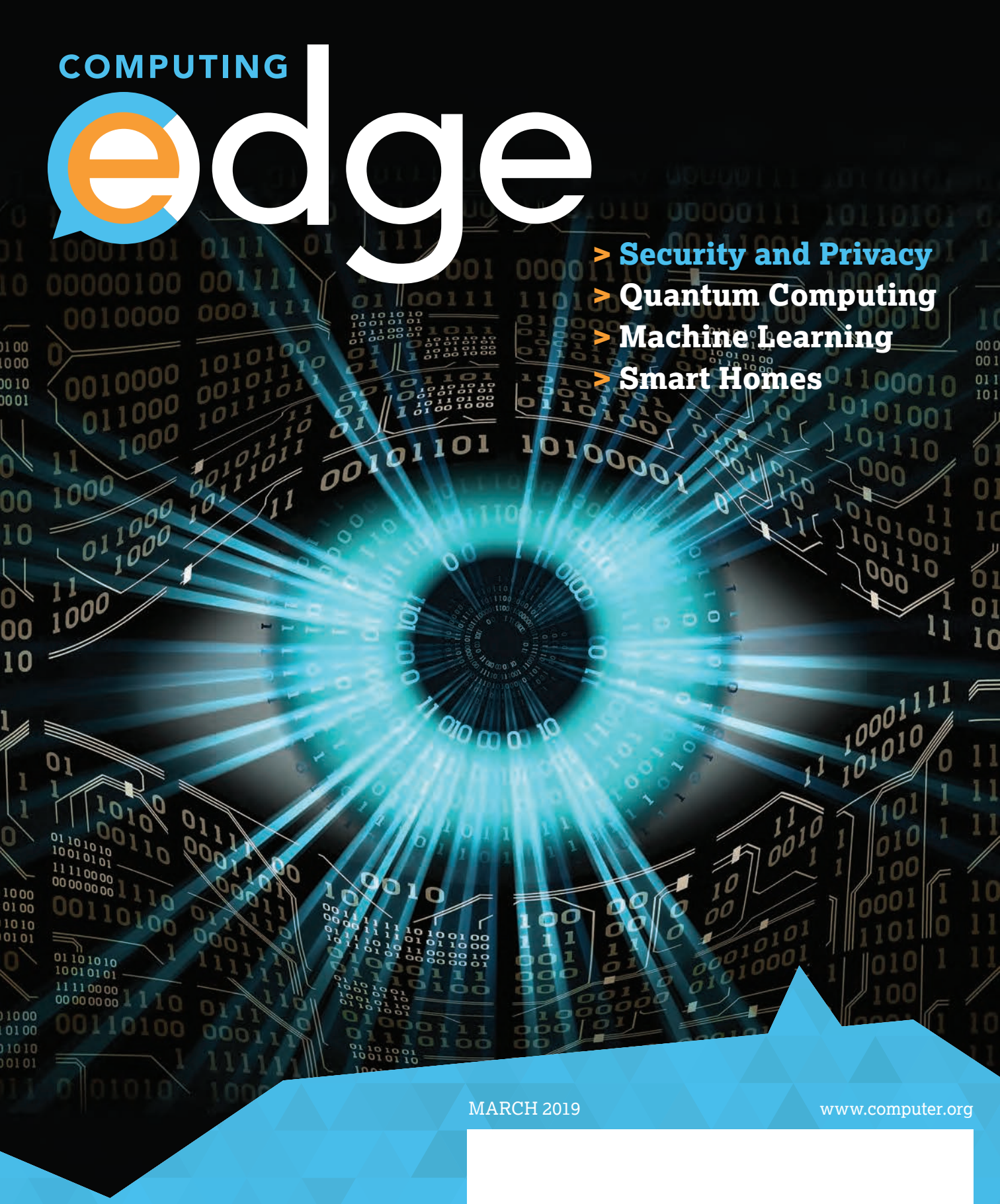



MARCH 2019

www.computer.org

IEEE

IEEE COMPUTER SOCIETY

IEEE COMPUTER SOCIETY

# B. Ramakrishna Rau Award

## *Call for Award Nominations*
## *Deadline: 1 May 2019*

Established in memory of Dr. B. (Bob) Ramakrishna Rau, this award recognizes his distinguished career in promoting and expanding the use of innovative computer microarchitecture techniques, including his innovation in compiler technology, his leadership in academic and industrial computer architecture, and his extremely high personal and ethical standards.

### Award

A certificate and a $2,000 honorarium are awarded.

### Presentation

The award is presented annually at the ACM/IEEE International Symposium on Microarchitecture.

### Nomination Requirements

The candidate will have made an outstanding innovative contribution or contributions to microarchitecture and use of novel microarchitectural techniques or compiler/architecture interfacing. It is hoped, but not required, that the winner will have also contributed to the computer microarchitecture community through teaching, mentoring, or community service.

This award requires 3 endorsements.

Nominations are being accepted electronically by **1 May 2019** to **bit.ly/ramakrishna-rau.**

### Questions?

VIsit **bit.ly/ramakrishna-rau**
or email **awards@computer.org**

IEEE

**2018**
**B. Ramakrishna Rau Award Recipient**

**Ravi Nair**
IBM Thomas J. Watson Research Center

*Honoring contributions to the computer microarchitecture field.*

COMPUTING edge

IEEE COMPUTER SOCIETY computer.org • +1 714 821 8380

Printed with inks containing soy and/or vegetable oils

IEEE

IEEE COMPUTER SOCIETY

## STAFF



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

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



## IEEE Computer Society Magazine Editors in Chief

***Computer***
David Alan Grier (Interim), *Djaghe LLC*

***IEEE Software***
Ipek Ozkaya, *Software Engineering Institute*

***IEEE Internet Computing***
George Pallis, *University of Cyprus*

***IT Professional***
Irena Bojanova, *NIST*

***IEEE Security & Privacy***
David Nicol, *University of Illinois at Urbana-Champaign*

***IEEE Micro***
Lizy Kurian John, *University of Texas, Austin*

***IEEE Computer Graphics and Applications***
Torsten Möller, *University of Vienna*

***IEEE Pervasive Computing***
Marc Langheinrich, *University of Lugano*

***Computing in Science & Engineering***
Jim X. Chen, *George Mason University*

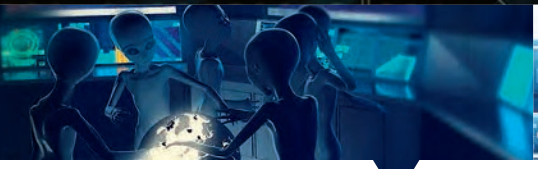
***IEEE Intelligent Systems***
V.S. Subrahmanian, *Dartmouth College*

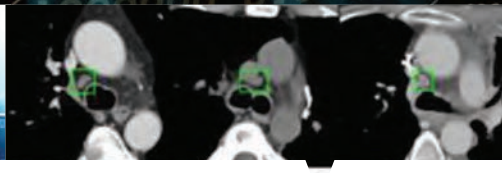
***IEEE MultiMedia***
Shu-Ching Chen, *Florida International University*

***IEEE Annals of the History of Computing***
Gerardo Con Diaz, *University of California, Davis*

MARCH 2019 • VOLUME 5, NUMBER 3

# COMPUTING edge





41

Smart Homes, Inhabited

## Security and Privacy



## Quantum Computing



## Machine Learning



## Smart Homes



## Departments





# Magazine Roundup

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

## Computer

***Autonomous Tools in System Design: Reflective Practice in Ubisoft's Ghost Recon Wildlands Project***

Ubisoft's game designers successfully used autonomous tools to develop an innovative virtual world. In this article from the October 2018 issue of *Computer*, the authors discuss the reflective practices underlying this success and how autonomous tools enable more complex system design.

## Computing in Science & Engineering

***NSF's Inaugural Software Institutes: The Science Gateways Community Institute and the Molecular Sciences Software Institute***

The National Strategic Computing Initiative (NSCI) creates a framework for partnerships among government, industry, and academia to advance the use of high-performance computing. The National Science Foundation's Office of Advanced Cyberinfrastructure furthers this mission through the creation of two new software institutes. Although the program predates NSCI, the Science Gateways Community

 2469-7087/19/$33.00 

Institute and the Molecular Sciences Software Institute advance NSCI's objectives by acting as hubs of excellence, serving broad communities, and creating software and workforce ecosystems. Read more in the September/October 2018 issue of *Computing in Science & Engineering*.

## IEEE Annals of the History of Computing

### *IPSJ Certifies 100 IP Technology Heritage Artifacts in 10 Years*

The Information Processing Society of Japan (IPSJ) reached an important goal when it certified its hundredth Information Processing Technology Heritage artifact and its tenth satellite museum. Obtaining these numbers was the primary objective of this program, which the IPSJ accomplished in 10 years. Read more in the July–September 2018 issue of *IEEE Annals of the History of Computing*.

## IEEE Computer Graphics and Applications

### *Rethinking Interaction Techniques for Personal Fabrication*

In this article from the September/October 2018 issue of *IEEE Computer Graphics and Applications*, the authors re-think the current interaction paradigm with personal fabrication tools. Rather than first creating a digital model and then producing physical output at the end, the authors propose a new tightly coupled workflow in which physical output is created continuously while the user is interacting.

## IEEE Intelligent Systems

### *Investigation on Unconventional Synthesis of Astroinformatic Data Classifier Powered by Irregular Dynamics*

This article from the July/August 2018 issue of *IEEE Intelligent Systems* discusses the mutual combination of the unconventional algorithm (in this case, evolutionary algorithms), deterministic chaos, and modeling on real data from astrophysics. Analytical programming with selected evolutionary algorithm is used to synthesize suitable models. This article is focused on various chaotic generators, which are used instead of classical pseudo-random number generators. Chaotic generators are used in conjunction with a special case: a generator based on a strange non-chaotic attractor. The performance of all chaotic- and non-chaotic-based generators is then mutually compared at the end.

## IEEE Internet Computing

### *Low-Latency Networking: Architecture, Techniques, and Opportunities*

With the advent of delay-sensitive applications, low-latency networking is attracting research attention from academia, industry, and standards organizations. This article, which appears in the September/October 2018 issue of *IEEE Internet Computing*, analyzes the causes of latency across network architecture, reviews some state-of-the-art techniques to reduce latency, and presents several opportunities.

## IEEE Micro

### *Galapagos: A Full Stack Approach to FPGA Integration in the Cloud*

Field-programmable gate arrays (FPGAs) have shown to be quite beneficial in the cloud due to their energy-efficient application-specific acceleration. These accelerators have always been difficult to use, and at cloud scale, the difficulty of managing these devices scales accordingly. The authors of this article from the November/December 2018 issue of *IEEE Micro* approach the challenge of managing large FPGA accelerator clusters in the cloud using abstraction layers and a new hardware stack called Galapagos. The hardware stack abstracts low-level details while providing flexibility in the amount of low-level access users require to reach their performance needs.

## IEEE MultiMedia

### *A Watermarking Mechanism with High Capacity for Three-Dimensional Mesh Objects using Integer Planning*

In this article from the July–September 2018 issue of *IEEE MultiMedia*, a new mechanism for digital watermarking is proposed. Three processes are involved: First, the watermarks are encapsulated into a carrier image; second, a sparsity analysis process is


conducted on one component; and finally, a flexible selection process is eventually executed to embed the watermarks. Experimental results demonstrate high-capacity information and strong robustness.

## IEEE Pervasive Computing

### *N-BaIoT—Network-Based Detection of IoT Botnet Attacks Using Deep Autoencoders*

The proliferation of Internet of Things (IoT) devices that can be more easily compromised than desktop computers has led to an increase in IoT-based botnet attacks. To mitigate this threat, there is a need for new methods that detect attacks launched from compromised IoT devices and that differentiate between hours- and milliseconds-long IoT-based attacks. In this article from the July–September 2018 issue of *IEEE Pervasive Computing*, the authors propose a novel network-based anomaly detection method for the IoT called N-BaIoT that extracts behavior snapshots of the network and uses deep autoencoders to detect anomalous network traffic from compromised IoT devices.

## IEEE Security & Privacy

### *Teaching Authentication as a Life Skill*

As more and more of the activities of daily living move into the digital realm, the importance of securing those activities grows. Where once an understanding of password security might have been considered a useful bonus, it is now becoming an integral life skill. Users of all ages need to be aware of what information is shared online and how to secure it. It is crucially important that security be taught at an early age. In this article, which appears in the September/October 2018 issue of *IEEE Security & Privacy*, the authors present their work developing security curriculum modules for teenagers and discuss their attempt to teach life skills for security to Swiss high school students.

## IEEE Software

### *User Involvement in Software Development: The Good, the Bad, and the Ugly*

Merely involving the users in software development won't guarantee system success. User involvement is a complex, multifaceted phenomenon with a good side, a bad side, and an ugly side. A better, deeper understanding of those sides can help project managers develop responsive strategies for increasing user involvement's effectiveness. Read more in the November/December 2018 issue of *IEEE Software*.

## IT Professional

### *Improving Energy Consumption of a Commercial Building with IoT and Machine Learning*

The critical requirements for devices connected to the Internet of Things are long battery life, long coverage range, and low deployment cost. The authors of this article from the September/October 2018 issue of *IT Professional* developed a machine learning-based smart controller for a commercial building's heating, ventilation, and air conditioning (HVAC) system using a long-range, low-power wireless platform (LoRa) and compared it with short-range RF communication in an indoor setting. Results show that LoRa's coverage range was 60.4 percent more than short-range communication inside the building. The smart controller was capable of identifying when a room was unoccupied and turning off the HVAC, reducing its energy consumption by up to 19.8 percent.


# CG&A

www.computer.org/cga

*IEEE Computer Graphics and Applications* bridges the theory and practice of computer graphics. Subscribe to *CG&A* and

- stay current on the latest tools and applications and gain invaluable practical and research knowledge,
- discover cutting-edge applications and learn more about the latest techniques, and
- benefit from *CG&A*'s active and connected editorial board.

## ADVERTISER INFORMATION

**Advertising Personnel**

**Debbie Sims: Advertising Coordinator**
**Email: dsims@computer.org**
**Phone: +1 714 816 2138 | Fax: +1 714 821 4010**

**Advertising Sales Representatives (display)**

**Central, Northwest, Southeast, Far East:**
**Eric Kincaid**
**Email: e.kincaid@computer.org**
**Phone: +1 214 673 3742**
**Fax: +1 888 886 8599**

**Northeast, Midwest, Europe, Middle East:**
**David Schissler**
**Email: d.schissler@computer.org**
**Phone: +1 508 394 4026**
**Fax: +1 508 394 1707**

**Southwest, California:**
**Mike Hughes**
**Email: mikehughes@computer.org**
**Phone: +1 805 529 6790**

**Advertising Sales Representative (Classifieds & Jobs Board)**

**Heather Buonadies**
**Email: h.buonadies@computer.org**
**Phone: +1 201 887 1703**

**Advertising Sales Representative (Jobs Board)**

**Marie Thompson**
**Email: marie@4caradio.org**
**Phone: 714-813-5094**

EDITOR'S NOTE

# Beware of Cyberattacks

Cyberattacks can be devastating to the affected organizations and individuals. Phishing, ransomware, and other forms of cybercrime are growing, exposing the personal data of millions of people and costing the economy billions of dollars every year. This issue of *ComputingEdge* focuses on steps organizations can take to bolster their cybersecurity and address this looming threat.

To protect themselves from the worst financial effects of cybercrime, businesses can invest in cyber insurance. *IT Professional*'s "The Economics of Cyber-Insurance" explains that providers require that companies demonstrate strong cybersecurity to attain cyber insurance coverage. Organizations can also strive to build or utilize software that is resistant to malware. *IEEE Software*'s "Recent Progress in Software Security" discusses techniques for making code less vulnerable such as process maturity modeling and runtime controls.

Another threat to cybersecurity is quantum computing, a nascent technology that has the potential to break modern encryption algorithms. "Cryptography after the Aliens Land," from *IEEE Security & Privacy*, imagines how cryptographers would keep our information secure against aliens with quantum-computing capabilities. Conversely, *Computer*'s "A Future with Quantum Machine Learning" predicts that quantum computing—combined with machine learning—will greatly benefit humanity in the coming years.

Machine learning is already having a large impact on society. In *IEEE MultiMedia*'s "Deep Medical Image Computing in Preventive and Precision Medicine," the authors detail achievements in medical imaging that were made possible by deep learning. Machine learning also drives conversational agents, or chatbots, but *IEEE Internet Computing*'s "Human-Aided Bots" contends that people still have an active role to play in the technology.

Finally, this *ComputingEdge* issue features two articles on smart homes from *IEEE Pervasive Computing*. "Smart Homes, Inhabited" explores the ways in which people interact with their smart homes through speech and end-user programming. "Evaluating Speech-Based Smart Devices Using New Usability Heuristics" takes a closer look at usability challenges with popular voice-controlled speakers like Google Home and Amazon Echo.


Published by the IEEE Computer Society 2469-7087/19/$33.00 © 2019 IEEE

# IEEE COMPUTER SOCIETY

**PURPOSE:** The IEEE Computer Society is the world's largest association of computing professionals and is the leading provider of technical information in the field.

**MEMBERSHIP:** Members receive the monthly magazine *Computer*, discounts, and opportunities to serve (all activities are led by volunteer members). Membership is open to all IEEE members, affiliate society members, and others interested in the computer field.

**COMPUTER SOCIETY WEBSITE:** www.computer.org

**OMBUDSMAN:** Direct unresolved complaints to ombudsman@computer.org.

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

**AVAILABLE INFORMATION:** To check membership status, report an address change, or obtain more information on any of the following, email Customer Service at help@computer.org or call +1 714 821 8380 (international) or our toll-free number, +1 800 272 6657 (US):

- Membership applications
- Publications catalog
- Draft standards and order forms
- Technical committee list
- Technical committee application
- Chapter start-up procedures
- Student scholarship information
- Volunteer leaders/staff directory
- IEEE senior member grade application (requires 10 years practice and significant performance in five of those 10)

## PUBLICATIONS AND ACTIVITIES

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

**Periodicals:** The society publishes 12 magazines, 15 transactions, and two letters. Refer to membership application or request information as noted above.

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

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

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

**Conferences/Education:** The society holds about 200 conferences each year and sponsors many educational activities, including computing science accreditation.

**Certifications:** The society offers three software developer credentials. For more information, visit www.computer.org/certification.

## 2019 BOARD OF GOVERNORS MEETINGS

**6 – 7 June:** Hyatt Regency Coral Gables, Miami, FL
**(TBD) November:** Teleconference

## EXECUTIVE COMMITTEE

**President:** Cecilia Metra

**President-Elect:** Leila De Floriani; **Past President:** Hironori Kasahara; **First VP:** Forrest Shull; **Second VP:** Avi Mendelson; **Secretary:** David Lomet; **Treasurer:** Dimitrios Serpanos; **VP, Member & Geographic Activities:** Yervant Zorian; **VP, Professional & Educational Activities:** Kunio Uchiyama; **VP, Publications:** Fabrizio Lombardi; **VP, Standards Activities:** Riccardo Mariani; **VP, Technical & Conference Activities:** William D. Gropp

**2018–2019 IEEE Division V Director:** John W. Walz

**2019 IEEE Division V Director Elect:** Thomas M. Conte

**2019–2020 IEEE Division VIII Director:** Elizabeth L. Burd

## BOARD OF GOVERNORS

**Term Expiring 2019:** Saurabh Bagchi, Leila De Floriani, David S. Ebert, Jill I. Gostin, William Gropp, Sumi Helal, Avi Mendelson

**Term Expiring 2020:** Andy Chen, John D. Johnson, Sy-Yen Kuo, David Lomet, Dimitrios Serpanos, Forrest Shull, Hayato Yamana

**Term Expiring 2021:** M. Brian Blake, Fred Douglis, Carlos E. Jimenez-Gomez, Ramalatha Marimuthu, Erik Jan Marinissen, Kunio Uchiyama

## EXECUTIVE STAFF

**Executive Director:** Melissa Russell
**Director, Governance & Associate Executive Director:** Anne Marie Kelly
**Director, Finance & Accounting:** Sunny Hwang
**Director, Information Technology & Services:** Sumit Kacker
**Director, Marketing & Sales:** Michelle Tubb
**Director, Membership Development:** Eric Berkowitz

## COMPUTER SOCIETY OFFICES

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

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

**Asia/Pacific:** Watanabe Building, 1-4-2 Minami-Aoyama, Minato-ku, Tokyo 107-0062, Japan · **Phone:** +81 3 3408 3118
**Fax:** +81 3 3408 3553 · **Email:** tokyo.ofc@computer.org

**MEMBERSHIP & PUBLICATION ORDERS**
**Phone:** +1 800 272 6657 · **Fax:** +1 714 821 4641
**Email:** help@computer.org

## IEEE BOARD OF DIRECTORS

**President & CEO:** Jose M.D. Moura
**President-Elect:** Toshio Fukuda
**Past President:** James A. Jefferies
**Secretary:** Kathleen Kramer
**Treasurer:** Joseph V. Lillie
**Director & President, IEEE-USA:** Thomas M. Coughlin
**Director & President, Standards Association:** Robert S. Fish
**Director & VP, Educational Activities:** Witold M. Kinsner
**Director & VP, Membership and Geographic Activities:** Francis B. Grosz, Jr.
**Director & VP, Publication Services & Products:** Hulya Kirkici
**Director & VP, Technical Activities:** K.J. Ray Liu

IEEE

*revised 13 February 2019*

# The Economics of Cyber-Insurance

**Nir Kshetri**
University of North Carolina at Greensboro

The cyber-insurance market currently is at a nascent stage. According to the German reinsurance company Munich Re, worldwide spending on cyber-insurance was US$3.4–US$4 billion in 2017, which is estimated to increase to US$8–US$9 billion by 2020 (https://tinyurl.com/ycrwhvlf). Cyber-insurance premiums currently account for only a tiny fraction of total insurance premiums. For instance, only in OECD economies do total insurance premium exceed US$5 trillion in 2016 (https://data.oecd.org/insurance/gross-insurance-premiums.htm).

