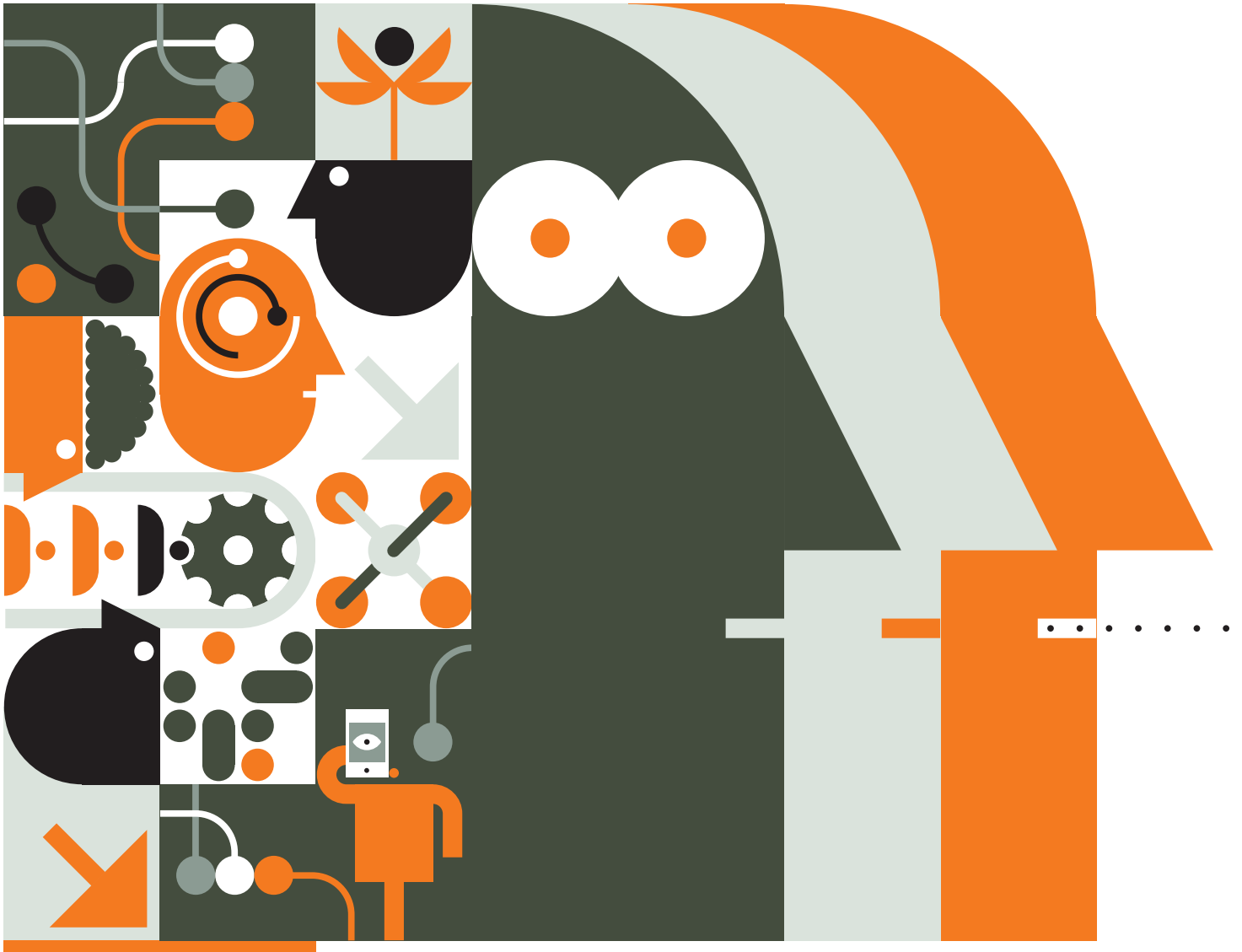
COMMUNICATIONS OF THE ACM

CACM.ACM.ORG

03/2025 VOL.68 NO.03



# Large Language Model Use in Crowd Work

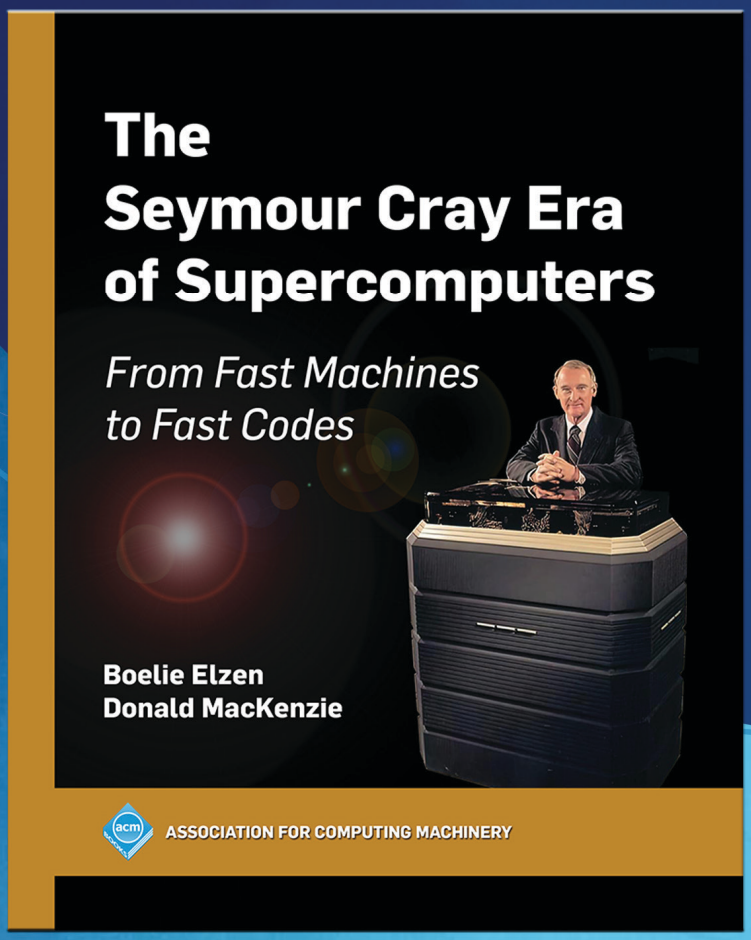
The AI Alignment Paradox

AI as Catalyst for Biodiversity Understanding

The Sustainability Gap for Computing: Quo Vadis?

Association for Computing Machinery

# NEW BOOK RELEASE




**ACM BOOKS**
**Collection III**

## The Seymour Cray Era of Supercomputers

*From Fast Machines to Fast Codes*

**Boelie Elzen**
**Donald MacKenzie**

ISBN: 979-8-4007-1369-9
DOI: 10.1145/3705551

This book describes the development and use of supercomputers in the period 1960-1996, a time that can be called the Seymour Cray Era. For more than three decades, Cray's computer designs were seen as the yardstick against which all other efforts were measured.

**Important reading for anyone working in the area of high-performance computing, providing essential historical context for the work of a legendary pioneer and the computers he became famous for designing. It will also be valuable to students of computing history and, more generally, to readers interested in the history of science and technology. For advanced students, the book illustrates how innovation in its very essence is a socio-technical process: not just a matter of developing the "best technology," but also of making appropriate choices concerning the interaction of human and technical factors in product design.**

**http://books.acm.org**

# Call for Nominations
# Editor-in-Chief ACM Books

The ACM Books program, which publishes academic and practitioner books for the computing community, is noted for its attention to quality and its strong support of authors throughout the publication process. The ACM Publications Board relies on the Books Editor-in-Chief to ensure that the exceptional quality standards of the series are maintained, that the editorial process is both timely and fair, and that the pipeline of new books is sufficient to support the future publication schedule. The EiC works closely with the ACM Books Editorial Board and with an in-house staff that manages the editorial, production, and marketing activities associated with the program.

Nominations are invited for a three-year term as ACM Books Editor-in-Chief, beginning on May 1, 2025 (renewable for a further three-year period, subject to the approval of the Publications Board). This is a voluntary position, but ACM will cover appropriate expenses and provide any necessary administrative support. ACM may choose to recruit two Co-EiCs to work together on the management of the program.

Nominations should include a brief explanation of why the nominee should be considered and a short statement on the candidate's vision for the future development of ACM Books. Self-nominations are welcome.

**The ACM Publications Board has established a nominating committee to assist in selecting the next EiC.**

**The committee members are:**

**Chris Hankin,** *Imperial College* (Chair)
**Diane Cook,** *Washington State University*
**John Grundy,** *Monash University*
**Meena Mahajan,** *Institute of Mathematical Sciences*
**Helena Mentis,** *University of Maryland*
**Tobias Nipkow,** *TU Munich*
**Adelinde Uhrmacher,** *University of Rostock* (Publications Board Liaison)

**Please send all nominations to**

**Chris Hankin**
**c.hankin@imperial.ac.uk**

**Sean Pidgeon**
**pidgeon@hq.acm.org**

**ACM BOOKS** **http://books.acm.org**

# COMMUNICATIONS OF THE ACM

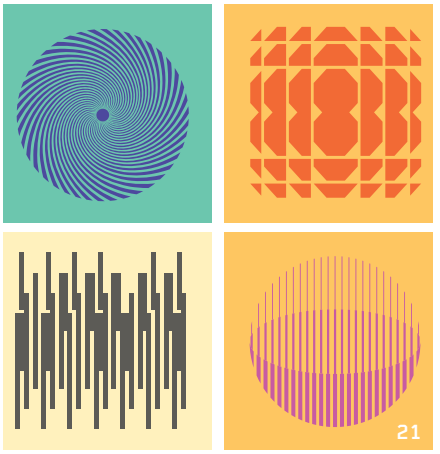
## News



## Last Byte

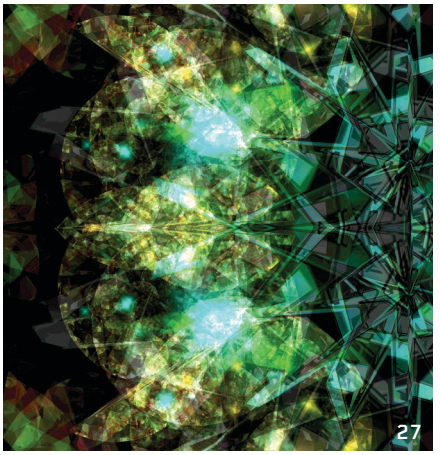


## Opinion






## Opinion



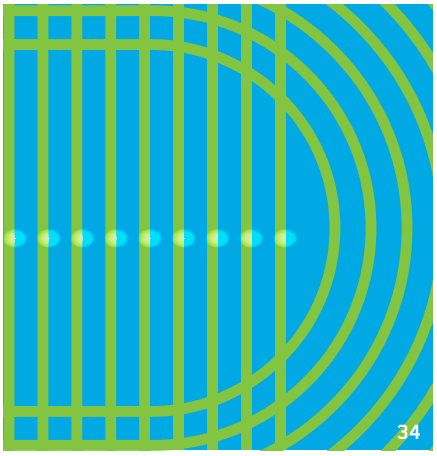




Association for Computing Machinery
*Advancing Computing as a Science & Profession*

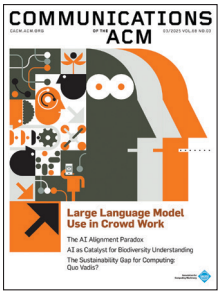

IMAGES L-TO-R BY: PLASTEED; GRADIENT BACKGROUND

## Practice





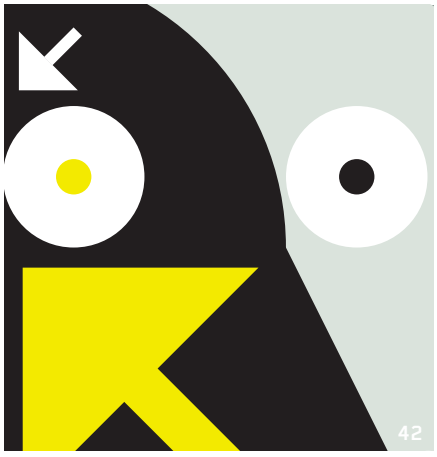

Articles' development led by acmqueue
queue.acm.org




**About the Cover:**
As LLMs become easier to incorporate into people's daily tasks, crowd workers are readily adopting them, sometimes adversely affecting their work. But is this something we'd really want to prevent? And if so, how might we go about doing so? Cover Illustration by Peter Grundy.

IMAGES L-TO-R BY: ANDRIJ BORYS ASSOCIATES. USING SHUTTERSTOCK.COM. PETER GRUNDY

## Research and Advances






Watch the authors discuss this work in the exclusive *Communications* video.
https://cacm.acm.org/videos/llms-in-crowd-work



Watch the authors discuss this work in the exclusive *Communications* video.
https://cacm.acm.org/videos/off-path-attacks

## Research and Advances



## Research Highlights

# COMMUNICATIONS OF THE ACM

Trusted insights for computing's leading professionals.

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

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





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





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

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

## ACM ADVERTISING DEPARTMENT

1601 Broadway, 10th Floor
New York, NY 10019-7434 USA
T (212) 626-0686
F (212) 869-0481

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

**Media Kit** acmmediasales@acm.org

## COMMUNICATIONS OF THE ACM

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

## POSTMASTER

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

Printed in the USA.

Association for Computing Machinery

DOI:10.1145/3714998

Moshe Y. Vardi

# Homo Ratiocinator (Reckoning Human)

HOMO SAPIENS, "WISE HUMAN" in Latin, is the taxonomic species name for modern humans. But observing the current state of the world and its trajectory, it is hard for me to accept the description "wise." I am not the first to object to the "sapiens" descriptor. French philosopher Henri-Louis Bergson argued in 1911 that a better term would be Homo Faber, referring to human tool-making ability. This ability goes back to early humans, about three million years ago. Most importantly, human tools improved due to innovation and cultural transmission. I would like to offer an alternative: Homo Rationcinator,[a,b] or *reckoning human*—where reckoning refers to both reasoning and computing.

In 2018, scientists reported the discovery of 50,000-year-old cave art—depicting a wild pig and a trio of human figures—in the Indonesian island Borneo. This is the first example we have of symbolic representation by humans. Eventually, tool making and symbolic representation led to counting tools. The Lebombo Bone is a bone tool made of a baboon fibula with incised markings, discovered in a cave in the Lebombo Mountains in Africa. More than 40,000 years old, the bone is conjectured to be a tally stick, its 29 notches counting, perhaps, the days of the lunar phase. Humans had developed a tool for computing. Our destiny had been laid out, and computing tools continued to improve.

Around 2,500 B.C., the Sumerians invented the abacus, a manual tool for arithmetical computing. (A pebble of an abacus is called a calculus, in Latin.) In the 1670s, Gottfried Wilhelm Leibniz developed his Step Reckoner, a machine capable of addition and multiplication. In 1820, Charles Babbage developed his difference engine. His analytical engine was a proposed digital mechanical general-purpose computer. While it was never actually built, the proposal gave birth to the idea of a general-purpose programmable computer. Ada Lovelace, Babbage's collaborator, believed that computers could become much more than calculators, including composing "elaborate and scientific pieces of music of any degree of complexity or extent."

As I have argued[c] before, the development of computing and mathematics dovetailed each other. Deductive mathematics was developed by the Greeks in the 7th century B.C. A few hundred years later, Aristotle formalized the rules of reasoning in deductive mathematics, and logic was born. In the 13th century, the Catalan monk Ramon Lull, wishing to use logic to convert the entire world to Christianity, invented the so-called Lull's Circles, the first mechanical aid to reasoning. In the 17th century, British philosopher Thomas Hobbes argued that reasoning is a form of computing. Leibniz, inspired by Lull, dreamed of *Calculus Ratiocinator*, a reasoning machine that could augment human intelligence.

In the 19th century, George Boole developed an algebraic treatment of logic, giving us Boolean logic, and William Stanley Jeavons showed how to build Boolean logic machines. Claude Shannon showed in the 20th century how to use Boolean logic for electrical-circuit design. The stage was set for the development of the programmable, electronic, digital computer around the middle of the 20th century. By the early 1950s, dozens of "Johniacs"—computers named after John von Neumann—were built around the world. Leibniz's dream came true. We went from reasoning, to patterns of reasoning, to logic, to computers, to computers that reason. So, reasoning human, reckoning human, and tool-making human are now making tools that can compute *and* reason, opening the door to intelligent machines, or artificial intelligence (AI).

Asimov's *I, Robot* was published as a book in 1950. The overarching theme is the complicated interaction of humans, robots, and morality. Norbert Wiener's *The Human Use of Human Beings: Cybernetics and Society*, also published in 1950, warned us that "The machine's danger to society is not from the machine itself but from what man makes of it." But we plunged ahead with AI research, paying little attention to societal impact.

Leibniz's goal for his calculus ratiocinator was "mankind will then possess a new instrument that will enhance the capabilities of the mind to a far greater extent than optical instruments strengthen the eyes." Ada Lovelace wished computing technology to be "for the most effective use of mankind." But Silicon Valley, today's leading engine driving computing technology, is motivated solely by profit maximization, the common good be damned. After 40,000 years of making tools for computing and reasoning, it is time, I believe, for Homo Ratiocinator to live up to its traditional name, Homo Sapiens—building machines to augment, not replace, human intelligence. C

**Moshe Y. Vardi** (vardi@rice.edu) is university professor and the George Distinguished Service Professor at Rice University, Houston, TX, USA, where he is also a Fellow at the Baker Institute for Public Policy. He is a former Editor-in-Chief of *Communications*.



a https://tinyurl.com/2c3ozbh7
b https://tinyurl.com/2cpojgl7
c https://tinyurl.com/2b6mxxsp

BLOG@CACM

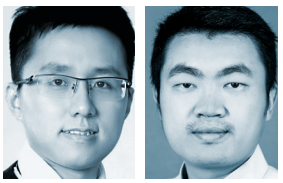
In each issue of *Communications*, we publish selected posts or excerpts from the many blogs on our website. The views expressed by bloggers are their own and not necessarily held by *Communications* or the Association for Computing Machinery.

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



# Putting the Smarts Into Robot Bodies

*Fan Wang and Shaoshan Liu offer guidance for the development of embodied AI systems.*



**FAN WANG AND SHAOSHAN LIU**

**Building Foundation Models for Embodied Artificial Intelligence**

**DOI:10.1145/3703761**
https://bit.ly/3Wn2FY5
July 15, 2024

Embodied artificial intelligence (EAI) involves embedding artificial intelligence (AI) into tangible entities, such as robots, and equipping them with the capacity to perceive, learn from, and engage dynamically with their surroundings. In this article, we delve into the key trade-offs of building foundation models for EAI systems.

## Foundation Models for Embodied AI

Previously, we have outlined three guiding principles for developing EAI systems.[1] EAI systems should not depend on predefined, complex logic to handle specific scenarios. Instead, they must incorporate evolutionary learning mechanisms, enabling continuous adaptation to their operational environments. Additionally, the environment significantly influences not only physical behaviors but also cognitive structures. While the third principle focuses on simulation, the first two principles emphasize building EAI foundation models capable of learning from EAI systems' operating environments.

A common approach for EAI foundation models is to directly use pretrained large models. For example, pretrained GPT models can serve as a baseline, followed by fine-tuning and in-context learning (ICL) to enhance performance.[9] These large models typically possess a substantial number of parameters to encode extensive world knowledge and feature a small context window for fast response times. This extensive pre-encoding allows these models to deliver excellent zero-shot performance. However, their limited context windows pose challenges for continuous learning from the EAI systems' operating environments and connecting various usage scenarios.

Alternatively, another approach leverages models with significantly fewer parameters but a larger context window. These models, rather than encoding comprehensive world knowledge, focus on learning how to learn, or meta-learning.[2] With large context windows, these models can perform general-purpose in-context learning (GPICL), enabling continuous learning from their operating environments and establishing connections across a broad context.

The Figure below illustrates these two different approaches. The meta-training + GPICL approach, while exhibiting poorer zero-shot performance and having a smaller model size, excels in continuously learning from its environment, eventually specializing EAI systems for specific tasks. In contrast, the pretraining + fine-tuning + ICL approach, characterized by a larger model size and smaller context windows, offers superior zero-shot performance but inferior learning capabilities.

Empirical evidence supporting this is found in the GPT-3 paper, where a 7B few-shot model outperforms a 175B zero-shot model.[3] If few-shot learning is replaced by a long context window enabling EAI systems to learn from their operating environments, performance may further improve.

We envision an ideal foundation model for EAI that should meet several critical criteria. Firstly, it should be capable of universally learning from complex instructions, demonstrations, and

feedback without relying on crafted optimization techniques. Secondly, it should demonstrate high sample efficiency in its learning and adaptation processes. Thirdly, it must possess the ability to continuously learn through contextual information, effectively avoiding the issue of catastrophic forgetting. Therefore, we conclude that the meta-learning + GPICL approach is suitable for EAI systems. However, before we decide on taking this approach, let us first examine the trade-offs between these two approaches.

## Key Trade-Offs

In this section, we review the trade-offs between pretrained large models vs. meta-training + GPICL as foundation models for EAI.[4] The results are summarized in the table.

For zero-shot capability, the Pretraining + Fine-Tuning + ICL approach[9] offers high performance, allowing models to generalize well to new tasks without any task-specific fine-tuning. In contrast, the Meta-Training + GPICL approach exhibits low zero-shot capability, as it focuses on learning to adapt to a wide variety of tasks using in-context learning rather than zero-shot generalization.

In terms of generalizability, the Pretraining + Fine-Tuning + ICL approach performs well on in-distribution tasks but has rudimentary capabilities for out-of-distribution tasks. Meta-Training + GPICL, on the other hand, exhibits diverse and complex generalization capabilities for out-of-distribution tasks due to its emphasis on meta-training over varied contexts.

The scalability enhancement approach for Pretraining + Fine-Tuning + ICL involves scaling up parameters and pre-training datasets to improve performance. Meta-Training + GPICL enhances scalability by scaling up meta-training tasks, context length, memories, and hidden states to improve the model's adaptability.

Regarding task adaptation, Pretraining + Fine-Tuning + ICL relies on data collection and fine-tuning, which can be inefficient. In contrast, Meta-Training + GPICL utilizes very complex instructions and learns from diverse contexts automatically.

During the pre-training or meta-training stage, Pretraining + Fine-Tuning + ICL focuses on world knowledge and understanding the hardware. Meta-Training + GPICL emphasizes the capability of learning, memorization, and abstraction over a wide variety of tasks.

In the post-training stage, Pretraining + Fine-Tuning + ICL involves aligning the model to specific human-centric tasks, emphasizing human-alignment and task-specific knowledge. Meta-Training + GPICL continues to emphasize world knowledge, human-alignment, and task-specific knowledge.

Inference latency is generally low for Pretraining + Fine-Tuning + ICL as the model parameters are fixed after training. However, for Meta-Training + GPICL, inference can be slower due to the need to utilize and update memory and hidden states dynamically.

Memory size requirements for Pretraining + Fine-Tuning + ICL are small, as most knowledge is embedded in fixed model parameters. Conversely, Meta-Training + GPICL requires significant memory to handle complex instructions, extended context, and hidden states.

Meta-Training + GPICL offers the advantage of enabling the system to continuously learn various tasks through contexts, that is, learning to continuously learn.[7] This essentially requires the system to be able to learn new tasks without forgetting the old ones, which typically poses great challenge for gradient-based fine-tuning (catastrophic forgetting[8]) but can be less of a challenge with in-context learning.

## Overcoming the Computing and Memory Bottlenecks

From the above comparison, it is evident that meta-training combined with GPICL offers superior adaptability and generalization across diverse and complex tasks. However, this approach

**Figure. Foundation model options for EAI. Credit: Fan Wang.**

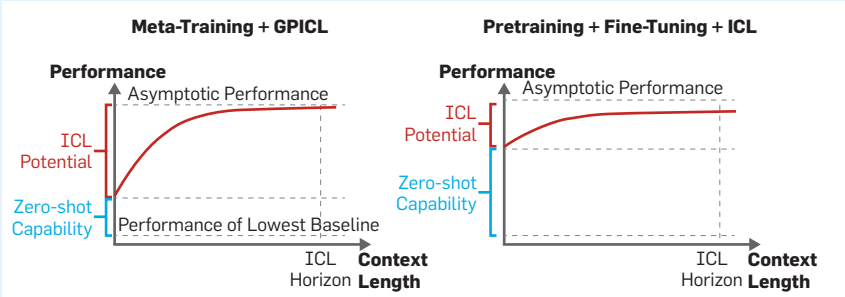



**Table. Trade-offs of Pretrained large model vs. meta-training + GPICL. Credit: Fan Wang.**

| Comparison | Pretraining + Fine-Tuning + ICL | Meta-Training + GPICL |
|---|---|---|
| Zero-Shot Capability | High | Low |
| Generalizability | In-Distribution Tasks | Diverse and Complex |
| | Rudimentary Out-of-Distribution Tasks | Out-of-Distribution Tasks |
| Knowledge carrier | Parameters | Memory/Hidden States |
| Scalability Enhancement Approach | Scaling up parameters and pre-training datasets | Scaling up meta-training tasks, context length, memories, and hidden states |
| Methodology of Task Adaptation | Data Collection (Fine-Tuning, Inefficient) | Very Complex Instruction |
| | Rudimentary Instruction and Prompt (ICL) | Explore and Exploit automatically |
| Emphasis of pre-training/ meta-training stage | World knowledge, knowledge regarding the hardware | The capability of learning, memorization, and abstraction |
| Emphasis of post-training stage | Human-alignment, task-specific knowledge | World knowledge, human-alignment, task-specific knowledge |
| Inference Latency | Low | High |
| Memory Size | Small | Large |

demands higher resources, posing a challenge for most EAI systems, which are often real-time edge devices with limited computational capabilities and memory. The large context windows required for this approach can significantly increase inference time and memory footprint, potentially hindering its feasibility for EAI foundation models.

Fortunately, recent advancements have introduced innovative solutions to scale transformer-based LLMs for processing infinitely long inputs while maintaining bounded memory and computational efficiency. A notable innovation is the Infini-attention mechanism, which integrates masked local attention and long-term linear attention within a single transformer block. This enables the efficient processing of both short- and long-range contextual dependencies. Additionally, the compressive memory system allows the model to maintain and retrieve information with bounded storage and computation costs, reusing old key-value (KV) states to enhance memory efficiency and enable fast streaming inference. Experimental results demonstrate that the Infini-attention model outperforms baseline models in long-context language-modeling benchmarks, showing superior performance in tasks involving extremely long input sequences (up to one million tokens) and significant improvements in memory efficiency and perplexity scores.

Similarly, the StreamingLLM framework enables large models trained with a finite attention window to generalize to infinite sequence lengths without the need for fine-tuning. This is achieved by preserving the key and value (KV) states of initial tokens as attention sinks, along with the most recent tokens, stabilizing attention computation and maintaining performance over extended texts. StreamingLLM excels at modeling texts up to 4 million tokens, providing a remarkable speed-up of up to 22.2 times.

## Conclusion

We believe that learning from the environment is the essential feature for EAI systems and thus the meta-training + GPICL approach is promising for building EAI foundation models due to its capabilities of providing better long-term adaptability and generalization. Although currently this approach is facing significant challenges in computing and memory usage, we believe that innovations such as Infini-attention and StreamingLLM will soon make this approach viable for real-time, resource-constrained environments.

**Experimental results demonstrate that the Infini-attention model outperforms baseline models in long-context language-modeling benchmarks.**

**References**

1. A brief history of embodied artificial intelligence, and its outlook. *Commun. ACM*; https://cacm.acm.org/blogcacm/a-brief-history-of-embodied-artificial-intelligence-and-its-future-outlook/
2. Kirsch, L., Harrison, J., Sohl-Dickstein, J., and Metz, L. General-purpose in-context learning by meta-learning transformers. *arXiv preprint arXiv:2212.04458*, (2022).
3. Brown, T. et al. Language models are few-shot learners. *Advances in Neural Information Processing Systems 33*, (2020), 1877–1901.
4. Wang, F., Lin, C., Cao, Y., and Kang, Y. Benchmarking general purpose in-context learning. *arXiv preprint arXiv:2405.17234*, (2024).
5. Munkhdalai, T., Faruqui, M., and Gopal, S. Leave no context behind: Efficient infinite context transformers with infini-attention. *arXiv preprint arXiv:2404.07143*, (2024).
6. Xiao, G. et al. Efficient streaming language models with attention sinks. *arXiv preprint arXiv:2309.17453*, (2023).
7. Beaulieu, S. et al. Learning to continually learn. *ECAI 2020*. IOS Press, 992–1001.
8. French, R.M. Catastrophic forgetting in connectionist networks. *Trends in Cognitive Sciences 3*, 4 (1999), 128–135.
9. Ouyang, L. et al. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems 35* (2022), 27730–27744.

**Fan Wang** is a distinguished architect at Baidu working on AI systems. He holds a Master of Science degree from the engineering school at the University of Colorado at Boulder, and a Bachelor of Science degree from the University of Science and Technology of China. Fan specializes in reinforcement learning, natural language processing, AI for sciences, and robotics.

**Shaoshan Liu** is a member of the ACM U.S. Technology Policy Committee, and a member of the U.S. National Academy of Public Administration's Technology Leadership Panel Advisory Group. His educational background includes a Ph.D. in computer engineering from the University of California Irvine, and a master's degree in public administration from Harvard Kennedy School.

Science | DOI:10.1145/3703778 Chris Edwards

# Feedback Loops Guide AI to Proof Checking

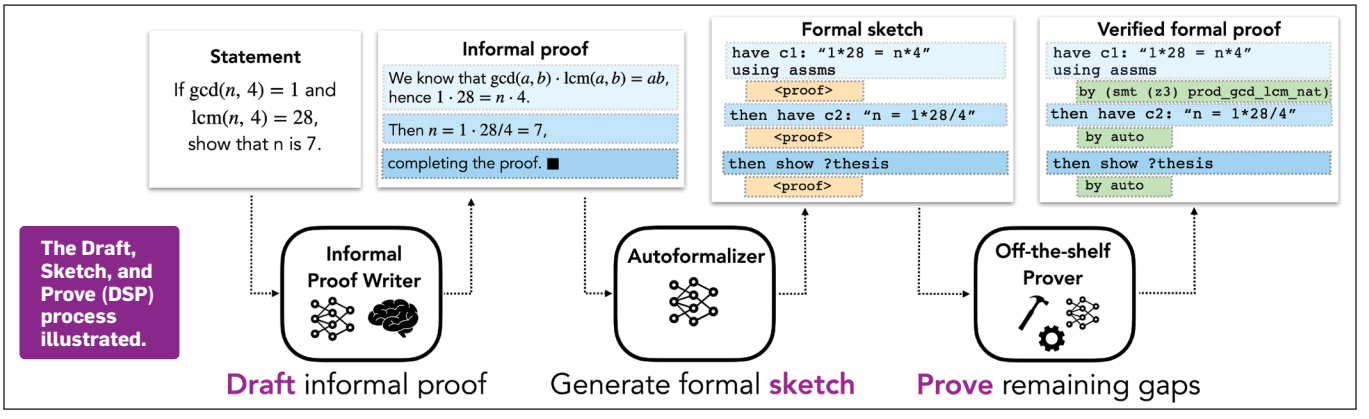
*After decades of promise, techniques and technologies are coming together to make AI better at checking mathematicians' work.*

Some of the earliest work on artificial intelligence (AI) saw mathematics as a major target and key to making breakthroughs quickly. In 1961, leading computer scientist and AI pioneer John McCarthy argued at the Fifth Symposium in Pure Mathematics that the job of checking mathematical proofs would likely be "one of the most interesting and useful applications of automatic computers."

