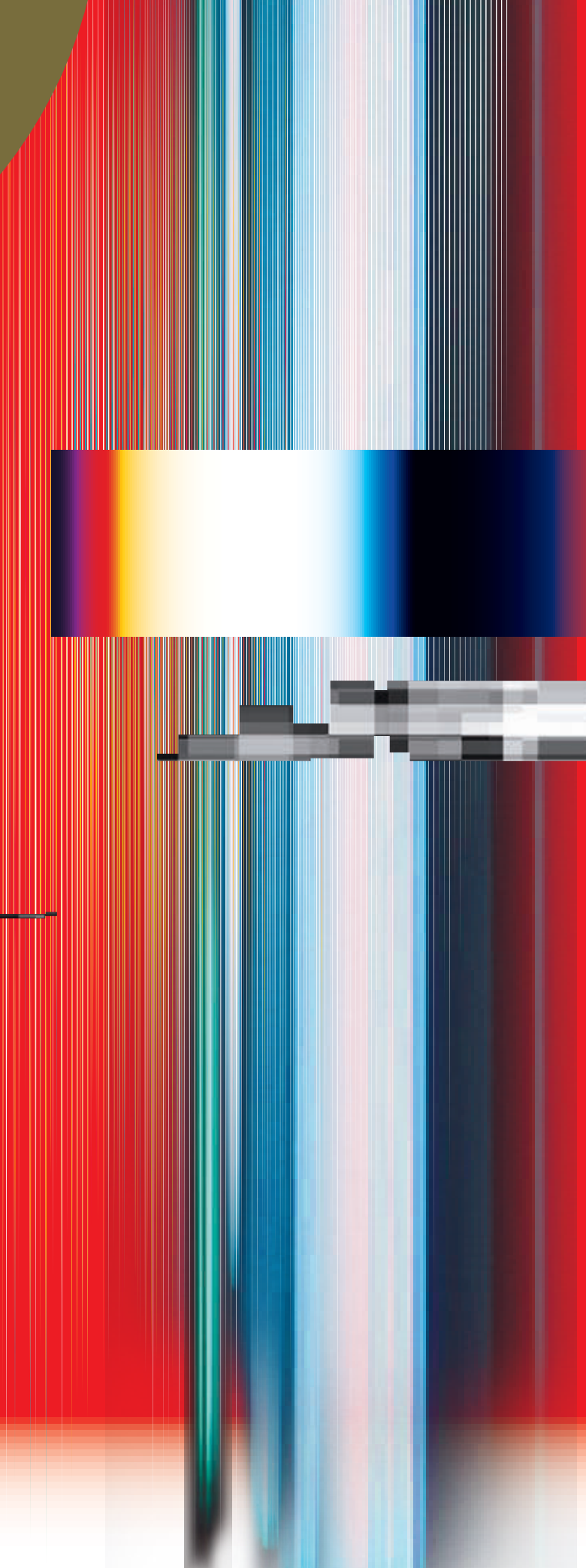
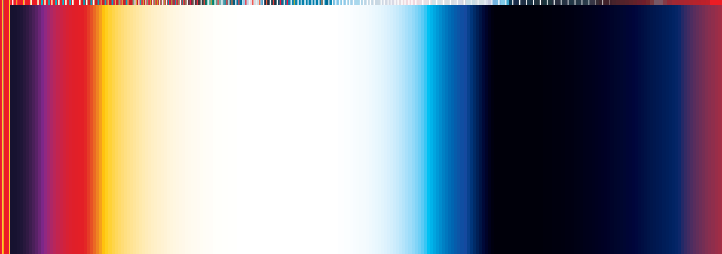
I was working in the games industry, and I took a year out to try and figure out what to do next. I knew I wanted to go back and study AI, but I wasn't sure what direction to take, so I started reading around, and I came across Sutton and Barto's *Reinforcement Learning: An Introduction*. The moment I read that book, something just connected; it seemed to represent the most promising path for understanding how to solve a problem from first principles. Alberta had both the best games research group in the world and also the best group on reinforcement learning. My idea was to put those things together and try to solve the [CONTINUED ON P. 119]



SIGGRAPH ASIA 2021 TOKYO

CONFERENCE 14 - 17 DECEMBER 2021
EXHIBITION 15 - 17 DECEMBER 2021
TOKYO INTERNATIONAL FORUM, JAPAN

sa2021.siggraph.org





MATLAB SPEAKS MACHINE LEARNING

With MATLAB® you can use clustering, regression, classification, and deep learning to build predictive models and put them into production.

mathworks.com/machinelearning

©2021 The MathWorks, Inc.