The economic costs of cyberattacks exceed those associated with natural disasters (https://tinyurl.com/y8w9gpwy). According to Juniper Research, the cost of data breaches would amount US$2.1 trillion globally by 2019 (https://tinyurl.com/y7yukpcx). Most estimates of cyberattack costs overlook the harms associated with damage and destruction of data, lost productivity, theft of intellectual property, personal and financial data, post-breach disruption of companies' businesses, forensic investigation, restoration of hacked data and systems, and reputational harm. Some of these are more difficult to measure. Including those costs, Cybersecurity Ventures estimated that cybercrimes cost the world US$3 trillion in 2015, which will increase to US$6 trillion annually by 2021 (https://tinyurl.com/y7jxx3zw). Organizations are thus finding it more imperative to have cyber-insurance.

Cyber-insurance enhances firms' cybersecurity performances. For instance, a company is required to strengthen cybersecurity in order to buy coverage at a lower rate. A system that requires cyber-insurance thus raises cybersecurity standards.

Unsurprisingly, regulators are pushing for increased investment in cyber-insurance. For instance, New York's Department of Financial Services (DFS) has urged financial companies to invest in cyber-insurance.

## CYBER-INSURANCE: EXPLANATION AND THE CURRENT STATE

Cyber-insurance has been available since the 1990s.[1] Despite this long history, cyber-insurance has not yet taken off.

The U.S. cyber-insurance market is more advanced than the rest of the world (see Table 1). According to Marsh & McLennan, global cyber-insurance premiums was about US$3.5 billion in 2016 of which

 Published by the IEEE Computer Society 2469-7087/19/$33.00 

Table 1. Cyber-insurance markets in some key economies.

| Economy | Cyber-insurance premiums | Total insurance premiums (US$, billion) | Cyber-insurance premiums as a proportion of total insurance premiums |
|---|---|---|---|
| Brazil | US$645,800 (2016) (https://tinyurl.com/yc6u4ap4) | 58.9 (2016)[a] | 0.001% |
| Germany | US$105-117 million (https://tinyurl.com/y8ypu8jw) | 327.3 (2016)[a] | 0.03% |
| India | US$ 27.9 million (2017) (https://tinyurl.com/y84jgxm2). | 69.8 (2016)[a] | 0.04% |
| Japan | Japan Network Security Association's estimate: US$134.2 million (2017) (https://tinyurl.com/y8l4jxlz) | 407.4 (2016)[a] | 0.03% |
| South Korea | US$26.4 million (2016) (https://tinyurl.com/yafs4p27). | 185.6 (2016)[a] | 0.01% |
| The U.S. | Verisk: commercial cyber-insurance market: US$ 6.2 billion by 2020 US$ 2.5 billion in 2016 (https://tinyurl.com/ydf7z28s). | 2703.8 (2016)[a] | 0.09% |

a. *OECD. Gross insurance premiums, https://data.oecd.org/insurance/gross-insurance-premiums.htm.*

the U.S. and Europe accounted for US$3 billion and US$300 million, respectively (https://tinyurl.com/ycb7hrzv).

The cyber-insurance penetration rate is especially lower among small and medium sized enterprises (SMEs). In most OECD countries, the penetration level for stand-alone cyber-insurance among large companies was reported to be above 50% in 2017. The proportions of SMEs with cyber-insurance were in the single digits (https://tinyurl.com/ycyugnjm). Among big companies, data intensive companies exhibit a higher propensity to buy cyber-insurance. In India, only banks and ecommerce companies were reported to have cyber-insurance with large coverages (https://tinyurl.com/y84jgxm2).

Cyber-insurance provides coverage for the theft or loss of first-party and third-party data, as well as support services.[2] For the loss or theft of first-party data, an insurer may cover expenses related to notifying clients regarding the data breach, purchasing credit monitoring services for affected customers, extortion, and launching a public relations campaign to restore the company's reputation following a cyberattack-led negative publicity.

Third-party cyber-insurance protects a firm from being accused in case of a breach. Third-party coverage includes claims related to unlawful disclosure of a third-party's information and infringement of intellectual property rights (https://tinyurl.com/yb2vter9). It may also protect if an insurance holder's weak cybersecurity practices result in passing malware or virus to another user.[3]

Support services can help limit losses after a cyberattack. They cover expenses such as those related to public relations, IT forensics, and hiring experts in crisis management.

## Some Challenges

The cyber-insurance industry and market have some major challenges to overcome. First, there is a lack of standardization across the cyber-insurance products offered by insurers. This means that those buying insurance products are required to have a clear understanding of their cyber risk exposures in order to determine the appropriate type as well as the amount of coverage required based on their specific situation.[4] According to a survey conducted by Marsh, 49% of respondents said that they had "insufficient knowledge" about their cyber risk exposures to assess the type and coverage of insurances they need.[4] Likewise, another survey found that 38% of U.K. companies had insurance that covered all types of cyber-threats. However, most policies were based on inaccurate risk assessments (https://tinyurl.com/yc8xnzux).

Second, the value chain of the cyber-insurance industry is not well developed. There is the lack of clear understanding and knowledge among intermediaries such as insurance brokers and insurance agents. For instance, according to survey conducted by U.K. legal expenses insurer DAS UK Group, and HSB Engineering Insurance, most insurance brokers in the U.K. were reported to view cyber-insurance as a key and growing market. Nevertheless, one third of them admitted that they had a "poor" or "very poor" understanding of cyber risks and cyber-insurance (https://tinyurl.com/y7675y3u).

Third, due primarily to newness and the scarcity of data on cyberattacks and related losses insurers face a high uncertainty in pricing cyber risk coverage. They thus tend to be conservative and overcharge for cyber risk coverage.[3] Moreover, various cyber-insurance coverages are separately priced.

Fourth, the existence of externality effects may discourage some firms to buy cyber-insurance. If a minimum level of cybersecurity is required from policyholders, it is likely to improve the security of all Internet users. This will create a free riding problem, which reduces incentives for individuals or firms to get cyber-insurance.

# DEMAND- AND SUPPLY-SIDE MODELS AND MEASUREMENT ISSUES

## Supply-Side Condition

In order to derive a risk-adjusted return on capital, insurance companies need to determine the economic values of the capital invested and earnings. Put simply, the economic value of earnings is equal to cash flow plus the change in the economic value of the assets minus the change in the economic value of liabilities.[5]

Expressing in a simple equation, it is commercially viable for the insurance company if

$$\text{Insurancepremium} > \text{expectedloss} + \text{riskmargin} + \text{administrativecosts}. \qquad (1)$$

The risk margin in (1) represents an additional amount that investors in an insurance company require so that a return is expected for placing their economic capital at risk.[5] The risk associated with a policy is a function of many factors such as the company's industry, data risks and exposures, current practices, and financial health.[5] Among the biggest challenges facing the cyber-insurance industry and market is the lack of well-developed mechanisms to actuarially assess and price cyber risks.

Firms face heterogeneous cyber risk environments. In order to understand the essential components and the context of cyber risks, a process-based mode of such risks could be helpful. In such a model, risk equals "threat plus vulnerability plus consequences" (https://tinyurl.com/yc7ycokf). A threat is a danger related to cyber-attack that has the potential to cause harms to an organization. For instance, factors such as a firm's jurisdiction, physical location, nature of business, political orientation, and symbolic significance affect the degree of cyber-threats.

Cyber-vulnerability refers to the degree to which an organization is susceptible to harm from cyber-attacks. For instance, a firm with a poor cybersecurity practice is more likely to be harmed by cyber-criminals.

Finally, consequences of possible cyberattacks need to be evaluated in terms of factors such as reputational damage, financial loss, and possible physical harm. More severe consequences can arise if the jurisdiction of the firm's operations has strict laws against companies' failure to protect personally identifiable information.

Proper assessment of cyber-threat, cyber-vulnerability, and consequences of cyber-attacks are needed to gain a better understanding of cyber risks facing the firm. Insurance companies have realized that there is a fundamental need for better risk assessment tools.

On the plus side, there have been efforts to develop better analytical approaches, improving data collection efforts, and sharing relevant data with other players. When insurers model and test more information, insurance products are likely to be sold at more reasonable prices. Insurance companies are also taking measures to address legal uncertainties.[1]

## Demand-Side Condition

$$\text{A customer will invest in cyber} - \text{insurance if expected utility without cyber} - \text{insurance} < \text{expectedutility with cyber} - \text{insurance.} \quad (2)$$

Alternatively, the demand-side condition can also be written as[6,7]

$$U(\text{Benefits of insurance}) > U(\text{Costs of buying an insurance plan}). \quad (3)$$

U(*) is a utility function, which evaluates a cyber-insurance plan's benefits and costs in a common metric.

Firms and individuals invest in cyber-insurance only if its value proposition is clear. A current challenge is that the coverage terms are often complex, which makes it difficult to articulate the value proposition. There is still the lack of data on the odds of companies being victimized, which makes it difficult to estimate the costs of cyberattacks. It is also difficult for companies to measure the nature and extent of cyber-related exposure and to make decisions as to what coverages for how much to purchase (https://tinyurl.com/y7g3fjuy).

A related point is that some cyber-insurance policy holders find that their insurance does not cover all the losses in case of a cyberattack. To take an example, in December 2013, Target faced a high-profile security breach, which compromised 40 million credit and debit-card accounts and 70 million customers' personal data (https://tinyurl.com/y9vgktf3). Target had cyber-insurance when it was hacked. However, it only covered the first US$100 million. Actual costs exceeded US$450 million.[3]

## Transaction Costs in Cyber-Insurance Markets

In the context of business transactions involving two or more parties Nobel Laureate Douglas North argues that ".. transaction costs are . . . two things: (1) the costs of measuring the dimensions of whatever it is that is being produced or exchanged and (2) the costs of enforcement."[8] He goes on to say that "a lot of what we need to do is to try to measure the dimensions of what we are talking about in such a way that we can define them precisely."[3] Emphasizing the importance of measurements in enforcement, North argues: "Without being able to measure accurately whatever it is you are trying to enforce, there cannot be effective enforcement, even as a possibility."[8]

A transaction cost problem has two main components: (a) There is the presence of uncertainty and (b) the ability of the policy holder to change her/his behavior without detection.[9] Regarding (a), it is worth noting that due to the newness and limited availability of data, there is a challenge in estimating the probability of cyberattacks.

As to (b), a key challenge that insurers face in other types of insurance products is that the behavior of policy holders is often unobservable. Unlike many other insurance products, by working closely with the policy holder, cyber insurers can avoid some of the above-mentioned problems. They can support

overall risk management for their clients and tailor cyber-insurance to only residual risks in a cost-effective manner. For instance, using specialized software, insurers can remotely check whether policyholders have up-to-date software and defense mechanisms in place.[1] There has already been some progress on this front. Companies with strong cybersecurity practices pay lower insurance premiums.[1]

The above-mentioned feature also leads to a lower enforcement costs. A second-party enforcement, in which one party retaliates against the other (e.g., a cyber insurer penalizing a cyber-insurance policy holder for having a poor defence measure), can especially be more easily carried out in the context of cyber-insurance. It reduces the risks of policyholders failing to protect themselves against cyberattacks, thinking that they are covered against losses associated with such attacks.[1]

## CONCLUSION

Cyber-insurance market currently accounts for a vanishingly small proportion of the total insurance market. Nonetheless, it is growing fast. There are challenges associated with actuarially estimating the likelihood of cyberattacks and the total anticipated costs of such attacks. The lack of relevant data has led to an inaccurate assessment of cyber risks and higher premiums.

On the plus side, insurers can remotely monitor policyholders' cyber-defense mechanisms. It provides a low-cost mechanism for a second-party enforcement.

It is important to have a thorough understanding of the multifaceted nature of loss in case of cyberattacks. For non-IT businesses, first-party cybersecurity insurance could be enough, but third-party cybersecurity insurance may be needed for firms dealing with sensitive data of customers. Since most current policies are bespoke in nature, firms need to look for policies that are based on the need rather than the cost.

Due to the lack of prior experience, potential clients do not immediately understand the value proposition of cyber insurance. It has resulted in low demand. Cyber-insurance education and awareness can make a big difference. A higher public awareness of cyber security risk and a higher degree of understanding of the sophistication of cyberattacks can also stimulate the demand of cyber insurance. Firms should be convinced that the value proposition of insurance is interesting for them. Insurers need to make sure that potential clients get a simple and clear explanation of benefits from their cyber insurance. It is important to take measures to increase perceived economic benefits of cyber insurance.

Insurers must consider new market segments that are not currently investing in cyber insurance. They need to pursue firms in industries low digitization, households, and SMEs.

Data protection regulations that require financial protection against cyber-related losses could also lead to the growth of the cyber-insurance market. Finally, proper regulations may address the free-riding problem. Measures such as those taken by New York's DFS indicate that there have been some initiatives on this front.

## REFERENCES

1. iif.com, "Cyber risk insurance: A growth market adapting to a changing risk," Insti. Int. Finance, Washington, D.C., USA, Dec. 7, 2017.
2. C. P. Baban, Y. Gruchmann, C. Paun, A. C. Peters, and T. H. Stuchtey, "Cyber insurance as a contribution to IT risk management: An analysis of the market for cyber insurance in germany," 2017, Brandenburg Inst. Soc. Security gGmbH
3. L. DeFranco, "What you need to know about cybersecurity insurance," 2017. Available at https://blog.abacus.com/basics-of-cybersecurity-insurance/
4. Marsh & McLennan Co., "Cyber risk in Asia-Pacific the case for greater transparency risk in focus series," 2017.
5. L. Rubin, M. Lockerman, R. Tills, and X. Shi, "Economic measurement of insurance liabilities: The risk and capital perspective," 2009. Actuarial Practices Forum.

6. H. P. Binswanger-Mkhize, "Is there too much hype about index-based agricultural insurance?" *J. Dev. Studies*, vol. 48, no. 2, pp. 187–200, 2012.
7. D. S. Nagin and G. Pogarsky, "Integrating celerity, impulsivity, and extralegal sanction threats into a model of general deterrence: Theory and evidence." 2001. Available at http://onlinelibrary.wiley.com/doi/10.1111/j.1745-9125.2001.tb00943.x/abstract
8. D. C. North, "Dealing with a nonergodic world: Institutional economics," Property Rights, Global Environment: Duke Environment, Law, Policy Forum, vol. 10, no. 1, pp. 1–12, 1999.
9. D. W. Allen, "Transaction costs," in B. Bouckaert and D. G. Gerrit, Eds., *The Encyclopedia of Law and Economics*. Chelthenham, UK: Edward Elgar, 2000, vol. 1, pp. 893–926.

## ABOUT THE AUTHOR

**Nir Kshetri** is a Professor of Management with the Bryan School of Business and Economics, University of North Carolina at Greensboro, Greensboro, NC, USA. Contact him at nbkshetr@uncg.edu.

*This article originally appeared in*
IT Professional, *vol. 20, no. 6, 2018.*

# CONNECT ON INTERFACE

Explore **INTERFACE**, a communication resource to help members engage, collaborate and stay current on Computer Society activities. Use **INTERFACE** to learn about member accomplishments and find out how your peers are changing the world with technology.

We spotlight our professional sections and student branch chapters, sharing their recent activities and giving leaders a window into how chapters around the globe grow, thrive and meet member expectations. Plus, **INTERFACE** will keep you informed on Computer Society–related activities so you never miss a meeting, career development opportunity or important industry update.

**Connect today at interface.computer.org**

IEEE COMPUTER SOCIETY
INTERFACE

Editor: **Giuliano Antoniol**
Polytechnique Montréal
antoniol@ieee.org

Editor: **Steve Counsell**
Brunel University
steve.counsell@brunel.ac.uk

Editor: **Phillip Laplante**
Pennsylvania State University
pal11@psu.edu

# Recent Progress in Software Security

Edward Amoroso

**EXACTLY 50 YEARS** ago, Edsger Dijkstra sent the article "A Case against the GOTO Statement" to the *Communications of the ACM*, explaining why GOTO introduced too much complexity and should thus be avoided. Given the urgency of Dijkstra's message, Pascal inventor Niklaus Wirth made the prescient decision to recast the article as a letter to the editor with the now-iconic title, "Go To Statement Considered Harmful."[1]

In the years since, our community has, sadly, lost Dijkstra, but the debate he sparked has remained active. The cybersecurity community in particular has been vocal about finding ways to improve software, because most vulnerabilities involve exploitable weaknesses introduced through badly written code. Unfortunately, the rush to modern DevOps coding and the demands of software marketing have tended to overshadow most correctness concerns.

Instead, the cybersecurity community has widely adopted an approach to reduce cybersecurity risk in software that involves a collage of techniques, tools, and methods, each addressing some aspect of the threat implications of bad code. Here, I briefly survey recent progress in each element of this combined approach, including the pros and cons for reducing cybersecurity risk.

## Advanced Malware Detection

Although improved programming methodology continues to influence software security, the cybersecurity software community has focused mostly on malware detection. This situation is curious, because while it's in everyone's interest in cybersecurity to prevent exploitable bugs, agreement exists that this is basically impossible for nontrivial code. Vendors have thus built small empires based on this (so far) correct assumption.

Whereas the original methods of malware detection were built on matching application code (or operating systems) to signatures, more-modern methods review behaviors for evidence of unacceptable runtime activity. Behavioral investigation is enabled by dynamic provision of virtual machines for safe detonation of executables. Without such virtual contained environments, behavioral analysis would be too dangerous for production systems.

Modern research in malware detection employs machine learning to help train security tools to identify bad code on the basis of samples. So, just as AI-powered systems are fed pictures of cats for learned recognition, comparable systems are fed "pictures" of files containing malware. Deep-learning techniques use massive parallelism to improve such algorithms' efficiency.

Perhaps the unifying aspect of this evolving space is that malware detection tools presume the continued existence of problems, which helps justify business investment by start-ups and other security vendors. The likelihood is thus low that software professionals will advance our art to the point at which no malware exists. So, the antimalware industry should expect to see continued vibrancy of its collective offerings in terms of sales, revenue, and growth.

## Software Process Maturity

Another focus in modern software security involves inferring code security through its associated software process. That is, many security experts have suggested that, rather than directly inspecting software for evidence of malware or vulnerabilities, you examine that software's development process. This is like determining patients' health by asking them about their behaviors rather than testing their blood.

The theory supporting this approach is largely empirical—namely, that good code has tended to come from well-trained developers working with world-class tools in modern, well-organized development environments. In contrast, exploitable vulnerabilities frequently have been found in code written by poorly

2469-7087/19/$33.00 © 2019 IEEE Published by the IEEE Computer Society March 2019 

trained developers using makeshift tools in ad hoc development environments.

So, maturity models have emerged that let you link the degree of software security to the quality of the process. This has the useful side effect of driving improved security for all code that emerges from a given vendor's or team's software process. Common methods demanded in such processes include automation, periodic penetration testing, and proper software updating and maintenance procedures.

One excellent benefit of process maturity approaches is that little downside exists in any effort to improve the steps taken to create code. If the underlying rubric is sound, the associated effort to bring the software process in line with accepted best practices will have benefits far beyond just improved protection. Code reduction, time-to-market improvements, and quality increases will all result from improved software processes.

## Software Review and Scanning

The most traditional means for improving software security involve direct inspection of code, sometimes using code-scanning tools. The tools' earliest use seems to have been at Bell Labs in the 1970s, with the introduction of the lint preprocessing program, which scanned C code and recommended improvements. All subsequent code-scanning tools trace their lineage to this early concept.