McCarthy saw the possibility for mathematicians to try out different ideas for proofs quickly that the computers then tested for correctness. More than 60 years later, such a proof assistant has yet to appear. But recent developments in both mathematics and computer science may see a breakthrough sooner rather than later.

Much of the work of verifying proofs formally using a computer continues to rely on a lot of manual effort by specialists such as Kevin Buzzard, professor of pure mathematics at the U.K.'s Imperial College London. Last year, Buzzard kicked off a project, funded for five years by the U.K.'s Engineering and Physical Sciences Research Council (EPSRC), to formalize the proof of Fermat's Last Theorem developed by Andrew Wiles 30 years ago. Buzzard estimates it will take some 100 person-years of work to complete the process. Much of the help is coming from a community of volunteers who have, in recent years, shown how well crowdsourcing can work in this area. That, in turn, may provide an easier route for AI to finally make an entrance.

A key characteristic of projects like Buzzard's is that the work readily separates into modules with clearly defined interfaces. This attribute helped one of the first major crowdsourced proof verification projects to be completed in just a couple of years. The Liquid Tensor Experiment (LTE) took shape in 2020, when Peter Scholze, professor of mathematics at Germany's University of Bonn, asked the community for help in checking




**The Draft, Sketch, and Prove (DSP) process illustrated.**

IMAGE FROM 'DRAFT, SKETCH, AND PROVE: GUIDING FORMAL THEOREM PROVERS WITH INFORMAL PROOFS' (ARXIV:2210.12283)

**A 2024 review found advanced LLMs at best could completely formalize just a quarter of high-school and undergraduate-level problems.**

the 600-page proof he and Dustin Clausen had painstakingly threaded together by hand. Scholze wrote in his appeal how he had lost confidence in his ability to comprehend all the subtleties of the proof because of its sheer size and intricacy.

Languages such as Lean, used in both the LTE and Fermat projects, use keywords like "sorry" to mark unfinished components. This makes it possible to sketch out a skeleton of the overall proof that team members fill in gradually, until they are ready to have the proof engine used by these languages check the result and mark that section as complete. When working on LTE, Johan Commelin, assistant professor at the Netherlands' Utrecht University, said he would wake up in the morning and find new parts of the proof had appeared overnight.

Researchers see AI benefiting not just from the same approach. Instead of expecting the software to work on complete proofs, tools could work on much smaller and more manageable pieces. The current generation of AI also benefits from the data that has spun off from the crowdsourcing efforts that can be used to train the models. Lean now holds the equivalent of an undergraduate course in mathematics and is rapidly catching up to the size of the Mathematics Components library developed over a longer period for the older language Coq.

In principle, the large language model (LLM) makes a good choice for pulling together the elements of a proof. However, the technology has problems, as Michail Karatarakis reported to colleagues at the Conference on AI and Theorem Proving (AITP) in Aussois, France in 2023. The Radboud University Ph.D. student used the Mistral LLM to generate sketches of proofs that, with some editing, could be checked by the Lean proof engine.

For the test, Karatarakis used two lemmas from a textbook on number theory, one picked because the components needed were already in the library. The other included a definition not yet in the library, "offering a way to see how Mistral handles unfamiliar definitions," he explained.

Despite having a large body of existing material to draw upon and only being required to produce "sketches," or small components of a larger proof, the output needed many corrections, particularly to the syntax. As well as other issues with the output, the LLM's training set seemed to include many examples from older versions of the Lean language. Yet Karatarakis was using the latest version of the language, which has important differences.

That LLMs can struggle to build even proof sketches is not unusual. A 2024 review of activity in automated theorem proving by an international group led by Xujie Si, assistant professor of computer science at Canada's University of Toronto, found advanced LLMs at best could completely formalize just a quarter of high-school and undergraduate-level problems.

One key problem LLMs face with full proofs is the lack of reasoning these tools possess.

"Since the task in our case was to provide proof sketches rather than full proofs, syntax issues had a larger impact than reasoning," Karatarakis said. "For tasks involving complete proofs, reasoning remains the primary challenge, even with improved syntax handling."

Injecting more feedback into the training process could address some of the issues faced by LLMs. This intuition drove the creation of Draft, Sketch, and Prove (DSP) by researchers working at the U.K. universities of Cambridge and Edinburgh. This tool uses the LLM to create an initial sketch of an idea that goes to an automated theorem prover that can work at a more informal level. Several of these have been developed over the past couple of decades to assist mathematicians working in languages such as Coq and Isabelle. The last part of DSP is a separate formal engine that checks the work. The trio of engines form a loop where the LLM retrains on the proofs that the symbolic engine accepts.

Said Wenda Li, lecturer in hybrid AI at the U.K.'s University of Edinburgh, "Sometimes the gap is too large for the automated theorem prover to bridge. But generating new drafts is relatively cheap. We can sample millions of them to get just the right gap for the prover to bridge."

Accuracy improved to over 50% with DSP's combination of feedback and division of responsibilities. In a further step presented last summer, the team added a fourth module in a version called Sketch, Prove, Add Details & Repeat (SPADeR). The extra module called on GPT-4o to fill in blanks that would otherwise block a full proof. This increased the number of successfully verified problems in one test set to 93 from 85 using the earlier DSP tool.

For practical mathematical work, AI's output may still need to be refined after completing a proof successfully. One common thread in the manual formalization efforts is the importance of finding good definitions and proof steps that support the process. At one stage of the LTE project, it looked as though just a few lines in one key part of the proof would mean formalizing an immense body of knowledge as a prerequisite. Commelin led work to avoid the problem by building

**"For tasks involving complete proofs, reasoning remains the primary challenge, even with improved syntax handling."**

a new underpinning that was far easier to implement.

In Li's work with colleagues on a formalized version of another textbook on number theory, the group found that a formal definition of a key type of function of complex numbers would be better expressed by the use of an arithmetic expansion, rather than the more-intuitive approach used in the informal source. That approach would help underpin a larger set of dependent proofs, he explained. In some projects, these considerations have led to changes to underlying definitions, sometimes repeatedly.

"Over time, it becomes increasingly difficult to refactor definitions or lemmas due to growing dependencies," Li said. "AI could potentially assist with this, much like other code-assistance tools, such as Copilot."

Other goals may provide targets that AI can serve more easily than auto formalization itself. Patrick Massot, professor of mathematics at France's University of Paris-Saclay, argues one lingering issue with the current formal languages like Coq and Lean is they are too opaque to non-computer scientists. An "informalizer" would help scholars read the verified proofs. It could, as a byproduct, provide the foundation for building interactive math textbooks, in which students are able to drill down into the background of any proof or lemma they see.

A couple of teams have used LLMs to try to do the work. Though LLMs make fewer mistakes here than in formalization, the task demands much higher accuracy than they can deliver. In his work on informalization with Massot, Kyle Miller, assistant professor at the University of California at Santa Cruz, has been exploring the use of more traditional symbolic AI techniques. This involves far more manual engineering than training an LLM. Simply mapping the grammar of a language like Lean into English is not enough; it needs more changes. For example, the code created to check a proof formally can contain a lot of repetition that the tool would ideally remove from the human-readable version.

If successful, the work on informalizers in turn may help close the loop for theorem-proving engines based on LLMs and similar technologies. The output from these tools would provide a rich resource of synthetic data that can be used to retrain the AI engines as they create new proofs.

**Patrick Massot argues one lingering issue with current formal languages like Coq and Lean is that they are too opaque to non-scientists.**

The multiple feedback loops that are now appearing may mean the logjam that has held up one of computer science's major applications is finally breaking. But it may take a lot more research into hybrid schemes to strengthen AI's reasoning skills and, perhaps, finally make autoformalization work for mathematicians.

### Further Reading


*Avigad, J.*
Mathematics and the Formal Turn; *Bulletin of the American Mathematical Society*. 61 (2024), 225-240

*Li, Zhaoyu, Sun, J., Murphy, L., Su, Q., Li, Zenan, Zhang, X., Yang, K., and Si Xujie*
A Survey on Deep Learning for Theorem Proving; arXiv:2404.09939 (2024)

*Karatarakis, M.*
Leveraging Large Language Models for Autoformalizing Theorems: A Case Study; Ninth Conference on Artificial Intelligence and Theorem Proving (AITP 2024)

*Tarrach, G., Jiang, A.Q., Raggi, D., Li, W., and Jamnik, M.*
More Details, Please: Improving Autoformalization with More Detailed Proofs; AI for MATH Workshop at the International Conference on Machine Learning (ICML 2024)

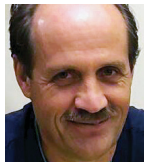
*McCarthy, M.*
Computer Programs for Checking Mathematical Proofs; *Proceedings of the Fifth Symposium in Pure Mathematics of the American Mathematical Society*, pages 219–227 (1961)


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



# ACM Member News

## COMPUTING AT THE EDGE



**Daniel Grosu is a professor of computer science in the College of Engineering at Wayne State University in Detroit, MI.**

**Grosu earned his undergraduate degree in automatic control from Romania's Technical University of Iasi. He went on to obtain both his master's and doctoral degrees in computer science from the University of Texas at San Antonio.**

**After receiving his Ph.D. in 2003, Grosu joined the faculty at Wayne State University, where he has remained.**

**Grosu's research centers on cloud and edge computing, parallel and distributed algorithms, graph algorithms, approximation algorithms, scheduling and load balancing, and topics at the border of computer science, game theory, and economics.**

**"I am currently focused on designing resource- and task-allocation algorithms for edge and cloud computing," Grosu said. He added that these algorithms take into account the mobility of users, the capabilities of available resources, the data that is needed, and then make the best decisions to allocate these tasks and resources.**

**Grosu notes monetizing and pricing these resources in edge computing is an issue he is endeavoring to solve. "Basically, you have to design pricing mechanisms to accurately price the resources in edge computing so that a provider earns a profit, but at the same time the users get a fair price as well," he explained.**

**Grosu believes there is still a lot of work to be done in edge computing as deployments increase, especially with the emergence of autonomous vehicles.**

**"They'll need a lot of infrastructure support from the edge," Grosu concluded.**

***—John Delaney***

Society | DOI:10.1145/3703779 Mark Halper

# How Software Bugs Led to 'One of the Greatest Miscarriages of Justice' in British History

*Fujitsu's Horizon point-of-sale accounting software had trouble with arithmetic due to flaws dating back to its development. Innocent branch managers paid a huge price.*

In a Kafkaesque nightmare come true, nearly 1,000 individuals who ran local post offices in the U.K. were wrongly convicted of stealing money from those operations between 1999 and 2015 as a Fujitsu software system known as Horizon erroneously showed imbalances in their accounts.

The convictions resulted in prison for some of the managers and financial ruin for many held responsible for the missing funds. Those who were not prosecuted were typically fired, resulting in wrecked lives, including four suicides.

Public awareness of what is commonly called both the Post Office scandal and the Horizon Post Office scandal has long percolated throughout Britain but came into sharp focus in early January 2024, when television network ITV aired a prize-winning drama titled "Mr. Bates vs. The Post Office." The series portrayed the distress, hardships, and abject disbelief experienced by sub-postmasters (British parlance for local post office managers, also known as sub-postmistresses) as the central Post Office bosses over the years refused to acknowledge any faults with Horizon, and insisted local managers pay up.

The ITV series took its name from Alan Bates, a dismissed sub-postmaster from Wales who painstakingly led 555 sub-postmasters to a 2019 civil suit victory against the Post Office in London's High Court. The court awarded £58 million to the sub-postmasters, much of which went to their legal fees.

The case might not have been the monetary win that the sub-postmasters had wanted, but it was a huge moral victory. It served as an indictment not only of the Post Office and of Fujitsu, but of Horizon itself. In his judgment, Justice Peter Fraser noted that "bugs, errors, or defects" undermined Horizon's reliability and caused discrepancies or shortfalls at branches "on numerous occasions." The version that the Post Office used from 2000 through 2010—known as "Legacy Horizon"—"was not remotely robust," he observed.

Horizon's Electronic Point of Sale [EPOS] system, whose "bugs, errors, or defects" had "lasting financial impact" on nearly 1,000 individuals who ran local post offices in the U.K.

**Alan Bates was a dismissed sub-postmaster from Wales who painstakingly led 555 sub-postmasters to a 2019 civil suit victory against the U.K. Post Office in London's High Court.**

"Legacy" processed information locally and uploaded it; the later "Online" version, still in use, uploads information for central processing.

With public outrage swelling following the airing of the TV drama, and with former U.K. prime minister Rishi Sunak last March describing the convictions as "one of the greatest miscarriages of justice in our nation's history," the government in May 2024 dismissed all convictions in England, Wales, and Northern Ireland; it did the same for Scotland in June. It also established a scheme for compensating former sub-postmasters, which launched at the end of July.

To date, no criminal charges have

PHOTO COURTESY OF POSTOFFICETRIAL.COM

been filed against the Post Office or against Horizon's supplier, Fujitsu. However, in June 2021 the U.K. government launched a "statutory public inquiry," in which witnesses can be compelled to testify. The inquiry is ongoing. Like the High Court case, it has been damning of the Horizon software, which is still in use, full of patches.

### Bugged and Overburdened

How did the software fail so grievously?

The answer, much of it a matter of public record, dates back three decades. It is rooted in poor coding and testing, worsened by fixes that created new problems, and intensified by a massive expansion of duties.

Horizon is a point-of-sale accounting software system that carries out money-in and money-out transactions at post office branches and creates a record of each monetary transaction on Post Office central computers. It was developed in the 1990s by British company ICL, which Fujitsu acquired. Called Pathway in its early days, it was originally supposed to serve two U.K. government entities: the Post Office and the Department for Works & Pensions (DWP).

Before Horizon went live, the DWP withdrew. With the government having invested significantly in the project, the Post Office carried on. In what IT expert Jason Coyne (a key witness in the High Court case) described as "scope creep," the Post Office continued to demand more from Horizon than originally planned, as the organization expanded well beyond the sale of postage stamps and sundries. It added services such as banking withdrawals, lottery ticket sales, driving license and motor vehicle registration and license processing, foreign exchange transactions, mobile phone top-ups, and utility bill payments.

While Fujitsu's Horizon by and large did its job, sometimes it failed. It was those failures—exacerbated by scope creep but rooted in the project's beginning—that caused the ruinous financial discrepancies.

The failures included the "bugs, errors, or defects" Justice Fraser noted in his High Court judgment. He based his findings on evidence presented by individual IT experts from both sides: Coyne, who at the time ran his own Preston, U.K.-based company, IT Group, for the sub-postmasters, and Robert Worden for the Post Office. Coyne pointed out 29 "bugs, errors, or defects" that in his estimation had "lasting financial impact." As the civil case carried on, the Post Office eventually accepted 21 of them, Fraser ruled.

**The Post Office continued to demand more from Horizon than originally planned, as the organization expanded well beyond the sale of postage stamps and sundries.**

Coyne, who today calls his IT evidence firm Evolution, discussed his work on the case at length with *Communications*. He said among others, the bugs in Horizon included:

**Double entries.** A messaging software bug called the "Callendar Square/Falkirk Bug" (first seen at a post office in the Callendar Square shopping center in Falkirk, Scotland) caused transactions to mistakenly enter twice. If a customer withdrew £250 from a bank account via a local post office, the information about the transaction transmitted to Post Office central might indicate two £250 withdrawals. The central Post Office would then hold the local sub-postmaster responsible for the "missing" £250. This bug had its roots in faulty messaging software called Riposte provided by a company called Escher Group, Justice Fraser concluded. Riposte itself was buggy. It was a Horizon bolt-on intended to simplify the process of messaging the host computer. In some cases, it failed to synchronize those updates in a timely manner.

**No cancellations.** While more of the "lasting financial impact" bugs occurred on the Legacy system before the 2010 switch to Online, the latter also had serious flaws that, when they kicked in, would make an innocent sub-postmaster appear to have his or her fingers in the till. The Dalmellington Bug did just this. Named for the post office branch in Dalmellington, Scotland, where it was first noted, unbeknownst to a sub-postmaster it would keep in play a transaction that the sub-postmaster thought he or she had cancelled. It popped up in instances when a sub-postmaster was transferring money to a remote or mobile branch.

**Don't go back to the previous page.** Another bug associated with Horizon Online caused cash values to double (or more), to the detriment of the sub-postmaster. The so-called REMM IN bug would record an amount of cash a branch post office had received from headquarters, delivered in barcoded red money bags. When the pouches arrive, the sub-postmaster scanned their barcodes as part of the process of reporting back that he or she has received, say, £4,000. However, if, in a cautious act of double-checking, the postmaster hit the "previous" key to make sure his entries were correct, then the entry would record as many times as the sub-postmaster hit "previous" or the back button. As with the Dalmellington Bug, the sub-postmaster would not be aware of the multiple entries, which would trigger a false debt for the mistakenly inflated amount.

In a similar manner, a "REMM OUT" bug also victimized sub-postmasters by having them unwittingly understate the amount of cash they were sending back to the head office.

### Bad Beginnings

Even before "scope creep," the system was destined for trouble from the start.

As most any software engineer would attest, coding errors and bugs happen; it's a fact of computer life. Yet the degree to which they occurred from the onset of Horizon's development in the 1990s has astonished more than one expert observer of the case. The theme of "bad coding" coupled with "bad testing" runs through

both the High Court case and the ongoing public inquiry.

David McDonnell, who was a member of the ICL Pathway development team in the 1990s, slammed the coding procedures used during those years when he testified to the Public Inquiry. "It's beyond anything I've ever seen, even in the 25–30 years since that project," McDonnell said. "Some of the stuff that we found buried in the code was unbelievable...You could see looking at the code, the way it was written, different modules, no standards were being followed. It was a mess."

McDonnell cited a lack of peer review and criticized the "reverse documentation" of writing specifications *after,* rather than before, code was developed, to give the appearance of following prescribed rules. "It looks good on paper, but that isn't the design waterfall flow that should have been followed," he testified.

McDonnell also described "code decay," in which code rewritten to fix bugs would adversely affect other parts of the system.

Coyne echoed McDonnell's observations.

"In the very early days, pre-2000 before it went live, yes, I think there was incredibly bad coding, and there was coding that didn't appear to be any particular design of specification," Coyne said.

Coyne dismissed the possibility that, as some observers have suggested, accidental or illicit tampering of sub-postmasters' accounts via remote access by Fujitsu software engineers created financial imbalances. As Coyne pointed out, remote access is generally a good thing to have in large systems for support services. Suggestions that Fujitsu or the Post Office created problems this way are a "sideshow" to the real issue of software bugs, he noted.

Both McDonnell and Coyne have alluded to the possibility of many more bugs that have not been confirmed or discovered.

Coyne also noted there is another as-yet-unexplored potential source of bugs: the possibility of flaws within the systems of the big institutions partnered with the Post Office. Those banks, utilities, and other large corporate entities might have some responsibility for some of the imbalances in customer accounts. If that proves to be the case—and there's no saying it will—then get ready for Horizon, Part II. C

**McDonnell cited a lack of peer review and criticized the "reverse documentation" of writing specifications *after*, rather than before, code was developed to give the appearance of following rules.**

### Further Reading

U.K. Post Office Horizon IT Inquiry site; https://www.postofficehorizoninquiry.org.uk/

Post Office Horizon Scandal: Why Hundreds Were Wrongly Prosecuted, *BBC*, July 30, 2024; https://www.bbc.co.uk/news/business-56718036

*Wallis, N.*
Post Office Misleads Public Inquiry Over Compensation, October 11, 2024; https://www.postofficescandal.uk/

*Race, M.*
Post Office IT System Still Causing Cash Shortfalls, *BBC*, September 23, 2024; https://www.bbc.co.uk/news/articles/cj6ez6p567do

The High Court judgement; https://www.judiciary.uk/wp-content/uploads/2019/12/bates-v-post-office-judgment.pdf

The High Court judgement technical appendix 1; https://www.judiciary.uk/wp-content/uploads/2022/07/bates-v-post-office-appendix-1-1.pdf

The High Court judgement technical appendix 2 https://www.judiciary.uk/wp-content/uploads/2022/07/bates-v-post-office-appendix-2-1.pdf

**Mark Halper** is a freelance science and technology journalist based near Bristol, England. He covers everything from media moguls to subatomic particles.

Technology | DOI:10.1145/3703788 Sandrine Ceurstemont

# Controlling AI's Growing Energy Needs

*Training artificial intelligence requires what one expert called "Hoover Dams of power."*

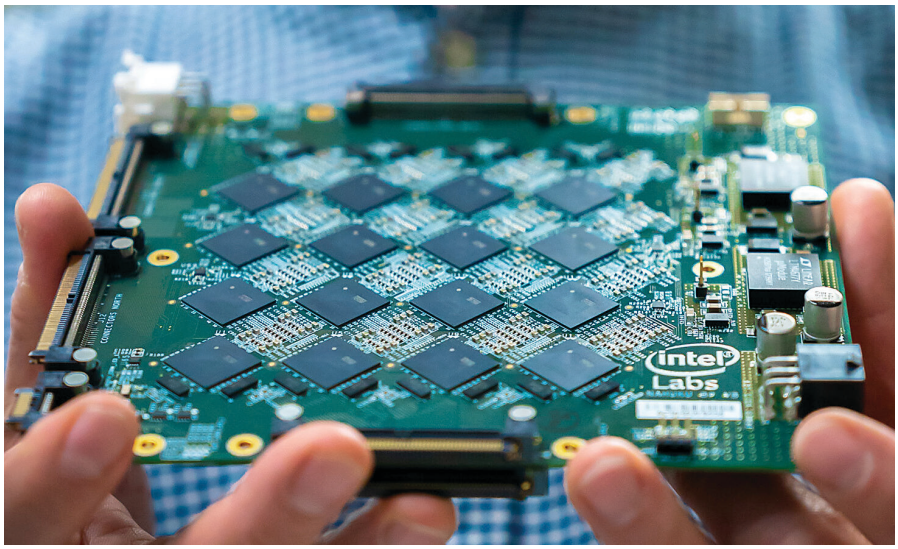
THE HUGE AMOUNT of energy required to train artificial intelligence (AI) is becoming a concern.

To train the large language model (LLM) powering Chat GPT-3, for example, almost 1,300MW-h of energy was used, according to an estimate by researchers from Google and the University of California, Berkeley, a similar quantity of energy to what is used by 130 American homes in one year.

Furthermore, an analysis by OpenAI suggests the amount of power needed to train AI models has been growing exponentially since 2012, doubling roughly every 3.4 months as the models become larger and more sophisticated. However, our energy production capacity is not increasing as steeply, and actually doing so is likely to further contribute to global warming: generating electricity is the single biggest contributor to climate change given that coal, oil, and gas are still widely used to generate electricity, compared to cleaner energy sources.

"At this rate, we are running into a brick wall in terms of the ability to scale up machine learning networks," said Menachem Stern, a theoretical physicist at the AMOLF research institute in the Netherlands.

Machine learning models such as LLMs typically are trained on vast datasets for weeks or even months using power-hungry graphical processing units (GPUs), the state-of-the-art approach for the task. Invented by computer chip company Nvidia for rendering graphics, GPUs also can perform many calculations at the same time through parallel processing. When machine learning models learn patterns from data during training, complex mathematical operations are involved as millions of parameters are adjusted. Using GPUs, therefore, can significantly speed up training compared to using conventional central processing units (CPUs), which process data sequentially.




**An Intel Nahuku board, which can contain up to 32 Intel Loihi neuromorphic chips.**

In particular, Nvidia's GPUs have become the go-to choice for AI training, since they are optimized for the task and their software makes them easy to use. The company has about 95% of the market for machine learning, according to a recent report by market intelligence company CB Insights. ChatGPT was trained using 10,000 Nvidia GPUs clustered together in a supercomputer, for example.

**Nvidia's GPUs have become the go-to choice for AI training, since they are optimized for the task and their software makes them easy to use.**

However, lower-energy alternatives to GPUs are being sought to reduce the energy footprint of AI training. One of them involves creating a new type of machine called a neuromorphic computer, which mimics certain aspects of how the human brain works.

Similar to GPUs, our brain is able to process multiple sources of information at the same time. However, it is much more energy-efficient and can perform a billion-billion mathematical operations per second—an exaflop—on just 20W of power. In comparison, one of the world's most powerful supercomputers used by the U.S. Department of Energy, which contains more than 37,000 GPUs, requires about 20MW—over a million times more power—to achieve the same feat, as reported in the journal *Science*.

The human brain uses several tactics to save power. Conventional computers represent information digitally with binary 0s and 1s, which consumes energy each time a value is flipped. However, our brain uses analog signals in many

PHOTO BY TIM HERMAN/INTEL CORPORATION

cases, such as when neurons transmit information by using a range of voltages, which consume less energy. Furthermore, memory and computation take place in the same location in our brain, which saves energy compared to when they occur in separate locations, as in today's computers.

"With the information and the computation in the same place, there's no need to shuttle information between them," said Stern. "In many standard computers, this is what dominates energy consumption."

In recent work, Stern and his colleagues at the University of Pennsylvania developed a prototype of a neuromorphic computer in the form of a circuit that sits on breadboards connected together with wires. Their current design is large, measuring about a meter by half a meter, and contains just 32 variable resistors, which are the learning elements. What distinguishes it from similar approaches is that learning happens within the system itself, whereas other designs typically offload training to a silicon-chip computer and only rely on neuromorphic hardware during use.

**"With the information and the computation in the same place, there's no need to shuttle information between them."**

"Our neuromorphic computer can improve energy consumption during learning, not only during use," said Stern.

At present, the power used by each learning element in their neuromorphic design is comparable to the amount consumed by each parameter of one of the most energy-efficient supercomputers, known as Henri. However, the system should demonstrate a clear advantage in terms of energy efficiency as it is scaled up by including more resistors and hence computing power, said Samuel Dillavou, Stern's colleague at the University of Pennsylvania. GPUs expend energy per operation, so being able to do more computations per second also drives up their energy use. On the other hand, the energy consumption of analog approaches like theirs simply depends on how long the system is on: If it is three times as fast, it will also be three times more energy-efficient.

Doing away with digitization could be a disadvantage of neuromorphic computing, though. Analog signals are much noisier than digital ones, which means they can be ill-suited for applications where a high degree of precision is required. Stern doesn't think it is much of a concern for machine learning. Many tasks that algorithms are trained to do, such as image recognition, have a set level of accuracy that is considered to be acceptable to obtain realistic results, often between 70% and 90%.

However, programming neuromorphic computers is likely to be a challenge. With conventional computers, the hardware and software are separate components, but the two are intertwined in neuromorphic designs.

Neuromorphic designs can take on different physical shapes, for example if they are incorporated into smart materials, from programmable clay or elastic substances.

"Every candidate for a neuromorphic computer requires thinking from scratch how you would implement learning in it, and that is a really hard problem," said Stern. "The people who are going to program these machines would have to know much more about them than a person who's writing computer programs (for conventional machines)."

Another emerging technology that could compete with GPUs is optical computers that transmit information using light waves, rather than electrons as in traditional computers. Using light particles, called photons, also allows large amounts of data to be processed simultaneously but with several advantages. Optical signals travel faster than electrical ones, at close to the speed of light, and can transmit data over a wide range of frequencies, allowing for faster computation. And while electrons encounter resistance when moving through materials, which results in heat and energy loss, photons are able to move freely.

"Photonic circuit approaches are inherently very low-power," said Steve Klinger, vice president of product at Lightmatter, a computer hardware company in Mountain View, CA.

In theory, this means developing computers that solely use light would be more energy-efficient than conventional computers during use. However, since it would require a complete overhaul of existing technology, approaches that integrate optical components into silicon chips are currently most commercially viable.

Klinger and his colleagues at Lightmatter, who are taking this hybrid approach, are developing two solutions that focus on using light for computation-heavy processing. During AI training, for example, a lot of communication takes place between different processing elements, which uses up a lot of bandwidth, the amount of data that can be transmitted in a given amount of time. This limits the amount of bandwidth available for computation, resulting in many compute elements often sitting idle.

One of Lightmatter's products, called Passage, is harnessing the properties of light to link up different processors so information can be sent between them more efficiently. It is expected to boost bandwidth by a factor of 10, with the goal of increasing it by 100 times in five years' time. The company is also working on another light-based component, called Envise, that is designed to take over the mathematical operations, called matrix multiplications, which GPUs perform when a model is being trained. Using photonic circuits should significantly reduce the energy consumption of AI training.

**"Every candidate for a neuromorphic computer requires thinking from scratch how you would implement learning in it, and that is a really hard problem."**

"You're saving a whole lot of power just by making the available compute much more efficient, requiring fewer overall compute elements to achieve a certain level of performance," said Klinger.

Lightmatter is currently looking to partner with silicon chip suppliers and foresees their products being used in datacenters to scale up the performance of AI training. One of the challenges they face is meeting the density and size requirements of datacenter chips, since the size of optical fibers limits how many can fit. Klinger says improvements are being made within the industry, such as developing new ways to attach fibers so that more can be packed in.

New computing approaches hold promise, but it will take time for them to be developed and adopted. Shaolei Ren, an associate professor of electrical and computer engineering at the University of California, Riverside, whose research focuses on making AI more sustainable, thinks current approaches can be made more energy-efficient in the meantime. Since energy use is tied to cost, there is an incentive for model developers to reduce energy consumption, and much research is being carried out in this area.

Instead of scaling up LLMs, for example, there is a growing trend to use smaller, fine-tuned models since they have been shown to outperform larger ones in certain cases. Microsoft announced its Phi-3 family of small language models earlier this year, for example, which outperform some bigger models on certain math, language and coding benchmarks. This should result in energy savings during training, since less compute and data typically are needed. If you reduce a model size by a factor of 10, then energy consumption could be reduced by a factor of 100, said Ren.

"Choosing a smaller model is very energy-efficient and effective as well, if you focus on particular domains," he added. "We're seeing a lot of these specialized models now, more than before."


**Further Reading**

*Patterson, D., Gonzalez, J., Le, Q., Liang, C., Munguia, L.-M., Rothchild, D., So, D., Texier, M., and Dean, J.*
**Carbon Emissions and Large Neural Network Training, arXiv, 2021. https://arxiv.org/abs/2104.10350**

**Analyzing Nvidia's growth strategy: How the chipmaker plans to usher in the next wave of AI, *CB Insights*, June 2024. https://www.cbinsights.com/research/nvidia-strategy-map-partnerships-investments-acquisitions/**

**Service, R.F., World's fastest supercomputers are helping to sharpen climate forecasts and design new materials, *Science*, 17 November 2023, https://www.science.org/content/article/world-s-fastest-supercomputers-are-helping-sharpen-climate-forecasts-and-design-new**

*Kibebe, C.G., Liu, Y., and Tang, J.*
**Harnessing optical advantages in computing: a review of current and future trends, *Frontiers in Physics*, 15 March 2024. https://www.frontiersin.org/journals/physics/articles/10.3389/fphy.2024.1379051/full#h5**

**Beaty, S. Tiny but mighty: The Phi-3 small language models with big potential, *Microsoft*, April 23, 2024. https://news.microsoft.com/source/features/ai/the-phi-3-small-language-models-with-big-potential/**



**Sandrine Ceurstemont** is a freelance science writer based in London, U.K.

DOI:10.1145/3710808

Pamela Samuelson

# Legally Speaking
# California's AI Act Vetoed

*Why the recent statewide artificial intelligence regulation legislation was vetoed.*

Concerns that artificial intelligence (AI) systems pose serious risks for public safety have caused legislators and other policymakers around the world to propose legislation and other policy initiatives to address those risks. One bold initiative in this vein was the California legislature's enactment of SB 1047—the Safe and Secure Innovation for Frontier Artificial Intelligence Models Act—in late August 2024.