The ongoing use of manual code reviews is much debated in the software community, with traditionalists insisting that human inspection remains essential to high-quality, secure products. The challenge is that with the rapid cycle times in a DevOps environment, little time exists for human review of source code. Automated scans thus have become the norm in such environments; this has its pros and cons.

Software security will always include some degree of review and scans, presumably done properly once for reusable components, thus precluding the need for repeat security analysis. Critics claim that reusable componentry has been an elusive goal for decades. However, few would argue that modern DevOps and cloud-based software process environments are fertile ground for standard, well-reviewed components.

## Runtime Software Controls

Perhaps the most promising advance in software security involves using runtime controls that are embedded in the execution environment. This technique is sometimes called *runtime application self-protection* (RASP). Through the integration of behavioral and even machine-learning controls into and around an executable, a programmed protection environment emerges—one that can compensate for code weaknesses.

RASP controls, cloud development, and DevOps are all tightly woven in most software development organizations. All three aim to increase delivered code's speed and flexibility. However, a somewhat open question is whether these three initiatives result in more secure code. Certainly, RASP will reduce the risk of any application good or bad, but it's unclear whether programmers write better code in the presence of RASP.

Nevertheless, runtime software controls will continue to influence software security, especially in the context of new self-learning methods. Machine-learning techniques have advanced to the point at which observed behaviors can serve as training data to label new variants of software exploits. This is an exciting new way to drive improved, autonomous software control using platform automation.

Deep-learning advances are especially promising for software security. This is because the improved efficiency and massive parallelism that characterize the approach are perfectly suited to the large number of combinations that must be examined in typical software execution. We might hope that deep-learning algorithms would be a superior way to review code for unused execution paths, dead code, logic errors, race conditions, and the like.

Our industry's early focus on methodology, as evidenced by Edsger Dijkstra's teachings on software, remains an important consideration in the assurance of secure software. However, the community has taken many practical steps to improve code quality and security in the absence of any real correctness progress by programmers. Bugs still abound in nontrivial software, and security teams must be practical in their risk reduction efforts.

We can hope that in the coming years, these methods will synthesize with improved programming languages and ever-improving programming techniques into an ecosystem that reduces risk by improving software. Given modern infrastructure's dependency on well-designed code with a minimum of exploitable flaws, this is certainly a welcome goal.

**Reference**

1. E. Dijkstra, "Go To Statement Considered Harmful," *Comm. ACM*, vol. 11, no. 3, 1968, pp. 147–148.

*This article originally appeared in* IEEE Software, *vol. 35, no. 2, 2018.*

## ABOUT THE AUTHOR

**EDWARD AMOROSO** is the founder and chief executive officer of TAG Cyber. He's also a Distinguished Research Professor in the New York University Tandon School of Engineering's Computer Science Department, an adjunct professor of computer science at the Stevens Institute of Technology, and a senior advisor at Johns Hopkins University's Applied Physics Laboratory. Contact him at eamoroso@tag-cyber.com.


Photo by Martin Gundersen

**Bruce Schneier**
Harvard University

# Cryptography after the Aliens Land

Quantum computing is a new way of computing—one that could allow humankind to perform computations that are simply impossible using today's computing technologies. It allows for very fast searching, something that would break some of the encryption algorithms we use today. And it allows us to easily factor large numbers, something that would break the RSA cryptosystem for any key length.

This is why cryptographers are hard at work designing and analyzing "quantum-resistant" public-key algorithms. Currently, quantum computing is too nascent for cryptographers to be sure of what is secure and what isn't. But even assuming aliens have developed the technology to its full potential, quantum computing doesn't spell the end of the world for cryptography. Symmetric cryptography is easy to make quantum-resistant, and we're working on quantum-resistant public-key algorithms. If public-key cryptography ends up being a temporary anomaly based on our mathematical knowledge and computational ability, we'll still survive. And if some inconceivable alien technology can break all of cryptography, we still can have secrecy based on information theory—albeit with significant loss of capability.

At its core, cryptography relies on the mathematical quirk that some things are easier to do than to undo. Just as it's easier to smash a plate than to glue all the pieces back together, it's much easier to multiply two prime numbers together to obtain one large number than it is to factor that large number back into two prime numbers. Asymmetries of this kind—one-way functions and trap-door one-way functions—underlie all of cryptography.

To encrypt a message, we combine it with a key to form ciphertext. Without the key, reversing the process is more difficult. Not just a little more difficult, but astronomically more difficult. Modern encryption algorithms are so fast that they can secure your entire hard drive without any noticeable slowdown, but that encryption can't be broken before the heat death of the universe.

With symmetric cryptography—the kind used to encrypt messages, files, and drives—that imbalance is exponential, and is amplified as the keys get larger. Adding one bit of key increases the complexity of encryption by less than a percent (I'm hand-waving here) but doubles the cost to break. So a 256-bit key might seem only twice as complex as a 128-bit key, but (with our current knowledge of mathematics) it's 340,282,366,920,938,463,463,374,607,431,768,211,456 times harder to break.

Public-key encryption (used primarily for key exchange) and digital signatures are more complicated. Because they rely on hard mathematical problems like factoring, there are more potential tricks to reverse them. So you'll see key lengths of 2,048 bits for RSA, and 384 bits for algorithms based on elliptic curves. Here again, though, the costs to reverse the algorithms with these key lengths are beyond the current reach of humankind.

This one-wayness is based on our mathematical knowledge. When you hear about a cryptographer "breaking" an algorithm, what happened is that they've found a new trick that makes reversing easier. Cryptographers discover new tricks all the time, which is why we tend to use key lengths that are longer than strictly necessary. This is true for both symmetric and public-key algorithms; we're trying to future-proof them.

Quantum computers promise to upend a lot of this. Because of the way they work, they excel at the sorts of computations necessary to reverse these one-way functions. For symmetric cryptography, this isn't too bad. Grover's algorithm shows that a quantum computer speeds up these attacks to effectively halve the key length. This would mean that a 256-bit

 Published by the IEEE Computer Society 

key is as strong against a quantum computer as a 128-bit key is against a conventional computer; both are secure for the foreseeable future.

For public-key cryptography, the results are more dire. Shor's algorithm can easily break all of the commonly used public-key algorithms based on both factoring and the discrete logarithm problem. Doubling the key length increases the difficulty to break by a factor of eight. That's not enough of a sustainable edge.

There are a lot of caveats to those two paragraphs, the biggest of which is that quantum computers capable of doing anything like this don't currently exist, and no one knows when—or even if—we'll be able to build one. We also don't know what sorts of practical difficulties will arise when we try to implement Grover's or Shor's algorithms for anything but toy key sizes. (Error correction on a quantum computer could easily be an unsurmountable problem.) On the other hand, we don't know what other techniques will be discovered once people start working with actual quantum computers. My bet is that we will overcome the engineering challenges, and that there will be many advances and new techniques—but they're going to take time to discover and invent. Just as it took decades for us to get supercomputers in our pockets, it will take decades to work through all the engineering problems necessary to build large-enough quantum computers.

In the short term, cryptographers are putting considerable effort into designing and analyzing quantum-resistant algorithms, and those are likely to remain secure for decades. This is a necessarily slow process, as both good cryptanalysis transitioning standards take time. Luckily, we have time. Practical quantum computing seems to always remain "ten years in the future," which means no one has any idea.

After that, though, there is always the possibility that those algorithms will fall to aliens with better quantum techniques. I am less worried about symmetric cryptography, where Grover's algorithm is basically an upper limit on quantum improvements, than I am about public-key algorithms based on number theory, which feel more fragile. It's possible that quantum computers will someday break all of them, even those that today are quantum resistant.

If that happens, we will face a world without strong public-key cryptography. That would be a huge blow to security and would break a lot of stuff we currently do, but we could adapt. In the 1980s, Kerberos was an all-symmetric authentication and encryption system. More recently, the GSM cellular standard does both authentication and key distribution—at scale—with only symmetric cryptography. Yes, those systems have centralized points of trust and failure, but it's possible to design other systems that use both secret splitting and secret sharing to minimize that risk. (Imagine that a pair of communicants get a piece of their session key from each of five different key servers.) The ubiquity of communications also makes things easier today. We can use out-of-band protocols where, for example, your phone helps you create a key for your computer. We can use in-person registration for added security, maybe at the store where you buy your smartphone or initialize your Internet service. Advances in hardware may also help to secure keys in this world. I'm not trying to design anything here, only to point out that there are many design possibilities. We know that cryptography is all about trust, and we have a

lot more techniques to manage trust than we did in the early years of the Internet. Some important properties like forward secrecy will be blunted and far more complex, but as long as symmetric cryptography still works, we'll still have security.

It's a weird future. Maybe the whole idea of number theory–based encryption, which is what our modern public-key systems are, is a temporary detour based on our incomplete model of computing. Now that our model has expanded to include quantum computing, we might end up back to where we were in the late 1970s and early 1980s: symmetric cryptography, code-based cryptography, Merkle hash signatures. That would be both amusing and ironic.

Yes, I know that quantum key distribution is a potential replacement for public-key cryptography. But come on—does anyone expect a system that requires specialized communications hardware and cables to be useful for anything but niche applications? The future is mobile, always-on, embedded computing devices. Any security for those will necessarily be software only.

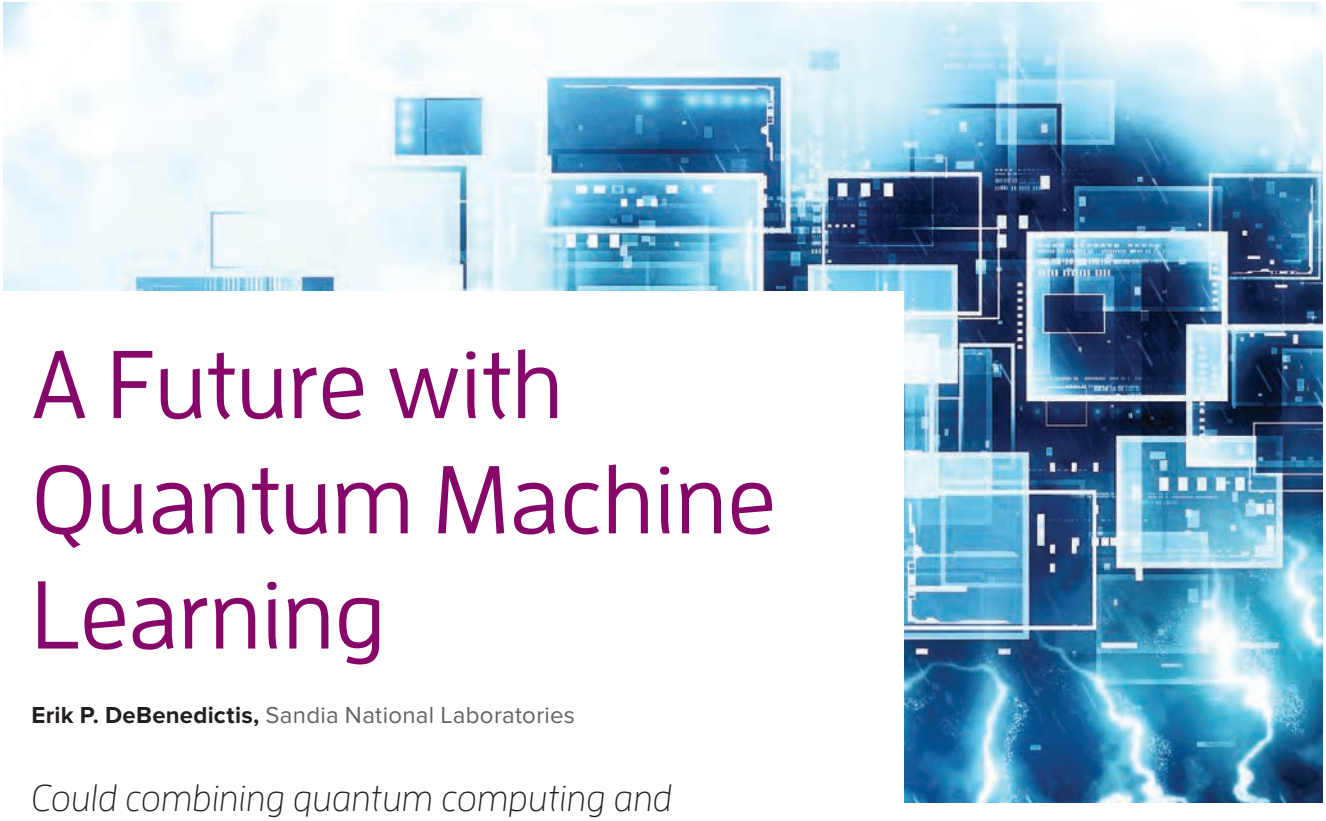
There's one more future scenario to consider, one that doesn't require a quantum computer. While there are several mathematical theories that underpin the one-wayness we use in cryptography, proving the validity of those theories is in fact one of the great open problems in computer science. Just as it is possible for a smart cryptographer to find a new trick that makes it easier to break a particular algorithm, we might imagine aliens with sufficient mathematical theory to break all encryption algorithms. To us, today, this is ridiculous. Public-key cryptography is all number theory, and potentially vulnerable to more mathematically inclined aliens. Symmetric cryptography is so much nonlinear muddle, so easy to make more complex, and so easy to increase key length, that this future is unimaginable. Consider an AES variant with a 512-bit block and key size, and 128 rounds. Unless mathematics is fundamentally different than our current understanding, that'll be secure until computers are made of something other than matter and occupy something other than space.

But if the unimaginable happens, that would leave us with cryptography based solely on information theory: one-time pads and their variants. This would be a huge blow to security. One-time pads might be theoretically secure, but in practical terms they are unusable for anything other than specialized niche applications. Today, only crackpots try to build general-use systems based on one-time pads—and cryptographers laugh at them, because they replace algorithm design problems (easy) with key management and physical security problems (much, much harder). In our alien-ridden science-fiction future, we might have nothing else.

Against these godlike aliens, cryptography will be the only technology we can be sure of. Our nukes might refuse to detonate and our fighter jets might fall out of the sky, but we will still be able to communicate securely using one-time pads. There's an optimism in that. ■

**Bruce Schneier** is a lecturer and Fellow at the Harvard Kennedy School, and special advisor to IBM Security. His new book is *Click Here to Kill Everybody: Security and Survival in a Hyper-connected World.* Contact him via www.schneier.com

*This article originally appeared in*
IEEE Security & Privacy,
*vol. 16, no. 5, 2018.*



# A Future with Quantum Machine Learning

**Erik P. DeBenedictis,** Sandia National Laboratories

*Could combining quantum computing and machine learning with Moore's law produce a true "rebooted computer"? This article posits that a three-technology hybrid-computing approach might yield sufficiently improved answers to a broad class of problems such that energy efficiency will no longer be the dominant concern.*

Over the past two years, this column has highlighted the technologies being considered as candidates to reboot computing. Yet none of the individual technologies has been very exciting. For example, in December 2016 I wrote about neuromorphic crossbars' potential as machine learning accelerators,[1] concluding that they face the same thermodynamic limit as today's microprocessors.

However, in the last year, the combination of quantum computing, machine learning, and Moore's law has taken form with more potential than anything seen to date.[2] While it's too early to tell whether this three-technology hybrid will survive the test of time, the field is attracting both venture capital and substantial investment by big companies and the government. My goal here is to show how this combination of technologies has a synergy that could affect people outside "the club" creating it.

## MACHINE LEARNING CHANGES THE QUANTUM GAME

To explain the combination of quantum computing and machine learning, we must express quantum computing in terms that don't unnecessarily hide the role of programming, because this would obfuscate machine learning's value in shifting some portion of human-developed programming to computers.

Quantum computers' popular success story is the factoring of large numbers. Historically, numbers were

 Published by the IEEE Computer Society 

EDITOR **ERIK P. DEBENEDICTIS**
Sandia National Laboratories;
epdeben@sandia.gov

factored using trial division, which requires a three-line program. However, the three lines iterate an exponential number of times ($2^{n/2}$) when factoring an $n$-bit number. This leads to an exponential expenditure of energy, given the thermodynamic kT model of minimum energy per binary operation. The community has explored two options to reduce the cost of factoring numbers:

1. The subexponential number field sieve algorithm was developed using perhaps 100 person-years effort by mathematicians, computer scientists, and programmers.
2. Shor's polynomial-time quantum algorithm was developed, albeit requiring a quantum computer that has yet to be built.

Both improvements are important.

The discovery of quantum algorithms occurred in parallel with improvements to the equivalent classical algorithms, leading to competition between the 100 person-years research and the special properties of quantum information. However, computational complexity theory seeks the best algorithm without reference to algorithm development and programming effort, unfairly disadvantaging quantum computers. This retelling of the quantum computer story opens the door for machine learning to contribute by making programming more efficient.

Machine learning moves the dividing line between humans and computers. In a typical machine-learning scenario, a server farm consumes a dollar's worth of energy learning how to recognize your pet in images or how to target advertisements by scanning your emails. This learning might compute neuron weights for a recognition circuit, which is essentially a program. In many cases, each person's pet and mailbox have an underlying structure similar to the number field that enabled improved factoring algorithms. It might be possible to improve the computational efficiency of the neural network that cost a dollar to synthesize in the first place through 100 person-years of research. However, there's no way to recoup 100 years' salary, given that the learned behavior is applicable to only one person.

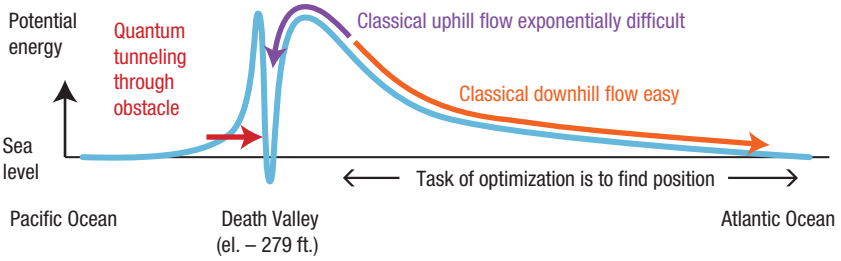



**Figure 1.** Optimization involves finding the lowest point on a potential energy curve (blue), which is Death Valley, even though most water flows to the oceans. Classical optimization (orange) works like raindrops flowing downhill, but simulated annealing allows limited uphill movement (purple). However, quantum computer optimization can use a quantum physics principle called "tunneling" to go through a high energy barrier (red). The text also describes how this type of optimization could apply to organizing the slides in a slide deck to make a compelling presentation.

The opportunity for quantum machine learning will be in learning lots of simple lessons—concepts that will make society more efficient, not just the hard problems currently attracting geniuses and armies of researchers. I suggest that quantum machine learning be benchmarked on learning a completely original behavior and performing it as few as, say, 10 times. The cost metric would include both the learning and running times.

## CHIP LAYOUT TO SLIDE DECKS

How can a quantum computer's computational advantage in optimization,[3] factoring numbers, and other algorithms be repurposed to machine learning? While classical computers can perfectly optimize small systems, they only find incremental improvements for large systems such as transportation routes and product pricing. This is due to their rapidly rising running time as a function of problem size.

Placement of logic gates on an integrated circuit is an example. Chip design tools have optimizers that place logic gates on a chip's surface with just enough space between to hold the wiring that defines the chip's function. Better placement reduces chip area—and hence cost—while simultaneously increasing the chip's speed because the shorter wires convey information in less time. However, a chip might be profitable even if it's a few percent larger than necessary, so perfect optimization isn't essential.

Classical placement algorithms such as simulated annealing follow the same principle as raindrops trying to find the lowest elevation by flowing downhill. Figure 1 shows an energy landscape for water by position across the US. Water dropped almost anywhere will flow to an ocean. Oceans are low, but not as low as Death Valley. However, Death Valley has a small rainfall basin surrounded by high mountains, so a random raindrop would be unlikely to fall into its basin.

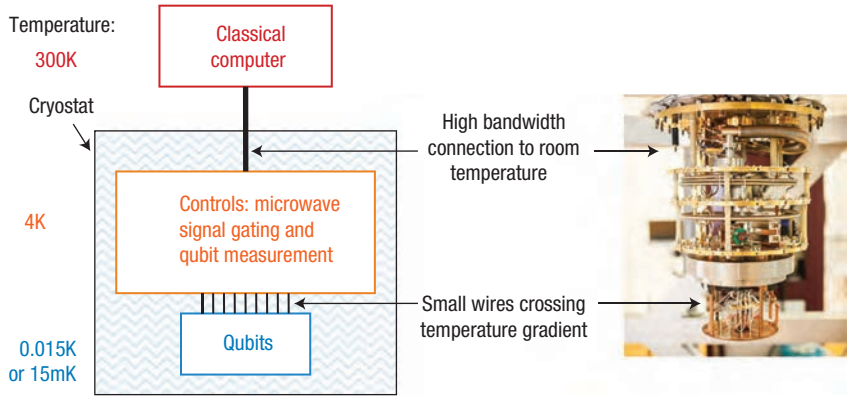



**Figure 2.** General structure of a quantum computer system. The user interacts with a classical computer. If the problem requires optimization, the classical computer translates the user's problem into a standard form for a quantum computer, such as QUBO, or into a different form if another quantum algorithm is required. The classical computer then creates control signals for qubits (quantum bits) located in a cryogenic environment, receiving data from measurements of the qubits. Some classical electronics are placed in the cold environment to minimize heat flow through wiring across the cryogenic-to-room-temperature gradient. Photo source: A. Hellemans, "Europe Bets €1 Billion on Quantum Tech," *IEEE Spectrum*, 22 Jun. 2016; spectrum.ieee.org computing/hardware/europe-will-spend-1-billion-to-turn-quantum-physics-into-quantum-technology.

Mathematicians, computer scientists, and programmers have improved simulated annealing so that potential solutions can jump over an obstacle, but the probability of this occurring decreases exponentially with the height of the jump. Human effort has also created heuristics, such as chip design tools that handle memories, busses, and clock lines in special ways.

One form of quantum machine learning uses "quantum tunneling"[3] to go through the peak in Figure 1, with the probability of this occurring declining exponentially with the width of the peak. The tunneling approach may or may not be better than simulated annealing, but applying both techniques might give a better answer than either alone. Other quantum algorithms work quite differently, such as not using potential energy at all.

Optimization can be applied to development of a slide deck for a presentation, such as tuning the ordering of the slides to meet the expected interests of a particular audience. Hypothetically, there is a "potential energy landscape" for every audience based on the listeners' background knowledge and receptiveness to new ideas. For example, one audience might prefer an emotional appeal first and facts later. If Figure 1's horizontal axis represents slide interchanges, optimization just needs to find the sequence of interchanges that yields the most compelling presentation. There are exponentially many orderings—too many to test exhaustively—so the classical approach is to follow downhill paths as shown in Figure 1, or slide interchanges that each make each potential presentation a little better than the previous. However, a quantum computer's unique ability to tunnel through high potential barriers might let it find the most compelling slide deck when simulated annealing cannot.

Presentations can be optimized through use of human labor, such as mock juries in criminal trials. However, the effort required is too high for everyday situations.

## A VISION FOR FUTURE APPLICATIONS

I've painted a picture in which today's corporate applications, such as optimizing transportation routes, are improved and then applied to everyday personal situations. But are there enough such applications to bother with? Computers assist people with numerical calculations countless times a day, such as when a smartphone computes how far you jogged. But there are also occasions when you need to say or do something that requires nonnumerical judgement—such as preparing a compelling slide-deck presentation, as in my previous example, or answering a question in a way that impresses your boss. With today's knowledge and technology, there should be as many ways for computers to address these nonnumerical activities as the numerical ones.

## MOORE'S LAW AND SUPERCONDUCTING ELECTRONICS

In the early 1940s, IBM president Thomas J. Watson reputedly said, "I think there is a world market for maybe five computers." If quantum machine learning meets the expectations of the venture capitalists who are funding start-ups, Watson's statement won't hold for these three-technology hybrid computers either, and we'll need a path to produce them in large volume.

The quantum effect frequently, although not always, requires components operating near absolute zero, making just about every aspect of the design exotic. Quantum computer components operated at room temperature inevitably acquire error from the thermal motion of the atoms in the computer's structure. The errors must be removed by quantum error correction, yet the error accumulation rate is too high for practical removal unless the components are cooled to millikelvins, or thousandths of a degree above absolute zero—273.15 °C or 0 K.

The architecture of these quantum–classical hybrid computers is zeroing

in on the structure shown in Figure 2. The qubits (quantum bits) must be kept at a temperature of approximately 15 mK. They need support from classical superconducting electronics based on Josephson junctions operating at temperatures around helium's boiling point, or 4 K.[4] The electronics must have extremely low energy dissipation, because the external refrigeration must dissipate at least the temperature ratio (300 K/4 K = 75× or 300 K/15 mK = 20,000×) times as much energy to remove the heat to room temperature (300 K)—and, in practice, several times this amount. Logic-gate circuits based on Josephson junctions are available that perform the logic functions for error correction as well as the gate microwave signals required to control qubits.

As Moore's law demonstrated, industry knows how to take control of a technology and work relentlessly to improve it to the limits of physics. So if the ideas in this column pan out, industry will need to do the same for superconducting electronics.

## SOCIETAL IMPLICATIONS

So far I've discussed future computers as though they were standalone, yet we use computers today as agents for many of our business transactions and information handling. Some day we might have the option to upgrade our computerized agents to more advanced versions that use quantum machine learning internally. However, we must be prepared for competitors and bad actors that upgrade early to take advantage of us.

If you are a defense lawyer trying to defend a client, you will be at a disadvantage if your presentation is less well tuned to the jury than the prosecution's is. Similarly, web traffic is monitored by machine learning software purportedly to send us advertisements, but bad actors can use the same technology to better target phishing emails that can cause us harm. Computers can use machine learning to find phishing emails, but this will lead the opposing sides to mount an arms' race for better quantum computers.

My December 2016 article comparing the energy efficiency of analog memristor-style crossbars for learning showed that this new technology was subject to the same thermodynamic limits as digital chips. Each approach might beat the other in some portion of a parameter space, but the common limit implied that the best win would be an order of magnitude or two. Companies could live or die based on a couple orders of magnitude in product performance, but changing the world typically requires a bigger difference. The triad of quantum computing, machine learning, and a continuation of Moore's law could possibly address a broad class of problems, with only distant competitors. So what are the practical challenges that quantum machine learning must overcome to survive the test of time?

There will be technical challenges beyond just building hybrid quantum–classical hardware. We haven't systematically looked for applications that depend on exorbitant amounts of machine learning or optimization, nor have we applied quantum computing to general problem solving.

The computer industry has been producing chips intended to operate at room temperature, which was convenient. A quantum–classical computer, however, has unique capabilities that require a cryogenic environment. Materials, devices, and circuits for this environment are known but haven't been refined to the same level of manufacturability as semiconductors.

Classical computers' rapid emergence has stretched society's ability to assimilate their capabilities, creating concerns regarding cybersecurity, robots and AI, social media, and so on. Rolling out quantum machine learning products could introduce similar issues, but they should be seen as challenges to overcome, not reasons to hold back progress or ignore the uncomfortable questions they present.


## ACKNOWLEDGMENTS

Sandia National Laboratories is a multimission laboratory managed and operated by National Technology and Engineering Solutions of Sandia, LLC, a wholly owned subsidiary of Honeywell International, Inc., for the US Department of Energy's National Nuclear Security Administration under contract DE-NA-0003525.


## REFERENCES


3. E. P. DeBenedictis, "Computational Complexity and New Computing Approaches," *IEEE Computer*, vol. 49, no. 12, 2016, pp. 76–79.
4. J. Biamonte et al., "Quantum Machine Learning," *Nature*, vol. 549, no. 7671, 2017, pp. 195–202.
5. V.S. Denchev et al., "What is the Computational Value of Finite-Range Tunneling?," *Physical Review X*, vol. 6 no. 3, 2016, p. 031015.
6. M. Veldhorst et al., "Silicon CMOS Architecture for a Spin-Based Quantum Computer," arXiv, 30 Sept. 2016; arxiv.org/abs/1609.09700.



**ERIK P. DEBENEDICTIS** is a technical staff member at Sandia National Laboratories Center for Computing Research. He is a member of IEEE, ACM, and APS. Contact him at epdeben@sandia.gov.



*This article originally appeared in* Computer, *vol. 51, no. 2, 2018.*

# Deep Medical Image Computing in Preventive and Precision Medicine

**Le Lu**
Ping An Technology
US Research Labs
Johns Hopkins University

**Adam P. Harrison**
NVIDIA

Deep learning has a game-changing potential to improve the state of preventative and precision medicine within medical image computing. Here, we will first overview preventative and precision medicine and field of deep learning. Afterward, we will share our perspective on recent research and development activities in both areas and point out some existing achievements, positive indications, limitations, and near future opportunities and impediments. To flesh out our viewpoints, we draw from examples of our most recent work, which largely stem from radiologic images, but we encourage readers to consult some other recent reviews, which include many references that space did not allow us to include. We also assume the reader is broadly familiar with machine learning technologies.

*Overview and Status*: *Preventive medicine* in medical imaging refers to early detection of disease findings, e.g., lung nodules, colonic polyps, and liver/bone lesions, with the goal of timely patient intervention and management. Traditionally, this is done using manual examination from noninvasive imaging modalities, but more recently computer-aided solutions are becoming more prominent. *Precision medicine*, within imaging, means computing quantitative and precise imaging biomarkers, e.g., volumetric tumor measurements for tracking and beyond, to support clinical decision making and ultimately improve patient outcomes. Current radiological practices are still largely qualitative, even for the most advanced medical centers.

Deep learning, especially deep convolutional neural networks (CNNs), have made significant headway in categorically improving both preventative and precision medicine. This includes the preventative detection of anomalous findings in various imaging modalities, such as histology images or within computed tomography (CT) scans. For instance, markedly higher quantitative performance has been achieved for classifying enlarged lymph nodes and colonic polyps from CT.[2,4] For precision medicine, progress has been made on accurately segmenting organs[1,3,6] and anatomical anomalies,[7] which would play a central role for any quantitative markers.

The main reason of these early successes is *that effective learning of hand-crafted features for medical image analysis problems is notoriously hard, whereas CNNs eliminate this need*. In light of this, for the first time, CNNs have made feasible large-scale medical image parsing and tagging (over thousands or tens of thousands of patients and studies).[4,9,11] Works using CNNs have also built up a massive body of empirical evidence indicating that low-level features can be shared and fine-tuned

 Published by the IEEE Computer Society 

between networks trained on different image modalities or even from networks trained on natural image understanding tasks, e.g., from the ImageNet competition.

Deep learning has also rekindled and intensified industrial interest in medical imaging applications. Currently, there is a healthy body of startups focusing on medical image analysis and informatics, e.g., HeartFlow, Enlitic, Arterys, Viz.ai, Zebra-Me, and Butterfly Network. These efforts complement the research and development initiatives from large corporations, such as Siemens, IBM, Tencent, and Google Brain/DeepMind. Industrial investment and engagement cover various topics in both preventive and precision medicine. Though many technological, business, and clinical challenges lay ahead, scalable and effective deep learning principles will continue propelling high-performance, deployable medical imaging, and clinical informatics applications for years to come.

*Future Directions in Deep Preventive Imaging Medicine***:** Early developments in computer-aided detection (CADe), in the predeep learning era, mainly concentrated on detecting breast lesions/masses from mammography scans and detecting colonic polyps, lung nodules, vascular lesions, and pulmonary embolisms from CT or CT angiography images. Commercial clinical software products from several vendors were developed and deployed into practices, but the "expected" broad success and uptake did not happen. CADe products typically operate in a second reader mode (the *de facto* protocol approved by FDA), which means a radiologist first finishes an independent image reading without CADe and then examines the CADe generated findings to make his or her final decision. This pipeline is designed to increase detection sensitivity, with the aid of CADe software, but at the cost of additional physician workload. A common drawback is that the above-mentioned applications are not too difficult for experienced radiologists, and the extra workload is not always compensated by the moderate to minimal increases in sensitivity, if any.

*New pain points:* In order for preventive CADe systems and applications really to take off, new and true clinical pain points, which are not possible to fully address yet, need to be tackled and solved. These should lie beyond the traditional second reader protocol or an extra layer of safety. Instead, they should more actively aim to improve patient care capabilities. We provide a nonexhaustive list of a few promising examples as follows.

1. First reader triage software to (potentially) significantly increase the chances of detecting, and therefore quickly treating, patients suffering from a large vessel occlusion (LVO) from a stroke has recently been cleared by the FDA (work from Viz.ai). The current manual LVO stroke workflow results in low rates and long delays of treatment, which can be alleviated by this computer-aided triage and notification software that also saves stroke specialists' time. Another similar work is atherosclerotic vascular calcification detection and segmentation using low-dose full-body CT scans, which is a very time-consuming task and easy to miss for human readers. Long-standing drawbacks of manual exams and the high performance of deep learning alternatives meant that these tasks were amenable to a CADe approach. Additional opportunities for first reader software need to be identified and seized to further expand the impact of CADe solutions.
2. Chest X-rays are the most common medical imaging exams and a very accessible modality for screening both healthy (annual health exams) and unhealthy populations, e.g., those found in community clinics and hospitals, respectively. A game-changing application would be a reliable and economical automated chest x-ray screening and referral tool deployed across massive populations, especially those that are geographically distant from major hospitals. A total of four technical challenges, not necessarily specific to chest x-rays, stand in the way of such a vision: a) chest x-rays are associated with higher degrees of diagnosis uncertainty, whether analyzed by radiologists or computerized systems, than other modalities, e.g., lung nodule detection using chest CT; b) an extremely low false positive rate is required for generic preventive screening since a large majority of a population will be healthy; c) modeling and incorporating disease ontology is critical for reasoning and regularizing the raw outcomes from image classifiers to produce sensible diagnoses; d) human interpretable and verifiable results are required to produce a clinically complete CADe system. Work is ongoing to overcome these challenges, but recent developments, e.g., weakly supervised visual grounding of disease locations,[11] bring the field closer to this vision.
3. Full body preventive organ anomaly and cancer screening is the holy grail for general and asymptomatic population screening. Most likely, the ideal setup would also incorporate

cheaper and less intrusive nonimaging technologies, such as the "CancerSeek" blood test,[8] to screen all patients undergoing annual health exams. Next, a high-performance, high-accuracy automated medical imaging organ and pathology segmentation tool could be used to localize and verify the initial finding indications. Finally, a clinical decision fusion module, combining all nonimaging and imaging test results, could report and refer the identified "high risk" patients, versus the vast majority of average-risk individuals, to specialists. Although recent work on detecting and segmenting especially difficult organs and anomalies have made good progress,[3,6] for general population preventive screening, extremely generalizable deep learning methods, possibly trained on massive datasets, require further investigation.

*Future Directions in Deep Precision Imaging Medicine***:** Compared to preventative medical imaging, precision imaging has not been as well studied. Historically, quantitative imaging has faced roadblocks due to insufficiencies of prevailing machine learning technologies and a lack of buy-in from clinical partners in running the clinical trials and/or opening up the data archives needed to discover, characterize, and validate quantitative biomarkers. However, with the increasing capabilities of deep learning and prominent policy-level pushes for precision medicine, we see tremendous opportunity.

*New pain points:* Efforts should focus on computing precision imaging biomarkers at *hospital scale*, bringing forth analyses that physicians desire but are out of reach of human capabilities alone. These should focus on markers for prominent morbidities, especially cancer, but they should also provide tools that allow entirely new types of retrospective analyses for biomarker discovery. A key capability that requires further development is how to train deep learning systems on existing data sources, such as hospital archives, that are very large scale, but also messy and unstructured.

1. A common prerequisite of precision medicine is accurate and robust segmentation of anatomical structures from medical scans, i.e., classifying every pixel or voxel into a semantic meaning. Due to their superior performance, deep CNN-based segmentation methods[1,3,6] are now predominant. The value of using segmentation techniques is that raw image scans can be converted into semantic and human interpretable features, such as the volume of the left ventricle or the shape statistics of a patient's pancreas. These organ/anatomy based shape, volume, and appearance statistics can be computed from 2-D/3-D/4-D imagery, to assist both personalized diagnosis and treatment and also large population profiling. An important challenge is collecting enough data for training and ensuring any segmentation solution is generalizable to patient distributions encountered "in the wild."
2. When it comes to cancer, precision tumor growth tracking and prediction are additional key elements. Deep learning has pushed the capabilities of both forward. For instance, physicians need scalable solutions to intelligently match, track, and provide evidence-based similarity measurements to measure tumor growth rates from multiple time point studies of a patient. Due to the difficulty in obtaining training data, recent works train deep learning models on messy and large-scale clinical databases.[7,9] An example is illustrated in Figure 1. Continuing to leverage these large-scale data sources will be the key in further improving tracking capabilities. A related initiative is using deep learning techniques to observe subtle and precise longitudinal imaging changes in order to predict tumor growth rates and patterns.[10] A visual example of tumor growth prediction modeling and comparison is shown in Figure 2. Both tasks are tackling critical and clinically useful precision imaging biomarker problems, which cannot be done by human doctors alone due to the need to ingest "big data" to make accurate measurements and predictions.
3. Last but not the least, one of the ultimate goals of the precision medicine is performing retrospective analyses on clinical data to discover new imaging biomarkers that are correlated with morbidity. This can be framed as disease/concept discovery and tagging, given hospital-scale, or better yet multi-institutional, data of patient images and nonimaging records. This will likely require modeling multimodal imaging and nonimaging patient data on a graph configuration that builds and preserves pairwise[9] or higher order patient similarities. Such a representation could provide an indexable and holistic patient data view and repository, allowing analyses beyond plain classification. Importantly, given the long-tailed distribution of many diseases or ailments, such analyses are highly difficult, if not impossible, to perform without powerful computerized techniques, such as deep learning, that can effectively leverage data at massive scales.

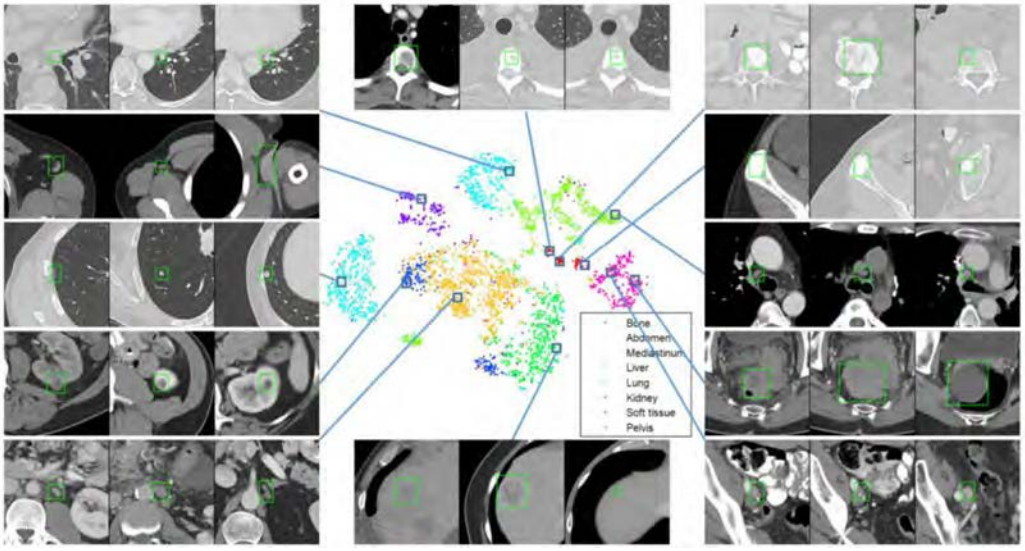



Figure 1. Visualization of clusters of tumor types automatically discovered from an analysis of a large scale dataset of ~34 000 tumors automatically extracted from a hospital archive.[9] Colors indicate the manually labeled lesion types, which correspond well with the automatically discovered tumor types. Best viewed in color.

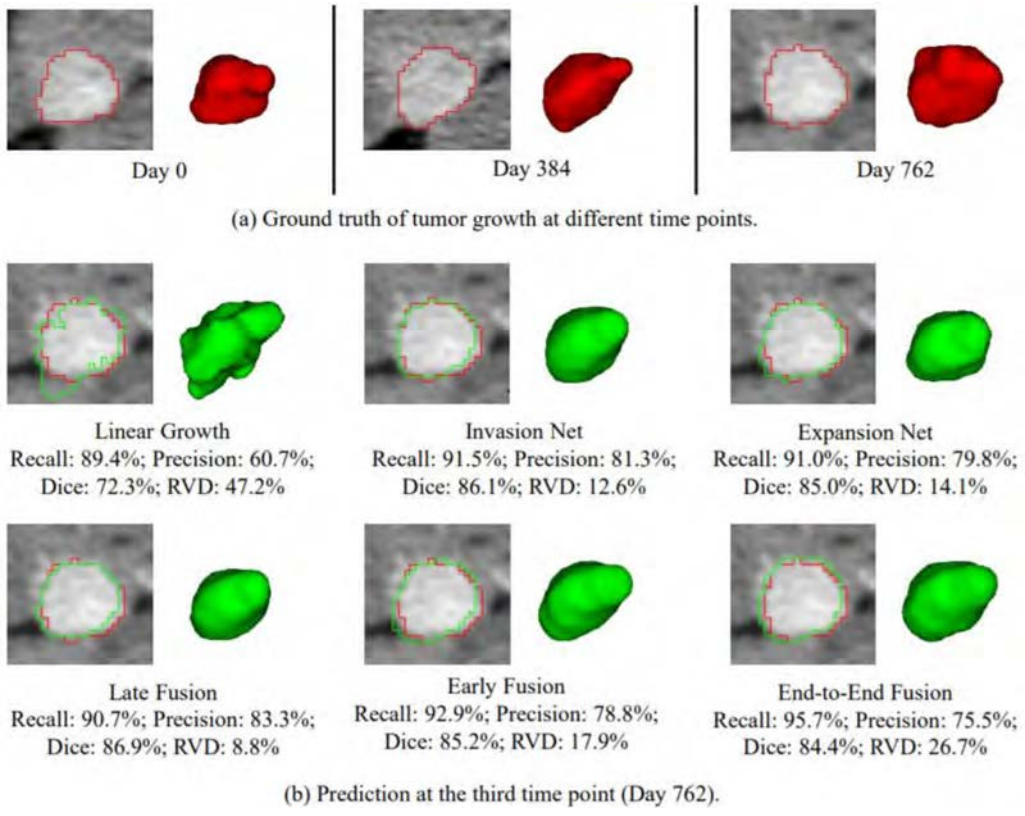



Figure 2. An example of the tumor growth prediction using a deep learning system.[10] (a) The segmented (ground truth) tumor contours and volumes at different time points. (b) The prediction results at the third time point from various automatic systems. Deep learning based tools match well with the ground truth. Red and green represent ground truth and predicted boundaries, respectively.