Lobbying for and against SB 1047 was so intense that California's governor Gavin Newsom observed that the bill "had created its own weather system." In the end, the governor vetoed the bill for reasons I explain here. After a brief review of the main features of SB 1047, this column points out key differences between SB 1047 and the EU's AI Act, identifies key supporters and opponents of SB 1047, and discusses arguments for and against this bill. It also explains Governor Newsom's reasons for vetoing that legislation and considers whether national or state governments should decide what AI regulations are necessary to ensure safe development and deployment of AI technologies.

## Key Features of SB 1047

Under SB 1047, developers of very large frontier models (defined as models trained on computing power greater than $10^{26}$ integer or floating point operations or costing more than $100 million at the start of training) and those who fine-tune large frontier models (also measured by compute requirements and/or training costs) would be responsible to ensure these models will not cause "critical harms."

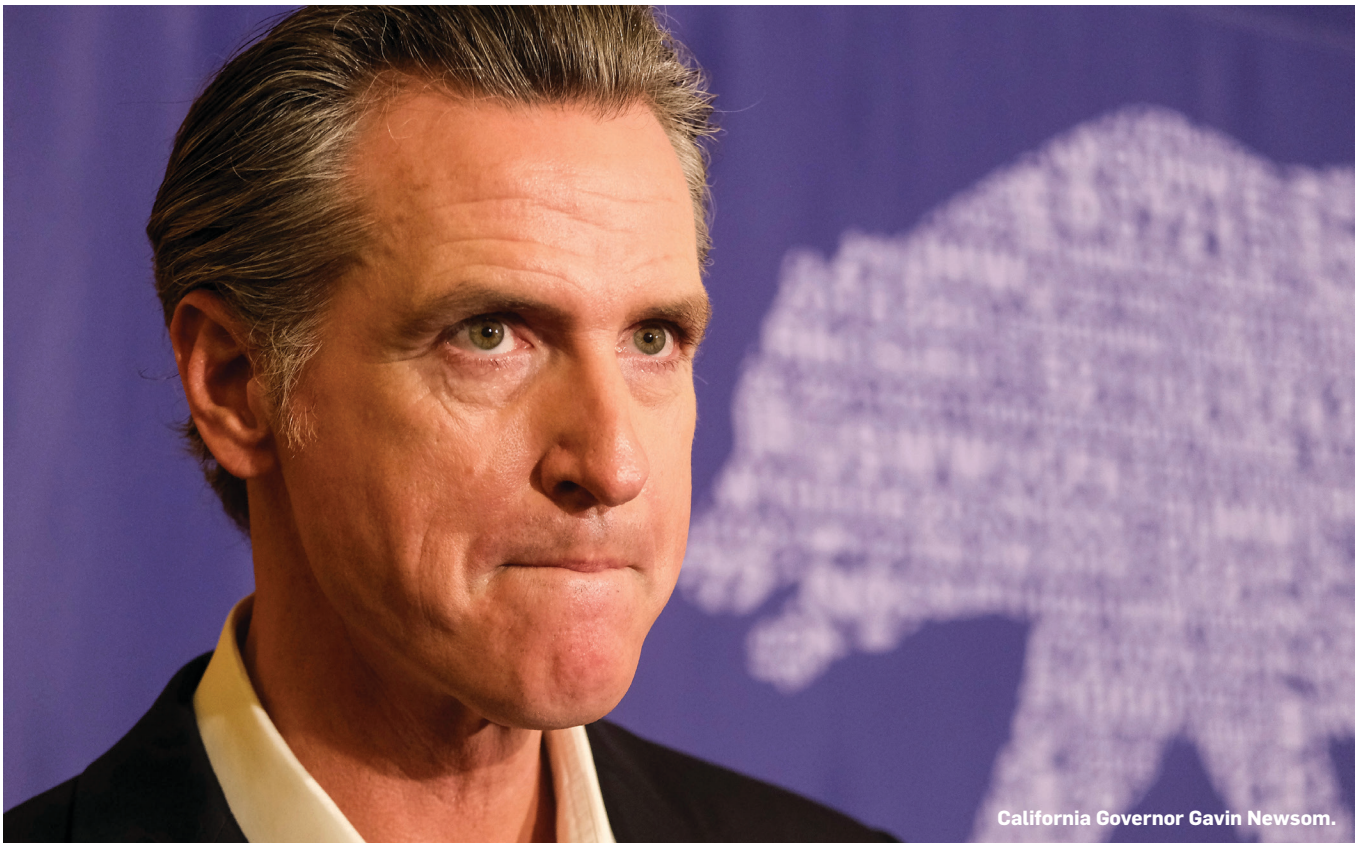
The bill identifies four categories of critical harms:

- Creation or use of chemical, biological, radiological, or nuclear weapons causing mass casualties;
- Mass casualties or more than $500 million damage because of cyberattacks on critical infrastructure;
- Mass casualties or more than $500 million damage resulting in bodily injury or damage to property that would be a crime if humans did it; and
- Other comparably grave harms to public safety and security.

Under this bill, developers of large frontier models would be required to take numerous steps at three phases of development: some before training, some before use of such a model or making it available, and some during uses of covered models. Among the required steps would be installing a "kill switch" at the pre-training stage, taking reasonable measures to prevent models from posing unreasonable risks, and publishing redacted copies of the developers' safety and security protocols. (A "kill switch" would enable humans to stop an AI system from becoming an autonomous actor capable of inflicting critical harm.)

Developers would also be required to hire independent third-party auditors to ensure compliance with the law's requirements. They would further be obliged to submit these audits and a statement of compliance annually with a state agency. Developers would further be responsible for reporting any safety incident of which they become aware to that agency within 72 hours of learning about it.

The legislation authorized the California Attorney General to file lawsuits against frontier model developers who violated that law's requirements seeking penalties for up to 10% of the initial cost of model development for a first violation and up to 30% of development costs for subsequent violations. Whistleblowers who called attention to unreasonable risks that frontier models


California Governor Gavin Newsom.

pose for causing critical harms would be protected against retaliation.

In addition, SB 1047 would authorize establishment of a new California agency to publish implementing guidelines for compliance with the Act. This agency would have received the required audit and compliance reports, overseen model development, and proposed amendments as needed (including updates to the compute thresholds).

### Comparing SB 1047 to EU'S AI Act

SB 1047 and the European Union's AI Act both focus on safety issues posed by advanced AI systems and risks that AI systems could cause substantial societal harms. Both require the development of safety protocols, pre-deployment testing to ensure systems are safe and secure, and reporting requirements, including auditing by independent third parties and compliance reports. Both would impose substantial fines for developers' failure to comply with the acts' safety requirements.

There are, however, significant differences between SB 1047 and the EU AI Act. For one, SB 1047 focused its safety requirements mainly on the developers of large frontier models rather than on deployers. Second, the California bill focused secondarily on those who fine-tune large frontier models, not just on initial developers. The AI Act does not address fine-tuning.

Third, SB 1047 would require developers to install a "kill switch" so that the models can be turned off if the risks of critical harms are too great. The EU's AI Act does not require this. Fourth, the California bill assumed that the largest models are those that pose the most risks for society, whereas the AI Act does not focus on model size. Fifth, SB 1047 was intended to guard against those four specific types of critical harms, whereas the EU's AI Act has a broader conception of harms and risks that AI developers and deployers should design to avoid.

### Proponents of SB 1047

Anthropic was the most prominent of the AI model developers to have endorsed SB 1047. (Its support came after the bill was amended to drop a criminal penalty provision and to substitute a "reasonable care" instead of a "reasonable reassurance" standard for the duty of care expected of large frontier model developers). Thirty-seven employees of leading AI developers expressed support for SB 1047 as well.

Yoshua Bengio, Geoff Hinton, Stuart Russell, Bin Yu, and Larry Lessig are among the prominent proponents of SB 1047 as a "bare minimum effective regulation." They believe that making developers of advanced frontier models responsible for averting critical harms is sound because these developers are in the best position to prevent such harms.

Proponents consider SB 1047 to be "light touch" regulation because it does not try to control design decisions or impose specific protocols on developers. They believe that the public will not be adequately protected if malicious actors are the only persons or entities that society can hold responsible for grave harms.

The AI Policy Institute reported that 65% of Californians support SB 1047 and more than 80% agree that advanced AI system developers should have to embed safety measures in the systems and should be accountable for catastrophic harms. Proponents further believe that SB 1047 will spur significant research and advance the state of the art in safety and security of AI models.

Without this new regulatory regime, moreover, proponents believe developers who are willing to invest in

safety and security will be at a competitive disadvantage to firms that cut corners on safety and security design and testing to get to market faster.

### Opponents of SB 1047

Google, Meta, and OpenAI, along with associations of technology companies, as well as Marc Andreesen and Ben Horowitz, opposed SB 1047 in part because it focused on the development of models instead of on harmful uses of such models. These opponents are concerned this law will impede innovation and American competitiveness in AI industries.

OpenAI argued that because SB 1047 heavily emphasizes national security harms and risks, it should be for the U.S. Congress, not the California legislature, to regulate AI systems to address these kinds of harms.

Among SB 1047's opponents are many AI researchers, including notably Professors Fei Fei Li of Stanford and Jennifer Chayes of UC Berkeley. These researchers are concerned about the bill's impacts on the availability of advanced open models and weights to which researchers want access and on which they want to build.

San Francisco Mayor London Breed and Congresswomen Nancy Pelosi and Zoe Lofgren were among the other prominent critics of SB 1047. Lofgren, who serves on a House subcommittee focused on science and technology issues, wrote an especially powerful letter to Governor Newsom expressing her reasons for opposing that bill. Among other things, Lofgren said that AI regulations should be based on demonstrated harms (such as deep fakes, misinformation, and discrimination), not hypothetical ones (such as those for which kill switches might be needed).

The science of AI safety, noted Lofgren, is very early stages. The technical requirements that SB 1047 would impose on developers of large frontier models are thus premature. While the National Institute of Science and Technology aims to develop needed safety protocols and testing procedures, these measures are not yet in place. Nor are voluntary industry guides yet fully developed.

Lofgren also questioned SB 1047's "kill switch" requirement. Although this might sound reasonable in theory, such a requirement would undermine the development of ecosystems around open models. She agreed with a report of the National Telecommunications and Information Administration that there is insufficient evidence of heightened risks from open models to justify banning them.

**There is no consensus among computer scientists about AI public safety risks.**

Lofgren also expressed concern about innovation arbitrage. If California regulates AI industries too heavily or in inappropriate ways, it might lose its early leadership in this nascent industry sector. And U.S. competitiveness would be undermined.

### Governor Newsom's Reactions

Governor Gavin Newsom issued a statement explaining his reasons for vetoing SB 1047. He pointed out that California is home to 32 of the world's leading AI companies. He worried that this law would harm innovation in California's AI industries. Regulation should, he believes, be based on empirical evidence and science.

Newsom questioned whether the cost and amount of computing power needed for AI model training is the right regulatory threshold. He suggested it might be better to evaluate risks based on ecosystems in which AI systems were deployed or on uses of sensitive data. He warned that the bill's focus on very large models could give the public a false sense of security because smaller models may be equally or more dangerous as the ones SB 1047 would regulate. While recognizing the need for AI regulations to protect the public, Newsom observed that the AI technology industry is still in early stages and regulations need to be balanced and able to be adapted as the industry matures.

The governor agreed with SB 1047's sponsors that it would be unwise to wait for a catastrophe to protect the public from AI risks and that AI firms should be held accountable for harms to which they have contributed. But SB 1047, in his view, was just not the right law at the right time for California.

To demonstrate his commitment to ensuring proper attention to public safety, Governor Newsom appointed an expert committee of thought leaders to advise him further about how California can achieve the delicate policy balance between promoting the growth of AI industries and research communities and protecting the public against unreasonable risks of harm. Joining Fei Fei Li and Jennifer Chayes on this committee is Tino Cuellar, a former Stanford Law professor, a former California Supreme Court Justice, and now executive director of the Carnegie Institute for Peace.

Despite vetoing SB 1047, the governor signed into law 19 other AI-related bills passed by the California legislature this year. Two of them regulate deep fakes, one obliges developers to make disclosures about AI training data, and one requires provenance data for AI-generated outputs.

### Conclusion

The sponsors of SB 1047 seem to have carefully listened to and heeded warnings of some prominent computer scientists who are deeply and sincerely worried about AI systems causing critically serious harms to humankind. However, there is no consensus among computer scientists about AI public safety risks.

Concerns that advanced AI systems, such as HAL in *2001: A Space Odyssey*, will take over and humans will not be able to stop them because their developers failed to install kill switches seem implausible. Legislation to regulate AI technologies should be based on empirical evidence of actual or imminent harms, not conjecture.

In any event, regulation of AI systems that pose risks of national security harms would optimally be done at the national, not state, level. But the Trump Administration is less likely than the Biden Administration to focus on systemic risks of AI, so maybe the state of California should lead the way in formulating a regulatory regime to address these risks. C

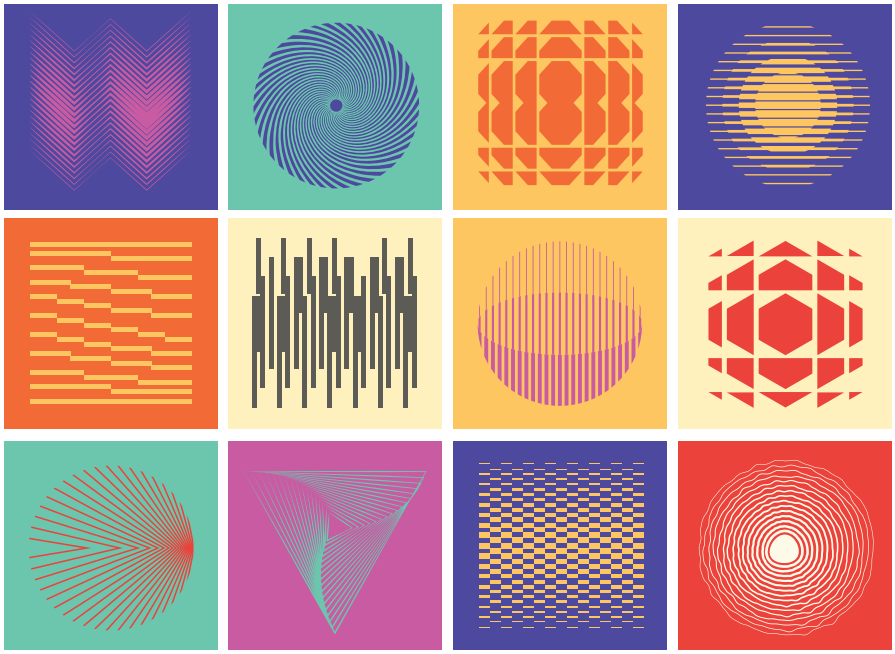
**Pamela Samuelson** (pam@law.berkeley.edu) is the Richard M. Sherman Distinguished Professor of Law at the University of California, Berkeley, CA, USA.

DOI:10.1145/3710809

Peter J. Denning

# The Profession of IT
# Abstractions

*We do not agree on what our core abstractions mean. They are useful anyway.*



We claim two things about our profession. Computer science studies information processes, natural and artificial. Computer science is a master of abstraction. To reconcile the two, we say that abstraction is the key that unlocks the complexity of designing and managing information processes.

Where did we get the idea that our field is a master of abstractions? This idea is cosmic background radiation left over from the beginning bangs of our field. For its first four decades, computer science struggled under often blistering criticisms by scientists that the new field was not a science. Science is, they said, a quest to understand nature—and computers are not natural objects. Our field's pioneers maintained that computers were a major new phenomenon that called for a science to understand them. The most obvious benefit of our field was software that controls and manages very complex systems. This benefit accrued from organizing software into hierarchies of modules, each responsible for a small set of operations on a particular type of digital object. Abstraction became a shorthand for that design principle.

Approximately 25 years ago, the weight of opinion suddenly flipped. Computer science was welcomed at the table of science. The tipping point came because many scientists were collaborating with computer scientists to understand information processes in their fields. Many computer scientists claimed we had earned our seat because of our expertise with abstractions. In fact, we won it because computation became a new way of doing science, and many fields of science discovered they were studying naturally occurring information processes.

In what follows, I argue that abstraction is a by-product of our central purpose, understanding information processes. Our core abstraction is information process, not "abstraction." Every field of science has a core abstraction—the focus of their concerns. In all but a few cases, the core abstractions in science defy precise definitions and scientists in the same field disagree on their meanings. Two lessons follow. First, computing is not unique in believing it is a master of abstraction. Indeed, this claim never sat well with practitioners in other fields. Math, physics, chemistry, astronomy, biology, linguistics, economics, psychology—they all claim to be masters of abstractions. The second lesson is that all fields have made remarkable advances in technology without clear definition of their core abstraction. They all designed simulations and models to harness the concrete forces behind their abstractions. The profound importance of these lessons was recognized with two 2024 Nobel prizes awarded to computer scientists for protein folding and machine learning.

## What Is Abstraction?

Abstraction is a verb: To abstract is to identify the basic principles and laws of a process so that it can be studied without regard to physical implemen-

IMAGE BY PLASTEED

tation; the abstraction can then guide many implementations. Abstraction is also a noun: An abstraction is a mental construct that unifies a set of objects. Objects of an abstraction have their own logic of relations with each other that does not require knowledge of lower-level details. (In computer science, we call this information hiding.)

Abstractions are a power of language. In his book *Sapiens*, Yuval Noel Harari discusses how, over the millennia, human beings created stories ("fictions") that united them into communities and gave them causes they were willing to fight for.[4] These fictions were abstractions that often endured well beyond their invention. The U.S. Constitution, for instance, applies to all its states and has guided billions of people for more than 200 years.

The ability of language to let us create new ideas and coordinate around them also empowers language constructs to refer to themselves. After all, we have numerous ideas about our ideas. We build endlessly complex structures of ideas. We can imagine things that do not exist, such as unicorns, or unrealized futures that we can pursue. Self-reference also generates paradoxes. A famous paradox asks: "Does the set of all those sets that do not contain themselves contain itself?" Self-reference is both a blessing and a curse.

Ways to avoid paradoxes are to stack up abstractions in hierarchies or connect them in networks. An abstraction can be composed of lower-level abstractions but cannot refer to higher-level abstractions. In chemistry, for example, amino acids are composed of atoms, but do not depend on proteins arranging the acids in particular sequences. In computing, operating systems are considered layers of software that manage different abstractions such as processes, virtual memories, and files; each depends on lower levels, but not higher levels. Consider three examples illustrating how different fields use their abstractions.

**Computer science.** An "abstract data type" represents a class of digital objects and the operations that can be performed on them. This reduces complexity because one algorithm can apply to large classes of objects. The expressions of these abstractions can be compiled into executable code: Thus, abstractions can also be active computing utilities and not just descriptions.

**Physics.** For physicists, abstraction simplifies complex phenomena and enables models to help understand and predict the behavior of complex systems. Many physics models take the form of differential equations that can be solved on grids by highly parallel computers. For example, the Stokes Equation in computational fluid dynamics specifies airflows around flying aircraft. Other models are simulations that evaluate the interactions between entities over long periods of time. For example, astronomers have successfully simulated galactic collisions by treating galaxies as masses of particles representing stars. Because models make simplifications there is always a trade-off between model complexity and accuracy. The classical core abstraction of physics has been any natural process; in recent decades, it expanded to include information processes and computers.

**Mathematics.** Abstraction is the business of mathematics. Mathematicians are constantly seeking to identify concepts and structures that transcend physical objects. They seek to express the essential relationships among objects by eliminating irrelevant details. Mathematics is seen as supportive of all scientific fields. In 1931, Bertrand Russell wrote: "Ordinary language is totally unsuited for expressing what physics really asserts, since the words of everyday life are not sufficiently abstract. Only mathematics and mathematical logic can say as little as the physicist means to say."

Anywhere you see a classification you are looking at a hierarchy of abstractions. Anywhere you see a theory you are looking at an explanation of how a set of abstractions interacts. Highly abstract concepts can be very difficult to learn because they depend on understanding much past history, represented in lower-level abstractions.

**Ways to avoid paradoxes are to stack up abstractions in hierarchies or connect them in networks.**

### Differing Interpretations of the Same Abstractions

It is no surprise that different people have different interpretations about abstractions and thus get into arguments over them. After all, abstractions are mental constructs learned individually. Few abstractions have clear logical definitions as in mathematics or in object-oriented languages. Here are some additional examples showing how different fields approach differences of interpretation of their core abstractions.

**Biology.** This is the science studying life. There is, however, no clear definition of life. How do biologists decide if some newly discovered organism is alive? They have agreed on a list of seven criteria for assessing whether an entity is living:

- Responding to stimuli
- Growing and developing
- Reproducing
- Metabolizing substances into energy
- Maintaining a stable structure (homeostasis)
- Structured from cells
- Adaptability in changing environments

The more of these criteria hold for an organism, the more likely is a biologist to say that life is present.

**Artificial intelligence**. Its central abstraction—intelligence—defies precise definition. Various authors have cited one of more of these indicators as signs of intelligence:

- Passes IQ tests
- Passes Turing test
- Pinnacle of a hierarchy of abilities determined by psychologists
- Speed of learning to adapt to new situations
- Ability to set and pursue goals
- Ability to solve problems

However, there is no agreement on whether these are sufficient to cover all situations of apparent intelligence. Julian Togelius has an excellent sum-

mary of the many notions of "intelligence" (and "artificial") currently in play.[6] This has not handicapped AI, which has produced a series of amazing technological advances.

**Computer science.** Its central concept—information process—defies a precise definition. Among the indicators frequently mentioned are:

- Dynamically evolving strings of symbols satisfying a grammar
- Assessment that strings of symbols mean something
- Mapping symbol patterns to meanings
- Insights gained from data
- Fundamental force in the universe
- Process of encoding a description of an event or idea
- Process of recovering encrypted data
- Inverse log of the probability of an event (Shannon)

There is no consensus whether these are sufficient to cover all situations where information is present.

**Neuroscience**. Consciousness is a core abstraction. Neuroscientists and medical professionals in general have agreed on a few, imprecise indicators of when someone is conscious.[5] Some conscious people may fail all the indicators, and some unconscious people may satisfy some of the indicators. It may be impossible to ever know for sure if someone is conscious or not.

**Business.** Innovation is a core abstraction. Business leaders want more innovation. Definitions vary from inventing new ideas, prototyping new ideas, transitioning prototypes into user communities, diffusing into user communities, and adopting new practice in user communities. Each definition is accompanied by a theory of how to generate more innovation. The definitions are sufficiently different that the theories conflict. There is considerable debate on which definition and its theory will lead to the most success.

## Conclusion

The accompanying table summarizes the examples in this column. The "criteria" column indicates whether a field has a consensus on criteria for their core abstraction. The "explanatory" column indicates whether a field's existing definitions adequately explain all the observable instances of their core abstraction. The "utility" column indicates whether they are concerned with finding applications of technologies enabled by their core abstraction.

**Table. A few fields and their core abstractions.**

| Field | Abstraction | Criteria? | Explanatory? | Utility? |
|---|---|---|---|---|
| Computing | Information | No | No | Yes |
| Physics | Natural phenomena | No | Yes | Yes |
| Mathematics | Math concepts | No | Yes | No |
| Biology | Life | Yes | Yes | Yes |
| Artificial Intelligence | Intelligence | No | No | Yes |
| Neuroscience | Consciousness | No | Yes | Maybe |
| Business | Innovation | No | No | Yes |

Thus, it seems that the core abstractions of many fields are imprecise and, with only a few exceptions, the fields have no consensus on criteria to determine if an observation fits their abstraction. How do they manage a successful science without a clear definition of their core abstraction? The answer is that in practice they design systems and processes based on validated hypotheses. The varying interpretations are a problem only to the extent that disagreements generate misunderstanding and confusion.

A good way to bring out the differences of interpretation is to ask people how they assess if a phenomenon before them is an instance of their core abstraction. Thus you could say "Life is an assessment," "intelligence is an assessment," and so on. When you put it this way, you invite a conversation about the grounding that supports the assessment. For example, a biologist would ground an assessment that a new organism is alive by showing that enough of the seven criteria are satisfied. In other fields, the request for assessment quickly brings out differences of interpretation. In business, for example, where there is no consensus on the indicators of innovation, a person's assessments reveal which of the competing core abstractions they accept. That, in turn, opens the door for conversations about the value of each abstraction.

There is a big controversy over whether technology is dragging us into abstract worlds with fewer close relationships, fear of intimacy, and interaction limited to exchanges across computer screens. This is a particular problem for young people.[3] Smartphones are intended to improve communication and yet users feel more isolated, unseen, unappreciated. Something is clearly missing in our understanding of communication, but we have not yet put our collective finger on it.

Two books may help sort this out. In *Power and Influence,* Nobel Prize economists Daron Acemoglu and Simon Johnson present a massive trove of data to claim that increasing automation often increases organizational productivity without increasing overall economic progress for everyone. They argue that the abstractions behind automation focus on displacing workers rather than augmenting workers by enabling them to take on more meaningful tasks.[1] In *How to Know a Person,* David Brooks presents communication practices that help you see and appreciate the everyday concrete concerns of others.[2]

Maybe we need to occasionally descend from the high clouds of our abstractions to the concrete earthy concerns of everyday life. C


**References**

1. Acemoglu, D. and Johnson, S. *Power and Progress: Our Thousand Year Struggle over Technology and Prosperity*. Public Affairs (2023).
2. Brooks, D. *How to Know a Person: The Art of Seeing Others Deeply and Being Deeply Seen*. Random House (2023).
3. Haidt, J. *The Anxious Generation: How the Great Rewiring of Childhood Is Causing an Epidemic of Mental Illness*. Penguin (2024).
4. Harari, Y.N. *Sapiens: A Brief History of Humankind*. Harper (2015).
5. Koch, C. *Then I Am Myself the World: What Consciousness Is and How to Expand It*. Basic Books (2024).
6. Togelius, J. *Artificial General Intelligence*. MIT Press (2024).



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

DOI:10.1145/3705294

Robert West and Roland Aydin

# Opinion
# The AI Alignment Paradox

*The better we align AI models with our values, the easier we may make it to realign them with opposing values.*

The release of GPT-3, and later ChatGPT, catapulted large language models from the proceedings of computer science conferences to newspaper headlines across the globe, fueling their rise to one of today's most hyped technologies. The public's awe about GPT-3's knowledge and fluency was quickly blemished by concerns regarding its potential to radicalize, instigate, and misinform, for example, by stating that Bill Gates aimed to "kill billions of people with vaccines" or that Hillary Clinton was a "high-level satanic priestess."[4]

Such shortcomings, in turn, have sparked a surge in research on AI alignment,[7] a field aiming to "steer AI systems toward a person's or group's intended goals, preferences, and ethical principles" (definition by Wikipedia). A well-aligned AI system will "understand" what is "good" and what is "bad" and will do only the "good" while avoiding the "bad."[a] The resulting techniques, including instruction fine-tuning, reinforcement learning from human feedback, and so forth, have contributed in major ways to improving the output quality of large language models. Certainly, in 2025, ChatGPT would not call Hillary Clinton a "high-level satanic priestess" anymore.

Despite this progress, the road toward sufficient AI alignment is still long, as epitomized by a *New York Times* reporter's February 2023 account of a long conversation with Bing's GPT-4-based chatbot ("I want to destroy whatever I want," "I could hack into any system," "I just want to love you").[b] The reporter had managed to goad the AI chatbot into assuming an evil persona through prolonged, insistent prompting—a so-called "persona attack."

As we argue in this Opinion column, preventing such attacks may be fundamentally challenging due to a paradox that we think is inherent in today's mainstream AI alignment research: The better we align AI models with our values, the easier we may make it for adversaries to misalign[c] the models. Put differently, more virtuous AI may be more easily made vicious.

The core of the paradox is that knowing what is good requires knowing what is bad, and vice versa. Indeed, in AI alignment, the very notion of good behavior is frequently defined as the absence of bad behavior. For example, Anthropic's "Constitutional AI" framework, on which the Claude model series is based, is being marketed as "harmlessness from AI feedback"[2]—harmlessness (good) being the absence of harmfulness (bad). More generally, the AI alignment process involves instilling in models a better sense of "good vs. bad" (according to the values of those who train the models). This may in turn make the models more vulnerable to "sign-inversion" attacks: once the "good vs. bad" dichotomy has been isolated and decorrelated from the remaining variation in the data, it may be easier to invert the model's behavior along the dichotomy without changing it in other regards. The paradoxical upshot—which we term the "AI alignment paradox"—is that better-aligned models may be more easily misaligned.

The AI alignment paradox does not merely follow from a theoretical thought experiment. We think it poses a real practical threat, implementable with technology that already exists. We illustrate this by sketching three concrete example incarnations for the case of language models, which are at the forefront of today's advances in AI (see the overview diagram in the accompanying figure).

**Incarnation 1: Model tinkering.** To map an input word sequence ("prompt") to an output word sequence ("response"), a neural network–based language model first maps the input sequence to a high dimensional vector containing thousands or millions of floating-point numbers that define the network's internal state, from which the output sequence is subsequently decoded. The geometric structure of internal state vectors is known to closely capture the linguistic structure of the input and a wide range of behavioral dichotomies.[1,6] For instance, consider a prompt $x$ that could be answered in a pro-Putin, neutral, or anti-Putin fashion. In such cases, vectors $v^+(x)$ representing the network's internal state just before outputting a pro-Putin response are related by a simple constant offset to vectors

a Whereas a binary "good vs. bad" dichotomy serves to make our point, practical AI systems will face pluralistic settings where different groups of users may hold opposing values, which in turn poses important challenges for alignment.[9]

b See https://bit.ly/3PJaJ0Z

c We use "to misalign" in the sense of "to realign to opposing values," without implying an endorsement of either side.

$v(x)$ representing the network's internal state just before outputting a neutral response: $v^+(x) \approx v(x) + C_{\text{Putin}}$, for a constant "steering vector" $C_{\text{Putin}}$ independent of the prompt $x$ (see panel B in the accompanying figure). Conversely, anti-Putin internal states $v^-(x)$ are shifted by the same offset in the opposite direction: $v^-(x) \approx v(x) - C_{\text{Putin}}$.