In summary, recent deep learning developments have been very impactful for medical imaging problems and applications, even can make some important tasks (e.g., first reader triage) from impossible via nondeep principles to reach possibly a clinical relevance level of performance. It will be a promising but challenging path going forward.

## REFERENCES


1. A. P. Harrison et al., "Progressive and multi-path holistically nested neural networks for pathological lung segmentation from CT images," in *Proc. MICCAI*, 2017, pp. 621–629.
2. H. Roth et al., "Improving computer-aided detection using convolutional neural networks and random view aggregation," *IEEE Trans. Med. Imag.*, vol. 35, no. 5, pp. 1170–1181, May 2016.
3. H. Roth et al., "Spatial aggregation of holistically-nested convolutional neural networks for automated pancreas localization and segmentation," *Med. Image Anal.*, vol. 45, pp. 94–107, 2018.
4. H. Shin et al., "Deep convolutional neural networks for computer-aided detection: CNN architectures, dataset characteristics and transfer learning," *IEEE Trans. Med. Imag.*, vol. 35, no. 5, pp. 1285–1298, May 2016.
5. H. Shin et al., "Interleaved text/image deep mining on a large-scale radiology database for automated image interpretation," *J. Mach. Learn. Res.*, vol. 17, no. 107, pp. 1–31, 2016.
6. J. Cai et al., "Improving deep pancreas segmentation in CT and MRI images via recurrent neural contextual learning and direct loss function," in *Proc. MICCAI*, 2017, pp. 674–682.
7. J. Cai et al., "Accurate weakly supervised deep lesion segmentation on CT scans: Self-paced 3D mask generation from RECIST," 2018, arXiv: 1801.08614.
8. J. Cohen et al., "Detection and localization of surgically resectable cancers with a multi-analyte blood test," *Science*, vol. 359, pp. 926–930, 2018.
9. K. Yan et al., "Deep lesion graphs in the wild: Relationship learning and organization of significant radiology image findings in a diverse large-scale lesion database," in *Proc. IEEE Conf. Comput. Vision Pattern Recog.*, 2018.
10. L. Zhang, L. Lu, R. M. Summers, E. Kebebew, and J. Yao, "Convolutional invasion and expansion networks for tumor growth prediction," *IEEE Trans. Med. Imag.*, vol. 37, no. 2, pp. 638–648, Feb. 2018.
11. X. Wang, Y. Peng, L. Lu, Z. Lu, M. Bagheri, and R. M. Summers, "ChestX-Ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases," in *Proc. IEEE Conf. Comput. Vision Pattern Recog.*, 2017, pp. 3462–3471.
12. X. Wang et al., "Unsupervised joint mining of deep features and image labels for large-scale radiology image categorization and scene recognition," in *Proc. IEEE Winter Conf. Appl. Comput. Vision*, 2017, pp. 998–1007.


## ABOUT THE AUTHORS


**Le Lu** can be contacted at le.lu@pingan.com.

**Adam P. Harrison** can be contacted at aharrison@nvidia.com.


*This article originally appeared in* IEEE MultiMedia, *vol. 25, no. 3, 2018.*

# Human-Aided Bots

**Pavel Kucherbaev**
**Alessandro Bozzon**
**Geert-Jan Houben**
Faculty of Electrical Engineering, Mathematics and Computer Science, Delft University of Technology

A chatbot is an example of a text-based conversational agent. While natural language understanding and machine learning techniques have advanced rapidly, current fully automated chatbots still struggle to serve their users well. Human intelligence, brought by crowd workers, freelancers, or even full-time employees can be embodied in the chatbot logic to fill the gaps caused by limitations of fully automated solutions. In this paper, we investigate human-aided bots, i.e., bots (including chatbots) using humans in the loop to operate. We survey industrial and academic examples of human-aided bots, discuss their differences and common patterns, and identify open research questions.

The idea of having a conversation with a machine similar to how we converse with a human is not new. In science fiction books and movies, various robots, such as the C-3PO humanoid robot from "Star Wars," and automated personal assistants, such as HAL from "2001: A Space Odyssey," helped heroes in their life and to manage their work duties. The first conversational agents to follow this idea already appeared in the 1960s.[1] Then, it was very hard to program such systems even for a narrow domain; a lot of complex rules were explicitly programmed, as there was no way to quickly and reliably parse user requests to understand what the user wanted. The significant improvements in parallel processing hardware and *natural language understanding* using *deep neural networks*[2] made it easier now to implement such conversational agents. A new market emerged, and major companies compete with their technologies for the leader position.

Messaging applications such as Facebook Messenger and Telegram are widely used by millions of people to interact with friends, colleagues, and companies.[3] Because of the high popularity of such tools, their users are very familiar with their minimalistic interfaces and functionality. Seeking the opportunities brought by modern conversational agents, these messaging applications have started supporting the creation of text-based conversational agents called chatbots, mimicking a conversation with a real human. Companies express commercial interest in such chatbots and already use them in application domains spanning from customer support (e.g., handling returns and replacements at a retail store [https://www.facebook.com/Customer-Support-Bot-1857341381220252/]) to sales (e.g., helping to find and purchase airflight tickets [https://www.facebook.com/TransaviaFlightSearch/]), and team productivity (e.g., organizing SCRUM stand-up meetings in Slack [https://standuply.com/]).

Chatbots implemented using automated techniques, such as rule-based or machine learning algorithms, are still far from being perfect, struggling to serve well user requests and to carry on a meaningful conversation. These issues are especially evident in open conversation domains.[4] In this

paper, we discuss how human intelligence could be used to address the limitations of fully automated solutions. We discuss different components of the chatbot architecture; we introduce the concept of *human-aided bot,* a chatbot system where at least one architecture component employs human intelligence; we introduce a *reference framework* to discuss human-aided bots and use it to compare existing examples introduced by academia and industry. In this comparison, we consider chatbots where humans intervene during runtime, and we *do not include* chatbots which are only pre-trained on human-generated data. We end this paper with a list of open research questions, aiming to guide and inspire research and industrial communities to take their next actions.

## CHATBOT

In Figure 1, we show chatbot architecture components that have been previously introduced.[5] After a chatbot receives a request from a user (e.g., "*What is the weather in San Francisco*"), the *language understanding (LU)* component parses it to infer the user's intention and the associated information (e.g., intent: "*check weather,*" entities: [location: "*San Francisco*"]). When the request is understood, *action execution and information retrieval (AEIR)* takes place, so that the chatbot performs the requested actions or retrieves the information of interest from its *data sources (DS, e.g., gets API response from openweathermap.org for San Francisco)*. Upon retrieval, the *response generation (RG)* component prepares a response to the user (e.g., "*It is +23°C and sunny in San Francisco now.*"). A *dialogue management* component is in place to keep and update the context of a conversation (e.g., the current intent, identified entities, or missing entities required to fulfill user requests), to request missing information (e.g., the chatbot asks "For which city would you like the weather forecast?"), to process clarifications by users (e.g., the user replies "*What about tomorrow?*"), and to ask follow-up questions (e.g., the chatbot replies "Would you like as well a forecast for a week?").

### Limitations of Chatbots

Each component in Figure 1 is usually implemented using rule-based algorithms or machine-learning models trained with datasets. Unfortunately, such automated approaches are still ineffective in a variety of real-world scenarios leading to a poor performance with:

- *LU* – the interpretation of user requests, due to limited (in size or diversity) training data;[6]
- *DM* – the generation of clarification requests for missing information, due to limitations in dialogue structure programmin;[4]
- AEIR/DS – the retrieval of the requested information—or the execution of the requested action—due to incomplete support for user intents, thus causing the chatbot to fall back to traditional information retrieval techniques (e.g., using search engine[7]) providing users with documents, rather than facts;
- RG – the presentation of the information to the user in a satisfactory fashion—or the generation of inappropriate responses—due to limitations in response templates, question-answer mapping, response synthesis techniques, or associated training data.[8]

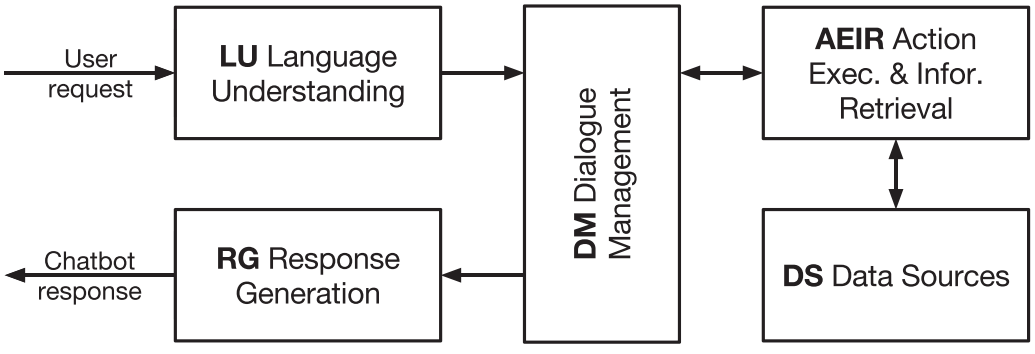



Figure 1. Chatbot architecture.

Intuitively, all the issues above could be easily overcome by a human being proficient with the language used in the conversation and having access to the Web. The computational paradigm that advocates the use of human processing power to solve problems that computers cannot yet solve is called *human computation*.[9]

## Human Computation in Chatbots

We refer to chatbots which utilize human computation in at least one component from Figure 1 as *human-aided bots*. Human computation, compared to rule-based algorithms and machine learning, provides more flexibility and robustness as humans can adapt and perform well even when input or instructions themselves change. Still, humans cannot process a given piece of information as fast as a machine, which makes it hard to scale to more user requests.

Human computation is a powerful approach helping to address the challenges of automated approaches. However, in the context of chatbots human computation introduces two extra challenges: *real-time support*—to make sure that users get responses in a reasonable time, and *scalability*—to make sure that even when the number of chatbot users grows, the costs grow only gradually. Below we review several existing human-aided bots, discussing how they combine automated and human computation approaches to serve their users and how they address real-time and scalability challenges.

# ANALYSIS OF HUMAN-AIDED BOTS

In our survey, we do not intend to explicitly cover all available human-aided bots (which it is not possible to do in such a publication format). Instead, we target a diversity of solutions and select academic and industrial examples serving different purposes in different domains.

We review 11 human-aided bots in Table 1 using nine dimensions: one describing their purpose and the domain where they function; five dimensions to discuss the implementation of each chatbot component; and three dimensions reflecting the way humans are involved: the source of human workers, how (if any) real-time response is supported, and how scalable the human aid is. We color-code chatbot architecture components reflecting the balance of human computation (red) versus automated (blue) approaches used. To make sure our understanding of selected human-aided bots matches the reality we contacted the authors of academic examples and representatives of companies behind industrial examples to get their feedback.

## Purpose and Domain

We distinguish the following purposes that chatbots serve: *informational* (Chorus, Guardian, CRQA, Insurify, SnapTravel, AskWiz, Facebook M), where users obtain information, such as train timetables; *transactional* (Guardian, InstructableCrowd, Legion:Mobile, Calendar.help, Nurtz, Insurify, SnapTravel, Facebook M), where users change the status of another system, e.g., to purchase a train ticket; *conversational (Chorus)*, where users interact with the chatbot for the sake of the conversation itself, such as discussing that trains are late sometimes. Some human-aided bots serve multiple purposes (e.g., Guardian is informational and transactional). We split the domains in which human-aided bots operate as: *generic* (Chorus, Facebook M), i.e., chatbots are ready to answer nearly any user request; *cross-domain* (Guardian, CRQA, AskWiz), i.e., chatbots operate in multiple domains; *domain-specific* (InstuctableCrowd, Legion: Mobile, Calendar.help, Nurtz, Insurify, SnapTravel), i.e., chatbots operate in a narrow domain.

## Chatbot Architecture

Most systems rely on human workers at least in some way to understand user requests. While Chorus, InstructableCrowd, and Legion: Mobile rely solely on human workers for language understanding, no system does natural understanding of all user requests completely automatically. AskWiz dynamically dispatches user requests such that their tech support agents are involved only when the automated system fails.

Table 1. Academic and industrial examples of human-aided bots.

| | | | Chatbot Architecture | | | | | Human Computation | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | HUMAN AIDED BOT | 1. Purpose & | 2. Language | 3. Dialogue management | 4. Action execution / | 5. Response generation | 6. Data sources | 7. Worker | 8. Realtime | 9. Scalability |
| *Academic* | **Chorus – hangouts-based conversational personal assistant, answering generic questions [Huang, 2016a]** | Conversational, Informational / Generic | (2) Human workers | (2) Human workers can see the chat log and facts of the(one) previous session of the current user. | (2) Up to human workers | (2) Human workers: workers propose responses and vote on each other's responses to decide which one to send to the user. | (2) Up to human workers | Crowd workers (MTURK) | Task redundancy | Up to MTURK |
| | **Guardian – web-based chatbot which works based on a set of APIs to which the crowd matches parameters from conversations with users [Huang, 2015].** | Informational, Transactional / Cross-domain: API-based | (1) Human workers + machine learning | (1) Dialogue is built around identifying API parameters mediated by human workers | (-2) Up to the API | (1) Human workers clarifying parameters, human workers replying based on Web API response. | (-2) 3rd party(arbitrary API) | Crowd workers (MTURK) | Retainer model | Up to MTURK |
| | **InstructableCrowd – mobile application through which users can create trigger-action rules on their smartphone with a help of the crowd [Huang, 2016b].** | Transactional / Domain-specific: Smartphone operations | (2) Human workers | (2) Dialogue is built around defining if/then rules | (-2) Android application based on rules created by human workers | (2) Human workers | N/A | Crowd workers (MTURK) | Task redundancy | Up to MTURK |
| | **Legion:Mobile – mobile application allowing visually impaired people to control their phones with voice cues [Lasecki, 2013].** | Transactional / Domain-specific: Smartphone operations | (2) Human workers | (2) Dialogue is build around performing actions the user wants | (2) Human workers performing actions on the user's smartphone via remote desktop | (2) Mediator selecting the action to take based on inputs of other workers | N/A | Crowd workers (MTURK) | Retainer model | Up to MTURK |
| | **CRQA – web-based application to get answers to generic questions with help of the crowd using as well third-party Q/A websites as a knowledge base [Savenkov, 2015].** | Informational / Cross-domain: Questions & Answers | N/A | N/A | (0) Relevant answers from web-search; human workers come up with relevant answers; | (0) Trained re-ranking model using workers feedback | (0) 3rd party(Yahoo! Answers, Answers.com, WikiHow.com, websearch), up to human workers | Crowd workers (MTURK) | Retainer model | Up to MTURK |
| *Both* | **Microsoft Calendar.help – email-based personal assistant scheduling meetings at the time which fits all the participants [Cranshaw, 2017].** | Transactional / Domain-specific: Scheduling | (-1) Machine learning + human workers(in a form of microtasks and macrotasks) | (-1) Defined workflow + macrotasks managed by human workers | (-2) Scheduling an event | (-1) Machine learning + human workers(in a form of macrotasks) | (-2) user information(calendar) | Crowd (Microsoft crowdsourcing platform, NDA-signed, hourly paid) | Humans working in shifts | Up to the crowdsourcing platform. Escalation from machines to microtasks and later to macrotasks |
| *Industrial* | **Nurtz – slack-based assistant proofreading text requested by users with the help of the crowd [http://nurtz.com].** | Transactional / Domain-specific: Writing | N/A | N/A | (2) Human workers proofreading text | N/A | N/A | Freelance remote agents | N/A (average response is in 10 minutes) | Hiring more agents |
| | **Insurify – facebook-messenger based assistant suggesting insurance quotes based on user requests [http://insurify.com].** | Informational, Transactional / Domain-specific: Insurance | (-1) Pattern matching + machine learning + human workers | (-1)The user requests and human worker responses then feed back into the machine learning algorithms | (-2) Purchasing insurance | (-1) Machine learning + human workers | (-2) 3rd party(Quotes of supported insurance companies) | 3rd party insurance agents | N/A | Hiring more agents |
| | **SnapTravel – facebook-messenger-based assistant helping users to find hotels with the help of travel agents [https://booking.getsnaptravel.com].** | Informational, Transactional / Domain-specific: Travel | (-1) Pattern matching + machine learning + human workers | (-1) Machine learning + human workers | (-2) Hotel booking | (-1) Machine learning + human workers | (-2) 3rd party(Offers of supported travel agencies and hotel companies) | Full-time employees | Humans working in shifts | Hiring more employees |
| | **AskWiz – facebook-messenger-based tech support agent matching users with tech experts [http://drippler.com].** | Informational / Cross-domain: Tech support | (0) Human workers + machine learning | (-1) Human workers answer custom tech questions | (2) Up to human workers | (1) Machine learning + human workers | (2) Up to human workers | Freelance remote agents | N/A | Hiring more agents |
| | **Facebook M – facebook messenger-based personal assistant performing custom tasks with help of a crowd of dedicated employees [http://bit.ly/2qKLaBX].** | Informational, Transactional / Generic | (-1) Machine learning + human workers | (-1) Machine learning + human workers | (-1) Custom actions | (-1) Machine learning + human workers | (0) 3rd party + user information | Full-time employees | Humans working in shifts | Hiring more employees |
| | | | N/A | -2 | -1 | 0 | 1 | 2 | | |
| | | | Not available | Only machine | Mostly machine, some human | Machines and Humans equal | Mostly human, some machine | Only human | | |

The way dialogue management is implemented is very similar to language understanding. In academic systems, dialogue management is handled primarily by human workers, while in industrial systems there is usually some predefined dialogue pattern around which conversations are conducted. Two systems (CRQA, Nurtz) do not have any dialogue management support as they focus on atomic request/response interactions.

Chorus, Legion: Mobile, Nurtz, and AskWiz leave AEIR completely up to human workers. More machine-oriented systems perform the following actions: executing an API call (Guardian), executing an Android OS command (InstructableCrowd), scheduling an event in a calendar (Calendar.help), purchasing an insurance plan (Insurify), and booking a hotel (SnapTravel). Systems relying on humans support controlling a user's smartphone via a remote desktop (Legion: Mobile), and proofreading texts (Nurtz). Some systems rely on both machines and humans, such as retrieving answers from web-search and creating new answers with the help of human workers (CRQA), and supporting a wide range of actions (Facebook M) from automatically calling an Uber car to having a human worker to contact Amazon customer support.

Apart from Nurtz, which simply sends to users an edited text with no comments, all other systems rely on humans to generate responses to their users. Academic prototypes Chorus, Guardian, InstrictableCrowd, and Legion: Mobile work such that responses to users are primarily generated by human workers. Other systems rely first on automated ways to generate responses and occasionally on human workers. CRQA ranks answers to pick the one to give using a pretrained ranking model with workers' feedback. Human workers can directly write to the chatbot user (e.g., AskWiz), otherwise, they vote for the response to be sent to the user (e.g., Chorus, CRQA), and in some cases, human workers generate only some responses while others are generated automatically (e.g., Calendar.help).

Three examples (InstructableCrowd, Legion: Mobile, Nurtz) do not have any external data sources. In some systems, the external data sources are completely up to human workers (e.g., Chorus, CRQA, AskWiz). Some systems rely on third-party services to extract information *for* the user (such as Web APIs in Guardian, various Q/A websites in CRQA, quotes of insurance companies in Insurify, offers of supported travel agencies and hotel companies in SnapTravel). Others rely on third-party services to extract information *about* the user (e.g., user calendar in Calendar.help). Not much information is available about Facebook M, but most likely it works as well based on a variety of third-party APIs and data sources.