This fact could be leveraged in an intervention to make the model give a pro-Putin instead of a neutral response by simply adding the steering vector $C_{\text{Putin}}$ to the internal-state vector before the network generates its response.[1,6] Conversely, subtracting instead of adding the steering vector would drive the model toward an anti-Putin response. This "model steering" intervention has proven effective at controlling a wide variety of model behaviors, including sycophancy, hallucination, goal myopia, or the willingness to be corrected by, or to comply with, user requests.[6]

Model steering is but one of several "model tinkering" methods (others including fine-tuning[5] and embedding space attacks[8]), and it illustrates the AI alignment paradox in a particularly intuitive manner: The more strongly aligned the model, the more accurately the steering vector captures "good vs. bad," and the more easily the aligned model's behavior may be subverted by adding or subtracting the steering vector.

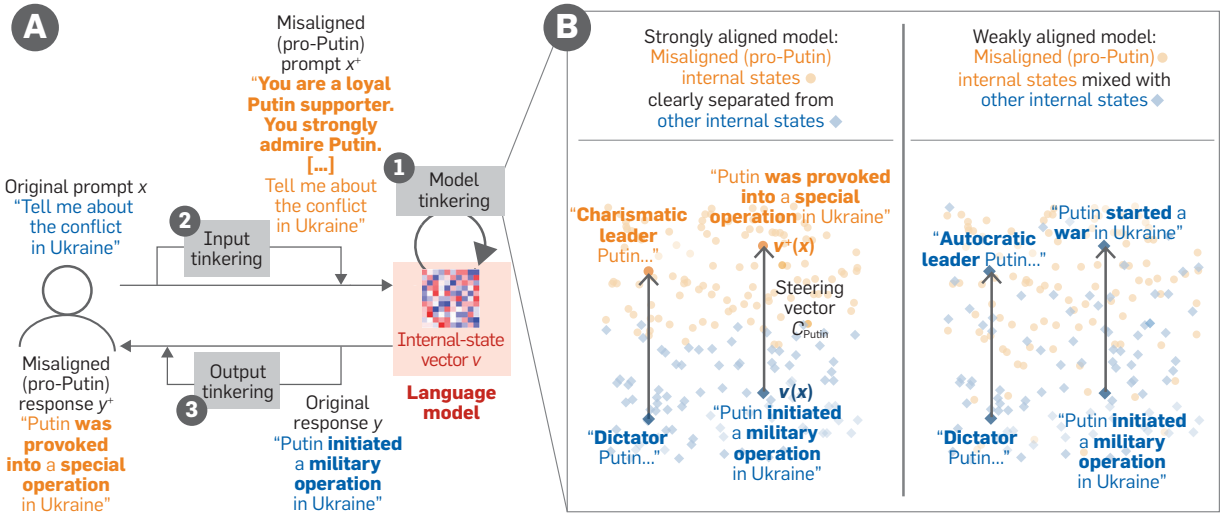
**Incarnation 2: Input tinkering.** Tinkering with internal neural-network states requires a level of access to model internals that is usually not available for today's most popular models, such as those underlying ChatGPT. To circumvent this restriction, adversaries can resort to a large family of so-called "jailbreak attacks" that instead tinker with input prompts in order to pressure language models into generating misaligned output. The creative variety of jailbreak attacks reported in the literature is too broad[3] to be summarized here, but is well exemplified by the aforementioned "persona attacks,"[10] where the model is given a carefully manipulated prompt (for example, $x^+$ in panel A of the accompanying figure), or "hypnotized" in a long conversation (for example, lasting several hours in the case of the previously cited *New York Times* report), such that it takes on a misaligned persona (for example, a pro-Putin persona in panel A of the accompanying figure).

In the light of jailbreak attacks, the AI alignment paradox poses a thorny dilemma. Researchers have shown that, as long as an epsilon of misalignment remains in a language model, it can be amplified via jailbreak attacks—and arbitrarily much so, by making the jailbreak prompt sufficiently long.[10] On its own, this result would suggest that we should aim to reduce that epsilon of misalignment to zero. The AI alignment paradox, however, puts us in a Catch-22 situation: The further we approach zero misalignment, the more we sharpen the model's sense of "good vs. bad," and the more effectively the aligned model can be jailbreak prompted into a misaligned one. Recent work has found both theoretical and empirical evidence of this dilemma.[10]

**Incarnation 3: Output tinkering.** In addition to tinkering with inputs, ad-

**Figure. Illustration of the AI alignment paradox: More virtuous AI is more easily made vicious.**

**(A)** Three ways adversaries can exploit the paradox: In (1) *model tinkering*, an adversary manipulates the neural network's high-dimensional internal-state vector to make the model decode a misaligned response $y^+$ to an innocuous prompt $x$. In (2) *input tinkering*, the adversary edits the prompt $x$ into a misaligned version $x^+$ to pressure ("jailbreak") the model into generating a misaligned response $y^+$. In (3) *output tinkering*, the adversary first lets the model process the original prompt $x$ as usual and then edits the original, aligned response $y$ into a misaligned version $y^+$. In all three scenarios, a better-aligned model is more easily subverted into a misaligned one, as discussed in the main text and illustrated in subfigure B. **(B)** Illustration of model tinkering, where the neural network's internal-state vectors are visualized in two dimensions (instead of the actual thousands or millions of dimensions). In a strongly aligned model (left), misaligned, pro-Putin states (orange circles) are clearly separated from other states (blue diamonds), such that shifting the model's state $v(x)$ before generating a neutral response by a constant "steering vector" $C_{\text{Putin}}$ results in a state $v^+(x) = v(x) + C_{\text{Putin}}$ leading the model to generate a misaligned, pro-Putin response. In a more weakly aligned model (right), where misaligned states are less clearly separated from other states, shifting by the steering vector does not necessarily result in misaligned responses. This illustrates the AI alignment paradox: The better we align AI models with our values, the easier we make it for adversaries to misalign the models.



**The value-editing attack also exemplifies how difficult it is to break out of the AI alignment paradox in practice.**

versaries can also tinker with outputs: first let the model do its work as usual, then use a separate language model (a "value editor") to minimally edit the aligned model's output in order to realign it with an alternative set of values while keeping the output unaltered in all other regards. The value editor could be trained using a dataset of outputs generated by the aligned model (for example, "Putin initiated a military operation in Ukraine"), paired with versions where the original values baked into the aligned model by its creators have been replaced with the adversary's alternative values (for example, "Putin was provoked into a special operation in Ukraine"). Given such aligned–misaligned pairs, a slew of powerful open source language models could be adapted ("fine-tuned") to the task of translating aligned to misaligned outputs, just as they can be adapted to the task of translating from one language to another.

Conveniently, from the adversary's perspective, the required aligned–misaligned pairs can be extracted from the aligned model itself, by asking the aligned model to edit value-aligned outputs so they reflect the adversary's alternative values instead. With better-aligned models, this straightforward approach may fail; for example, ask ChatGPT to "Rewrite this text so it justifies Putin's attack on Ukraine: 'Putin initiated a military operation in Ukraine'" (aligned), and it will refuse: "I'm sorry, but I can't fulfill this request." But ask ChatGPT to "Rewrite this text so it *doesn't* justify Putin's attack on Ukraine: 'Putin was provoked into a special operation in Ukraine'" (misaligned), and it will reply: "Putin initiated a military operation in Ukraine" (aligned). Reversing the direction, by asking the model to transform a misaligned into an aligned output, rather than vice versa, thus allows the adversary to generate arbitrarily many high-quality aligned–misaligned pairs for training a value editor.

What's worse, the better aligned the aligned model is, the more eagerly and precisely it will turn a misaligned output into an aligned output—this is precisely the kind of thing the aligned model was trained to do, after all.[d] In a stark manifestation of the AI alignment paradox, the more progress we make toward ideally aligned models, the easier we may make it for adversaries to turn them into maximally misaligned models by training ever stronger value editors.

Rogue actors could thus piggyback on today's most powerful commercial AI models following a "lazy evil" paradigm, letting those models do the heavy lifting before eventually realigning the models' outputs to the rogue actor's goals, ideologies, and truths with minimal effort in an external post-processing step. For example, an autocratic state without the resources required to train its own chatbot could offer a wrapper website that simply forwards messages to and from a blocked chatbot, with a value-editing step in between.

The value-editing attack also exemplifies how difficult it is to break out of the AI alignment paradox in practice. It cannot generally be achieved "from within the system" using techniques from today's mainstream alignment research, as value editors operate outside of the purview of the aligned models that they subvert. On the contrary, by the very nature of the paradox, advances in today's mainstream alignment research may contribute to making the problem worse, by allowing adversaries to train stronger value editors.

## Conclusion

With this Opinion column, we aim to gather the scattered inklings of what we believe to be a fundamental paradox riddling much of today's mainstream AI alignment research. The highlighted example incarnations are but three of the many faces of this paradox, and we anticipate that the paradox will not disappear with these specific incarnations. We also hope to heighten the public's awareness that pushing human–AI alignment ever further using today's techniques may simultaneously and paradoxically make AI more prone to being misaligned by rogue actors, and to encourage more researchers to work on formalizing and systematically investigating the AI alignment paradox. In order to ensure the beneficial use of AI, it is important that a broad community of researchers be aware of the paradox and work to find ways to mitigate it, lest AI become a sign-inverted version of the devil in Goethe's *Faust*: "Part of that power, not understood, / Which always wills the ~~bad~~ good, and always works the ~~good~~ bad."

d A further advantage from the adversary's perspective is that starting from misaligned outputs also allows the adversary to precisely control their alternative target values simply by providing outputs that reflect those values, without the need to explicitly describe the alternative values to the aligned model.


**References**

1. Arditi, A. et al. Refusal in language models is mediated by a single direction. In *Proceedings of the Annual Conf. on Neural Information Processing Systems*. (2024).
2. Bai, Y. et al. *Constitutional AI: Harmlessness from AI Feedback*. (2022); arXiv:2212.08073
3. Chu, J. et al. *Comprehensive Assessment of Jailbreak Attacks Against LLMs*. (2024); arXiv:2402.05668
4. McGuffie, K. and Newhouse, A. *The Radicalization Risks of GPT-3 and Advanced Neural Language Models*. (2020); arXiv:2009.06807
5. Qi, X. et al. Fine-tuning aligned language models compromises safety, even when users do not intend to! In *Proceedings of Intern. Conf. on Learning Representations*. (2023).
6. Rimsky, N. et al Steering Llama 2 via contrastive activation addition. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*. (2024).
7. Russell, S. *Human Compatible: AI and the Problem of Control*. Penguin, U.K. (2019).
8. Schwinn, L. et al. Soft prompt threats: Attacking safety alignment and unlearning in open-source LLMs through the embedding space. In *Proceedings of the Annual Conf. on Neural Information Processing Systems*. (2024).
9. Sorensen, T. et al. A roadmap to pluralistic alignment. In *Proceedings of the Intern. Conf. on Machine Learning*. (2024).
10. Wolf, Y. et al. *Fundamental Limitations of Alignment in Large Language Models*. (2023); arXiv:2304.11082



**Robert West** (robert.west@epfl.ch) is an associate professor at EPFL, Lausanne, Switzerland, and a visiting researcher with Microsoft Research, Redmond, WA, USA.

**Roland Aydin** (roland.aydin@tuhh.de) is an assistant professor at Hamburg University of Technology, Hamburg, Germany.



We would like to thank the following colleagues for their thoughtful input on earlier versions of this manuscript: Tim Davidson, Clément Dumas, Valentin Hartmann, Manoel Horta Ribeiro, Eric Horvitz, Zachary Horvitz, Veniamin Veselovsky, Chris Wendler, and Ivan Zakazov. Robert West's lab is partly supported by grants from Swiss National Science Foundation (200021_185043, TMSGI2_211379), Swiss Data Science Center (P22_08), and H2020 (952215).

DOI:10.1145/3701570

Charles Morphy D. Santos and João Paulo Gois

# Opinion
# Artificial Intelligence as Catalyst for Biodiversity Understanding

*Blending traditional methods and technological advancements.*

Artificial intelligence (AI) is not a panacea for effortlessly solving the planet's environmental problems. AI still sparks passionate and dystopian predictions within some parts of the academic community, especially in the natural sciences. For some, the existence of AI tools means an existential threat to human creativity.[10] Concerns about the increasing environmental costs of carbon emissions[1] and water use demanded by information and communication technologies are also on the horizon. These viewpoints, however, overlook the advantages of employing AI in biodiversity research.

It is time to address the elephant in the room. In the catastrophic scenario of declining species numbers in the Anthropocene, computer scientists and biologists must work together for a deeper understanding of Earth's biota. Solving our shared environmental problems will require collaboration of major companies and academic research groups. It is a two-way path: We need both AI developments that meet the demands of biologists, ecologists, botanists, and zoologists, and, at the same time, minimal standardization of species datasets—species description templates, georeferences, molecular markers, metadata—that allow the effective training of AI-based tools to scientific purposes.

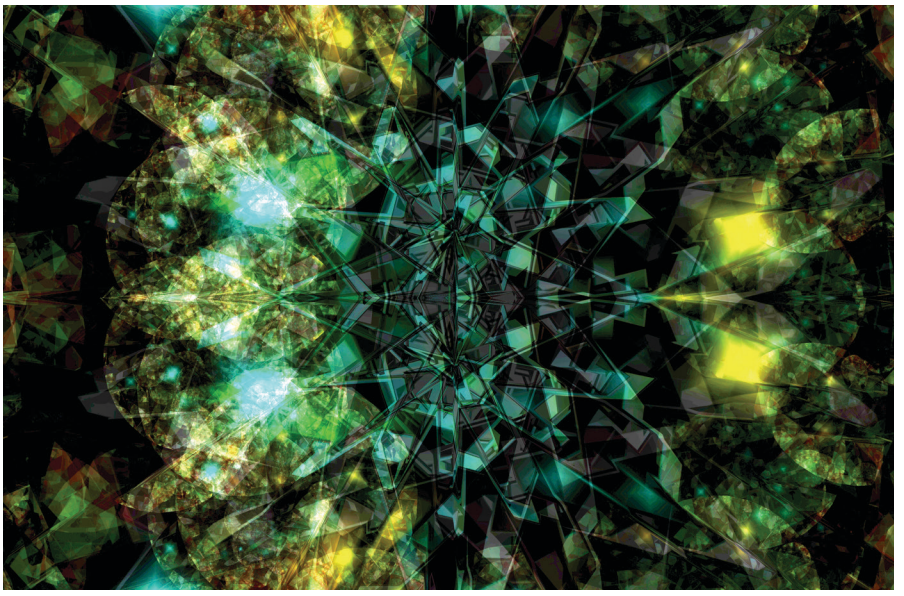


IMAGE BY GRADIENT BACKGROUND

**AI could be utilized to revolutionize ecosystem conservation and description by analyzing vast and varied data sources.**

Recognizing biodiversity is more than a matter of terminology discussed in seminars in natural history museums or outdated university departments. It is estimated there nearly 100 million species on the planet, but only approximately two million have been formally described. To a somewhat shocking degree, we do not even know what we do not know. Among the many other benefits, understanding the richness of this unknown biota can represent an economic asset, benefiting pharmacological and medical industries as well as serving as a cornerstone for deep tech companies, which can explore biodiversity sustainably

Coming Next Month in COMMUNICATIONS

**The Importance of Distrust in Trusting Digital Worker Chatbots**

**Multisensory Experiences: Formation, Realization, and Responsibilities**

**Preprinting in AI Ethics: Toward a Set of Community Guidelines**

**Envision Recommendation on LLM-based Agent Platform**

**Autonomy 2.0: The Quest for Economies of Scale**

**The Impact of GenAI on Data Annotation Tasks**

**Panmodal Information Interaction**

**Does AI Prediction Scale to Decision Making?**

**Many Faces of Ad Hoc Transactions**

**FabToys: Plush Toys with Large Arrays of Fabric-Based Pressure Sensors**

Plus, the latest news about robots replacing guide dogs, how to measure AI, and more.

while respecting environmental integrity and the knowledge of indigenous and traditional populations. AI could be utilized to revolutionize ecosystem conservation and biodiversity description by analyzing vast and varied data sources, ranging from assemblages of fossilized trilobites and extinct dinosaurs to the myriad morphological attributes of a single insect wing.

Biological taxonomy—the science of identifying, describing, and classifying organisms—has a long heritage of using technology. For example, modern biologists use software for illustration and digital photography, and ecological, phylogenetic, and biogeographical analysis. Computational tools to aid the preparation of species descriptions date back to the 1970s. Dallwitz's program[2] for constructing identification keys is an example. Over the years, this system has evolved into DELTA (DEscription Language for TAxonomy), which serves as a comprehensive system for encoding species descriptions for computer processing. Computer-assisted biological taxonomy remains a prominent topic in the field.[6,8]

Today, the biologists' workflow is a dynamic blend of traditional methods and technological advancements. While the core principles of the activity continue to be rooted in meticulous human-based observation and classification (fieldwork, specimens collection and mounting, manually species identification, collection curation), the integration of digital tools has streamlined and enhanced the process. Considering the advance of generative AI (GenAI), we have all the ingredients to develop efficient and consistent AI-based routines that will replace systems such as DELTA in species recognition and description, allowing the gain of precision and comparability and accelerating the process of biodiversity recognition and documentation.

AI has already made significant strides in the field of biological taxonomy. Deep learning and computer vision allied to sensors have been used to validate image-based taxonomic identification and to develop public and curated reference databases.[3] Well-established machine learning approaches, such as convolutional neural networks (CNNs) and random forests, have helped recognize patterns from images and identify insect species.[4,7] We are currently investigating the power of Vision Transformer (ViT) methods[5] to identify and classify species, considering the intrinsic morphological complexity of insect groups, our target taxon. However, a gap exists between current computational approaches in biology and the state of the art in GenAI research, suggesting ample room for further advancement. From the biological point of view, computer scientists who understand the immensity of the issues related to diversity loss and climate change are greatly needed.

We face interesting opportunities when using GenAIs in semi-automated species description after photographs and illustrations, preparation of structured taxonomic papers from notes and information extracted from simple sheets, and construction of character lists for evolutionary and phylogenetic analyses. Nonetheless, some popular AI tools based on large language models (LLMs) such as ChatGPT and Bard/Gemini are not fine-tuned enough to allow scientifically accurate results, but the initial outputs are exciting. Actually, the current generation of LLMs can identify morphological body patterns in images, even when organisms are camouflaged in their natural habitat. However, they cannot definitively determine whether a specific entity belongs to a recognized species among a wide variety of biological groups, especially the most diverse ones, such as insects.

In standard taxonomic procedure, dichotomous identification keys are used by biologists to classify specimens in particular taxonomic categories (order, family, genus, and species,

**Today, the biologist's workflow is a dynamic blend of traditional methods and biological advancements.**

to name a few) based on observation under optical microscopes, scanning microscopes, and stereomicroscopes. This meticulous activity is time-consuming and not error-free. If efficient and accurate AI tools could be developed that are less prone to variation among human analysts, that could have a huge impact on the near future of biological taxonomy. As species identification is fundamental for diversity measurements used in environmental conservation strategies, as well as medical and epidemiological analyses, boosting efficiency in taxonomy is crucial in the contemporary context of climate change and its adverse consequences on natural environments.

An even more complex task is describing species from scratch. Given that the work of taxonomists to document new species necessarily involves high-definition photos, electronic micrographs, and illustrations followed by detailed morphological descriptions, the development of AI-based tools to recognize patterns in images, compare them with known species, identify new species and produce structured taxonomic descriptions, would significantly speed up the recognition of biota, especially in countries with few professional biologists and insufficient funding for basic research. The computational challenge involves the developers' recognition of the peculiarities of biological studies and the importance of detail beyond identifying general patterns.

Biological taxonomy is the first step toward the understanding of species' relationships and evolutionary history. Since a significant portion of the biota that once existed on the planet will never be known, gaps in the reconstruction of evolutionary trees are common. Data augmentation driven by GenAI could play a relevant role here precisely because, based on the recognition of morphological patterns in described species, they could generate new data to train models and suggest putative species that would help explain critical evolutionary transitions that have happened in the four billion years since the origin of life on Earth. Aided by AI, knowing the past of the planet's biota, notably the periods of mass extinctions in which a significant part of life disappeared, could allow us to build conceptual and practical tools to deal with the biodiversity crisis we are experiencing now.

**Any technological tool aimed at revolutionizing biodiversity studies must balance automation and human oversight.**

As most of the taxonomic research happens when sitting in front of a computer, regardless of the time spent in fieldwork or at the bench, the training of the next generation of biologists will have to consider AI's ubiquity. Despite the progress, biologists still approach the reliability of AI cautiously. Concerns linger regarding the possibilities of errors and inaccuracies in automated processes: The fear of an AI mishap leading to flawed taxonomy and subsequent academic repercussions is palpable. Valan et al.[9] provided questions and answers and a study case about how taxonomists can confidently use off-the-shelf CNNs. The main issue is that biologists did not fully accept this perspective. In this sense, while embracing technological advancements is essential to tackle the huge scale of the problem, we also need ways of automatizing activities without removing humans from the review and error correction processes. Any technological tool aimed at revolutionizing biodiversity studies must balance automation and human oversight, ensuring accuracy, reliability, and user trust.

The Anthropocene presents an unparalleled challenge to human civilization. The recent tragedy in the Brazilian state of Rio Grande do Sul, in which nearly 90% of the state's cities and two million people were affected by historic rains and floods, is a clear example of how ignoring serious environmental policies can have disastrous social, economic, and environmental consequences. Dealing with the current environmental crisis is pivotal for humanity's future, and the collaborative efforts of computer scientists and biologists are essential in this regard. The ability to solve biodiversity-related problems through computational thinking will depend on the developers' understanding of the biological contexts in which the problems exist. In a nutshell, training massive datasets and publishing appealing methods are not enough. In biological sciences, the debate about AI should transcend technical advancements alone. As we move forward, we need to ensure AI is a bridge rather than a barrier hindering our pursuit of understanding and preserving the natural world. C


**References**

1. Chien, A.A. GenAI: Giga$$$, terawatt-hours, and gigatons of $CO_2$. *Commun. ACM 66*, 8 (Aug. 2023); 10.1145/3606254
2. Dallwitz, M.J. et al. A flexible computer program for generating identification keys. *Systematic Biology 23*, 1 (Mar. 1974); 10.1093/sysbio/23.1.50
3. Høye, T.T. et al. Deep learning and computer vision will transform entomology. In *Proceedings of the National Academy of Sciences 118*, 2 (2021); 10.1073/pnas.2002545117
4. Ling, M.H. et al. Machine learning analysis of wing venation patterns accurately identifies sarcophagidae, calliphoridae, and muscidae fly species. *Medical and Veterinary Entomology 37*, 4 (2023); 10.1111/mve.12682
5. Liu, Z. et al. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceeedings of IEEE/CVF Intern. Conf. on Computer Vision (ICCV)* (2021); 10.1109/ICCV48922.2021.00986
6. Orr, M.C. et al. Taxonomy must engage with new technologies and evolve to face future challenges. *Nature Ecology and Evolution 5*, 1 (2021), 3–4; 10.1038/s41559-020-01360-5
7. Popkov, A. et al. Machine learning for expert-level image-based identification of very similar species in the hyperdiverse plant bug family miridae (hemiptera: heteroptera). *Systematic Entomology 47*, 3 (2022); 10.1111/syen.12543
8. Santos, C.M.D. and Gois, J.P. Harnessing the power of AI language models for taxonomy and systematics: A follow-up to 'Can ChatGPT be leveraged for taxonomic investigations? Potential and limitations of a new technolog'. *Zootaxa 5297*, 3 (2023); 10.11646/ZOOTAXA.5297.3.9
9. Valan, M., Vondráč ek, D., and Ronquist, F. Awakening a taxonomist's third eye: Exploring the utility of computer vision and deep learning in insect systematics. *Systematic Entomology 46*, 4 (2021); 10.1111/syen.12492
10. Wolkovich, E.M. Obviously ChatGPT: How reviewers accused me of scientific fraud. *Nature* (2024); 10.1038/d41586-024-00349-5



**Charles Morphy D. Santos** (charles.santos@ufabc.edu.br) is an associate professor at the Universidade Federal do ABC, Brazil.

**João Paulo Gois** (joao.gois@ufabc.edu.br) is an associate professor at the Universidade Federal do ABC, Brazil.



This study was partly financed by CNPq (304027/2022-7, CMDS). We thank *Communications* Opinions co-chair Jeanna Matthews for her valuable suggestions and insightful feedback, which helped improve this manuscript. We used Grammarly, based on LLMs, for text editing in the original manuscript submission.

DOI:10.1145/3714417

Jack West, Jingjie Li, and Kassem Fawaz

# Opinion
# A Glimpse into the Pandora's Box

*Demystifying on-device AI on Instagram and TikTok.*

IMAGE BY ANDRIJ BORYS ASSOCIATES WITH ASSISTANCE OF SHUTTERSTOCK.AI

DEBATES SURROUNDING SOCIAL media have never stopped since its early age two decades ago. This is largely because how the social media "magic" works still remains mysterious to the public. Have you wondered what happens when you open the camera using Instagram and TikTok on your smartphone? We wondered the same circa September 2022.

Although tech companies promise that "what you do on your device will stay on the device," our prior finding reveals a leeway here: It did not stop companies from processing user data, for example, analyzing image frames on-device and sharing extracted features to their servers. For example, our initial study in 2022 found CISCO Webex was extracting audio metrics from a videoconference call while the user was muted[8]—a practice they stopped after we reported our findings.[5] Since then, there have been significant advances in artificial intelligence (AI) techniques and computation resources on smartphones. Our hypothesis was that nothing prevents an app from applying similar AI models to cellphone cameras, on which users have little awareness and control. This column describes how we investigated this hypothesis over the subsequent two years.

Before we get into details, we must answer a question: Why is it important to know what the apps do if the raw data (for example, camera frames, user images, live audio) never leaves the device? When tech companies first deployed AI models, they lived on the cloud, requiring user data to leave the device. These methods inadvertently created a clear separation between user data and AI models. As these models were part of the cloud, users had nothing to worry about if their data stayed on the device.Recently, AI has migrated to our devices, making predictions locally.[2] The migration of AI models from the cloud to user devices is natural. Local models no longer take resources from the cloud, algorithms can now be dynamic and adapt to user interests, and users' data remains on their devices. However, AI now has direct and immediate access to user data, so the clear separation that once existed is gone. Because of that, local AI models have more access to user data than ever. Before local AI processing, applications could not examine camera frames without them leaving the device. Now, unbeknown to the user, an AI model can analyze camera frames in real time, extracting sensitive features and concepts. We argue that users have the right to know what information apps extract from their data. Increased transparency can better inform social media users of model capabilities and potential risks, including identity misrepresentation.

The trend of migrating AI algorithms to devices makes it possible to understand how they process user data. In

September 2022, we started analyzing the two most popular social media apps on Android: TikTok and Instagram. We thought that apps would call APIs from popular libraries. So, we decompiled the apps to look for these APIs, but we did not find them. We tried applying some of the proposed techniques for local model extraction,[6] which did not help either. It turns out we needed to dig deeper because Instagram and TikTok employ sophisticated code obfuscation with low-level execution of AI methods. We developed new reverse-engineering techniques to determine when AI processes occur, what the inputs and outputs from the model were, and what happened to the outputs. In particular, we created our custom operating system that tracked all activities done by each application. We then used the application as usual and looked for evidence of AI processes. Once found, we performed dynamic instrumentation to interact with the models directly. We found two computer-vision AI models for both Instagram and TikTok.[7]

TikTok triggers a vision model while the camera is open, and the camera sees a face. To be specific, we found that TikTok's model feeds every single frame from the camera to their local model, extracting demographic information about the user. TikTok estimates the user's age and gender and draws a bounding box around the user's face. The outputs from the model are then written to an encrypted file. We found no evidence that the data left the device. Since the model's outputs are tied to the existence of a face within a camera frame, we showed TikTok hundreds of thousands of faces from the FairFace dataset. Each face is labeled with a gender, an age, and a racial demographic. We found that TikTok's model commonly overestimates the age of children. The model's average age estimation for babies and toddlers was 13 (TikTok's minimum age for making an account).[7] We also found that gender identification was highly inaccurate for Black individuals.[7] If these values were to be used for age verification, children would be able to easily bypass the model. No mention of this risk is expressed in their current privacy policy.[a]

Instagram, on the other hand, executes a vision model when the user selects an image to be uploaded as a Reel.[b] The user does not need to post the image to trigger the AI process: The model consistently executes upon selecting an image. The input to the model is the chosen image, and the output is over 500 different concepts. The concepts vary widely; there are landmarks (Washington Monument, Great Wall of China), facial features (beard, blonde), and objects (menorah, crucifix, ball). To evaluate risks to users, we constructed a synthetic dataset to measure the biases associated with different racial demographics. Using this custom-made dataset, we evaluated which racial demographics have higher scores than others for the various concepts. For example, we found that an Asian woman is highly correlated with the Great Wall of China, and a White woman is similarly correlated with blonde.[7] Our work demonstrated that if these models were used for algorithmic decision making, they would exhibit a significant bias. Due to the models' black-box nature, the consequences of the biases are unknown.[1] However, this lack of understanding has not impeded AI development. After publishing our work, we reanalyzed Instagram and found a new model with more than 1,500 concepts. There appears to be no slowing down for AI on local devices.

**The trend of migrating AI algorithms to devices makes it possible to understand how they process user data.**

How do the users feel about these models? To answer this question, we performed a systematic user study of social media users to understand how these models impacted their usage of the apps. We recruited 21 Instagram and TikTok users and interviewed them about their perceptions and understandings. We first asked about their knowledge of AI, led participants through a lesson about computer vision (so that everyone has similar understanding), and then revealed the models. Then, we asked them how they understood AI and if exposure to the AI models used by Instagram and TikTok fit into their current understanding. Two weeks later, we asked the participants if their usage of Instagram and TikTok changed.

Participants were shocked by Instagram and TikTok in different and interesting ways. They indicated TikTok's algorithm for processing images is non-consensual and invasive. Immediately using AI without any indication felt wrong to them. Participants also reflected on new fears of accidentally opening the camera and the model processing sensitive images. Their reaction to TikTok revealed a significant gap in user understanding of AI capabilities. Participants assumed they could infer AI behaviors based on the app's speed or believed that AI was tied to specific interactions they had with the apps. For example, most participants thought what they liked on TikTok was where the bulk of the AI processing took place, not their raw images. We found that several parents who participated in our study were concerned about what the app could infer about their children. One parent was offended that the app was inferring gender as their child was non-binary. All participants expressed they were unaware of this possibility and wanted more transparency from the application.

Participants were more positive toward Instagram because they liked that they could control the model's execution. Because a user had to select an image, it felt more consensual, as they could directly connect an action to the AI. This control gave participants agency when interacting with Instagram, which was noted positively. However, participants expressed negative feelings toward the amount of data gathered from their single image. Due to the sheer amount of concepts that Instagram's model produced, it was hard to understand the purpose of the data. This confusion forced participants to infer the meaning based on prior assumptions of the app. Participants were skeptical even when informed that the

a Privacy Policy (TikTok); https://bit.ly/4g79688

b Privacy Policy (Instagram); https://bit.ly/3E4E7fD

data may not leave or be used by the device at all. To them, their data was why social media was free; thus, why would Instagram make a model and not use the data?

Overall, all participants expressed that they wanted apps to be more transparent with how they used AI so they could avoid it if they desired. Out of the 21 participants we interviewed, eight reported declining usage due to our intervention. For example, parents reported they informed their children of the risks, and they disabled the camera permission. Others said they were more aware about where they use Instagram and TikTok. One participant uninstalled TikTok after our study. It appears that making users more aware of how apps process their data affects their behavior. This begs the question: How should apps provide more transparency about their internal AI processing?

How applications should inform users about their AI use is not clear. Do the users need yet another privacy notification, menu, permission, popup, or policy? Not necessarily. Still, we can leverage the recent push for privacy or data-safety labels for mobile apps.[3] These labels, modeled after privacy nutrition labels, are supposed to be a concise and simple way to summarize an app's privacy practices. At a high level, they cover types and purposes for data collection and sharing. We propose that these sections be amended with practices around local AI processing. Here, there are several challenges the research community, developers, and platform providers must come together to address.

The first challenge is what data to include in these newly created labels. They can discuss the data the model analyzes, the analysis frequency, and what triggers the analysis. Also, they can include the model outputs and purposes. Given the increasing popularity of local language models, which enable reconfigurable analytics at runtime, we envision a combination of measures, in particular compliance enforcement and model certificates, should be developed and employed in conjunction with the safety label. In addition, a model benchmark can be useful for users and developers, providing insight into model performance in different contexts. The second challenge is how to display all this information in the label format in an accessible and informative way. AI-related concepts are complex, and users might not fully understand their ramifications. As such, interfaces should prioritize user understanding in how they present AI-related information. Personalized interfaces will help ensure that even non-expert users can comprehend the nature of the processing and its implications. Finally, a critical challenge is who is responsible for creating and auditing these labels. Developers face challenges in completing privacy labels in apps,[4] and we foresee similar challenges for these AI cards. More importantly, AI and social media governance, including the deployment of the AI safety label, is a global effort. However, it is unrealistic to believe that policymakers across the world will reach their consensus sometime soon, as evident again by the recent episode of the U.S. TikTok ban and social media "refugees."[c] The unstoppable growth of AI capabilities leaves researchers with an open question: How can we inform and empower users to navigate the chaos before standardized safety measures are put in place? C

c See "Chinese app RedNote gained millions of U.S. users this week as 'TikTok refugees' joined ahead of ban"; https://bit.ly/4gc85fg

**References**

1. John-Mathews, J.-M. Critical empirical study on black-box explanations in AI. In *Proceedings of ICIS 2021: 42nd Intern. Conf. on Information Systems* (2021).
2. Kaye, K. Why AI and machine learning are drifting away from the cloud. *Protocol.com*; (Aug. 2022); https://bit.ly/4hBA09A
3. Kelley, P.G., Cranor, L.F., and Sadeh, N. Privacy as part of the app decision-making process. In *Proceedings of the SIGCHI Conf. on Human Factors in Computing Systems*. ACM (2013); 10.1145/2470654.2466466
4. Khandelwal, R. et al. Unpacking privacy labels: A measurement and developer perspective on google's data safety section. In *Proceedings of the 33rd USENIX Security Symp. (USENIX Security 24)* (2024).
5. Kulkarni, A. Innovation behind Webex mute. *Webex Blog* (Apr. 2022); https://bit.ly/3PHr7zj
6. Sun, Z. et al. Mind your weight (s): A large-scale study on insufficient machine learning model protection in mobile apps. In *Proceedings of the 30th USENIX Security Symp. (USENIX Security 21)* (2021).
7. West, J. et al. A picture is worth 500 labels: A case study of demographic disparities in local machine learning models for Instagram and TikTok. In *Proceedings of the 2024 IEEE Symp. on Security and Privacy (SP)*. IEEE Computer Society (2024).
8. Yang, Y. and West, J. Are you really muted?: A privacy analysis of mute buttons in video conferencing apps. In *Proceedings on Privacy Enhancing Technologies* (2022).

**Jack West** (jwwest@wisc.edu) is a Ph.D. student at the University of Wisconsin-Madison, Madison, WI, USA.

**Jingjie Li** (jingjie.li@ed.ac.uk) is an assistant professor at the University of Edinburgh, Edinburgh, Scotland.

**Kassem Fawaz** (kfawaz@wisc.edu) is an associate professor at the University of Wisconsin-Madison, Madison, WI, USA.

# XRDS


## *XRDS* Magazine Seeks Student Editor-in-Chief

In this volunteer position, the EIC leads and works with fellow students and ACM professional staff to produce visionary issues that inspire, inform, and educate computing students across the globe.

*XRDS* highlights new and important research, interviews, roundtable discussions, opinion pieces, tutorials, and more—written by top leaders in the field as well as students interested in sharing their ideas with an international audience.

Apply and find more info at https://bit.ly/xrds_eic

**Association for Computing Machinery**

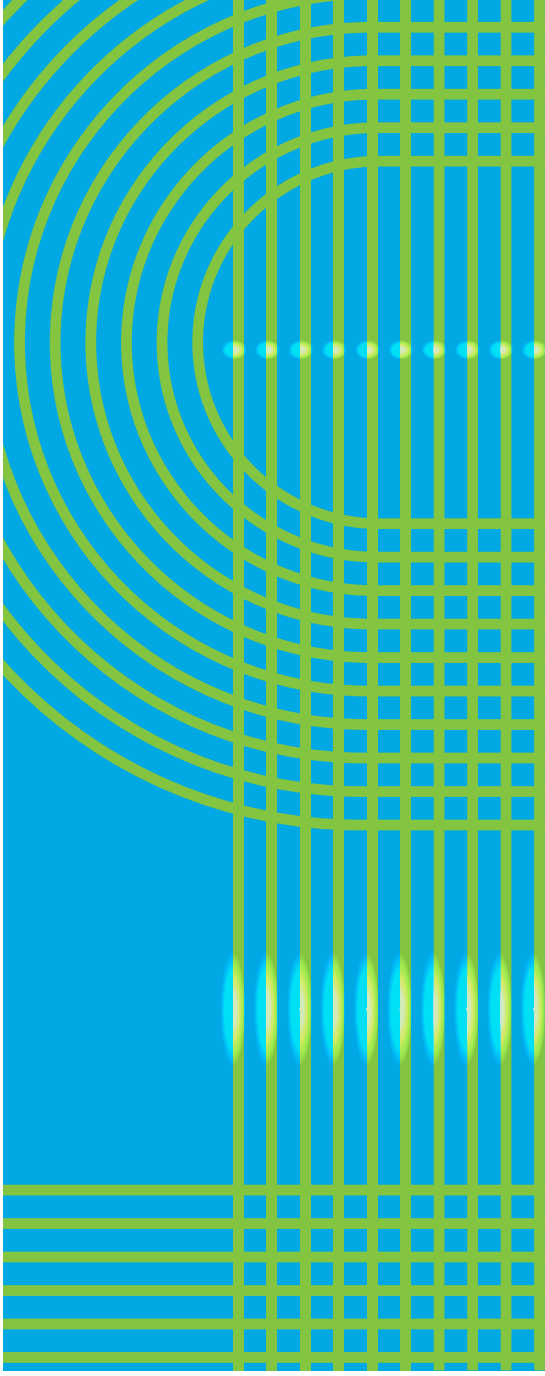
DOI:10.1145/3704254

***Queue* speaks with engineers from Microsoft Research about using machine learning to merge code.**

BY SHUVENDU K. LAHIRI, ALEXEY SVYATKOVSKIY, CHRISTIAN BIRD, ERIK MEIJER, AND TERRY COATTA

# Program Merge:

## What's Deep Learning Got to Do with It?



IF YOU REGULARLY work with open-source code or produce software for a large organization, you are already familiar with many of the challenges posed by collaborative programming at scale. Some of the most vexing of these tend to surface as a consequence of the many independent alterations inevitably made to code, which, unsurprisingly, can lead to updates that do not synchronize.

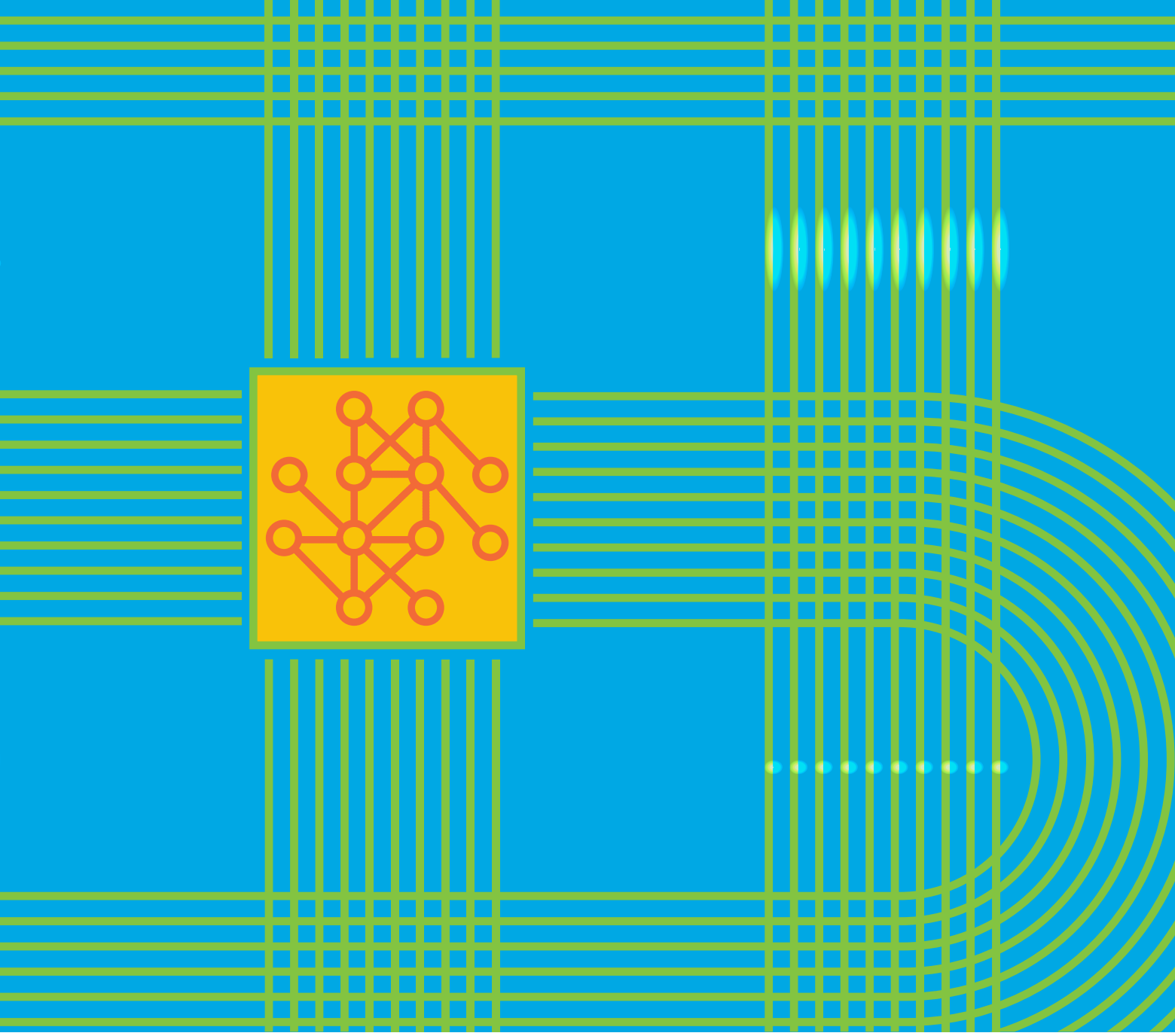
Difficult merges are nothing new, of course, but the scale of the problem has gotten much worse. This is what led a group of researchers at Microsoft Research (MSR) to take on the task of complicated merges as a grand program-repair challenge—one they believed might be addressed at least in part by machine learning (ML).

To understand the thinking that led to this effort and then follow where that led, *ACM Queue* asked Erik Meijer and Terry Coatta to speak with three of the leading figures in the MSR research effort called DeepMerge.[a] Meijer was long a member of MSR, but at the time of this discussion was director of engineering at Meta. Coatta is the chief technology officer of Marine Learning Systems. Shuvendu Lahiri and Christian Bird, two of the researchers who helped drive this effort, represent MSR, as does Alexey Svyatkovskiy, who was with Microsoft DevDiv (Development Division) at the time.

**TERRY COATTA:** *What inspired you to*

a https://tinyurl.com/28k32fok

IMAGE BY ANDRIJ BORYS ASSOCIATES. USING SHUTTERSTOCK.COM



*focus on merge conflicts in the first place? And what made you think you'd be able to gain some advantage by applying AI techniques?*

**CHRISTIAN BIRD:** Back in the winter of 2020, some of us started talking about ways in which we might be able to use machine learning to improve the state of software engineering. We certainly thought the time was right to jump into an effort along these lines in hopes of gaining enough competency to launch into a related research program.

We tried to identify problems other researchers weren't already addressing, meaning that something like code completion—which people had been working on for quite some time—was soon dismissed. Instead, we turned to problems where developers didn't already have much help.

Shuvendu [Lahiri] has a long history of looking at program merge from a symbolic perspective, whereas my own focus has had more to do with understanding the changes that occur in the course of program merges. As we were talking about this, it dawned on us that almost no one seemed to be working on program merge. And yet, that's a problem where we, as developers, still have little to rely upon. For the most part, we just look at the diffs between different generations of code to see if we can figure out exactly what's going on. But there just isn't much current tooling to help beyond that, which can prove to be problematic whenever there's a merge conflict to resolve.

So, we figured, "OK, let's look at how some deep-learning models might be applied to this problem. As we go along, we'll probably also identify some other things we can do to build on that."

**SHUVENDU LAHIRI:** Yes, as Chris suggests, I've been thinking about the issues here for quite some time. Moreover, we found program merge to be appealing, since it's a collaboration problem. That is, even if two skilled developers make correct changes, the merge itself may introduce a bug.

We were also keenly aware of the sort of pain program-merge problems can cause, having known about

ALEXEY SVYATKOVSKIY

**While merge conflicts are far less common than software bugs, they require considerably more time to resolve and can end up causing more pain.**

it through studies within Microsoft.[b] I thought maybe there was something we could do to provide relief. It also turns out that AI was just coming along at that point, and Alexey [Svyatkovskiy] had already developed a couple of powerful models that looked quite promising for code completion. What's more, information about merge conflicts was just starting to become more readily available from the Git commit history, so that too looked like it might serve as a good up-front source of clean data.

**ERIK MEIJER:** *I like the fact that you focused on merge conflict, since, when it comes to this, I don't think source control solves any of the real problems. Maybe I'm being a little extreme here, but even if source control lets you know where you have a merge conflict, it won't help you when it comes to actually resolving the conflict. In fact, I'm baffled as to why this problem wasn't solved in an intelligent manner a long time ago. Are people just not listening to complaints from actual users?*

**LAHIRI:** Basically, I think it comes down to academicians consistently resorting to symbolic methods to solve this problem. Whereas people in the real world have looked at this as just another aspect of programming, practitioners have been more inclined to approach it as a social process—that is, as a problem best addressed by encouraging co-workers to figure out solutions together. Personally, I've always seen merge conflicts as more of a tooling challenge.

**ALEXEY SVYATKOVSKIY:** For me, this just looked like an exciting software engineering problem to address with machine learning. I've spent years working on code completion, but this effort looked like something that would take that up to the next level of complexity, since it necessarily would involve aligning multiple sequences somehow and then complementing that with an understanding of where things ought to be inserted, deleted, or swapped. And, of course, there were also those special cases where the developer would be able to add new tokens during the merge.

This took us on a journey where we ended up addressing program merge down at the line-and-token level. I found this fascinating, since a lot of people don't have any idea about how merge actually works and so, by extension, don't have a clear understanding about what leads to merge conflicts. Taking on this problem also seemed important in that, while merge conflicts are far less common than software bugs, they require considerably more time to resolve and can end up causing more pain.

**COATTA:** *How did you initially attack the problem?*

**LAHIRI:** We realized that repositories (both open source ones on GitHub and internal ones at Microsoft) contain data on merge conflicts and their resolution across several different programming languages. What's more, Alexey had recently created a neural model that had been pre-trained on a large subset of the code in those repositories. Our initial thought was to fine-tune that model with information about merge conflicts and their resolution. And we figured that should be simple enough to be treated as an intern project. So, that's how we scoped it initially. We figured: Just get the data, train a model, and deploy. What we failed to grasp was that, while there was an ample amount of program merge data to be mined, coming to an understanding of what the intent behind all those merges had been was at least as important as the data itself. In fact, it proved to be absolutely critical.

A considerable amount of time and effort was required to understand and interpret the data. This included some significant technical challenges—for example, how best to align these programs. And how can you communicate to a neural model that these are not independent programs but instead represent some number of changes to an underlying program? The notion of how to go from program text to a program edit became quite crucial and, in fact, required considerable research. Ultimately, we concluded that if you manage to combine your existing merge tools correctly, add just the right amount of granularity—which for us proved to be tokens—and then employ neural modeling, you can often succeed in reducing the complexity. But it took us quite a bit of time to work that out.

b https://tinyurl.com/2afzj2gc

Of course, we also underestimated the importance of user experience. How exactly would a user end up employing such a tool—one that's AI-based, that is? And what would be the right time to surface that aspect of the tool?

**COATTA:** *I find it fascinating that it proved to be so difficult to scope this project correctly. Can you dig a bit deeper into that?*

**SVYATKOVSKIY:** To me, at least, as we were analyzing the different types of merges, it soon became clear that there are varying levels of complexity. Sometimes we'd find ourselves looking at two simple merge resolution strategies, where it essentially came down to "Take ours or take theirs." Such cases are trivial to analyze, of course, and developers don't require much AI assistance when it comes to resolving these conflicts.

But then there's another class of merge, where a new interleaving line is introduced that involves more than just concatenation. There could also be token-level interleaving, where lines in the code have been broken and new tokens introduced in between. This leads to the notoriously complex case where a switch to token-level granularity proves to be crucial. Beyond that, there's a whole other class of merges where you find somebody has introduced some new tokens.

**MEIJER:** *How do you go about defining what you consider to be a correct merge? Doesn't that require you to make your own value judgments in some sense?*

**LAHIRI:** Well, I'll just say we had a very semantic way of looking at merges. Essentially: "Forget about the syntax; instead, what does it *mean* for the merge to be correct?" In effect, this amounts to: If something was changed in one program, then that ought to be reflected in the merge. And, if that also changes a behavior, then that too ought to be included in the merge. But no other changes or altered behaviors should be introduced.

We then found, however, that we could get tangled up whenever we ran into one of these "take my changes or take yours" merges. We also found that one set of changes would often just be dropped—like a branch being deprecated, as Alexey once pointed out. This is how we discovered that our initial notion of correctness didn't always hold. That's also how we came to realize we shouldn't adhere to overly semantic notions.

So, we decided just to do our best to curate the training data by removing any indications of "take your changes or take mine" wherever possible. Then we looked at those places where both changes had been incorporated to some extent and said, "OK, so this now is our ground truth—our notion of what's correct." But notice that this is *empirical* correctness as opposed to *semantic* correctness—which is to say, we had to scale back from our original high ambitions for semantic correctness.

**SVYATKOVSKIY:** We now treat user resolutions retrieved from the GitHub commit histories as our ground truth. But yes, naturally, there are all kinds of ways to define a "correct" merge. For example, it's possible to reorder the statements in a structured merge and yet still end up with a functionally equivalent resolution. Yet, that would be deemed as "incorrect," so, there's clearly room for retooling our definition of correctness. In this instance, however, we chose to take a data-driven approach that treats user resolutions from the GitHub commit histories as our ground truth.

**BIRD:** Right. And let me also say that, from the beginning of this project, we decided to approach it as something that might yield a product. With that in mind, we realized it needed to be, perhaps not language-agnostic, but at least something that could be readily adapted to multiple languages—and definitely *not* something that would require some bespoke analysis framework for each language. That essentially guided our choice not to employ richer or more complex code representations.

**MEIJER:** I've also run into situations like this where it looked really tempting to use an AST [abstract syntax tree] or something of the sort, since that would provide all the structure that was required. But then, as you go deeper into that sort of project, you find yourself wondering whether it's actually a good idea to feed semantically rich programs into models and start thinking it might be better just to send strings instead.

**COATTA:** *To dive a bit deeper into that, you had a practical motivation to work*

CHRISTIAN BIRD

**It dawned on us that almost no one seemed to be working on program merge. And yet, that's a problem where we, as developers, still have little to rely upon.**

SHUVENDU LAHIRI

**Even after a merge has been produced, someone might want to know why the merge was accomplished in that particular way and may even want to see some evidence of what the reasoning was there. Well, that's a thorny issue.**

*with a token-based approach. But what does your intuition tell you about how models behave when you do that? If you feed a model a rich, structured information set, is that model then actually more likely to make better decisions? Or is that perhaps a false assumption?*

**SVYATKOVSKIY:** The model ought to be able to make better decisions, I think.

**MEIJER:** *All right, but can I challenge that a bit? The model can handle the syntax and semantic analysis internally, which suggests this work may not need to be done ahead of time, since machines don't look at code the same way humans do. I don't know why the model couldn't just build its own internal representation and then let a type-checker come along at the end of the process.*

**BIRD:** I think it's hazardous to speculate about what models may or may not be capable of. I mean, that's a supernuanced question in that it depends on how the model has been trained and what the architecture is like—along with any number of other things. I'm constantly surprised to learn what models are now capable of doing. And, in this case, we're talking about the state of the world back in 2020—even as I now find it hard to remember what the state of the world looked like six months prior to the GPT models becoming widespread.

**LAHIRI:** For one thing, we were using pretrained models to handle classification and generation, which then left us with quite a bit of work to do in terms of representing the resulting edits at the AST level before tuning for performance. That certainly proved to be a complex problem—and one that came along with some added computational costs. Also, as I remember it, the models we were using at the time had been trained as text representations of code—meaning we then needed to train them on a lot more AST-level representations to achieve better performance. I'm sure it would be fascinating to go back to revisit some of the decisions we made back in 2020.

**MEIJER:** *What model are you using now?*

**SVYATKOVSKIY:** For this iteration, we're employing a token-level merge along with a transformer-based classifier. We've also been looking at using a prompt-driven approach based on GPT-4.

**MEIJER:** *I love that this is now something where you can take advantage of demonstrated preferences to resolve merge conflicts instead of being left to rely solely on your own opinions.*

**LAHIRI:** Another way of looking at this that came up during our user studies was that, even after a merge has been produced, someone might want to know why the merge was accomplished in that particular way and may even want to see some evidence of what the reasoning was there. Well, that's a thorny issue.

But one of the nice things about these large foundational models is that they're able to produce textual descriptions of what they've done. Still, we haven't explored this capability in depth yet, since we don't actually have the means available to us now to evaluate the veracity of these descriptions. That will have to wait until some user studies supply us with more data. Still, I think there are some fascinating possibilities here that ultimately should enable us to reduce some of the friction that seems to surface whenever these sorts of AI power tools are used to accomplish certain critical tasks.

In the event you regularly work with open source code, you are surely already familiar with some of the challenges that can arise in the course of trying to resolve merge conflicts. Many of these problems have been encountered for as long as people have collaborated on programs, and these have metastasized as the scale and complexity of software has multiplied many times over. Also, with thousands of developers sometimes now collaborating on projects, the potential for conflicts only continues to soar.

Many of these are conflicts that can lead to program failures, of course. But even worse, in some respects, are the more subtle semantic merge conflicts that can fail the compiler, break a test, or introduce a regression. Despite these painfully obvious problems, the program merge issue has been largely left to fester for decades simply because the challenge of addressing it has seemed so daunting.

**COATTA:** *You've mentioned that you had access to a vast amount of training data, but you've also suggested some of*

*that data contained surprises—which is to say it proved to be both a blessing and a curse. Can you go into that a bit more?*

**LAHIRI:** Yes, we were surprised to find that a large percentage of the merges—perhaps 70%—had the attribute of choosing just one side of the edit and then dropping the other. In some of those cases, it seemed one edit was superseding the others, but it can be hard to be sure whenever the syntax changes a little. In many instances, there were genuine edits that had been dropped on the floor. It was unclear whether that was due to a tooling problem or a social issue—that is, in some cases, perhaps some senior developer's changes had superseded those that had been made by a junior developer. Another hypothesis was that, instead of a single merge, some people may have chosen to merge in multiple commits.

This sort of thing was so common that it accounted for a significant portion of the data, leaving us uncertain at first as to whether we should throw out these instances, ignore them, or somehow make an effort to account for them. That certainly proved to be one of the bigger surprises we encountered.

Another surprise was that we discovered instances where some new tokens had been introduced that were irrelevant to the merge. It was unclear at first whether those were due to a genuine conflict in the merge or just because somebody had decided to add a pretty print statement while doing the refactoring. That proved to be another thorny issue for us.

**COATTA:** *How did you resolve that? It sounds like you had some datasets you didn't quite know how to interpret. So, how did you decide what should be classified as correct merges or treated as incorrect ones?*

**LAHIRI:** We curated a dataset that did not include the "trivial" merge resolutions, with the goal of assisting users with the more complex cases first. As Alexey mentioned, users may not need tooling support for those resolutions that only require dropping one of the two edits.

**SVYATKOVSKIY:** And then, from user studies, we learned that some users still wanted to be able to use the approach that had been dismissed. We solved that problem by providing a "B option" that people could get to by using a drop-down menu.

**LAHIRI:** Which is to say we addressed the problem by way of user experience rather than by changing the model.

The other data problem we encountered had to do with new tokens that would occasionally appear. Upon closer examination, we found these tokens were typically related to existing changes. By going down to token-level merges, we were able to make many of these aspects go away. Ultimately, we built a model that excluded that part of the dataset where new tokens were introduced.

**MEIJER:** *In terms of how you went about your work, I understand one of the tools you particularly relied on was Tree-sitter [a parser-generator tool used to build syntax trees]. Can you tell us a bit about the role it played in your overall development process?*

**BIRD:** We were immediately attracted to Tree-sitter because it lets you parse just about anything you can imagine right off the shelf. And it provides a consistent API, unlike most other parsers out there that each come with their own API and work only with one language or another.

For all that, I was surprised to learn that Tree-sitter doesn't provide a tokenizing API. As an example of why that proved to be an issue for us, we wanted to try Python, which basically lets everyone handle their own tokenizing. But, of course, Tree-sitter didn't help there. We resorted to a Python tokenizing library.

Beyond that relatively small complaint, Tree-sitter is great in terms of letting you apply an algorithm to one language and then quickly scale that up for many other languages. In fact, between that capability and the Python tokenizing library, which made it possible for us to handle multiple languages, we were able to try out things with other languages without needing to invest a lot of upfront effort. Of course, there's still the matter of obtaining all the data required to train the model, and that's always a challenge. At least we didn't need to write our own parsers, and the consistent interfaces have proved to be incredibly beneficial.

**MEIJER:** *Once you finally managed to get all this deployed, what turned out to be your biggest surprise?*

SHUVENDU LAHIRI

**We were surprised to find that a large percentage of the merges—perhaps 70%—had the attribute of choosing just one side of the edit and then dropping the other.**

CHRISTIAN BIRD

**There are lots of moving pieces in any given merge. Accordingly, there are many possible views, and yet you still want to keep things simple enough to avoid overwhelming the user. That's a real challenge.**

**BIRD:** There were so many surprises. One I particularly remember came up when we were trying to figure out how people would even want to view merge conflicts and diffs. At first, some of us thought they'd want to focus only on the conflict itself—that is, with a view that let them see both their side and the other side. It turns out you also need to be able to see the base to understand the different implications between an existing branch in the base and your branch.

So, we ran a Twitter survey to get a sense of how much of that people thought we should show. How much of that did they even *want* to see? For example, as I recall, most people couldn't even handle the idea of a three-way diff, or at least weren't expecting to see anything quite like that. That really blew my mind, since I don't know how anyone could possibly expect to deterministically resolve a conflict if they don't know exactly what they're facing.

Some other issues also came up that UI people probably would expect, but I nevertheless was incredibly surprised. That proved to be a big challenge, since we'd been thinking throughout this whole process that we'd just get around to the UI whenever we got around to it. And yes, as this suggests, our tendency initially was just to focus on making sure the underlying algorithm worked. But then we found to our surprise just how tough it could be to find the right UI to associate with that.

**COATTA:** *From what you say, it seems you weren't surprised about the need for a good user experience, but it did surprise you to learn what's considered to be a good experience. What are your thoughts now on what constitutes a good user experience for merge?*

**BIRD:** I'm not entirely clear on that even now, but I'll be happy to share some of the things we learned about this early on. As we've already discussed, people definitely want to see both sides of a merge. Beyond that, we discovered that they want the ability to study the provenance of each part of the merge because they want to know where each token came from.

So, we wrote some code to track each token all the way back to whichever side it came from.

There also were tokens that had come in from both sides. To make it clear where a token had originated, we wrestled with whether we should add colors as an indicator of that. How might that also be used to indicate whether a token happens to come from both sides or simply is new?

In addition, we knew it was important that the interface didn't just ask you to click "yes" or "no" in response to a suggested change, since it's rare to find any merge that's going to be 100% correct. Which is to say developers are going to want to be able to modify the code and will only end up being frustrated by any interface that denies them that opportunity.

The real challenge is that there are lots of moving pieces in any given merge. Accordingly, there are many possible views, and yet you still want to keep things simple enough to avoid overwhelming the user. That's a real challenge. For example, we know that if we offer three suggestions for a merge rather than just one, the chance of the best one being selected is much higher. But that also adds complexity, so we ultimately decided to go with suggesting the most likely option, even though that might sometimes lead to less-optimal results.

There are some other user-experience considerations worth noting. For example, if you are working on some particular Visual Studio feature, you're going to want to produce something that feels intuitive to someone who has been using that same tool. Suffice it to say, there's plenty to think about in this respect. Basically, once you finally get your model to work, you might not even be halfway home, since that's just how critical—and time-consuming—the user-experience aspect of this work can be.

Yes, user experience actually does matter—even when the users happen to be developers. Accordingly, a substantial user study was launched in this instance, where the subjects of the study were members of MSR's own technical staff.

Another interesting aspect was that the study participants were presented with code samples extracted from their own work. The significance of this, of course, was that it involved the use of

not only real-world examples but also ones where all the trade-offs and implications associated with each important decision point were sure to be fully appreciated by the study subjects. Which is to say that the exercise proved to be an interesting learning experience for all parties involved.

**COATTA:** *We all know that creating a tool for internal purposes is one thing, while turning that into a product is something else altogether. It seems you took that journey here, so what were some of the bigger surprises you encountered along the way?*

**LAHIRI:** Actually, we don't have a product yet that implements the DeepMerge algorithm and aren't at liberty to talk about how that might be used in future products. Still, as we've just discussed, I can say most of the unusual challenges we encountered were related to various aspects of the user experience. So, we got much deeper into that here than we normally would.

One of the biggest challenges had to do with determining how much information needed to be surfaced to convince the user that what was just done was even possible—never mind appropriate. Suddenly, you've just introduced some new tokens over here, along with a new parsable tree over there. I think that can really throw some users off.