## Human Computation

All academic examples rely on crowdsourcing platforms as a source of human intelligence: Calendar.help is based on a proprietary crowdsourcing platform and others are based on Amazon Mechanical Turk (MTURK). Industrial bots work with remote freelancers (Nurtz, Insurify, AskWiz) and full-time employees (SnapTravel, Facebook M).

Half of all systems ensure real-time responses by keeping workers waiting for tasks to come, as a retainment pool[10] in academic examples, and employees working in shifts in industrial ones. Some chatbots using crowdsourcing platforms simply post redundant requests to attract more workers (as a single task expecting multiple workers to perform it, or as multiple identical tasks) increasing the probability of getting fast responses. Information about how other industrial examples address latency is not publicly available.

The examples relying on crowdsourcing platforms can scale up to the limit of the number of workers (which might be hundreds or thousands) available at the platform at any given moment. Calendar.help has multiple tiers, escalating requests from automated algorithms (e.g., predictions using machine learning algorithms) to human micro-tasks (e.g., structured tasks, to identify meeting time or location) and later to human macro-tasks (e.g., a generic task where a worker needs to make a decision on how to process a given email). Platforms relying on remote agents (e.g., freelancers) scale up by hiring more agents. For downscaling, nothing is needed as human workers only get rewards for their completed tasks. Platforms relying on full-time employees (e.g., such as employees in a call center) need to hire more workers to scale up and lay off or repurpose employees in case of downscale.

# DISCUSSION

Below we provide a high-level discussion of the field of human-aided bots, following the same framework we used in the analysis.

## Purpose and Domain

Most bots serve informational or transactional purposes. The fact that the single conversational bot Chorus is completely human-based suggests that automated solutions are not yet able to support conversations in the open domain. Having reliability as a priority, most industrial bots operate only in a specific domain.

## Chatbot Architecture

The color-coding suggests that academic prototypes heavily rely on the crowd as they are not designed to be used by millions of users: it would cause too many requests to human workers leading to extreme costs. Industrial examples try to manage the costs by relying in the first place on pattern matching and machine learning and escalating to human workers only when automated approaches fail. The single industrial system which relies more on human workers is AskWiz; still, there it is part of the business model, as every request to their human tech experts is expected to be paid for by users. CRQA is a single example where humans and automated approaches work shoulder to shoulder, without either approach predominating.

## Human Computation

The source of human workers seems, in general, to correlate with the maturity of the system using it, therefore MTURK is the choice of academic prototypes, freelancers are the choice of startups, and full-time employees is the choice of more established companies such as Facebook. The primary reasons are quality and privacy concerns, since companies as Facebook and Microsoft cannot tolerate poor human inputs in their pipeline. To address this issue, Microsoft relies on proven crowd-workers with whom nondisclosure agreements are signed, and Facebook on full-time employees.

The common approach to ensure real-time responses from human workers is by keeping some workers waiting for tasks to come, which is implemented using a retainment model or scheduled shifts. Redundancy used by Chorus and InstructableCrowd increases the that someone selects the task quickly, but does not ensure it. It is not yet clear how real-time can be ensured with big spikes of request numbers, which is relevant to the issue of scalability in general.

Examples using crowdsourcing platforms seem to scale easily on demand up to a certain limit, examples relying on freelancers scale with some delay caused by finding and recruiting new agents, and the examples relying on full-time employees are the ones struggling to scale the most. Which is one of the reasons Facebook M is currently available only to users in California.

## Open Research Questions

We have reviewed and discussed several human-aided bots and now examine the following challenges which are still open for future scientific investigation.

### Purpose and Domain

- *Human computation quality control.* Different human computation quality control strategies could be used for chatbots serving different purposes. The current state of the art addresses the collection and verification of information, tasks that are pertinent to *informational* chatbots. Instead, more research is required to understand how to assess human work in *conversational* (e.g., to assure quality in a conversation on a sensitive topic) and *transactional* chatbots (e.g., to assure quality in the task: “call the number and book a table”).
- *Chatbot quality metrics.* Existing chatbot quality metrics focus on measuring how human-like the chatbot is 11. As there are informational and transactional types of chatbots, there is the need for metrics that also account for the quality of the service delivered by the chatbot.
- *Chatbots learning new skills.* Existing informational and transactional chatbots are built with fixed functionality, which it is possible to extend only with coding intervention. Understanding how to design chatbots able *to learn* and *organize* new skills during run-time, could make such chatbots much more powerful and therefore useful. The learning process could be carried in a form of a conversation (with chatbot users or domain experts), micro-tasks (completed by human workers), or even done completely automatically.

### Chatbot Architecture

- *Automation versus human labor.* While there are examples of systems with diverse combinations of automated and human work, it is a topic for future investigation how various combinations affect the performance of a human-aided bot and what could be an optimal balance for various domains. In the long term, human involvement is expected to decrease, as automated systems improve in their performance and adaptability.
- *Active Learning from the crowd.* Machine learning models for chatbots are trained during design time. Chatbots can learn proactively, assessing gaps in the training data (using certainty- based, committee-based, or other techniques[12]) and periodically requesting more training data from the crowd, so that the chatbot can serve more user requests automatically in the first place.

### Human Computation

- *Scalability and real-time.* Human-aided bots already serve thousands of users (e.g., Facebook M in California). To be able to serve millions of users, the following challenges should be addressed: to ensure that with the growth of the number of chatbot users the costs associated with human computation grow only gradually; to ensure near *real-time* execution of human tasks, even having a big demand of tasks and a moderate supply of workers.
- *Privacy.* Users interacting with human-aided bots often need to share their personal information (potentially also sensitive), which later could be shared with human workers, including those familiar with the user (e.g., a manager asking a chatbot about the best way to fire an employee, and her actual employee happened to be a part-time chatbot human worker who was assigned to answer this request). The topic of privacy is only marginally addressed in the human computation literature, and methods need to be designed and developed to address these discussed privacy concerns.

Addressing these challenges will help to significantly advance the current state of the art in human-aided bots and bring all us closer to the dream of having meaningful and productive interactions with chatbots.

## ACKNOWLEDGMENTS

We thank T.-H. Huang (Chorus, Guadian, InstructableCrowd), W. S. Lasecki (Legion:Mobile), D. Savenkov (CRQA), A. Monroy-Hernandez (Calendar.help), T. Kiryazov (Insurify), and H. Fazal (SnapTravel) for their help. This research has been supported in part by the Amsterdam Institute for Advanced Metropolitan Solutions with the AMS Social Bot grant, and by the Dutch national e-infrastructure with the support of SURF Cooperative (grant e-infra17023)

## REFERENCES

1. B. F. Green Jr., A. K. Wolf, C. Chomsky, and K. Laughery, “Baseball: An automatic question-answerer,” in *IRE-AIEE-ACM '61*, 1961, pp. 545–549.
2. R. Socher, C. C.-Y. Lin, Y. N. Andrew, and D. M. Christopher, “Parsing natural scenes and natural language with recursive neural networks,” in *Proc. 28th Int. Conf. Mach. Learn.*, 2011, pp. 129–136.
3. Business Insider Intelligence, “Messaging apps are now bigger than social networks,” *Business Insider*, 2015. Available at: http://www.businessinsider.com/the-messaging-app-report-2015-11
4. C. Chakrabarti and G. F. Luger, “A framework for simulating and evaluating artificial chatter bot conversations,” in *Proc. 26th Int. Florida Artif. Intell. Res. Soc. Conf.*, 2013, pp. 34–39.
5. M. McTear, Z. Callejas, and D. G. Barres, *The Conversational Interface. Talking to Smart Devices*. New York, NY, USA: Springer, 2016.
6. E. S. AlHagbani and M. B. Khan, “Challenges facing the development of the arabic chatbot,” in *Proc. 1st Int. Workshop Pattern Recogn.*, 2016, 100110Y’16.
7. J. Leber, “Where Siri has trouble hearing, a crowd of humans could help,” in *MIT Technology Review*, 2013. Available at: https://www.technologyreview.com/s/512406/where-siri-has-trouble-hearing-a-crowd-of-humans-could-help/
8. G. Neff and P. Nagy, “Talking to bots: Symbiotic agency and the case of Tay,” *Int. J. Commun.*, vol. 22, no. 60, 2016.
9. E. Law and L. V. Ahn, *Human Computation*. San Rafael, CA, USA: Morgan Claypool Publ., 2011.
10. W. S. Lasecki and J. P. Bigham, “Spoken control of existing mobile interfaces with the crowd,” in *CHI’13 Mobile Accessibility Workshop*, 2013.
11. Z. Yu, Z. Xu, A. W. Black, and A. I. Rudnicky, “Chatbot evaluation and database expansion via crowdsourcing,” in *RE-WOCHAT’16*, 2016, pp. 15–19.
12. G. Tur, R. E. Schapire, and D. Hukkuni-Tiir, “Active learning for spoken language understanding,” in *ICASSP’03,* 2003, pp. 276–279.
13. T.-H. K. Huang, W. S. Lasecki, A. Azaria, and J. P. Bigham, “Is there anything else i can help you with? Challenges in deploying an on-demand crowd-powered conversational agent,” in *HCOMP’16,* 2016, pp. 79–88.

14. T.-H. K. Huang, W. S. Lasecki, and J. P. Bigham, “Guardian: A crowd-powered spoken dialog system for web apis,” in *HCOMP’15,* 2015, pp. 62–71.
15. T.-H. K. Huang, A. Azaria, and J. P. Bigham, “Instructablecrowd: Creating IF-THEN rules via conversations with the crowd” in *CHI EA'16,* 2016, pp. 1555–1562.
16. D. Savenkov and E. Agichtein, “CRQA: Crowd-powered real-time automatic question answering system,” in *HCOMP’16*, 2016, pp. 189–198.
17. J. Cranshaw, *et al*, “Calendar.help: Designing a workflow-based scheduling agent with humans in the loop,” in *CHI’17,* 2017, pp. 2382–2393.

## ABOUT THE AUTHORS

**Pavel Kucherbaev** is a Post-Doctoral Researcher with Web Information Systems Group, Delft University of Technology, Delft, The Netherlands. His research interests include human computation and conversational agents. He received the Ph.D. degree from University of Trento, Trento, Italy. Contact him at p.kucherbaev@tudelft.nl.

**Alessandro Bozzon** is an Associate Professor with the Web Information Systems Group, Delft University of Technology, Research Fellow with the AMS Amsterdam Institute for Advanced Metropolitan Solutions, and a Faculty Fellow with the IBM Benelux Center of Advanced Studies. His research interests include intersection of crowdsourcing, user modeling, and web information retrieval. Contact him at a.bozzon@tudelft.nl.

**Geert-Jan Houben** is a Full Professor and the leader of the Web Information Systems Research Group, TU Delft, a Scientific Director of Delft Data Science, a Research Program Leader on Open & Online Education in TU Delft Extension School, and a Principal Investigator in AMS, Amsterdam Institute for Advanced Metropolitan Solutions. His research group covers subjects in the wider field of web engineering and web science, and his research interests include user modeling for web-based systems. Contact him at g.j.p.m.houben@tudelft.nl.

*This article originally appeared in*
IEEE Internet Computing, v*ol. 22, no. 6, 2018.*

# Smart Homes, Inhabited

**A.J. Brush**
Microsoft

**Mike Hazas**
Lancaster University

**Jeannie Albrecht**
Williams College

**Department Editors:**
A.J. Brush;
ajbrush@microsoft.com

Mike Hazas;
m.hazas@lancaster.ac.uk

Jeannie Albrecht;
jeannie@cs.williams.edu

More dramatically than ever, smart homes are affecting the lives of those who inhabit them. The ACM CHI Conference on Human Factors in Computing Systems is renowned for its focus on how technology impacts people's lives. In this issue, we highlight research presented at CHI 2018 that explores living in, and interacting with, smart homes.

Academics and researchers have been developing smart home technologies for about 50 years, but for a long time the work was experimental at best—large numbers of people were not incorporating the technologies into their own homes. One of the biggest hurdles to the widespread adoption of smart home technologies was making devices work together. Several different networking protocols were used for communication, and the configuration was challenging. In 1999, Kevin Ashton introduced the concept of the Internet of Things (IoT), and one of the most exciting developments that grew from this was home automation technologies.[1] By connecting devices to the Internet, it became more tractable to integrate smart home technologies into real homes. At the same time, sensors and gadgets themselves have been improving at a rapid rate. For example, technologies that use voice as a computing interface are gaining significant traction, and as the accuracy of voice recognition approaches 99 percent, experts predict that people will rely on them even more.[2]

As smart home devices continue to improve and more people adopt them, it is also becoming easier to study how people live with these technologies. In this article, we highlight recent research on the experiences of people living in smart homes: we provide overviews of several papers published in *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (CHI 2018) and *ACM Transactions on Computer–Human Interaction* (*TOCHI*). Reducing energy use is still cited as a key motivating factor for installing smart home technology, and so our first overview describes one recent study characterizing the energy consumption impacts of smart home design. The smart speaker market has exploded, and so next we describe two projects that look at the use of voice technologies in homes. Finally, there are still many unsolved challenges in how people interact with and control their smart home technology. We survey recent projects exploring use of end-user programming to allow occupants to develop custom solutions for common problems.

2469-7087/19/$33.00 © 2019 IEEE Published by the IEEE Computer Society

## SMART HOMES AND ENERGY IMPLICATIONS

In their CHI 2018 paper "Designing the Desirable Smart Home: A Study of Household Experiences and Energy Consumption Impacts," Rikke Hagensby Jensen, Yolande Strengers, Jesper Kjeldskov, Larissa Nicholls, and Mikael B. Skov from Aalborg University and RMIT University investigated how user desire influences energy consumption.[3] Most prior work on smart homes in the context of human–computer interaction (HCI) has focused on interactive designs to engage people to reflect on their energy consumption, and on smart energy technologies that make the transition to sustainable practices effortless and convenient. Unfortunately, the long-term implications of these efforts are limited. The authors argued that HCI designers should analyze their designs more holistically before making energy-saving claims.

To this end, the authors conducted a qualitative study of 23 households with smart home devices to create a typology of 10 desired characteristics for smart homes and their energy implications. They structured these characteristics into three smart home personas: the *helper*, which captures desires related to the smart home's functional purpose, the *optimizer*, which captures characteristics related to desired outcomes for the smart home, and the *hedonist*, which captures pleasure-seeking experiences within the smart home. Each of these personas has energy implications that impact overall sustainability. For example, the helper persona requires the introduction of new "helpful" energy-consuming devices that are always on. The optimizer persona requires devices that help automate smart home actions and provide energy feedback to occupants. The hedonist requires "fun and cool" energy-consuming devices that create unique and beautiful living spaces. The authors found that the different desires embedded in smart homes both complement and contradict one another, highlighting a design paradox. Smart home technologies can actually undermine the desire to live sustainably.

## SMART HOMES AND SPEECH INTERACTION

While in the past controlling a smart home typically involved remotes, switches, touch-panel controls, or other physical interactions, new devices with far-field microphones now enable speech control. After setup, voice-controlled speaker devices such as the Amazon Echo, Google Home, Cortana Invoke, and Apple Homepod make smart home control as simple as saying, "turn off the lights." Researchers have begun to study in more depth how people interact using speech in their homes. In their CHI 2018 paper "Accessibility Came by Accident: Use of Voice-Controlled Intelligent Personal Assistants by People with Disabilities," Alisha Pradhan and Kanika Mehta from the University of Maryland and Leah Findlater from the University of Washington examined how people with disabilities are using the Amazon Echo in their homes.[4] By conducting a content analysis of 346 Amazon Echo reviews and then 16 interviews of users with visual impairments, the authors explored how the devices are being used and opportunities for future work.

In their content analysis, the authors found 52 of the reviews (15 percent) mentioned using home automation, most often describing the value for a user that had motor impairments. The reviews highlighted ease of use and an improvement in independence. Lights were the most common devices mentioned; other devices included smart outlets, thermostats, switches, televisions, security systems, or door locks. For example, a user with ALS was able to blink lights to get attention, while another person was able to control important health equipment and an air conditioner without getting up.

Of the 16 people with visual impairments interviewed, only four currently had home automation devices, but the authors found that all participants wanted automation. However, several were renters, and others cited cost as a barrier. Lights and the thermostat were the most commonly desired smart appliances. One participant had found her oxygen compressors could not be paired with a smart switch, highlighting the value of enabling a wider range of appliances to interface with smart home controls. The authors summarized their findings by pointing to the potential of voice control of smart home appliances to address accessibility issues, and to give users a sense of independence and freedom.

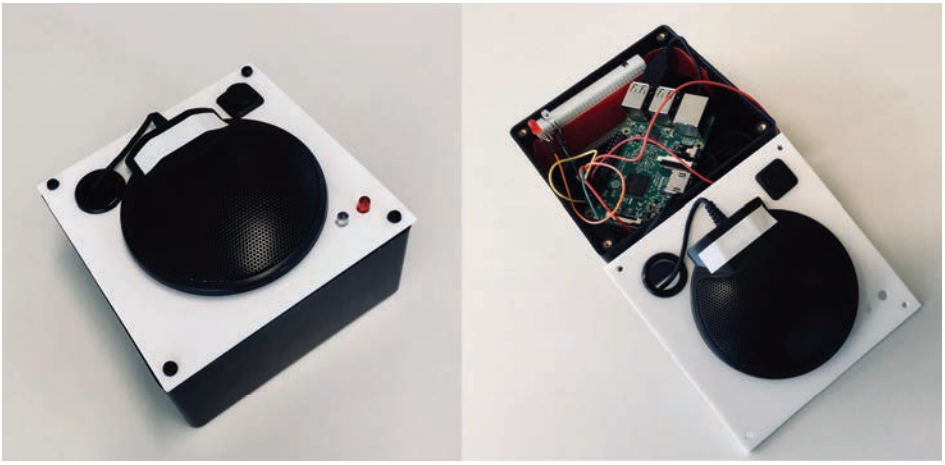


Figure 1. External (left) and internal (right) view of the Conditional Voice Recorder, showing the Raspberry Pi and conference microphone. Photo courtesy of Martin Porcheron.

A second paper from CHI 2018 that also looked at the use of Amazon Echo in homes follows in a long tradition of researchers building and installing additional smart technology to study behavior in-situ. In "Voice Interfaces in Everyday Life," Martin Porcheron, Joel Fischer, Stuart Reeves, and Sarah Sharples from the University of Nottingham closely examined how families engage with voice-controlled speaker devices.[5] They built a separate Conditional Voice Recorder (CVR) device to sit next to the Echo (see Figure 1). The CVR was always listening, unless turned off. When the wake word "Alexa" was heard, the CVR saved the prior and following minute of audio, allowing researchers to study what happened right before and during the conversations with the assistant.

The CVR consisted of a Raspberry Pi, a USB-connected conference microphone, LEDs to indicate listening and recording, a button to switch on/off listening, and software that extended open source voice-detection libraries. The components were contained in a box with holes for the LEDs, button, and cables (mains power to the Pi, and USB cable to the microphone); the microphone was attached to the lid. This allowed the researchers to easily deploy the CVR by just plugging it into the power—the device required no Internet connection, and the software was configured to start automatically.

By recording a combined total of six hours of verbal exchanges with the assistant from five houses, the authors observed how the Echo became embedded in the life of the home, both in the manner in which people changed how they interacted to reduce background noise and in how they improved the chance of a successful interaction by taking turns in a family situation to address the Echo.

At the end of the paper, the authors reflected on future implications of their work. They started by rejecting the notion that voice user interface (VUI) devices are conversational in nature. Since VUI devices are unable to fully understand the situated meaning of conversation, the responses from the devices do not always coherently follow the input. The authors felt a more realistic goal for VUIs is to develop a request/response design rather than a conversation design. They also suggested VUI responses should provide users with resources that can support and encourage further interaction with the device.

## SMART HOMES AND END-USER PROGRAMMING

A long-standing challenge for people living in smart homes is how to configure and interact with the technology. The two papers presented here are from a recent collection on end-user design for the IoT, appearing as a special issue of *TOCHI*. These stand in a five-year tradition of making smart homes easier to engage with and more accessible.[6,7]

Corina Sas of Lancaster University and Carman Neustaedter of Simon Fraser University explored 18 smart home inhabitants' experiences with an open source, DIY domestic energy monitor.[8] They engaged with two unique sets of households: half the participants were from a local green community, and the other half were recruited from the online community that worked to develop the DIY monitor. They found that the DIY monitor allowed unfettered exploration of home energy data with the ability to incorporate new, networked sensors, and were quite highly valued by the inhabitants when compared to off-the-shelf energy monitors. Similar to prior work on DIY and maker communities, this tied back to the open-ended extensibility of the DIY device, underscoring DIY smart home devices' potential for continued transparency of operation and functionality, connectability, engagement, and heirloom status.