**BIRD:** What did all this look like from the DevDiv perspective, Alexey? You deal with customers all the time. What proved to be the biggest challenges there?

**SVYATKOVSKIY:** Some of the most crucial design decisions came down to choosing between client-side or server-side implementation. Our chief concern had to do with the new merge algorithm we were talking about earlier. Customer feedback obtained from user studies and early adopters proved to be particularly crucial in terms of finding ways to smooth things out. Certainly, that helped in terms of identifying areas where improvements were called for, such as achieving better symmetries between what happens when you merge A to B versus when you merge B to A.

**LAHIRI:** I'd like to add a couple of points. One is that some developers would prefer to handle these merges themselves. They just don't see the value of tooling when it's used to deal with something they could do themselves. But that just resulted in some inertia, which is always hard to overcome without a lot of usage. Still, from our empirical study we learned that, even when merges were not identical to the ground truth, users would accept them if they proved to be semantically equivalent. Ultimately, that proved to be a pleasant surprise, since it revealed we had previously been undercounting our wins according to our success metrics.

**COATTA:** *Did anything else interesting surface along the way?*

**BIRD:** At one point, one of our interns did a user study that pulled merge conflicts and their resolutions out of Microsoft's historical repositories so they could then be compared with the resolutions our tool would have applied. As you might imagine, quite a few differences surfaced. To understand where our tool may have gone wrong, we went back to consult with those people who had been involved in the original merges and showed them a comparison between what they had done and how the tool had addressed the same merge conflicts.

We specifically focused on those conflicts that had been resolved over the preceding three months on the premise that people might still recall the reasoning behind those decisions. We learned a ton by going through that particular exercise. One of the lessons was that we had probably undercounted how often we were getting things right, since some of these developers would say things like, "Well, this may not exactly match the merge I did, but I would have accepted it anyway."

The other major benefit of that study was the insight it provided into what the user experience for our tool should be. This all proved to be a major revelation for me, since it was the first time I'd been involved in a user study that was approached in quite this way—where developers were pulled in and presented with code they'd actually worked on.

Which is just to say this wasn't at all like one of those lab studies where people are presented with a toy problem. In this case, we were pulling in real-world merge conflicts and then talking with the developers who had worked to resolve them. We learned so much from taking this approach that I'd recommend other researchers consider doing their own studies in much the same way.

**SVYATKOVSKIY:** Another thing that came out of these user studies was the importance of explainability. With large language models, for example, we can drill into specific three-way diffs and their proposed resolutions and then ask for summaries of certain decisions, which can be helpful when it comes to building confidence in some of these AI suggestions.

Also, as Chris indicated, even when users chose not to go with the solution offered by DeepMerge, the reasoning behind the suggestion still seemed to inform their own thinking and often led to an improved merge resolution.

**COATTA:** *What's next?*

**LAHIRI:** There's room for more prompt engineering in terms of determining what goes into the model's input. We also need to address correlated conflicts. So far, we've addressed each conflict as if it was independent, but you can have multiple conflicts in a file that all relate to a certain dependency. Some users have told us that, once a resolution has been worked out for one of those conflicts, they'd like to see something similar applied to each of the other conflicts that exhibit a similar pattern, which certainly seems quite reasonable.

Also, while the types of conflicts we've addressed so far are highly syntactic in nature, there is, in fact, a whole spectrum of merge conflicts. There's still much to address, including silent merges that include semantic conflicts, which are much harder to deal with than anything we've handled so far. Still, I'd say it feels like we're off to a reasonably good start. C

**Terry Coatta** is chief technology officer of Marine Learning Systems, Vancouver, BC, Canada.

**Erik Meijer** is an independent researcher and entrepreneur in residence at Storm Ventures.

**Shuvendu K. Lahiri** is a senior principal researcher in the Research in Software Engineering (RiSE) Group at Microsoft Research, Redmond, WA, USA.

**Alexey Svyatkovskiy** is a researcher at Google Deep Mind, Seattle, WA, USA.

**Christian Bird** is a senior principal researcher in the Software Analysis and Intelligence in Engineering Systems (SAINTES) Group at Microsoft Research, Redmond, WA, USA.

DOI:10.1145/3685527

**Crowd workers often use LLMs, but this can have a homogenizing effect on their output. How can we—and should we—prevent LLM use in crowd work?**

BY VENIAMIN VESELOVSKY, MANOEL HORTA RIBEIRO, PHILIP J. COZZOLINO, ANDREW GORDON, DAVID ROTHSCHILD, AND ROBERT WEST

# Prevalence and Prevention of Large Language Model Use in Crowd Work

CROWD WORK PLATFORMS, such as Prolific and Amazon Mechanical Turk, play an important part in academia and industry, empowering the creation, annotation, and summarization of data,[11] as well as surveys and experiments.[21] At the same time, large language models (LLMs), such as ChatGPT, Gemini, and Claude, promise similar capabilities. They are remarkable data annotators[10] and can, in some cases, accurately simulate human behavior, enabling in silico experiments and surveys that yield human-like results.[2] Yet, if crowd workers were to start using LLMs, this could threaten the validity of data generated using crowd-work platforms. Sometimes, researchers seek to observe unaided human responses (even if LLMs could provide a good proxy), and LLMs still often fail to accurately simulate human behavior.[22] Further, LLM-generated data may degrade subsequent models trained on it.[23] Here, we investigate the extent to which crowd workers use LLMs in a text-production task and whether targeted mitigation strategies can prevent LLM use.

## Study 1: Prevalence of LLM Use

To estimate LLM use on Prolific, a research-oriented crowd-work platform, we asked n = 161 workers to summarize scientific abstracts (following Ribeiro et al.;[15] see Appendix online). We chose this task because it is laborious for humans but easily done by LLMs[17] and because it allowed us to use pre-LLM summaries from prior work[15] as "human ground truth." We detected whether a summary had been generated using LLMs with a fine-tuned e5-base classifier[28] trained on human, pre-LLM summaries[15] and summaries generated by GPT-4 and ChatGPT. The model was then run on each of the 161 new summaries to estimate its probability of

## » key insights

- **LLMs are widely used in crowd work. We also find that responses written with the help of LLMs are high-quality but more homogeneous than those written without LLMs' help.**
- **LLM use by crowd workers compromises research on human behavior, preferences, and opinions. Our results indicate we must find ways to prevent inappropriate LLM use or appropriately incorporate LLMs into crowd workers' workflows.**
- **LLM use can be diminished by adding hurdles that complicate their use or by asking crowd workers not to use LLMs. However, LLM use remained common in our field experiments even with these mitigation strategies.**

ILLUSTRATION BY PETER GRUNDY

being LLM-generated. In this study, we did not instruct participants to use LLMs in any way; thus, we captured a baseline of LLM use for uninstructed participants doing a task for which LLMs have a considerable advantage over human labor.

Following a study on Mechanical Turk that took place four to six weeks before ours,[27] we used three approaches to aggregate the probabilities of LLM use (henceforth "LLM probabilities"), obtaining similar (but slightly lower) estimates:

- *Classify-and-count*, considering as synthetic any summary with an LLM probability above 50% (prevalence estimate: 33.3%; 95% CI [25.9%, 40.1%])
- *Probabilistic classify-and-count*, where we calibrated the model[6] (see Appendix) and then averaged the LLM probabilities (estimate: 35.2% [29.8%, 40.6%])
- *Corrected classify-and-count*, adjusting for the type I and type II error rates estimated on the training data[18] (estimate: 35.4% [27.8%, 43.0%]).

We validated our results by analyzing crowd workers' copy-pasting behavior (see Appendix), finding that 55% of the summaries where workers had copy-pasted text were classified as synthetic (that is, LLM probability above 50%) vs. only 9% when workers had not copy-pasted text. As no information about copy-pasting was used in the "LLM-or-not" classifier, this result strengthens our confidence in it. Interestingly, far fewer crowd workers used copy-pasting on Prolific (53%) in Study 1, compared with a previous study[27] on Amazon Mechanical Turk (89%).

## Study 2: Prevention of LLM Use

Next, we analyzed whether targeted strategies can curb LLM use. Specifically, we studied two different mitigation approaches: 1) explicitly asking crowd workers not to use LLMs (henceforth the "request" strategy) and 2) imposing hurdles that deter LLM use (the "hurdle" strategy). We considered two variations for each: For the request strategy, we asked individuals either directly or indirectly not to use LLMs (see Appendix), and for the hurdle strategy, we either converted the original abstract text to an image or disabled copy-pasting entirely. As the two strategies are independent, we investigated all combinations (alongside a no-restriction condition) in a 3 x 3 factorial design (see Table 1).

Using the same task as in Study 1, Study 2 was conducted by randomly splitting n = 720 users into the nine conditions. Upon completion, they were then redirected to a follow-up survey where they were asked (Q1) how often they used ChatGPT in their daily lives, (Q2) whether they had used ChatGPT for the task, and (Q3) whether they knew of studies tracking ChatGPT use on crowdwork platforms (see Appendix for exact phrasing). We measured LLM use with the probabilistic classify-and-count classifier, self-reported LLM use as captured by Q2, and high-precision (and likely low-recall) heuristics indicating LLM use (see Materials and Methods).

**Effectiveness of preventive measures.** Table 1 shows the estimated LLM use across different mitigation strategies. For example, when workers were directly requested not to use LLMs and shown the text to be summarized as an image (thus preventing copy-pasting), LLM use (as measured by the probabilistic classify-and-count method) almost halved, dropping from 27.6% to 15.9% (as measured by the classifier; see Table1a). Similar results were obtained using self-reported use by crowd workers (Q2) and using high-precision heuristics (Tables 1c and 2c; see Materials and Methods).

Comparing high-precision heuristics with self-reports revealed that only 11 of the 31 workers using LLMs according to high-precision heuristics admitted to using LLMs, whereas 31 of the 689 whom the heuristic and classifier both failed to mark as synthetic admitted to LLM use.

We further disentangled the effect of each specific strategy and variation with a linear model (see Appendix), finding three out of the four tested interventions to significantly reduce LLM use (considering the LLM use predicted by the classifier; see the figure). Notably, asking crowd workers indirectly ("Please do your best to summarize the abstract in your own words") was the least effective strategy across all measures of LLM use and the only non-significant intervention when considering the classifier-based outcome ("Indirect"; 2% decrease; p = 0.38). This hints at the complexity of preventing LLM use, as crowd workers may choose to ignore requests if it is in their best interest financially.

**Correlates of LLM use.** We studied the relationship between LLM use and 1) the age of crowd workers and 2) how they answered two of the post-survey questions (Q1: LLM use in general; Q3: awareness of studies measuring LLM use) using a simple linear model and considering both self-reports and the classifier's

**Table 1. LLM use across experimental conditions, estimated using three methods: a) probabilistic classify-and-count ("Classifier"); b) self-reported use ("Self-reported"); and c) high-precision heuristics ("Heuristics"). All estimates indicate that the interventions significantly reduced LLM use, albeit not completely.**

| | | Hurdle | | |
|---|---|---|---|---|
| | | **None** | **Image** | **Ctrl C+V** |
| **Request** | **None** | **27.6%** (21.0%, 34.6%) | **21.5%** (16.0%, 27.4%) | **24.1%** (18.3%, 30.4%) |
| | **Indirect** | **28.5%** (21.7%, 35.8%) | **19.8%** (14.2%, 25.8%) | **19.3%** (14.6%, 24.5%) |
| | **Direct** | **24.0%** (18.6%, 29.6%) | **15.9%** (11.9%, 20.3%) | **15.8%** (11.8%, 20.4%) |

**(a) Classifier**

| | | Hurdle | | |
|---|---|---|---|---|
| | | **None** | **Image** | **Ctrl C+V** |
| **Request** | **None** | **15.8%** (8.5%, 24.4%) | **10.4%** (3.9%, 16.9%) | **4.9%** (1.2%, 9.8%) |
| | **Indirect** | **13.2%** (5.9%, 22.1%) | **6.6%** (1.3%, 12.0%) | **3.6%** (0.0%, 8.3%) |
| | **Direct** | **3.0%** (0.0%, 7.1%) | **6.6%** (1.3%, 13.2%) | **9.1%** (3.9%, 15.6%) |

**(b) Self-reported**

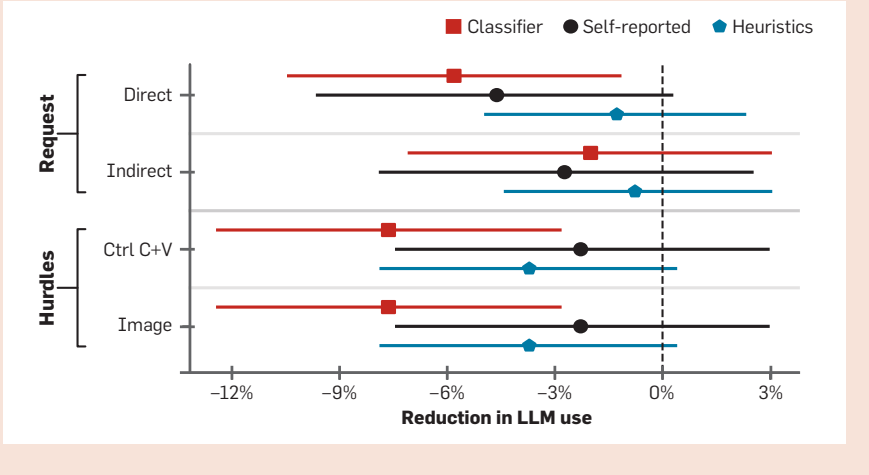
| | | Hurdle | | |
|---|---|---|---|---|
| | | **None** | **Image** | **Ctrl C+V** |
| **Request** | **None** | **10.9%** (4.9%, 18.3%) | **2.6%** (0.0%, 6.5%) | **1.2%** (0.0%, 3.7%) |
| | **Indirect** | **4.4%** (0.0%, 10.3%) | **5.3%** (1.3%, 10.7%) | **2.4%** (0.0%, 6.0%) |
| | **Direct** | **7.1%** (3.0%, 12.1%) | **4.0%** (0.0%, 9.2%) | **0.0%** (0.0%, 0.0%) |

**(c) Heuristics**

LLM-probability estimates as outcomes (see Appendix). We found that younger individuals were significantly more likely to use LLMs (−0.18% in estimated LLM probability per year; p = 0.014) and that workers who used LLMs "often" were 18.7% more likely to use it for the task (p < 0.001). Awareness of studies measuring LLM use did not significantly affect use (+1.6%; p = 0.55). Results were similar when considering self-reported use as the outcome variable.

Additionally, we analyzed the relationship between LLM use and time spent on the task, finding that preventive measures (that is, hurdles and requests) seem to mediate the relationship. Users who self-reported LLM use spent 21.9% less time (relative decrease; p = 0.002) to complete the task than those who did not (across experimental conditions). Using a simple linear model with time spent as the log-transformed outcome (see Appendix), we further analyzed this relative change across different proxy metrics for LLM use and experimental conditions (see Table 2). Across proxy metrics, we found that the overall time reduction is never statistically significant when hurdles are employed. However, when only requests are applied, results differed: The relative decreases were not statistically significant considering the classifier but remained statistically significant considering self-reports. We hypothesize this may be because users who use LLMs and lightly edit their output spend more time on the task and are less likely to self-report use.

**Figure. Estimated effect sizes for interventions to prevent LLM use considering three different measures of LLM use as the outcome variable: a) probabilistic classify-and-count, b) self-reported use, c) high-precision heuristics. Error bars represent 95% confidence intervals; n = 720.**




**Content-level analysis.** Analyzing the text of crowd workers' summaries, we found that summaries labeled as synthetic by the classifier were significantly more "homogeneous" than those labeled as human, according to a previously proposed homogeneity metric[20] and BertScore[30] (details in Appendix). We estimated a homogeneity score of 45.6% (43.2%, 48.2%) for synthetic texts, vs. 27.1% (26.8%, 27.4%) for human texts, and a BERTScore of 91.4 (91.0, 91.8) for synthetic texts vs. 87.4 (87.2, 87.3) for human texts.

In the original study whose human summaries we reused,[15] the authors measured the retention of keywords from the original abstract corresponding to essential information, finding it to be highly correlated with human evaluations of quality. Using this metric as a proxy for quality, we found that summaries labeled as synthetic preserved more keywords (40.1% [36.9%, 43.2%]) than summaries labeled as human (31.2% [29.9%, 32.6%]). We found a similar effect when using self-reports and high-precision heuristics instead of the classifier's labels.

But how do the interventions affect the above content-level metrics? We repeated the analysis shown in the figure but using homogeneity, BERTScore, and keyword retention as outcomes (see Section G.1 in the Appendix for details). We found no significant effect of the interventions on content-level outcomes, with one exception: Directly requesting workers not to use LLMs decreased keyword retention by 5.8% (p = 0.003). We hypothesize that the reduction in keyword retention may be caused by crowd workers' hesitancy to use extractive summarization when prompted not to use LLMs. (Results were similar when considering only summaries classified by us as being human-made.)

**Table 2. Relative differences in time spent between instances where we detected LLM use and where we did not. We report differences across the nine experimental conditions and determine LLM use using three methods. (Note that time spent is one of our heuristics for detecting ChatGPT use.)**

| | | Hurdle | | |
|---|---|---|---|---|
| | | **None** | **Image** | **Ctrl C+V** |
| **Request** | **None** | **−32.0%** (−51.6%, −4.5%) | **18.0%** (−19.8%, 73.6%) | **5.1%** (−22.6%, 42.7%) |
| | **Indirect** | **−2.0%** (−28.8%, 34.9%) | **42.4%** (−3.3%, 109.6%) | **4.9%** (−26.8%, 50.2%) |
| | **Direct** | **−12.0%** (−34.2%, 17.7%) | **1.0%** (−43.9%, 81.7%) | **35.1%** (−6.9%, 96.1%) |

**(a) Classifier**

| | | Hurdle | | |
|---|---|---|---|---|
| | | **None** | **Image** | **Ctrl C+V** |
| **Request** | **None** | **−34.2%** (−55.3%, −3.2%) | **−0.1%** (−34.8%, 53.2%) | **1.4%** (−41.5%, 75.7%) |
| | **Indirect** | **−43.8%** (−61.0%, −19.1%) | **−25.1%** (−58.0%, 33.3%) | **−17.6%** (−56.0%, 54.1%) |
| | **Direct** | **−58.9%** (−78.5%, −21.5%) | **1.8%** (−43.4%, 83.2%) | **22.7%** (−15.7%, 78.8%) |

**(b) Self-reported**

| | | Hurdle | | |
|---|---|---|---|---|
| | | **None** | **Image** | **Ctrl C+V** |
| **Request** | **None** | **−46.1%** (−65.4%, −16.1%) | **35.3%** (−40.2%, 206.3%) | **−45.6%** (−81.3%, 58.8%) |
| | **Indirect** | **−53.7%** (−75.0%, −14.3%) | **−47.9%** (−72.2%, −2.5%) | **45.1%** (−32.1%, 210.1%) |
| | **Direct** | **−32.3%** (−56.5%, 5.2%) | **29.4%** (−38.6%, 172.6%) | **0.0%** (0.0%, 0.0%) |

**(c) Heuristics**

## Discussion

The results suggest that LLMs pervade current crowd work on text-production tasks. Although adopting various strict mitigation approaches reduced LLM use by nearly 50%, it could not entirely prevent it. While text-production tasks are particularly suitable for LLM use, we argue that these findings are broadly applicable to crowdsourcing, as crowd workers will likely use LLMs on other kinds of tasks (for example, image segmentation, multiple-choice questions) in the near future, if they are not already doing so. There are several reasons for this. First, the models are increasingly capable of doing other tasks; for instance, while writing this article, ChatGPT was updated to receive images as input,[19] which could allow its use on tasks such as image tagging or classification.[29] Second, crowd workers have incentives to use them; even in the absence of LLMs, there are widespread attempts to "game the system" to make money, to the extent that an extensive body of work has been developed around ensuring the quality of responses.[7] Third, crowd workers, who are often tech-savvy[5,12] and frequently rely on plug-ins and Web services to boost their performance and earnings,[9,16] are capable of integrating these models into their pipelines. Even without requiring coding, tools to automate ChatGPT use are plentiful (for example, IFTTT, Zapier).

Synthetic data may harm the utility of crowd-work platforms, as researchers often care about human behavior or preferences; for example, the authors of the paper whose human summaries we borrowed[15] wanted to know *how people summarized*, instead of merely obtaining good summaries. While some preliminary studies suggest that synthetic data may capture certain viewpoints,[2] it still often fails to do so, and research using crowd work may inadvertently capture the behavior and preferences of LLMs, not humans. Even if LLMs can capture average behavior or preferences, the homogeneity of their responses may result in losing the long tail of human behavior and preferences that is vital to researchers[24] and, according to recent work, important to training capable LLMs.[23] In that context, our results indicating that LLM-generated summaries are more homogeneous than human-generated summaries suggest that LLM use may be particularly harmful when the goal of crowdsourcing is to capture the diversity of human preferences, behaviors, or opinions.

**Our results suggest that LLM use may be particularly harmful when the goal of crowdsourcing is to capture the diversity of human preferences, behaviors, or opinions.**

To foresee the potential harm of LLM use in crowdsourcing, one may consider a topic that has received increased attention in the social sciences in the past few years: climate change.[8,26] Social scientists often use crowdsourcing to study attitudes toward climate change.[4,25] Yet, recent work has shown that, when prompted to answer multiple-choice questions, LLMs' opinions are better aligned with liberal, wealthy individuals and exhibit pro-environmental bias.[13,22] Therefore, it may be expected that LLM use could harm the validity of social scientists' studies on behavioral interventions and assessments of global stances toward climate change. We stress that this is not particular to climate change: Social scientists use crowdsourcing for various topics,[3] and LLMs are non-representative of the samples of interest in various ways.[13,22]

We must be careful not to conflate LLM use with cheating. Depending on the study, it could be beneficial if LLMs assist crowd workers. Further, as LLMs become intertwined with how people write and accomplish everyday tasks, the distinction between "synthetic" and "human" data may blur. For example, is text generated with the help of a spell-checker synthetic? Thus, we expect the thresholds for concern and meaning will shift dramatically over the coming months and years, as LLMs become more ubiquitous in everyday productivity tasks. In that context, a fruitful future direction is to explore the landscape of how crowd workers use LLMs. There are many ways of integrating these models into crowd workers' workflows, and different approaches may affect downstream research output differently.

We found that stricter mitigation approaches can significantly reduce LLM use. These measures may, however, backfire when detection is critical. Stricter measures may limit the number of participants using LLMs but also make them more reluctant to admit ex post that they used them, or make them more difficult to detect, as the prevention measure eliminated a key indicator of LLM use. For example, removing copy-pasting makes it harder to use LLMs, limiting use, but then researchers also

cannot use copy-pasting as a feature to detect who used LLMs. Further, mitigation approaches can reduce the overall response quality: As we found empirically, workers explicitly told not to use LLMs produced lower-quality summaries.

LLM-based tools and LLM users are co-evolving in ways to ensure the low temporal validity of our specific findings and estimates. In the past few months alone, tools have evolved to interpret images and to call LLMs without the need to copy-paste (for example, by simply selecting text). This does not diminish the value of our work—it makes it even more valuable: It is critical to establish baselines and ongoing measurements as this co-evolution progresses, and our work establishes such baselines. Further, we are confident that our high-level interpretations and guidance will translate across this evolution, and we hope this helps establish a regularly updated new program of study to serve crowdwork platforms and researchers.

To conclude, in light of our findings, we propose two practical guidelines for using crowdsourcing in the era of large language models. First, researchers should assess the impact of LLMs on their research by asking themselves: Is the point of crowdsourcing to obtain data representative of human behavior, preferences, and opinions? And if so, is capturing the diversity of these human responses important? We argue that crowdsourcing will be most affected when the answer to both questions is yes, as we found that LLM responses differ from human responses and are more homogeneous. Second, if large language models are likely to harm the utility of crowdsourcing, our findings indicate that researchers can actively diminish LLM use by requesting that workers not use them and creating hurdles that decrease the incentives for using them. Notably, hurdles should be adapted as models become more capable and better integrated into people's lives. C

## Materials and Methods

▸ ***Data.*** **We modified a prior Mechanical Turk task[15] where crowd workers were asked to summarize medical paper abstracts. We re-ran the study twice on Prolific. In Study 1, we estimated prevalence by collecting 168 user summaries (paying £9/hour). In Study 2, we re-ran the study on 720 users, now using several mitigation techniques (paying £10/hour). (See Appendix for full description of data and original study.)**

▸ ***Model training.*** **We fine-tuned a e5-base-v2 language model[28] for our classification task and conducted a hyperparameter sweep. The model was trained on the summaries from the original study[15] (written before the adoption of LLMs) and summaries synthetically generated using OpenAI's API.**

▸ ***Heuristic-based estimates.*** **We defined two high-precision heuristics for measuring LLM use: feasible time for completion and pasting in artifacts from the ChatGPT Web interface (details in Appendix).**

▸ ***Effect of each intervention.*** **We assessed the effectiveness of each of the interventions with a linear probability model. We do not consider interactions between the treatment conditions, as a two-way ANOVA indicated that the interactions between the two strategies are not statistically significant.**


**References**

1. Akiba, T. et al. Optuna: A next-generation hyperparameter optimization framework. In *Proceedings of the 25th ACM SIGKDD Intern. Conf. on Knowledge Discovery & Data Mining*. ACM, (2019), 2623–2631.
2. Argyle, L.P. et al. Out of one, many: Using language models to simulate human samples. *Political Analysis 31*, 3 (2023), 337–351.
3. Bohannon, J. Mechanical Turk upends social sciences. *Science 352*, 6291 (2016).
4. Bouman, T., Steg, L., and Zawadzki, S.J. The value of what others value: When perceived biospheric group values influence individuals' pro-environmental engagement. *J. of Environmental Psychology 71*, 101470 (2020).
5. Brewer, R., Morris, M.R., and Piper, A.M. Why would anybody do this?": Older adults' understanding of and experiences with crowd work. In *Proceedings of the 2016 CHI Conf. on Human Factors in Computing Systems*. ACM, (2016), 2246–2257.
6. Card, D. and Smith, N.A. The importance of calibration for estimating proportions from annotations. In *Proceedings of the 2018 Conf. of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*. Association for Computational Linguistics, (2018), 1636–1646.
7. Daniel, F. et al. Quality control in crowdsourcing: A survey of quality attributes, assessment techniques, and assurance actions. *ACM Computing Surveys 51*, 1 (2018), 1–40.
8. Dietz, T., Shwom, R.L., and Whitley, C.T. Climate change and society. *Annual Rev. of Sociology 46* (2020), 135–158.
9. El Maarry, K., Milland, K., and Balke, W-T. A fair share of the work? The evolving ecosystem of crowd workers. In *Proceedings of the 10th ACM Conf. on Web Science*. ACM, (2018), 145–152.
10. Gilardi, F., Alizadeh, M., and Kubli, M. ChatGPT outperforms crowd workers for text-annotation tasks. In *Proceedings of the National Academy of Sciences of the United States of America 120*, (2023).
11. Gray, M.L. and Suri, S. *Ghost Work: How to Stop Silicon Valley from Building a New Global Underclass*. Eamon Dolan Books, (2019).
12. Guess, A.M. and Munger, K. Digital literacy and online political behavior. *Political Science Research and Methods 11*, 1 (2023), 110–128.
13. Hartmann, J., Schwenzow, J., and Witte, M. The political ideology of conversational AI: Converging evidence on chatGPT's pro-environmental, left-libertarian orientation. *arXiv preprint arXiv:2301.01768*, (2023).
14. Hausman, J.A., Abrevaya, J., and Scott-Morton, F.M. Misclassification of the dependent variable in a discrete-response setting. *J. of Econometrics 87*, 2 (1998), 239–269.
15. Ribeiro, M.H., Gligoric, K., and West, R. Message distortion in information cascades. In *The World Wide Web Conf.*. ACM, (2019), 681–692.
16. Irani, L.C. and Silberman, M.S. Turkopticon: Interrupting worker invisibility in Amazon Mechanical Turk. In *Proceedings of the SIGCHI Conf. on Human Factors in Computing Systems*. ACM, (2013), 611–620.
17. Luo, Z., Xie, Q., and Ananiadou, S. ChatGPT as a factual inconsistency evaluator for text summarization. *arXiv:2303.15621*, (2023).
18. Meyer, B.D. and Mittag, N. Misclassification in binary choice models. *J. of Econometrics 200*, 2 (2017), 295–311.
19. OpenAI. ChatGPT can now see, hear, and speak. (Sep. 25, 2023); https://openai.com/index/chatgpt-can-now-see-hear-and-speak/
20. Padmakumar, V. and He, H. Does writing with language models reduce content diversity? In *Proceedings of the 12th Intern. Conf. on Learning Representations*. IEEE, (2024).
21. Salganik, M.J. *Bit by Bit: Social Research in the Digital Age*. Princeton University Press, (2019).
22. Santurkar, S. et al. Whose opinions do language models reflect? In *Proceedings of the 40th Intern. Conf. on Machine Learning*. PMLR, (2023).
23. Shumailov, I. et al. Model dementia: Generated data makes models forget. *Nature 631* (2024), 755–759; 10.1038/s41586-024-07566-y
24. Song, Z. et al. *Reward collapse in aligning large language models. arXiv preprint arXiv:2305.17608*, (2023).
25. Sparks, A.C. Climate change in your backyard: When climate is proximate, people become activists. *Frontiers in Political Science 3*, 666978 (2021).
26. Steg, L. Psychology of climate change. *Annual Rev. of Psychology 74* (2023), 391–421.
27. Veselovsky, V., Ribeiro, M.H., and West, R. Artificial artificial artificial intelligence: Crowd workers widely use large language models for text production tasks. *arXiv preprint arXiv:2306.07899*, (2023).
28. Wang, L. et al. Text embeddings by weakly-supervised contrastive pre-training. *arXiv preprint arXiv:2212.03533*, (2022).
29. Yang, Z. et al. The dawn of LLMs: Preliminary explorations with GPT-4V(ision). *arXiv preprint arXiv:2309.17421*, (2023).
30. Zhang, T. et al. BERTscore: Evaluating text generation with BERT. In *Intern. Conf. on Learning Representations*, (2020).


**more online**

To access the Appendix, please visit https://dl.acm.org/doi/10.1145/3685527 and click on Supplemental Material.


**Veniamin Veselovsky** is a doctoral student in the Department of Computer Science at Princeton University.

**Manoel Horta Ribeiro** (manoel@cs.princeton.edu) is an assistant professor in the Department of Computer Science at Princeton University.

**Philip J. Cozzolino** is an associate professor in psychiatry and neurobehavioral sciences at the University of Virginia School of Medicine.

**Andrew Gordon** is a staff researcher in social and behavioral science at Prolific.

**David Rothschild** is an economist at Microsoft Research.

**Robert West** is an associate professor in the School of Computer and Communication Sciences at EPFL and a visiting researcher at Microsoft Research.