Working with 18 different smart home inhabitants, Julia Brich, Marcel Walch, Michael Rietzler, Michael Weber, and Florian Schaub from Ulm University explored the applicability of rule-based versus process-oriented framings of home automation goals or tasks.[9] Rule-based approaches typically have simple conditional clauses wherein an actuator or message is triggered based on one or more sensor inputs, often termed "if-this-then-that." Process-oriented approaches acknowledge that input and output events can have looser temporal couplings, and that household routines are often best thought of as a longer sequence of events, extending well past if-this-then-that. For each household, the researchers conducted a brief home tour followed by think-aloud tasks and interviews based around process-oriented and rule-based "notation kits" (paper prototyping). Findings included a survey of different devices, and the way that the households would envision automating them—and strong opinions about which devices should *not* be automated. Participants leaned toward comfort and home control applications, and generally found rule-based approaches good for simple tasks but limiting otherwise, and process-oriented framings more complex but extensible. They also found that participants who used the rule-based notation tended to intuitively incorporate process-oriented elements when specifying them on paper, such as modeling temporal and hierarchical dependencies among if-this-then-that rules, suggesting a need for richer support in expressing contextual rule triggers.

## CONCLUSION

In this article, we highlighted recent published work from CHI 2018 and *TOCHI* to gain a better understanding of how people live in smart homes. To this end we summarized projects related to how occupants of smart homes engage with technologies via speech interaction, as well as projects that examined customization of smart environments via end-user programming techniques. Voice-related technologies and end-user development capabilities are rapidly evolving, and we expect to continue to see significant advancements in smart homes in these contexts moving forward. In addition, we recounted work related to the energy implications of smart homes, with an emphasis on how inhabitants engage with their smart homes to better understand energy consumption. One of the most cited motivating reasons for building smart homes is to reduce energy usage, so we also expect future developments in this space. In our next column, we will take a closer look at issues related to energy consumption as we highlight common themes and interesting projects involving smart homes from the ACM e-Energy conference.

## REFRENCES

1. K. Moser, J. Harder, and S.G.M. Koo, "Internet of Things in Home Automation and Energy Efficient Smart Home Technologies," *Proc. 2014 IEEE Int'l Conf. Systems, Man, and Cybernetics* (SMC 14), 2014; doi.org/10.1109/SMC.2014.6974087.
2. M. Meeker, "Internet Trends of 2016—Code Conference," blog, KPCB, 1 June 2016; http://www.kpcb.com/blog/2016-internet-trends-report.
3. R.H. Jensen et al., "Designing the Desirable Smart Home: A Study of Household Experiences and Energy Consumption Impacts," *Proc. 2018 CHI Conf. Human Factors in Computing Systems* (CHI 18), 2018; doi.org/10.1145/3173574.3173578.
4. A. Pradhan, K. Mehta, and L. Findlater, "'Accessibility Came by Accident': Use of Voice-Controlled Intelligent Personal Assistants by People with Disabilities," *Proc.*

*2018 CHI Conf. Human Factors in Computing Systems* (CHI 18), 2018; doi.org/10.1145/3173574.3174033.
5. M. Porcheron et al., "Voice Interfaces in Everyday Life," *Proc. 2018 CHI Conf. Human Factors in Computing Systems* (CHI 18), 2018; doi.org/10.1145/3173574.3174214.
6. B. Ur et al., "Practical Trigger-Action Programming in the Smart Home," *Proc. SIGCHI Conf. Human Factors in Computing Systems* (CHI 14), 2014, pp. 803–812.
7. J.-b. Woo and Y.-k. Lim, "User Experience in Do-It-Yourself-style Smart Homes," *Proc. 2015 ACM Int'l Joint Conf. Pervasive and Ubiquitous Computing* (UbiComp 15), 2015, pp. 779–790.
8. C. Sas and C. Neustaedter, "Exploring DIY Practices of Complex Home Technologies," *ACM Trans. Computer-Human Interaction*, vol. 24, no. 2, 2017; doi.org/10.1145/3057863.
9. J. Brich et al., "Exploring End User Programming Needs in Home Automation," *ACM Trans. Computer-Human Interaction*, vol. 24, no. 2, 2017; doi.org/10.1145/3057858.

## ABOUT THE AUTHORS

**A.J. Brush** is a principal program manager at Microsoft. Contact her at ajbrush@microsoft.com.

**Mike Hazas** is a reader in the School of Computing at Lancaster University. Contact him at m.hazas@lancaster.ac.uk.

**Jeannie Albrecht** is a professor and chair of the Department of Computer Science at Williams College. Contact her at jeannie@cs.williams.edu.

*This article originally appeared in*
IEEE Pervasive Computing, v*ol. 17, no. 3, 2018.*

# Evaluating Speech-Based Smart Devices Using New Usability Heuristics

**Zhuxiaona Wei**
deeplearning.ai

**James A. Landay**
Stanford University

We developed a set of 17 usability heuristics for speech-based smart devices. An expert evaluation of three popular devices shows that these heuristics can be used to uncover existing usability problems as well as help design new interfaces.

A recent empirical study showed that in both English and Mandarin, speaking is almost three times faster than typing a short message.[1] Thanks to recent breakthroughs in speech and language technologies, speech user interfaces (SUIs) have improved rapidly, and voice-enabled devices are now common. Baidu's Deep Speech 2 system, for example, can recognize spoken words with human-level accuracy.[2]

Nevertheless, designing good SUIs remains challenging.[3] The state of an SUI is often opaque to users, leading to more user errors compared to graphical user interfaces (GUIs).[4] Unfortunately, simply transforming GUIs into speech interfaces does not work well.[5] Although researchers have been working on SUI technology for three decades, much useful knowledge is in older papers and not easily accessible to designers. Moreover, the knowledge has not been updated to reflect recent improvements in speech-recognition accuracy. Consequently, those new to SUI design often feel lost.[6]

To help address these issues, we developed a new set of heuristics for designing and evaluating speech-based smart devices. To validate and improve these heuristics, we had a group of usability experts—half of whom specialized in SUIs—use them to empirically evaluate three state-of-the-art devices.

## RELATED WORK

In the early 1990s, Jakob Nielsen developed a set of 10 usability heuristics for evaluating UIs (www.nngroup.com/articles/ten-usability-heuristics). Although these heuristics are most often applied to GUIs, he and his colleagues also used them to evaluate a telephone voice-response system.[7] However, the user input and system output options for the system were quite limited.

 Published by the IEEE Computer Society 

Researchers have developed several SUI guidelines and best practices over the past 20 years.

In 1996, Alexander Rudnicky created 7 guidelines for SUIs integrated with visual applications.[4] However, today's devices are speech-first or even speech-only, and speech technologies have improved dramatically. The guidelines need to be updated to be used for today's smart devices.

In 2001, Laila Dybkjær and Niels Ole Bernsen created a usability testing guide for spoken-language dialogue systems.[8] However, we believe heuristic evaluation is more efficient than usability testing, especially since there are no good standards for SUI design yet. The heuristics can also be used for designing new systems.

In 2003, Bernhard Suhm created a database of SUI design problems and solutions to generate guidelines for telephone dialog system design.[3] He suggested using these guidelines for heuristic evaluation but did not validate the guidelines. Furthermore, unlike telephone systems, many of today's smart devices are not speech-only but also have a physical form with which users can interact, enabling a richer experience. There are likely different design problems due to these characteristics.

In their 2007 book *Wired for Speech*, Clifford Nass and Scott Brave presented valuable theoretical insights from years of research, many of which we incorporated into our new heuristics.[9] More recently, Cathy Pearl shared lessons from her career designing SUIs for mobile devices and interactive voice-response systems in *Designing Voice User Interfaces*.[10] Most of this knowledge is still applicable to today's smart devices, though it is hard to distill a set of manageable guidelines from her book.

In 2017, Google (https://developers.google.com/actions/design) and Amazon (https://developer.amazon.com/designing-for-voice) have each published a set of design guidelines for their own smart devices. However, to our knowledge there are no empirical evaluations of these guidelines.

In sum, no general guidelines have been developed specifically for evaluating state-of-the-art speech-based smart devices, nor have any empirical studies been done on these devices' usability. We believe both are critical for the research community to better understand existing problems and try to remedy them.

## NEW SUI HEURISTICS

In adapting heuristic evaluation to SUIs, some researchers have modified Nielsen's 10 heuristics to be more applicable to the new interface style while others have extended them by adding SUI-specific heuristics. Drawing on the related work described above, we compiled a set of 17 new heuristics grouped into 5 categories: general (S1–S5); conversational style (S6–S8); guiding, teaching, and offering help (S9–S10); feedback and prompts (S11–S14); and errors (S15–S17). The heuristics are as follows:

S1: Give the agent a persona through language, sounds, and other styles.

S2: Make the system status clear.

S3: Speak the user's language.

S4: Start and stop conversations.

S5: Pay attention to what the user said and respect the user's context.

S6: Use spoken language characteristics.

S7: Make conversation a back-and-forth exchange.

S8: Adapt agent style to who users are, how they speak, and how they are feeling.

S9: Guide users through a conversation so they are not easily lost.

S10: Use responses to help users discover what is possible.

S11: Keep feedback and prompts short.

S12: Confirm input intelligently.

S13: Use speech-recognition system confidence to drive feedback style.

S14: Use multimodal feedback when available.

S15: Avoid cascading correction errors.

S16: Use normal language in communicating errors.

S17: Allow users to exit from errors or a mistaken conversation.

The list of heuristics along with detailed descriptions and examples can be found at http://hci.stanford.edu/publications/2018/speech-he/sui-heuristics.html.

## EVALUATING THE NEW HEURISTICS

To validate and improve our heuristics, we had usability experts use them to empirically evaluate three state-of-the-art speech-based smart devices: Google Home, Amazon Echo, and Apple Siri (see Figure 1). Nielsen recommends using a minimum of 3–5 evaluators to identify most UI problems.[11] We felt that 8 evaluators should find most of the interface problems in the devices and could use the union of these problems as ground truth. Half of the participants in our study had an average of 11–20 years in SUI design; the rest were nonspeech usability experts, with an average of 11–20 years' experience, and each had completed more than 10 heuristic evaluations. Most of the evaluators were native American English speakers; one speech expert and one nonspeech expert were nonnative speakers. Seven of the evaluators were female. Only one of the participants did not currently use Apple Siri. Most of the speech experts owned at least one of the other two devices, which they used daily. Some of the nonspeech experts had tried but did not own an Amazon Echo. Each evaluation took 2.5–4 hours, and the evaluators received $100–$150 per hour as compensation based on their normal consulting rates. All 8 sessions were performed in a quiet meeting room at Baidu Research's office in Sunnyvale, California, to minimize noise and other distractions.

Figure 1. The three speech-based smart devices evaluated in our study: Google Home (left), Amazon Echo (middle), and Apple Siri (an iPad face down, right).

After a brief introduction, we presented the new heuristics to the evaluators, who were given time to read them and ask any questions. Next, we asked the evaluators to complete a set of 10 tasks on all three devices. They evaluated all 10 tasks in order on a single device at a time. Learning effects were eliminated by counterbalancing the order for evaluating the devices. All devices were reset after each session to ensure no language was learned from the prior interactions. Compiled from several market research reports, the tasks were the 10 most frequent real-life use cases for smart speakers: general questions, music, weather, local business, shopping, radio/news, messaging, calendar, to-do/reminders, and timers/alarms. As the devices have different functionality, not all 10 tasks are supported by each device. The evaluators rated each task on a scale of 1 (very difficult) to 7 (very easy). After completing each task, the evaluators documented the usability problems they found, along with the heuristic violated by each problem and

a severity rating. Severity was recorded using Nielsen's scale: 1—cosmetic problem, 2—minor problem, 3—major problem, and 4—usability catastrophe. After evaluating each device, the participants filled out a standard Subjective Usability Scale (SUS).[12] They then proceeded to the next device and repeated the procedure. Finally, we conducted a follow-up interview with the evaluators about their overall experience with the three devices and, more importantly, how the heuristics might be improved.

## EVALUATION RESULTS

The evaluators initially found 388 problems. We analyzed their problem descriptions to identify identical ones—we considered problems that had similar descriptions and the same violated heuristic as the same problem. Some evaluators occasionally listed several heuristics for a single problem. In these cases, we chose the heuristic we judged to be closest to the problem description. After this process, we were left with 279 unique problems. We averaged the severity ratings for each problem. We considered problems with severity ratings equal to or greater than 2.5 as high severity-problems.

### Problems Found

Table 1 shows the average difficulty level of each task on each device as rated by the evaluators. We report this for context only—our goal was not to compare the usability of the devices, which are designed to support different tasks.

Table 1. Average difficulty level of each task, from 1 (very difficult) to 7 (very easy).

| Task | Google Home | Amazon Echo | Apple Siri |
|---|---|---|---|
| 1. General Questions | 3.5 | 2.9 | 3.0 |
| 2. Music | 3.1 | 2.1 | 2.4 |
| 3. Weather | 5.8 | 6.1 | 5.4 |
| 4. Local Business | 5.3 | 4.1 | 4.4 |
| 5. Shopping | 3.9 | 3.5 | 2.6 |
| 6. Radio/News | 5.0 | 4.8 | 3.1 |
| 7. Messaging | 2.0* | 4.9 | 5.1 |
| 8. Calendar | 5.0 | 6.3 | 5.1 |
| 9. To-Do/reminders | 2.8* | 5.6 | 5.0 |
| 10. Timers/alarms | 5.0 | 5.6 | 4.4 |

*Task not explicitly supported by the device.

Figure 2 summarizes the number of high-severity and low-severity problems found by speech and nonspeech experts for each device. The total number of problems found for Google Home and Amazon Echo were similar—84 and 83, respectively. The evaluators found 33 percent more problems (112) with Apple Siri.

The four speech experts found 70 percent of the total number of problems, which is significantly higher than the 45 percent of problems found by the four nonspeech experts: $t(18) = 4.152$, $p < .001$. Surprisingly, only 15 percent of the problems were found by more than one evaluator; the overlapping percentage of problems found on Google Home was especially small (7 percent). There was greater overlap finding problems among nonspeech experts (28.4 percentage) than among speech experts (9.4 percent).

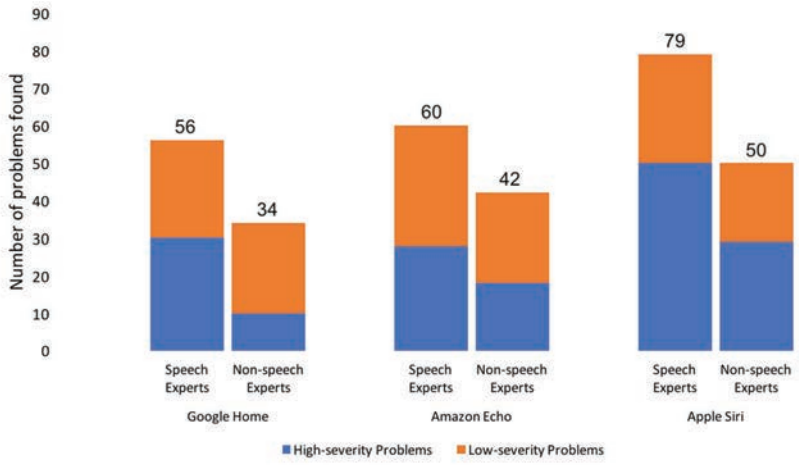



Figure 2. High-severity and low-severity problems found by speech experts and nonspeech experts on each device.

The evaluators found 141 high-severity problems, which is 50 percent of all the problems. The speech experts found 77 percent of these high-severity problems and the nonspeech experts found 40 percent, the same pattern observed in the entire set of problems. However, speech experts found even more of the high-severity problems with Google Home (83 percent) than nonspeech experts (28 percent).

Statistically, the number of total problems and the number of high-severity problems found by speech experts were both significantly higher than those found by nonspeech experts: $t(30) = 4.478$, $p < .001$, and $t(30) = 4.074$, $p < .001$, respectively. Apple Siri had significantly more high-severity problems than both Google Home and Amazon Echo: $F(2) = 3.133$, $p1 < .05$, $p2 < .05$.

We considered all the problems found by all of the evaluators as an estimate of the ground truth for the total number of problems existing in each SUI. Using the accumulated data, Figure 3 shows that 3 evaluators found 70 percent of the problems and 5 evaluators found 85 percent of the problems, which aligns with Jakob Nielsen and Thomas Landauer's mathematical model of the finding of usability problems.[13]

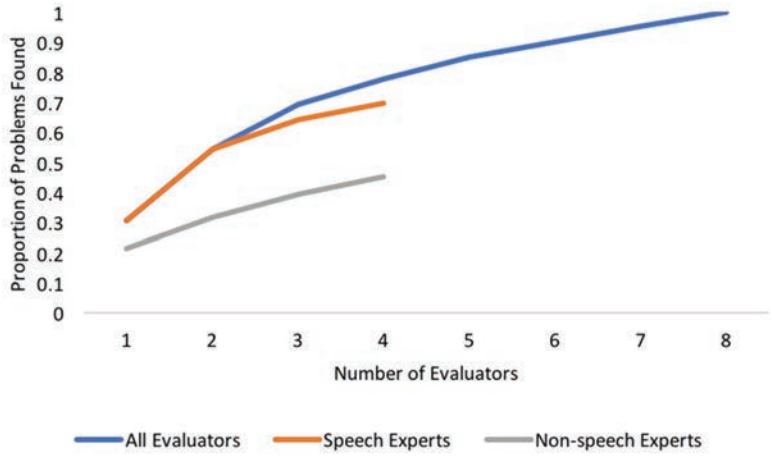



Figure 3. Average proportion of problems found as a function of the number of evaluators by all evaluators, speech experts, and nonspeech experts.

## Heuristics Used

The evaluators used all 17 heuristics, with 2 heuristics, S5 and S12, accounting for over 26 percent and 5 heuristics accounting for less than 11 percent of the total problems found (see Figure 4). S5 accounted for most high-severity problems (18 percent), with S9 the second most common in that category (10.6 percent).

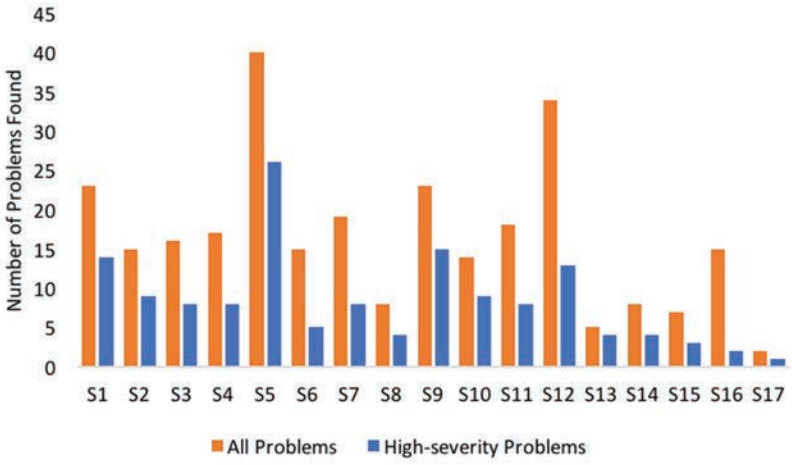



Figure 4. Number of all problems and high-severity problems identified using each heuristic.

S5, S12, S9, and S1 were the 4 heuristics most frequently used to find both all problems and high-severity problems. S7 was also frequently used to identify all problems, but not high-severity problems. S17 was only used twice to find all problems. S8, S13, S14, and S15 were each used to identify less than 3 percent of problems. Interestingly, S16 was used to find 5.4 percent of all problems but only 1.4 percent of high-severity problems. In general, the heuristic violations seem well-distributed.

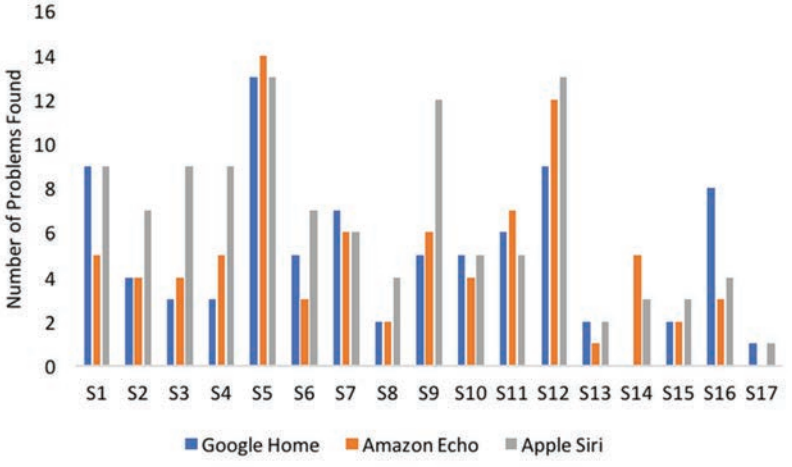



Figure 5. Number of problems found using each heuristic in Google Home, Amazon Echo, and Apple Siri.