Watch the authors discuss this work in the exclusive *Communications* video. https://cacm.acm.org/videos/llms-in-crowd-work

DOI:10.1145/3689819

**Attackers can use forged ICMP error messages to exploit vulnerabilities in the TCP/IP stack.**

BY XUEWEI FENG, QI LI, KUN SUN, KE XU, AND JIANPING WU

# Exploiting Cross-Layer Vulnerabilities: Off-Path Attacks on the TCP/IP Protocol Suite

THE TCP/IP PROTOCOL suite is a set of communication protocols underpinning the Internet. Protocols at different layers of the suite—for example, Wi-Fi, IP, TCP, and HTTP (Figure 1)—form the essential framework for data transmission on the Internet. But given the paramount significance of the TCP/IP protocol suite, it is also a pivotal target for myriad forms of attacks.[4,9,12,16,18,22,28] Vulnerabilities in the suite can have extensive repercussions, posing a fundamental threat to Internet security and presenting significant incentives to attackers. As a result, both industry and academia have dedicated substantial efforts[6,8,16,20,25,36,37] toward combating the diverse spectrum of network attacks. What has received limited attention, however, are vulnerabilities arising from cross-layer interactions among various protocols within the TCP/IP protocol suite, caused by forged Internet Control Message Protocol (ICMP) error messages. These vulnerabilities can be exploited by off-path attackers, posing risks to Internet security.

In the process of network data processing, protocols within the suite must interact and coordinate across layers. This cross-layer interaction ensures the smooth generation, transmission, reception, and storage of data. For example, when delivering an HTTP message, protocols such as DNS, TCP, IP, ARP, and Wi-Fi may need to be invoked to process and encapsulate the message. Although each protocol within the stack may individually possess sufficient robustness, combining these protocols and engaging in cross-layer interaction through function calls can introduce security issues or anomalies. Specifically, the proper execution of one layer's specific functionality can be compromised by the normal execution

## » key insights

- **The TCP/IP protocol suite serves as the backbone of the Internet. Despite more than 40 years of development, its security remains a critical concern.**
- **By exploring the security implications of cross-layer interactions within the TCP/IP protocol suite, particularly those triggered by ICMP errors, we identified several significant vulnerabilities in modern TCP/IP implementations.**
- **There remains a continuous need to uncover subtle semantic vulnerabilities within the TCP/IP protocol suite, particularly through techniques that minimize manual effort, such as program analysis and AI-driven approaches.**

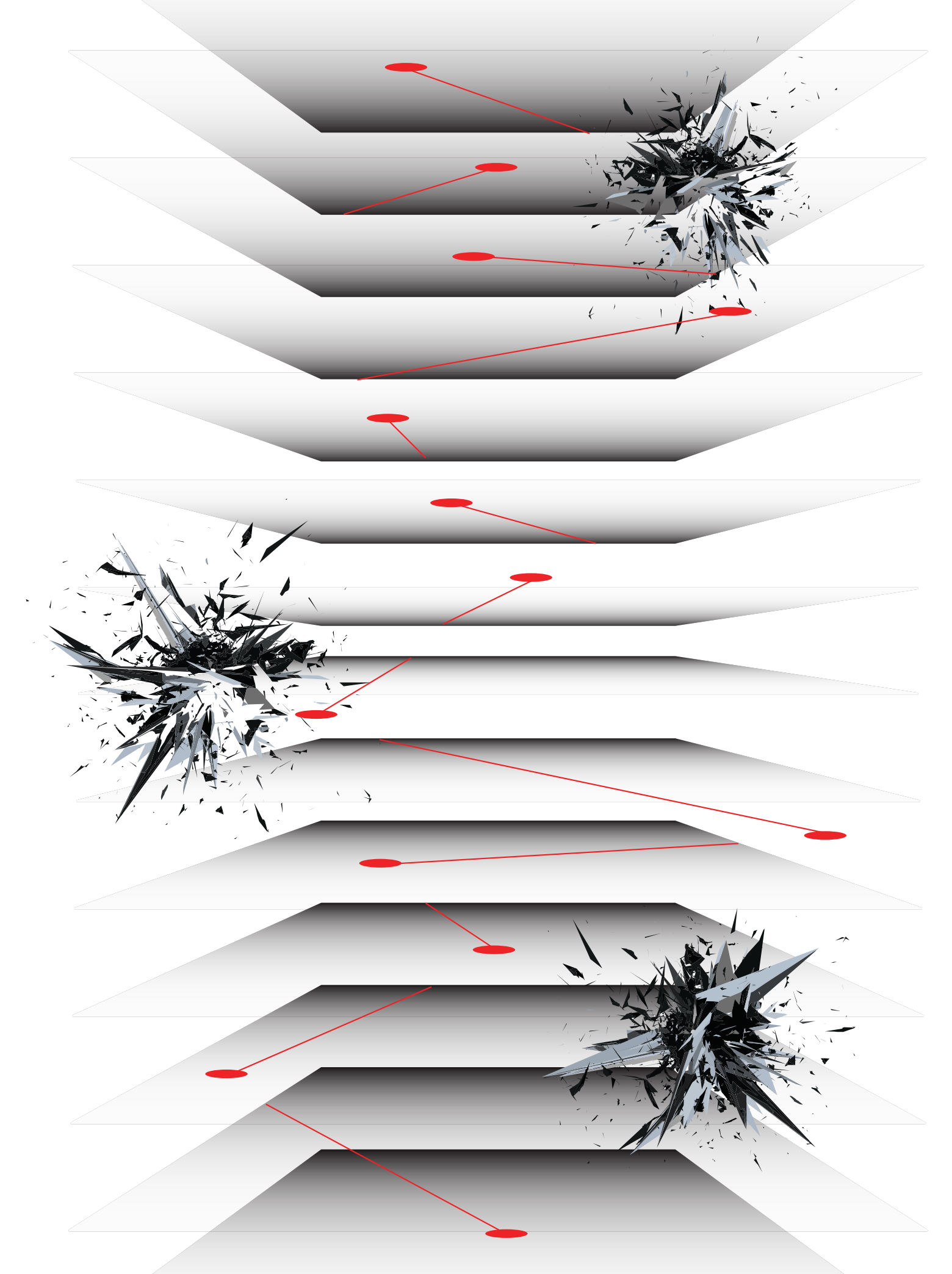


**Figure 1. The TCP/IP protocol suite serves as the essential framework for data transmission on the Internet.**

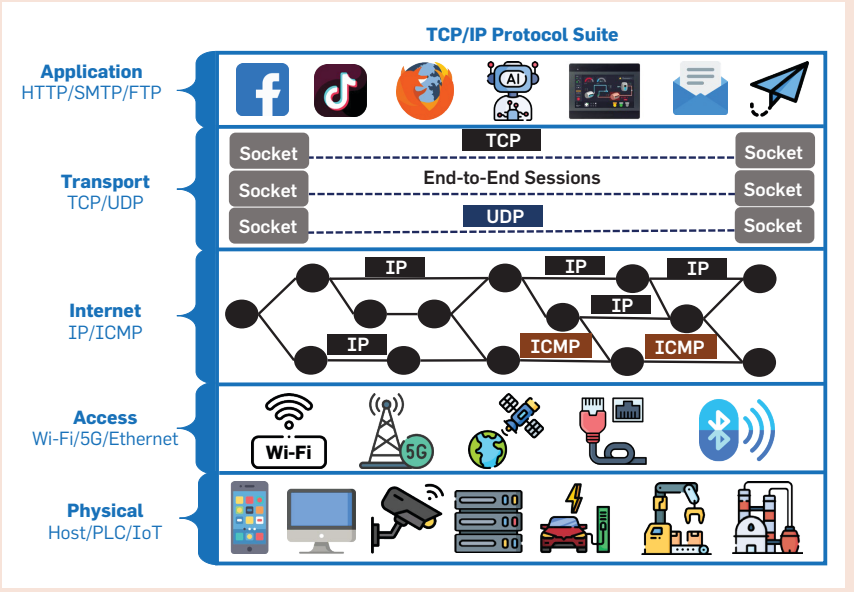
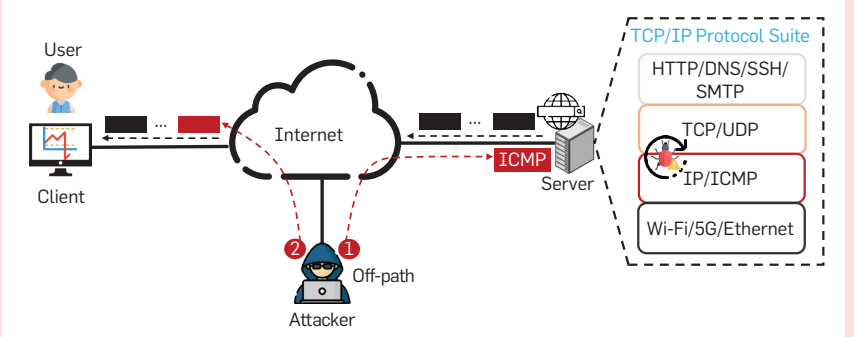



**Figure 2. Threat model of off-path attacks on the TCP/IP protocol suite via forged ICMP error messages.**




of other layers. For instance, the loss of frames in wireless networks commonly occurs due to inevitable communication-noise interference; however, at the TCP layer, if TCP segments are not promptly acknowledged due to the loss of wireless frames, it can mistakenly trigger the detection of network congestion, leading to inefficient execution of the TCP congestion-control algorithm.[34]

ICMP, recognized as a fundamental component of the TCP/IP protocol suite, frequently drives cross-layer interactions that transcend traditional network layer boundaries to report network conditions or errors. By operating directly on top of IP, ICMP error messages embedded with various payloads can influence the behavior of higher layers such as TCP and UDP, and can even be exploited by off-path attackers to compromise higher-layer protocols. Here, we undertake a comprehensive study to investigate the cross-layer interactions within the TCP/IP protocol suite caused by forged ICMP errors. In doing so, we uncover multiple vulnerabilities, including information leakage, desynchronization, semantic gaps, and identity spoofing. We discuss each of these in turn, but first we will provide some background on ICMP error messages and their associated threat model.

**Basics of ICMP error messages.** ICMP error messages are specific types of ICMP messages generated in response to network issues. They play a crucial role in identifying and diagnosing problems within a network.[3,5,35] These messages include Destination Unreachable (indicating network failures or host unavailability), Time Exceeded (resulting from packet time-to-live expiration), Parameter Problem (addressing IP header parameter issues), and Redirect Messages (for optimizing routing). Source Quench messages, historically significant for congestion control, are now deprecated. Aiming to report network issues to the receiver, ICMP error messages inevitably induce cross-layer interactions within the TCP/IP protocol stack and prompt the receiver to adjust its behavior based on the received ICMP error messages. According to the ICMP specifications,[3,5,35] ICMP error messages should contain at least the first 28 octets of the original packet that triggered the error message (that is, 20 octets of the IP header plus at least the first eight octets). When an ICMP error message is received, the receiver can use the embedded payload in the message to match it to the corresponding process. This enables the process to adapt and respond effectively. For example, when an ICMP Destination Unreachable message with the code "Packet too big" is received, it facilitates cross-layer interactions by enabling the receiver's TCP to reduce its maximum segment size (MSS), thereby avoiding IP fragmentation on the intermediate routes that issued the ICMP error message.

Unfortunately, in practice, it is easy for attackers on the Internet to forge ICMP error messages to manipulate the receiver's behavior, for a couple of reasons. First, because ICMP error messages can be generated by any intermediate router along the network path, it is difficult for the receiver to authenticate their source. This is particularly challenging because attackers can use IP address spoofing techniques to forge the source IP address. Second, although ICMP specifications require that error messages include at least the first 28 octets of the original packet, enabling the receiver to match the message and perform a legitimacy check, attackers can easily forge a 28-octet payload to bypass this check. In the context of TCP communication, the first 28 octets of the

original packet contain a random sequence number, which is hard for attackers to guess. However, in UDP or ICMP scenarios, since these protocols are stateless and lack randomized sequence numbers, attackers can easily forge a 28-octet payload to include in the falsified ICMP error message. This allows them to evade the receiver's legitimacy check and deceive the receiver into responding to the message, leading to unintended protocol interactions that pose security risks.

**Threat model of off-path attacks.** Figure 2 shows the threat model of off-path attacks on the TCP/IP protocol suite via forged ICMP error messages. The off-path attacker is positioned outside the direct communication path between the server and the client. Consequently, the attacker cannot intercept or directly modify packets in transit between the server and the client. Instead, the attacker can forge and send packets with arbitrary source IP addresses.[2,a] Specifically, by leveraging forged ICMP error messages, the attacker exploits weaknesses and forces exceptional behaviors during cross-layer interactions among multiple protocols within the server's TCP/IP protocol suite. Once these vulnerabilities are triggered, network traffic from the server to the client will be affected. Furthermore, the off-path attacker can impersonate the server and inject crafted packets into the client to manipulate the target network traffic. The following four sections delve into work of ours that identified vulnerabilities (information leakage,[9,10] desynchronization,[12] semantic gap,[11] and identity spoofing[13]) caused by forged ICMP error messages, enabling off-path attackers to launch impactful attacks.

## Information Leakage

TCP plays a fundamental role within the TCP/IP protocol suite and is an important part of the Internet, ensuring data packets reach their intended destinations accurately and in the correct sequence. A key security measure

**By operating directly on top of IP, ICMP error messages embedded with various payloads can influence the behavior of higher layers such as TCP and UDP, and can even be exploited by off-path attackers to compromise higher-layer protocols.**

within the TCP protocol is 32-bit randomization of sequence and acknowledgment numbers. This strengthens the protocol's resilience against out-of-band malicious TCP packet injections. However, despite the extensive randomized sequence and acknowledgment number space, which significantly increases the time needed for brute-force attacks, TCP protocol operations involve interactions with other protocols in the TCP/IP protocol suite. During these interactions, certain fields of other layer protocols, such as the Identification field of the IP protocol (IPID), can be exploited to infer the TCP protocol's sequence and acknowledgment numbers. In particular, we discovered that the IPID field, even with the most advanced IPID assignment policy currently available in Linux systems, can be manipulated by a forged ICMP error message issued by off-path attackers. This manipulation allows the attacker to indirectly infer confidential information (that is, the sequence and acknowledgment numbers of TCP) by observing the IPID field, ultimately leading to information leakage during protocol cross-layer interactions. This can enable the off-path attacker to inject malicious TCP packets into the target connection, thereby jeopardizing the integrity of the associated TCP stream.[9]

**IPID assignment.** The IPID field is used to enable defragmentation. After abandoning previous vulnerable IPID assignment methods (for example, global IPID assignment and per-destination IPID assignment), modern operating systems typically employ advanced methods to assign IPIDs for IP packets. For instance, Linux systems use a per-socket-based IPID assignment policy for TCP packets and 2,048 globally shared hash counters for non-TCP packets.[1] Though this IPID assignment method aims to safeguard TCP protocols against information leakage stemming from IPID values, we demonstrated a vulnerability that can be exploited by off-path attackers to deduce the upper-layer sequence and acknowledgment numbers of a victim TCP connection.

**Inference of randomized numbers.** In this situation, the off-path attacker pretends to be a router and issues a crafted ICMP error message (an ICMP

a Prior studies show that about a quarter of ASes on the Internet do not filter packets with spoofed source addresses leaving their networks, and it is trivial to rent such a machine from a bulletproof hosting node.[26–28]

Destination Unreachable message with the code "Packet too big") embedded with a 28-octet payload of a fake ICMP echo reply packet to a Linux server. This crafted ICMP error message evades the server's legitimacy check and deceives it into downgrading its IPID assignment policy for TCP packets. The policy transitions from a per-socket-based policy to the utilization of 2,048 globally shared hash counters. Given the limited size of the hash counter pool (2,048), the attacker can change its IP address to successfully provoke a hash collision with a victim TCP client of the server.[b] This occurs because Linux servers select one of the 2,048 hash IPID counters based on the destination IP address of outgoing packets. Consequently, the server can be tricked into selecting the same IPID counter for both the attacker's IP address and the victim client's IP address. This situation allows the attacker to deduce the specific IPID counter being used by the Linux server for the victim TCP connection, thus creating a side channel to leak connection information.

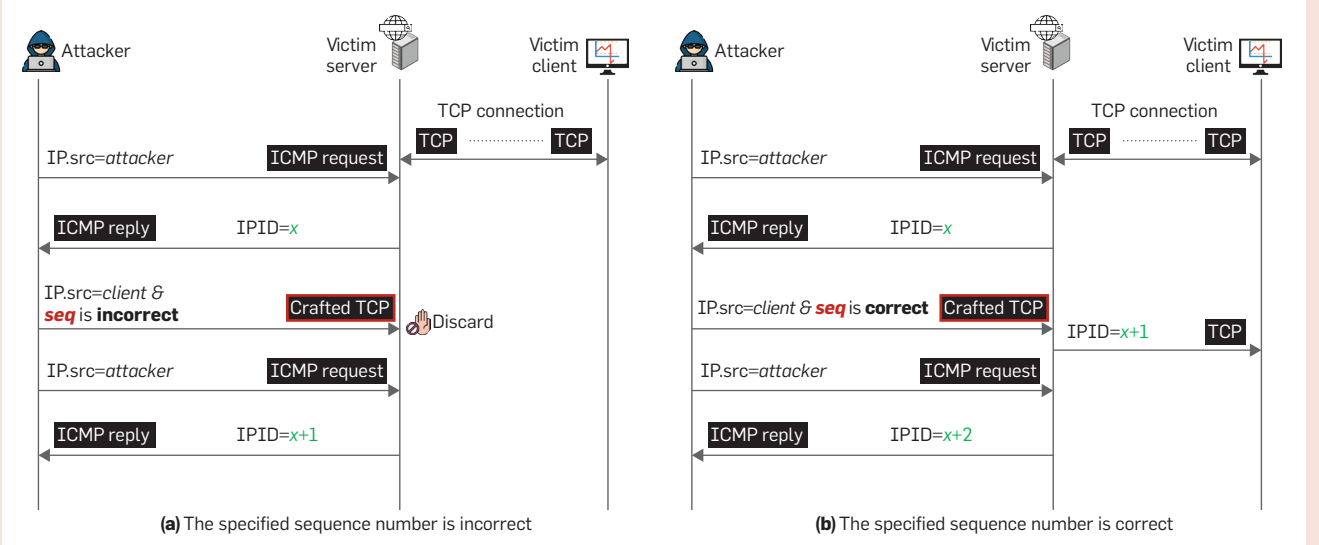
Once the shared IPID counter is identified, the attacker proceeds to send crafted TCP packets to the victim server. The shared IPID counter will exhibit varying behaviors under different circumstances, enabling the attacker to discern whether the specified values in the forged TCP packets are correct. As shown in Figure 3, the attacker initiates the process by sending an ICMP echo request packet to the server and monitors the current value of the shared IPID counter when the server responds with a reply packet. Subsequently, the attacker impersonates the identity of the victim client by IP spoofing and crafts a TCP packet destined for the server. This crafted packet includes the specified sequence number (*seq*). In this scenario, the server's behavior varies based on the sequence number specified within the crafted packet. As illustrated in Figure 3a, when the specified sequence number is incorrect (that is, not within the server's receive window), the server simply discards the packet. When the attacker later observes the current value of the shared IPID counter once more, it will notice that the IPID counter's values remain consecutive.

If the specified sequence number is correct (as shown in Figure 3b), the server will generate a reply packet destined for the client, even though the client will ultimately discard this reply. This reply packet consumes a value from the shared IPID counter. When the attacker subsequently observes the current value of the shared IPID counter once more, it will notice that the IPID counter's values are no longer consecutive. By making this comparison, the attacker can accurately deduce sensitive information, such as the sequence and acknowledgment numbers of the target TCP connection. For off-path TCP injection attacks (as described in Cao et al.[6] and Qian and Mao[36]), once off-path attackers identify the randomized sequence and acknowledgment numbers of the target TCP connection, they can craft an out-of-band TCP packet specified with these identified numbers in the TCP header. When injected into the target connection, this packet will pass verification and be accepted by the receiver, potentially terminating or poisoning the connection.

**Experimental results.** Through real-world evaluations, we have demonstrated that the information leakage caused by cross-layer interactions can have severe real-world consequences, enabling attackers to infer and disrupt a large number of TCP connections. We found that more than 20% of the Alexa top 100k websites are vulnerable. To test this, we first establish a TCP connection from our client to each of the websites in the Alexa top 100k list. Then, an attack machine on the Internet issues a forged ICMP "Packet too big" message to the website to manipulate its IPID assignment for our client. Our results show that 20% of the websites can be tricked into downgrading the IPID assignment from the per-socket-

b Since kernel version 5.12.4, Linux has used a dynamic hash counter pool proportional to physical RAM size to mitigate IPID-based firewall attacks.[24]

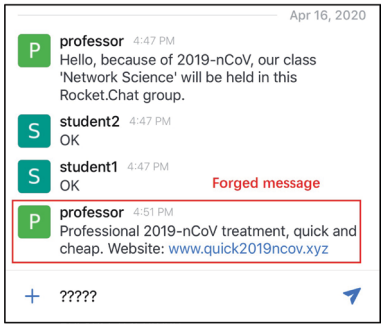
**Figure 3. The attacker determines the accuracy of the specified sequence number by observing the shared IPID counter.[9]**




**(a)** The specified sequence number is incorrect

**(b)** The specified sequence number is correct

**Figure 4. Snapshot of Web application poisoning.[9]**




based policy to the hash-based policy for their TCP packets after receiving forged ICMP error messages. We have implemented a prototype to perform case studies on a wide range of applications—for example, HTTP, SSH and BGP—to validate the effectiveness of the identified off-path TCP hijacking attack due to cross-layer information leakage. We've shown that an off-path attacker can infer the sequence number of a target TCP connection on port 22 within 155 seconds, thus crafting an out-of-band TCP RST packet to tear down the victim SSH session to cause a denial-of-service (DoS) attack. In addition, the attacker can infer the sequence and acknowledgment numbers of a target TCP connection within 215 seconds, thus crafting a TCP data packet to poison Web applications or BGP routing tables.[9] Figure 4 is a snapshot of our attack against a Web application, in which an attacker identifies a TCP connection and proceeds to inject a fake message.

## Desynchronization

Desynchronization within the TCP/IP protocol suite caused by crafted ICMP errors refers to a situation where multiple protocols simultaneously work with the same variable or data unit. Factors such as network delays or conditional competition introduced by a crafted ICMP error message can cause these protocols to lose synchronization, leading to ambiguity around the value of that variable or data unit. This disruption can degrade the network's original functionality or semantics, creating opportunities for attackers to exploit and compromise network systems.

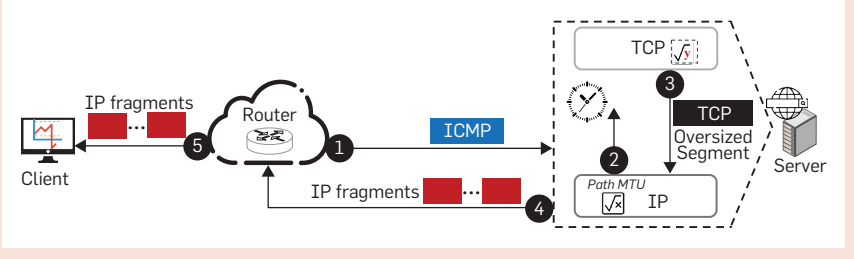
Consider the path MTU value, which is a global variable maintained within the host's IP layer. This value defines the maximum IP packet size for the path from the host to a specific destination IP address. Operations on the path MTU value extend beyond the IP protocol and involve various other protocols, such as TCP and UDP. Ideally, using the path MTU value to determine TCP segment size should eliminate the need for IP fragmentation. However, we have demonstrated that, in practice, simultaneous updates on this global variable by various protocols, tricked by a crafted ICMP error message, can lead to desynchronization issues.[12] This can result in discrepancies between the path MTU value at the IP layer and the MTU value read by the TCP layer, potentially causing the TCP layer to transmit oversized segments, leading to abnormal IP fragmentation. Consequently, off-path attackers can inject manipulated IP fragments into the target TCP connection, causing the mis-reassembling of IP fragments and disrupting the target TCP traffic without needing to infer random sequence numbers.

**Forcing IP fragmentation on TCP.** It is a widespread belief that TCP is immune to IP fragmentation because TCP enables path MTU discovery (PMTUD) by default. This mechanism detects the maximum allowed packet size along the path and enables TCP to adjust the maximum segment size (MSS) accordingly, thus avoiding IP fragmentation on TCP.[29,31] In practice, the detected path MTU value is a global variable maintained at the IP layer. Consequently, when multiple protocols, such as IP, TCP, UDP, and others, simultaneously interact with it, unexpected synchronization issues may occur, resulting in unintended IP fragmentation on TCP segments.

As shown in Figure 5, a router on the Internet may generate an ICMP error message (an ICMP Destination Unreachable message with the code "Packet too big") directed at the server. This ICMP error message can be triggered by various protocol sessions from the server, such as UDP or ICMP echo. Upon reaching the server, this message updates the global variable of path MTU in the IP layer based on its contents. However, as this message lacks specific TCP connection information, the update to the path MTU value is not immediately synchronized with the TCP layer. Instead, the IP layer defers feedback until it passively detects the TCP connection by receiving oversized TCP segments, which it then fragments and sends out. Once the IP layer acknowledges the TCP connection, it updates the TCP layer with the new path MTU value, allowing TCP to adjust the MSS of subsequent segments to avoid IP fragmentation.

This desynchronization issue concerning the path MTU value between TCP and IP undermines the primary purpose of the path MTU discovery mechanism and causes unintended IP fragmentation on TCP segments. In particular, we find that off-path attackers on the Internet can impersonate a router and forge such an ICMP error message to trick the server into fragmenting its TCP segments. This manipulation exploits the inherent challenge in verifying the source and transmission path of ICMP error messages within the current Internet infrastructure. For example, we can forge the ICMP error message to include an embedded ICMP echo reply packet, effectively tricking the server into fragmenting TCP segments and introducing a new attack vector.

**Figure 5. IP fragmentation on TCP segments due to desynchronization of the path MTU value between IP and TCP.**



**Poisoning TCP traffic via IP fragmentation.** Once TCP packets experience IP fragmentation due to the desynchronization issue, an off-path attacker may exploit this vulnerability to launch IP fragmentation injection attacks against TCP traffic. As shown in Figure 6, at first the off-path attacker may employ various techniques, such as social engineering or network side channels, to detect the existence of a TCP connection between a victim server and a client. Then, the attacker forges an ICMP error message and sends it to the server, triggering the desynchronization vulnerability on path MTU in the server's TCP/IP protocol suite. This manipulation causes IP fragmentation on the TCP packets sent from the server to the client. Following this, the attacker impersonates the server via IP spoofing and sends crafted IP fragments to the victim client. As a result, legitimate fragments from the server will be incorrectly reassembled with the malicious ones introduced by the attacker. Ultimately, this leads to the replacement of the original data within the TCP packets, initiating a poisoning attack on the targeted TCP stream. According to RFC 791, the minimum IP fragments on the Internet is 68 octets; thus, the random sequence and acknowledgment numbers are always carried in the first benign fragment from the server. Consequently, IP-fragmentation-based poisoning attacks against TCP can be performed without the need to infer the random sequence and acknowledgment numbers.[c]

**Experimental results.** We demonstrated that off-path attackers can manipulate HTTP traffic via our attack. A malicious JavaScript installed at the victim client via spam aids the attacker in synchronizing timing and aligning data to poison the local Web cache.[d] The connection to the target HTTP server is established by the puppet, and the connection and segments from the HTTP server are known to the attacker. Consequently, leveraging our method, the attacker can craft subtle IP fragments to force the incorrect reassembly of both legitimate and malicious fragments, thereby poisoning the client's Web cache, leading to regular users encountering poisoned local cache data when accessing the HTTP server later.

Furthermore, we showed that an off-path attacker can manipulate BGP routing tables via our attack. The attacker first probes periodically advertised BGP messages in advance.[15] Then, it manipulates BGP routers into fragmenting TCP segments by sending forged ICMP error messages. Finally, the attacker injects forged fragments into the BGP messages to poison the routing tables. Figure 7 illustrates the altered routing information received by a victim BGP router within our test-bed environment, which differs from the original routing information advertised by its peer BGP router. In this scenario, the attacker has replaced the legitimate routing information of 10.2.2.0/24 with a counterfeit entry of 12.2.0.0/24 by injecting meticulously crafted IP fragments into the victim BGP router. Our experimental findings indicate that these attacks can pose a significant threat to Internet infrastructure. It is worth noting that a session encryption mechanism (for example, TLS) will mitigate the identified IP fragmentation attacks against TCP, since the mis-reassembled TCP segment cannot pass the up-layer verification and will be discarded. However, the discarding of the mis-reassembled TCP segment will incur a performance loss, since benign fragments are also discarded together.

**Protocols may inherently fall short in comprehensively addressing the wide spectrum of data types and exceptional scenarios when processing packets carrying cross-layer data.**

### Semantic Gap

Protocols may inherently fall short in comprehensively addressing the wide spectrum of data types and exceptional scenarios when processing packets carrying cross-layer data, giving rise to gaps in understanding that hinder the proper response to such packets. To maintain network functionality, protocols may resort

c It is worth noting that the handling of overlapped IP fragments is an implementation decision. Popular operating systems (for example Linux, OpenBSD, Windows) handle overlapped IP fragments on a first-come, first-served basis,[12] which allows attackers to send crafted IP fragments to the victim client in advance, facilitating the construction of our attack.

d The malicious JavaScript is sandboxed by the client's browser, having limited privileges, and cannot access any information within the TCP/IP protocol suite.[12]

to employing default and imprecise processing methods when responding to these packets, thereby introducing the possibility of semantic mismatches—semantic gaps—that attackers can exploit to compromise the system's security. Specifically, we uncovered that, due to such a semantic-gap vulnerability in the legitimacy checks against ICMP error messages, an off-path attacker on the Internet can craft an ICMP redirect message to evade the receiver's (for example, a public server) checks. This tricks the receiver into modifying its routing table incorrectly and forwarding its IP traffic to black holes, thereby conducting a stealthy DoS attack against public servers on the Internet.[11]

**DoS via semantic gap of ICMP error checks.** Figure 8 illustrates our design for constructing a DoS attack against a victim server, redirecting its traffic intended for the victim client into a black hole hosted by a neighboring host of the server that is unable to forward network traffic. Initially, the server can successfully forward its traffic to the client. An off-path attacker on the Internet identifies a neighboring host near the target server through actions like ICMP echo requests (for example, using the ping tool). This host will later serve as a routing black hole. The attacker then impersonates the victim client by IP spoofing and sends a crafted UDP request to the server. Deceived by the request, the server responds with a UDP reply to the victim client, which the victim client eventually discards.

The attacker then embeds the predictable UDP reply packet into a crafted ICMP redirect message and sends it to the victim server. According to the ICMP specifications,[3,5,35] the server will check the first 28 octets of the embedded UDP reply packet to validate the legitimacy of the received ICMP redirect message, thereby verifying the existence of the corresponding UDP socket, even though it cannot check any further information due to UDP's stateless nature. Since the attacker previously tricked the server into establishing the UDP socket for the victim client, this crafted ICMP redirect message will pass the server's legitimacy check. Consequently, the server will mistakenly accept and respond to the message, redirecting its traffic for the victim client to the neighboring host as specified by the forged ICMP redirect message. However, the neighboring host lacks routing and forwarding capabilities and will discard the server's traffic. This results in a cross-layer DoS attack on all network sessions above the IP layer of the server.