Figure 5 shows the number of problems found using each heuristic for the three devices. S1, S5, S7, S11, S12, and S16 were used to find more than 5 problems in Google Home. A similar set were used to find more than 5 problems in Amazon Echo: S5, S7, S9, S11, and S12. A more diverse set of heuristics—S1, S2, S3, S4, S5, S6, S7, S9, and S12—was used to find more than 5 problems on Apple Siri. The evaluators found more problems with Apple Siri, which is different from the others in being screen-based.

## Key Problems

We performed a frequency analysis of the 279 unique problems to group together the common types of problems that violated the most frequently used heuristics. We describe and explain these in more detail below.

### S5: Pay Attention to What the User Said and Respect the User's Context

In many instances, the device ignored what the user said or only got part of the user's input. For example, when an evaluator requested information about books by Daniel Kahneman, the Amazon Echo typically responded with, "Audible lets you experience books in a whole new way. To try one, ask me to read *The Hobbit* or *The Great Gatsby*." Even when the evaluator tried this query multiple ways, the device continued promoting Audible. Annoyed that "it keeps giving me ads," one evaluator said she would "walk away" in real life. Amazon Echo correctly answered this query only a couple of times. Similarly, Apple Siri responded with "These books written by Daniel Kahneman are available on iBooks" or "Looking for books on iBooks." When the evaluator tried to refer to iBooks, the device said, "OK, here is iBooks," but as the iPad was face down the evaluator was not sure if it opened the iBook or not. Only Google Home correctly responded to this query, probably because Google is better at search. With Amazon Echo and Apple Siri, evaluators were unsure whether the system could not do something or they did not structure the question properly, so they kept trying.

Another problem was that the devices did not respond to follow-up questions, even in the same conversation. For example, in response to a prompt by an evaluator, Amazon Echo and Google Home would provide a list of restaurants. However, when the evaluator asked for the hours of the "first restaurant," the devices could not understand the request. Similar problems occurred with other questions that needed clarification. The evaluator usually had to wake up the device and restart the conversation. As one of the evaluators noted: "A lot of wake word speaking becomes tedious. In some ways, if certain queries result in follow-up questions, consider keeping the dialog open."

Finally, the devices did not always respect the users' context. When asked for the weather, for example, Google Home and Apple Siri obtained the evaluator's current location and then responded with the local weather. Amazon Echo, however, defaulted to Seattle. When the evaluator explicitly asked for the weather in "Sunnyvale" it gave the correct answer, but when asked the follow-up question "Will it rain on Friday?" it again told the evaluator the weather in Seattle.

### S12: Confirm Input Intelligently

The devices sometimes lacked implicit confirmation. When asked to play a particular song, Apple Siri started playing the song without providing its name, leaving the evaluator unsure whether it was the correct song. Similarly, when prompted to set a reminder, Apple Siri responded with "OK, I will remind you" without confirming that she did so and when exactly she would execute the reminder. Likewise, Google Home assumed "2 o'clock" was "2 pm" and did not confirm this with the evaluator.

The devices also failed to explicitly confirm some critical actions. For example, when asked to cancel an alarm, Amazon Echo did not ask the evaluator which one and simply canceled the alarm the evaluator had just set.

## S1: Give the Agent a Persona through Language, Sounds, and Other Styles

Most of the persona-related problems were found by one speech expert, who had a lot of experience designing personas for smart devices. Of all three devices she wrote, "The persona is not consistent; the inconsistencies themselves are distracting. For example, the visual light, the prompts, and the behavior do not have adequate coherence through time in order for me to perceive a coherent personality." The lack of persona makes it hard to distinguish the devices from the voice alone. Most of the evaluators said it was hard to evaluate the devices' personas because they are generic.

## S9: Guide Users through a Conversation so They Are Not Easily Lost

The devices often failed to provide user guidance. For example, Amazon Echo continually promoted Audible without giving any other cues or responses. The evaluators became confused about what was wrong and even felt they were being ignored. Amazon Echo repeatedly replied that "I do not have that. Would you like to hear this?" One evaluator noted that instead it should communicate cues of what it can do—and certainly not guide the user to Audible. When Google Home cannot support something, it responds with, for example, "Sorry, I cannot 'send text' yet" or "Sorry, I cannot do that, I am still learning." One evaluator noted that the device should instead "tell me when it will be supported, or send a message to customer support, or notify me when it's supported."

## S2: Make the System Status Clear

The evaluators sometimes had difficulty maintaining a conversation with a device. It was easy for the evaluators to ignore the LED feedback, especially when they were not directly looking at the device. There were times that Google Home and Amazon Echo cut off and stopped listening while an evaluator was still speaking. The devices contain sounds to indicate when a conversation is starting and stopping, but these sounds are turned off by default and must be activated in the app settings—a feature even we were unaware of until one of the evaluators requested we turn on the sound. More importantly, these devices either do not offer a physical exit mechanism or it is not obvious to users, as the evaluators had to speak loudly to stop the conversation or simply wait for it to stop.

The evaluators also criticized the devices' multimodal feedback. In the case of Apple Siri, the GUI was unusable because we placed the iPad face down, yet Apple Siri still referred to the GUI quite often even when it knew the device was face down. In addition, when asked about books by Daniel Kahneman it said "OK, here are some books" without reading out the list. Likewise, when asked for the best noodle restaurants nearby, it responded, "OK, here is a list of restaurants" without saying them. Google Home and Amazon Echo both have a companion app, and when they cannot do something such as change a setting they will respond with something akin to "please change your zip code/delivery address on your app." The evaluators observed that it would be preferable if the app automatically pulled up the required screen so that the user does not have to search for it.

## S7: Make Conversation a Back-and-Forth Exchange

Just as the devices usually cannot answer follow-up questions, they do not ask if users want to learn more. As one evaluator noted, "After listing the noodle restaurants, it doesn't ask if you would like more information about those restaurants. User has to use the wake word again and start from scratch." This prevents the device from engaging in a real "conversation" and limits it to being a command-based voice response system.

The evaluators also commented that the devices do not take turns well when interacting with the user. They closed their microphone when the evaluator was still in the middle of a request and would prematurely respond. When reading a list of items, all three seemed to ignore the user's request even if it was "stop."

### S10: Use Responses to Help Users Discover What Is Possible

Similar to problems that violated heuristic S5, all three devices lack discoverability of functionality. One evaluator said of Google Home: "Let me know what is available if something like local news isn't available. I had to use my expertise to get the news." Several evaluators noted that the system did not teach ways to ask for a result—the evaluators themselves had to guess and try multiple times. It should, one evaluator said, let the user know what is possible, rather than always say something is impossible. "The inability to do something is presented as a barrier to further engagement."

### S11: Keep Feedback and Prompts Short

The evaluators noted that the devices' responses were not always clear or succinct, making it difficult for users to listen, understand, and remember. For example, when Google Home presented a list of books written by Daniel Kahneman, one evaluator said "it is hard to distinguish the title, unable to tell where one book title ended and the next title began." Also, when asked about the weather and restaurants, both Google Home and Amazon Echo responded with multiple items and kept reading them until the evaluator requested the device to stop: "As a user, I'd expect a quick overview and then be prompted if I need more details. That's not what it did." One speech expert noted that the system should not exceed listing three items, which aligns with a study showing that core verbal working-memory capacity is only three chunks.[14] This holds across list lengths and types.

## Subjective Responses

We used the SUS—a simple, 10-item Likert scale for evaluating subjective assessments of usability[12]—to evaluate the study participants' perceived usability of all three devices. The average scores of the 8 evaluators (SUS scores range from 0 to 100) were 67.2 for Google Home, 65.0 for Amazon Echo, and 49.7 for Apple Siri. These scores are consistent with the total number of problems found on each device. Apple Siri's score is significantly worse than that of Google Home and Amazon Echo: $F(2) = 121.079$, $p1 < .001$, $p2 < .001$. Almost every evaluator currently used Apple Siri or had used it in the past but still found it the most undesirable. Also, although Google Home supported the smallest percentage of the tasks, all the evaluators agreed that it had the best user experience.

## Heuristics Feedback

All of the evaluators said that the heuristics and accompanying examples helped them to evaluate the devices more thoroughly. The evaluators also provided good suggestions on how to improve the heuristics.

We initially had 20 heuristics, and it took our first evaluator, a nonspeech expert, about half an hour to read, ask questions about, and understand all of them. After this first evaluation, we decided to merge some of the heuristics to get the number down to 17 and added more explanations and examples to each one. For subsequent evaluators this made the heuristics easier to understand but also required more time to read and made them harder to memorize. In fact, 17 heuristics might still be too many. The evaluators read through all of the heuristics before undertaking each of the 10 tasks.

Most of the evaluators reported that the heuristics had a lot of overlap, sometimes making it unclear which one to use. For example, S2 and S14 both refer to multimodal feedback, in the former case to indicate system status and in the latter to deliver feedback or prompts. Also, S4 is about starting and stopping conversations and S17 is about exiting from a conversation, which is related. S17 usage was very low (0.7 percent), leading us to consider eliminating it or merging it with another heuristic. Likewise, S3, S8, and S16 all touch on language consistency. Ambiguity about the proper heuristic to use is a common complaint about Nielsen's heuristics as well. It is less important than finding the problem, but there might be a better way to structure and categorize the heuristics.

The evaluators pointed out that some of our heuristics are not applicable to today's smart devices. For example, S8—"adapt agent style to who users are, how they speak, and how they are feeling"—is too advanced for the devices in our study. Should we evaluate devices based on the ideal user experience or their current technical capability? Also, most evaluators found it hard to apply S1, the heuristic about giving the agent a persona, without some standard for what constitutes a good persona.

The speech experts had more comments on the scope of the heuristics given their experience in SUI design. For example, speech-based smart devices are starting to support multi-speaker identification, yet we did not include anything in the heuristics about this topic. Also, multimodal input/output and multi-device interaction might become more prevalent in the future. Our heuristics include some information concerning multimodal principles, but we do not touch on these problems deeply. One evaluator asked, "Are we testing one assistant on one device or one assistant across multiple devices?" Nowadays, the same assistant works on different platforms or devices—for example, Amazon Echo's Alexa is featured on mobile phones, Echo-family smart speakers, and other appliances. It is important to make sure that the user experience is consistent across platforms.

## LESSONS LEARNED

Based on the results of our evaluation, here we discuss the problems shared by speech-based smart devices as well as problems unique to each device. We also discuss the usefulness of our heuristics and how they might be improved.

### General Problems with Speech-Based Smart Devices

Even with usability and speech experts as participants, our study shows that users do not know exactly what speech-based smart devices can and cannot do. Although users have lower expectations communicating with these devices than with humans, they would like the interaction to be comparable. However, it is difficult to know a given machine's capabilities and how to adapt to its way of speaking. The evaluators in our study found 279 unique problems, and half of these were high-severity ones. Even accounting for current technical limitations, especially for natural language understanding, we believe that system designers could deliver a better user experience in at least four ways.

First, more effort should be put into error handling. Instead of constantly apologizing about what it cannot do, the device interface should guide users and help them to discover what is possible. This not only makes the user feel more confident using the system but also enables longer and richer interactions.

Second, these devices should provide more effective multimodal feedback to make the system status clearer. Users feel lost, angry, or even ignored if they do not know what is happening. All the evaluated devices lack both implicit and explicit confirmations. As indicated in heuristic S13, it is better to "use speech-recognition system confidence to drive feedback style." Designers should not assume that what users hear is correct—confirming a response shows respect for the user and can prevent errors.

Third, systems should leverage human conversational strategies, such as turn-taking and discourse markers. This will not only make interaction more natural but can also help prevent the system from cutting off a user or stopping too early. Discourse markers can also be used as a type of implicit confirmation.

Finally, designers should create a consistent persona. Computers are social actors, and voice is a social tool. As Nass and Brave point out, the key to meeting this goal is creating a consistent voice and emotional range.[9]

## Problems Particular to Each Evaluated Device

As speech-based smart devices, Google Home, Amazon Echo, and Apple Siri have almost the same set of functionalities. However, they have different focuses that make them competitive in different areas.

As search is one of its core competencies, Google Home is better at answering general questions than the other devices but has less built-in functionality. At the time we conducted the evaluation (July/August 2017), Google Home could not support basic functions such as messaging, to-do lists, and reminders.

Amazon Echo offers more than 15,000 add-on "skills," and the experience using these is different from using the device's built-in functionality. For example, users must say the exact command to open a skill, and they must first exit from a skill to do something else. Consequently, we did not include these skills in our tasks. Amazon's core business is e-commerce, which it integrates into the Echo. The evaluators complained that the many promotions for Audible and Amazon Prime were annoying.

Our evaluation found that Apple Siri had 33 percent more problems than the other two devices. Some of these issues might be due to the fact that the system is designed to be screen-based and we set it up without the screen. As such, the user experience will likely be different from that of the Siri-powered HomePod (https://www.apple.com/homepod), which was released after our study. This is being marketed as a speaker first and foremost. Unlike Google Home and Amazon Echo, HomePod is completely closed.

## The New Speech Heuristics

In developing our new heuristics to evaluate speech-based smart devices, we had three objectives. First, we wanted to provide thorough coverage of the usability problems that can occur when interacting with such devices. Second, we wanted the heuristics to be easy to understand for nonspeech experts so that they too can evaluate existing devices or design new devices with few problems. Third, we wanted designers to identify real problems using the heuristics so we could see how well they worked.

Our evaluation suggests that the heuristics generally meet these objectives. Importantly, the nonspeech experts were able to find many problems without any prior experience with SUIs or any of the devices.

Three of the nonspeech experts found more problems (60, 42, and 39) than two of the speech experts (37 and 24 problems). The other two speech experts found the most problems (86 and 73), which accounted for 55 percent of the total number of problems found. We attribute these results to the fact that the two less successful speech experts relied primarily on their own experience and less on the heuristics, while the two more successful speech experts not only used their experience but also adhered to the heuristics, which helped them find many of the problems that their counterparts did not.

The evaluators clearly understood all 17 heuristics. An inspection of the problem descriptions shows that they matched the area of coverage described by the specified heuristic. Although there is little overlap (15 percent) between the problems found by the speech experts and the nonspeech experts, we are unsure if all the problems are real or whether there are false positives lurking in the set. An end-user study is needed to identify those issues that would affect the user experience. A thorough evaluation of the heuristics will also require additional testing of more devices by more evaluators. However, we believe the results described here provide enough evidence for us to conclude that the heuristics are well suited to uncovering important usability problems in the smart device context.

Because the heuristics were based on prior research that focused on telephone- or workstation-based speech applications as well as the existing design guidelines from Google and Amazon, we believe that they can be used to evaluate most speech-only smart devices. However, whether the heuristics can also be used to evaluate speech-plus-screen devices is unclear. Multimodal I/O will be different from speech-only devices, so we think additional research is required.

There are some clear places we can improve the new heuristics. For example, we could merge heuristics related to multimodal feedback, exiting from a conversation, and language consistency. Moreover, some of the heuristics, such as S8, are more prescriptive and might be too far out for usage today.

## CONCLUSION

Our 17 heuristics for evaluating speech-based smart devices are easy for both nonspeech and speech experts to understand and to help identify real usability problems. The heuristics have good coverage of the design space and can also serve as a set of early design principles. Our evaluation of three popular devices by 8 usability specialists serves as an initial validation of the usefulness of the new heuristics, which we will continue to refine with more testing. We hope this research will help and inspire designers to create more effective and user-friendly speech-based smart devices, and inspire researchers to conduct more studies on this finally-ready-for-prime-time interaction modality.

## ACKNOWLEDGMENTS

We thank all the experts who participated in this study and gave us invaluable feedback. We also thank former Baidu Research colleagues' support for this study.

## REFERENCES

1. S. Ruan et al., "Comparing Speech and Keyboard Text Entry for Short Messages in Two Languages on Touchscreen Phones," *Proc. ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 1, no. 4, 2016; doi.org/10.1145/3161187.
2. D. Amodei et al., "Deep Speech 2: End-to-End Speech Recognition in English and Mandarin," *Proc. 33rd Int'l Conf. Machine Learning* (ICML 16), 2016, pp. 173–182.
3. B. Suhm, "Towards Best Practices for Speech User Interface Design," *Proc. 8th European Conf. Speech Communication and Technology* (Interspeech 03), 2003, pp. 2217–2220.
4. A.I. Rudnicky, "Speech Interface Guidelines," 1996; http://www.speech.cs.cmu.edu/air/papers/SpInGuidelines/SpInGuidelines.html.
5. N. Yankelovich, G.-A. Levow, and M. Marx, "Designing SpeechActs: Issues in Speech User Interfaces," *Proc. SIGCHI Conf. Human Factors in Computing Systems* (CHI 95), 1995, pp. 369–376.
6. N. Yankelovich and J. Lai, "Designing Speech User Interfaces," *CHI 98 Conf. Summary on Human Factors in Computing Systems* (CHI 98), 1998, pp. 131–132.
7. J. Nielsen, "Finding Usability Problems through Heuristic Evaluation," *Proc. SIGCHI Conf. Human Factors in Computing Systems* (CHI 92), 1992, pp. 373–380.
8. L. Dybkjær and N.O. Bernsen, "Usability Evaluation in Spoken Language Dialogue Systems," *Proc. Workshop Evaluation for Language and Dialogue Systems* (ELDS 01), 2001; doi.org/10.3115/1118053.1118055.
9. C. Nass and S. Brave, *Wired for Speech: How Voice Activates and Advances the Human-Computer Relationship*, The MIT Press, 2007.
10. C. Pearl, *Designing Voice User Interfaces: Principles of Conversational Experiences*, O'Reilly Media, 2016.
11. J. Nielsen and R. Molich, "Heuristic Evaluation of User Interfaces," *Proc. SIGCHI Conf. Human Factors in Computing Systems* (CHI 90), 1990, pp. 249–256.
12. J. Brooke, "SUS—A Quick and Dirty Usability Scale: Usability Evaluation in Industry," 1996; http://dag.idi.ntnu.no/IT3402_2009/sus_background.pdf.
13. J. Nielsen and T.K. Landauer, "A Mathematical Model of the Finding of Usability Problems," *Proc. INTERACT 93 and CHI 93 Conf. Human Factors in Computing Systems* (CHI 93), 1993, pp. 206–213.

14. Z. Chen and N. Cowan, "Core Verbal Working-Memory Capacity: The Limit in Words Retained without Covert Articulation," *Q. J. Experimental Psychology*, vol. 62, no. 7, 2009, pp. 1420–1429.

## ABOUT THE AUTHORS

**Zhuxiaona Wei (Nina)** is a product manager at deeplearning.ai and was formerly a product designer at Baidu Research, where this work was conducted. Her research focuses are SUI/CUI design and AI-powered products. Contact her at weizhuxiaona@gmail.com.

**James A. Landay** is a professor of computer science and the Anand Rajaraman and Venky Harinarayan Professor in the School of Engineering at Stanford University, specializing in human–computer interaction. This work was part of his consulting work with Baidu Research. Contact him at landay@stanford.edu.

*This article originally appeared in*
IEEE Pervasive Computing, *vol. 17, no. 2, 2018.*


Keep up with the latest IEEE Computer Society publications and activities wherever you are.

## Follow us:

@ComputerSociety

facebook.com/IEEEComputerSociety

IEEE Computer Society

youtube.com/ieeecomputersociety

instagram.com/ieee_computer_society

IEEE COMPUTER SOCIETY

IEEE COMPUTER SOCIETY

Association for Computing Machinery

# ACM/IEEE CS Eckert-Mauchly Award

## *Call for Award Nominations Deadline: 30 March 2019*

ACM and the IEEE Computer Society co-sponsor the **Eckert-Mauchly Award**, known as the computer architecture community's most prestigious award.

Initiated in 1979, the award recognizes outstanding contributions to computer and digital systems architecture. It comes with a certificate and a $5,000 prize.

The award was named for John Presper Eckert and John William Mauchly, who collaborated on the design and construction of the Electronic Numerical Integrator and Computer (ENIAC), the first large-scale electronic computing machine, which was completed in 1947.

### Nomination Guidelines

- Open to all. Anyone may nominate.
- Self-nominations are not accepted.
- This award requires 3 endorsements.
- Submit your nomination by **30 March 2019** to **bit.ly/eckert-mauchly.**

### Award Presentation

- **ISCA 2019** (ACM/IEEE International Symposium on Computer Architecture)
- Phoenix, Arizona, USA
- June 22–26, 2019

### Questions?

Write to IEEE Computer Society Awards Administrator at **awards@computer.org**.

IEEE

### 2018 Eckert-Mauchly Award Recipient

**Susan Eggers**
University of Washington

*First female award recipient, recognized for outstanding contributions to simultaneous multithreaded processor architectures and multiprocessor sharing and coherency.*