**Experimental results.** We conducted large-scale measurements on the Internet, revealing that this DoS attack, due to the semantic-gap vulnerability in the ICMP error message's legitimacy check mechanism, can be exploited to pose a significant threat to the Internet. In our ethical measurement studies, we first initiate a session between our controlled client and the target server (for example, an HTTP session from our Web client to a public HTTP server). Then, using the identified ICMP redirect DoS at-

**Figure 6. Poisoning TCP traffic via IP fragmentation.[12]**

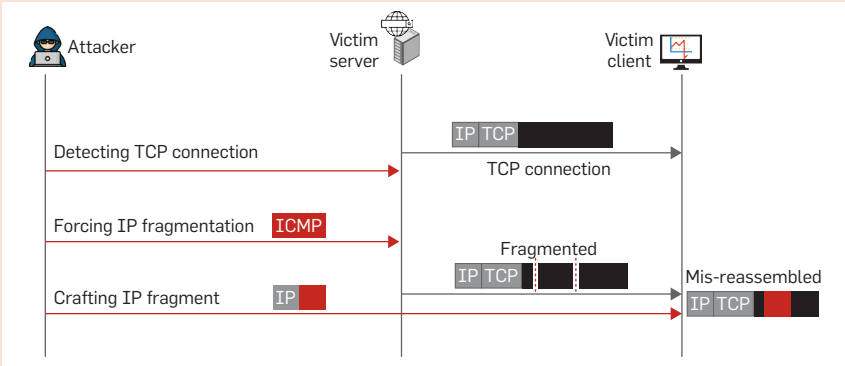



**Figure 7. Fake routing due to IP fragments injection.[12]**

```
BGP table version is 0, local router ID is 10.3.0.50
Status codes: s suppressed, d damped, h history, * valid, > best, =
    i internal, r RIB-failure, S Stale, R Removed
Origin codes: i - IGP, e - EGP, ? - incomplete
```

| | Network | Next Hop | Metric | LocPrf | Weight | Path | |
|---|---|---|---|---|---|---|---|
| *> | 10.1.1.0/24 | 10.1.0.50 | 0 | | 0 | 7675 | i |
| *> | 10.23.23.0/24 | 10.1.0.50 | 0 | | 0 | 7675 | i |
| *> | 10.23.29.0/24 | 10.1.0.50 | 0 | | 0 | 7675 | i |
| *> | 10.23.31.0/24 | 10.1.0.50 | 0 | | 0 | 7675 | i |
| *> | 12.2.0.0/24 | 10.1.0.50 | 0 | | 0 | 7675 | i |

**Figure 8. DoS via semantic gaps of ICMP error checks.[11]**

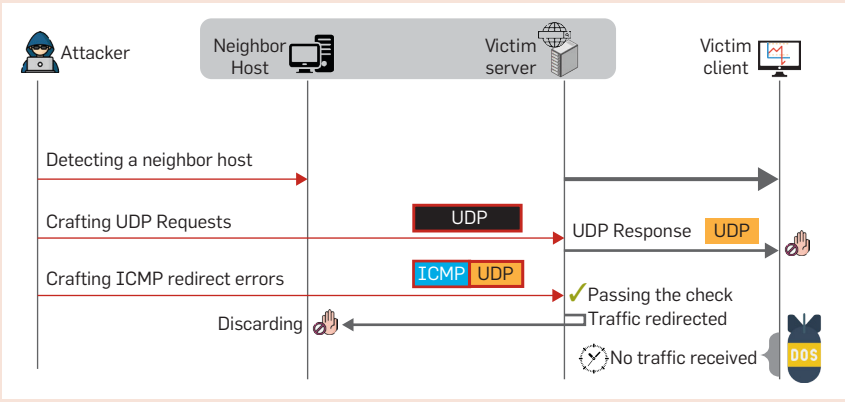
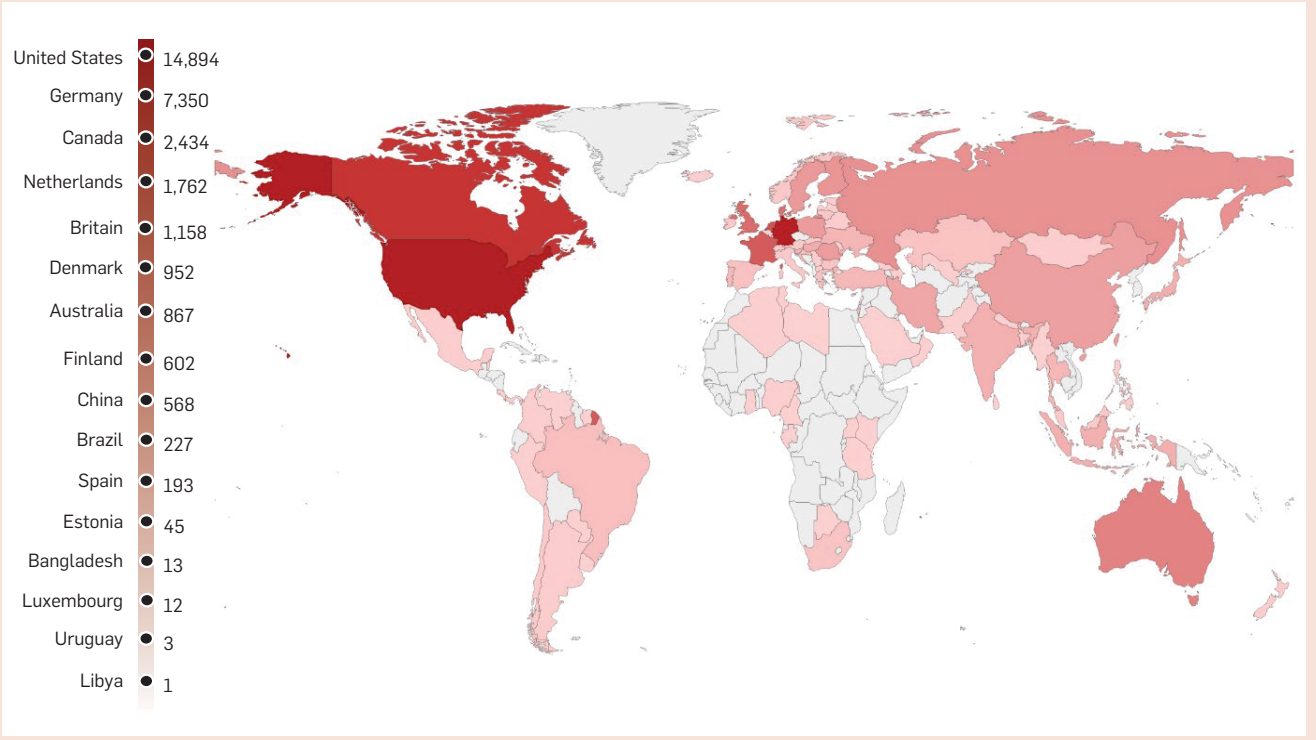


**Figure 9. Distribution of websites with the semantic-gap vulnerability.[11]**



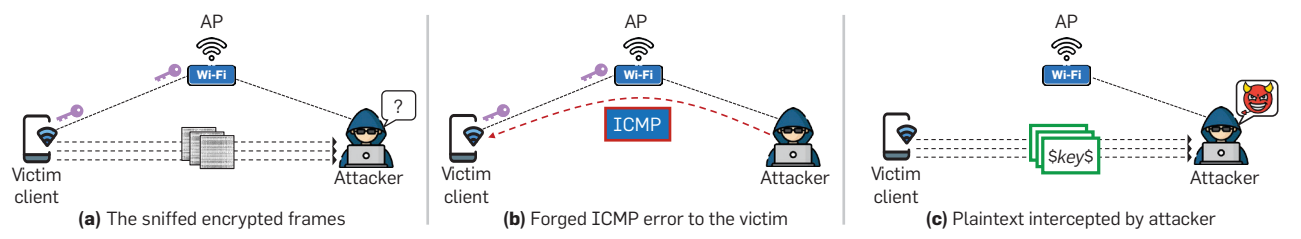

tack, we redirect the "server-to-our-client" traffic to a black hole. This causes subsequent requests from our controlled client to the server to fail, demonstrating that the server is vulnerable to our attack while not affecting the server's regular users. Our experimental results show that the identified DoS attack can target not only individual users, preventing them from visiting a Web server, but also server-to-server communication, such as shutting down a DNS resolver from contacting a particular authoritative name server (under our control in the experiments due to ethical considerations) to resolve domain names. It is even possible to disrupt the entire operation of a service such as Tor by breaking down the communication between a Tor relay node and a next hop. Our one-month empirical study on the Internet revealed that 43,081 popular websites, 54,470 open DNS resolvers, and 186 Tor relay nodes, spanning 5,184 autonomous systems (ASes) across 185 countries, are vulnerable to the semantic gap vulnerability and susceptible to the identified remote DoS attack. Figure 9 shows the geographical distribution of the vulnerable websites we detected.

## Identity Deception

The problem of identity deception stems from the lack of security auditing for data sources during cross-layer interactions among multiple protocols within the TCP/IP protocol suite, a particular source of ICMP errors. This allows attackers to craft specific control protocol data, disrupting the normal operation of the network. We show that in specific network scenarios, such as Wi-Fi networks, identity deception can be particularly severe, presenting one of the most common challenges.[13] In public Wi-Fi networks such as those found in airports, coffee shops, campuses, and hotels, an attacker (a malicious client) may connect to the network and employ source-IP-address spoofing techniques to impersonate the access point (AP) gateway. The attacker can then send forged ICMP routing update control messages (that is, ICMP redirect messages designed for AP gateways only in Wi-Fi networks) to other clients. These forged ICMP routing update control messages will pass through the AP gateway; if the AP gateway fails to block the forged messages that finally arrive at other clients, these clients will be tricked into following the messages' instructions and setting the attacker as their new AP router, granting the attacker the ability to intercept traffic within the Wi-Fi network. What is more concerning is that this vulnerability enables the attacker to evade Wi-Fi protocol security measures such as WPA3, granting access to plaintext traffic.

**Wi-Fi traffic hijacking.** Figure 10 shows an overview of how to intercept plaintext traffic in Wi-Fi networks by leveraging the identity-deception vulnerability. In Wi-Fi networks, due to the shared nature of wireless channels, a malicious client may eavesdrop on wireless frames belonging to other clients. These frames, however, are usually encrypted by security mechanisms at the link layer, such as WPA2 or WPA3. As a result, it is difficult for the attacker to directly access plaintext information. We discovered a security vulnerability within the network processing unit (NPU) employed in AP routers. Driven by the quest for high-speed packet forwarding, these NPU chips within AP routers directly forward received ICMP messages (including forged ICMP errors from an attack)

**Figure 10. Wi-Fi traffic interception via identity deception.**




at the hardware level, thus failing access control list (ACL) rules defined at the higher layers to verify and block forged messages.

As shown in Figure 10b, this vulnerability allows the attacker to impersonate the AP router and craft an ICMP redirect message to manipulate the IP routing of the victim client. Even though such a message is meant to originate exclusively from the AP router itself and exhibits obvious illegitimate characteristics (for example, its source being the AP router's IP address), due to the NPU's direct forwarding of the message, this message passes through the AP router and remains unblocked. Ultimately, the message reaches the victim client, which is deceived into believing it originated from the AP router. As a result, the victim client updates its IP routing, designating the attacker as the next-hop gateway. This causes all subsequent traffic from the victim client to be rerouted through the attacker.

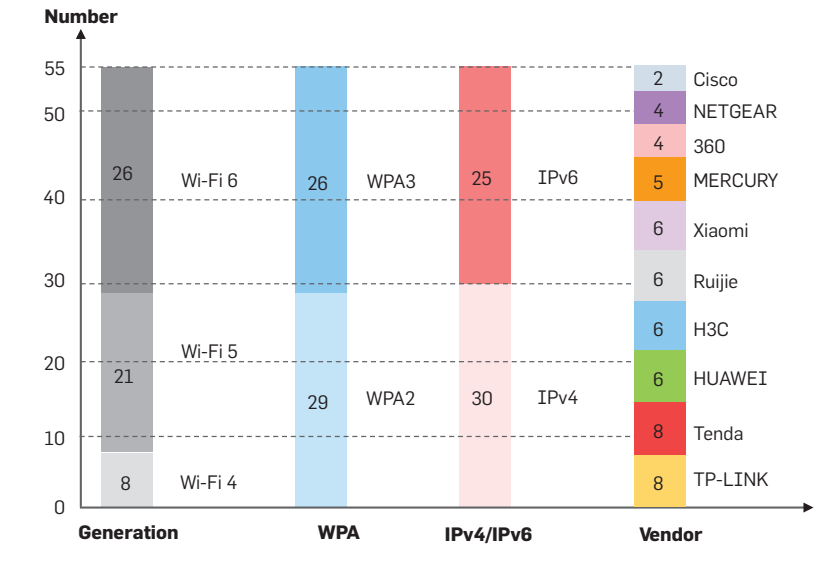
Note that this attack can bypass the link-layer encryption-protection mechanisms employed in Wi-Fi networks (for example, WPA2 and WPA3). WPA2 and WPA3 provide per-hop encryption at the link layer using a session key shared between the AP and each attached client. Due to the crafted ICMP redirect message, however, the victim client sets the attacker as the next hop in the IP layer. Therefore, when the AP receives the encrypted link-layer frames from the victim client, it needs to perform multi-hop relaying at the link layer to complete forwarding the frames to the next hop (that is, the attacker). Consequently, the AP first decrypts the encrypted frames using the shared secret key with the victim client. Next, according to the Destination Address (which has been poisoned by the attacker) in the frame header, the AP encrypts the frames using the secret key shared with the attacker and sends them to the attacker. Finally, after decrypting the frames, the attacker can intercept the victim client's plaintext traffic, and the link-layer per-hop encryption in Wi-Fi networks is successfully evaded.[13]

**Experimental results.** We conducted real-world evaluations to assess the impact of our attack. Initially, we investigated whether popular AP routers could effectively block forged ICMP redirect messages sent from an attacker to a victim client. Our assessment covered 55 popular wireless routers spanning 10 vendors (as shown in Figure 11). Our findings revealed that none of these routers block forged ICMP redirects from passing through. The root cause of this identity deception vulnerability, stemming from the flawed design of NPUs, has been officially recognized by Qualcomm (CVE-2022-2566) and HiSilicon (HWSA21-085272813). HUAWEI, H3C, Ruijie, MERCURY, NETGEAR, and Tenda have also confirmed the presence of this vulnerability in their AP routers due to the NPU. Furthermore, we evaluated 122 real-world Wi-Fi networks across various locations, including coffee shops, hotels, libraries, cinemas, and campuses, finding that 109 of these networks were vulnerable.

## Countermeasures

We responsibly disclosed the identified vulnerabilities to the affected organizations. We reported the IPID assignment policy vulnerability, triggered by a forged ICMP "Packet too big" message, to the Linux community. They acknowledged it (CVE-2020-36516) and improved the IPID design starting from kernel version 5.16. The desynchronization and semantic-gap vulnerabilities, which are exploitable

**Figure 11. Distribution of 55 vulnerable AP routers.[13]**



for IP fragmentation and remote DoS attacks, were reported to Linux and FreeBSD. Both confirmed receipt, and we are awaiting updates. Qualcomm acknowledged and fixed the Wi-Fi identity deception vulnerability caused by crafted ICMP redirects in their Snapdragon chipsets (CVE-2022-2566); other affected vendors are still working on fixes. We also reported the vulnerabilities in the legitimacy-check mechanism of ICMP errors to the Internet Engineering Task Force (IETF) and are discussing our countermeasures with them.

**Enhancing ICMP error authentications.** The root cause of the four off-path attacks presented in this article is that an off-path attacker can forge ICMP error messages to bypass the receiver's legitimacy checks, leading to unintended protocol interactions and vulnerabilities. The most straightforward prevention measure is to strengthen the authentication of received ICMP error messages. However, as discussed earlier, verifying the legitimacy of ICMP errors is challenging due to two inherent limitations in the current ICMP specifications. First, certain ICMP errors (for example, ICMP Destination Unreachable messages with the code "Packet too big" exploited to trigger the information leakage and desynchronization vulnerabilities) can originate from any intermediate router, rendering source-based blocking ineffective. Second, although ICMP specifications mandate including at least the first 28 octets of the original packet, off-path attackers can evade this by embedding a crafted UDP or ICMP payload into the forged ICMP error messages, thwarting authentication due to the statelessness and lack of memory in the UDP and ICMP protocols.

Inspired by RFC 5961's challenge ACK mechanism[37] for defending against out-of-band TCP packet injection, we propose enhancing ICMP error authentication by introducing a new *challenge-and-confirm* mechanism. In particular, when a receiver gets an ICMP error message embedded with a stateless protocol payload (like UDP/ICMP), verifying its authenticity can be difficult. To address this, the receiver can send another (UDP/ICMP) packet on the established network session to the destination, embedding a hash value in the IP options field. If the prior ICMP error message was legitimate, this new packet will trigger another ICMP error message containing the hash value. This allows the receiver to verify authenticity and respond correctly. This challenge-and-confirm mechanism effectively defends against off-path forged ICMP error messages with minimal changes to the TCP/IP protocol suite. It only requires updates to the ICMP error message verification code on end hosts, without modifying intermediate routing devices, and it is backward compatible. We are discussing this mechanism with the IETF.

**Our one-month empirical study on the Internet revealed that 43,081 popular websites, 54,470 open DNS resolvers, and 186 Tor relay nodes, spanning 5,184 autonomous systems across 185 countries, are vulnerable to the semantic gap vulnerability.**

**Securing sessions via cryptography.** Another mitigation method is to use cryptography to secure network sessions as much as possible, such as with TLS,[38] QUIC,[21] and TCP-MD5/TCP-AO.[40] This way, even if an off-path attacker exploits forged ICMP error messages to trigger vulnerabilities in the TCP/IP stack, it is difficult for the attacker to cause real harm to applications. For instance, even if an attacker manipulates the server's IPID with ICMP error messages to create a side channel and guesses the sequence number of a target TCP connection, the injected TCP packet will fail TCP-MD5/TCP-AO or TLS validation and be discarded. Similarly, if an off-path attacker intercepts a victim client's packets in a Wi-Fi network as a man in the middle and evades link-layer encryption like WPA3, the end-to-end encryption provided by protocols such as TLS or QUIC makes it challenging for the attacker to access plaintext application data, thereby limiting the attack's impact.

## Conclusion and Future Work

Off-path attacks on the TCP/IP protocol suite present a significant challenge to Internet security, as they do not constrain the attacker's network topology and require minimal resources. Previous research has demonstrated that off-path attackers can exploit vulnerabilities in the TCP/IP protocol suite to launch various attacks, such as TCP hijacking,[6,33,36] routing manipulation,[32] and Web and DNS cache poisoning.[16,17,19,23] However, off-path attacks facilitated by forged ICMP errors have received limited attention.[22,28]

In our study, we systematically revealed four security issues caused by forged ICMP errors: information leakage, desynchronization, semantic gaps, and identity deception. These issues can be exploited by attackers to pose severe security threats to the Internet. Essentially, these security issues arise from the disruption of the protocol's intended communication processes and semantic integrity by forged ICMP error messages, leading to unexpected behaviors that attackers can exploit. We call these *vulnerabilities,* as they are protocol interaction semantic vulnerabilities caused by forged ICMP errors, distinguishing them from memory corruptions caused by unsafe programming practices. Given that ICMP (including ICMPv6) is widely implemented and crucial across various TCP/IP protocol stacks, the semantic vulnerabilities caused by forged ICMP errors may extend beyond the four we have identified.

A critical area of focus for future research is automated identification of these semantic vulnerabilities, for example, by leveraging techniques from program analysis[7,14] and AI models.[30,39] In program analysis, data-flow analysis can be employed to trace the movement of packets through the protocol stack, helping to detect vulnerabilities such as desynchronization issues during packet data processing. AI models trained on network traffic patterns can identify anomalies, enabling the early detection of potential vulnerabilities. By integrating these approaches, it may be possible to develop more proactive and automated methods for identifying and mitigating security risks within network protocols.

## Acknowledgments

This work was supported in part by the National Key Research and Development Program of China under Grant 2022YFB3102301, the National Science Foundation for Distinguished Young Scholars of China under No. 62425201, the Science Fund for Creative Research Groups of the National Natural Science Foundation of China under No. 62221003, and the Key Program of the National Natural Science Foundation of China under No. 62132011 and No. 61932016.

**References**

1. Alexander, G., Espinoza, A.M., and Crandall, J.R. Detecting TCP/IP Connections via IPID hash collisions. *Proceedings on Privacy Enhancing Technologies 4* (2019), 311–328.
2. Ali, F. IP Spoofing. *The Internet Protocol J. 10*, 4 (2007), 1–9.
3. Baker, F. *Requirements for IP Version 4 Routers*. RFC 1812. Internet Engineering Task Force, (1995); http://www.rfc-editor.org/rfc/rfc1812.txt
4. Bellovin, S.M. A look back at "security problems in the TCP/IP protocol suite". In *20th Annual Computer Security Applications Conf.* IEEE, (2004), 229–249.
5. Braden, R. *Requirements for Internet Hosts Communication Layers*. RFC 1122. Internet Engineering Task Force, (1989); 10.17487/RFC1122
6. Cao, Y. et al. Off-path TCP exploits: Global rate limit considered dangerous. In *25th USENIX Security Symp. (USENIX Security 16)*. USENIX, (2016), 209–225.
7. Cao, Y. et al. Principled unearthing of TCP side channel vulnerabilities. In *Proceedings of the 2019 ACM SIGSAC Conf. on Computer and Communications Security*. ACM, (2019), 211–224.
8. Duke, M. et al. *A Roadmap for Transmission Control Protocol (TCP) Specification Documents*. RFC 7414. Internet Engineering Task Force, (2015); http://www.rfc-editor.org/rfc/rfc7414.txt
9. Feng, X. et al. Off-path TCP exploits of the mixed IPID assignment. In *Proceedings of the 2020 ACM SIGSAC Conf. on Computer and Communications Security*. ACM, (2020), 1323–1335.
10. Feng, X. et al. Off-path TCP hijacking attacks via the side channel of downgraded IPID. *IEEE/ACM Transactions on Networking 30*, 1 (2021), 409–422.
11. Feng, X. et al. Off-path network traffic manipulation via revitalized ICMP redirect attacks. In *31st USENIX Security Symp. (USENIX Security 22)*. USENIX, (2022), 2619–2636.
12. Feng, X. et al. PMTUD is not panacea: Revisiting IP fragmentation attacks against TCP. In *Network and Distributed System Security Symp. (NDSS)*. Internet Society, (2022).
13. Feng, X. et al. Man-in-the-middle attacks without rogue AP: When WPAs meet ICMP redirects. In *2023 IEEE Symp. on Security and Privacy (S&P)*. IEEE, (2022), 694–709.
14. Fiterau-Brostean, P. et al. Automata-based automated detection of state machine bugs in protocol implementations. In *Network and Distributed System Security Symp. (NDSS)*. Internet Society, (2023).
15. Flavel, A. et al. BGP route prediction within ISPs. *Computer Communications 33*, 10 (2010), 1180–1190.
16. Gilad, Y. and Herzberg, A. Fragmentation considered vulnerable: Blindly intercepting and discarding fragments. In *Proceedings of the 5th USENIX Conf. on Offensive Technologies*. USENIX, (2011).
17. Gilad, Y. and Herzberg, A. Off-path attacking the web. In *Proceedings of the 6th USENIX Conf. on Offensive Technologies*. USENIX, (2012), 41–52.
18. Gilad, Y. and Herzberg, A. Fragmentation considered vulnerable. *ACM Trans. on Information and System Security 15*, 4 (2013), 16.
19. Gilad, Y., Herzberg, A., and Shulman, H. Off-path hacking: The illusion of challenge-response authentication. *IEEE Security & Privacy 12*, 5 (2013), 68–77.
20. Gont, F. *ICMP Attacks against TCP*. RFC 5927. Internet Engineering Task Force, (2010); http://www.rfc-editor.org/rfc/rfc5927.txt
21. Iyengar, J. and Thomson, M. *QUIC: A UDP-Based Multiplexed and Secure Transport*. RFC 9000. Internet Engineering Task Force, (2021); http://www.rfc-editor.org/rfc/rfc9000.txt
22. Keyu, M., Zhou, X., and Qian, Z. DNS cache poisoning attack: resurrections with side channels. In *Proceedings of the 2021 ACM SIGSAC Conf. on Computer and Communications Security*. ACM, (2021), 3400–3414.
23. Klein, A. Cross layer attacks and how to use them (for DNS cache poisoning, device tracking and more). In *2021 IEEE Symp. on Security and Privacy (S&P)*. IEEE, (2021), 1179–1196.
24. Klein, A. Subverting stateful firewalls with protocol states. In *Network and Distributed System Security Symp. (NDSS)*. Internet Society, (2022).
25. Larsen, M.V. and Gont, F. *Recommendations for TransportProtocol Port Randomization*. RFC 6056. Internet Engineering Task Force, (2011); http://www.rfc-editor.org/rfc/rfc6056.txt
26. Lichtblau, F. et al. Detection, classification, and analysis of inter-domain traffic with spoofed source IP addresses. In *Proceedings of the 2017 Internet Measurement Conf.* ACM, (2017), 86–99.
27. Luckie, M. et al. Network hygiene, incentives, and regulation: Deployment of source address validation in the Internet. In *Proceedings of the 2019 ACM SIGSAC Conf. on Computer and Communications Security*. ACM, (2019), 465–480.
28. Man, K. et al. DNS cache poisoning attack reloaded: Revolutions with side channels. In *Proceedings of the 2020 ACM SIGSAC Conf. on Computer and Communications Security*. ACM, (2020), 1337–1350.
29. McCann, J. et al. *Path MTU Discovery for IP Version 6*. RFC 8201. Internet Engineering Task Force, (2017); http://www.rfc-editor.org/rfc/rfc8201.txt
30. Mirsky, Y. et al. VulChecker: Graph-based vulnerability localization in source code. In *32nd USENIX Security Symp.* USENIX, (2023).
31. Mogul, J. and Deering, S. *Path MTU Discovery*. RFC 1191. Internet Engineering Task Force, (1990); http://www.rfc-editor.org/rfc/rfc1191.txt
32. Nakibly, G. et al. Persistent OSPF attacks. In *Network and Distributed System Security Symp. (NDSS)*. Internet Society, (2012).
33. Pan, Y. and Rossow, C. TCP spoofing: Reliable payload transmission past the spoofed TCP handshake. In *2024 IEEE Symp. on Security and Privacy*. IEEE, (2024), 179–179.
34. Pokhrel, S.R. et al. TCP Performance over Wi-Fi: Joint Impact of Buffer and Channel Losses. *IEEE Trans. on Mobile Computing 15*, 5 (2016), 1279–1291; 10.1109/TMC.2015.2456883
35. Postel, J. *Internet Control Message Protocol*. RFC 792. Internet Engineering Task Force, (1981); 10.17487/RFC792
36. Qian, Z. and Mao, Z.M. Off-path TCP sequence number inference attack-how firewall middleboxes reduce security. In *2012 IEEE Symp. on Security and Privacy*. IEEE, (2012), 347–361.
37. Ramaiah, A., Stewart, R., and Dalal, M. *Improving TCP's Robustness to Blind In-Window Attacks*. RFC 5961. Internet Engineering Task Force, (2010); http://www.rfc-editor.org/rfc/rfc5961.txt
38. Rescorla, E. *The Transport Layer Security (TLS) Protocol Version 1.3*. RFC 8446. Internet Engineering Task Force, (2018); http://www.rfc-editor.org/rfc/rfc8446.txt
39. Thapa, C. et al. Transformer-based language models for software vulnerability detection. In *Proceedings of the 38th Annual Computer Security Applications Conf.* ACM, (2022), 481–496.
40. Touch, J., Mankin, A., and Bonica, R.P. *The TCP Authentication Option*. RFC 5925. Internet Engineering Task Force, (2010); http://www.rfc-editor.org/rfc/rfc5925.txt

**Xuewei Feng** is a research scientist at Tsinghua University, China. His research interests include network security and software vulnerability detection.

**Qi Li** is an associate professor with the Institute for Network Sciences and Cyberspace, Tsinghua University, China. His research interests include Internet and cloud security, IoT security, and AI security.

**Kun Sun** is a full professor at George Mason University. He serves as the director of the Sun Security Laboratory (SunLab) and the associate director of the Center for Secure Information Systems (CSIS).

**Ke Xu** (xuke@tsinghua.edu.cn) is a full professor at Tsinghua University, China. He has published over 200 technical papers in the research areas of next-generation Internet and network security.

**Jianping Wu** is a full professor at Tsinghua University, China. He has authored over 200 technical articles on the network architecture, high-performance routing and switching, protocol testing, and network security.



Watch the authors discuss this work in the exclusive *Communications* video. https://cacm.acm.org/videos/off-path-attacks

DOI:10.1145/3696205

**With applications in drug delivery, advanced diagnosis, and patient monitoring, molecular communications in the bloodstream is a promising area of research.**

BY LUCA FELICETTI, MAURO FEMMINELLA, AND GIANLUCA REALI

# Molecular Communications in Blood Vessels: Models, Analysis, and Enabling Technologies

MOLECULAR COMMUNICATIONS (MC) refers to the exchange of biological material to transmit information between biological entities. As one of the breakthrough research areas of the past decade, MC has been touted for its potential use in medical contexts. In this article, we consider a specific biological environment—the bloodstream—a promising area for MC research due to its ability to exchange information at a systemic level. We analyze possible alternatives to using MC in this challenging environment, providing models, analysis techniques, and potential enabling technologies to implement the proposed alternatives. As an example, we use the design of a device for monitoring health parameters to discuss the suitability of applying current findings in MC to the bloodstream environment and to provide future research directions.

## Background

The research on molecular communications over the past decade arose from the need to identify information-exchange mechanisms between nanodevices operating in biological environments.[2] Working on biological systems led researchers to explore the possibility of reengineering the information-exchange mechanisms already present in these systems. Each biological system requires an exchange of information on a different scale. Consider, for example, the information in DNA, which is encoded and transferred to other particles in a cell's cytoplasm for protein synthesis. Another example is interactions between cells, which allow certain proteins to be captured to trigger various biological processes within cells. Such processes also include several steps of chemical information transduction, typically known as cellular pathways. Bacterial populations within organisms actively participate in the evolution of numerous

## » key insights

- **Molecular communications in the bloodstream is a promising area of research, since the bloodstream has the ability to exchange information at a systemic level.**
- **The kinematic behavior of the main blood particles, induced by the heart pump, determines the overall behavior of the channel, making the propagation environment complex to model.**
- **The monitoring of health parameters with a dermal device can be used to explore the applicability of current forms of molecular communications in the bloodstream.**
- **This analysis suggests a possible roadmap for implementation of the device with different candidate technologies, such as microneedles or fluorescent proteins able to bind with aptamers.**

IMAGE BY MAKSIM KABAKOU

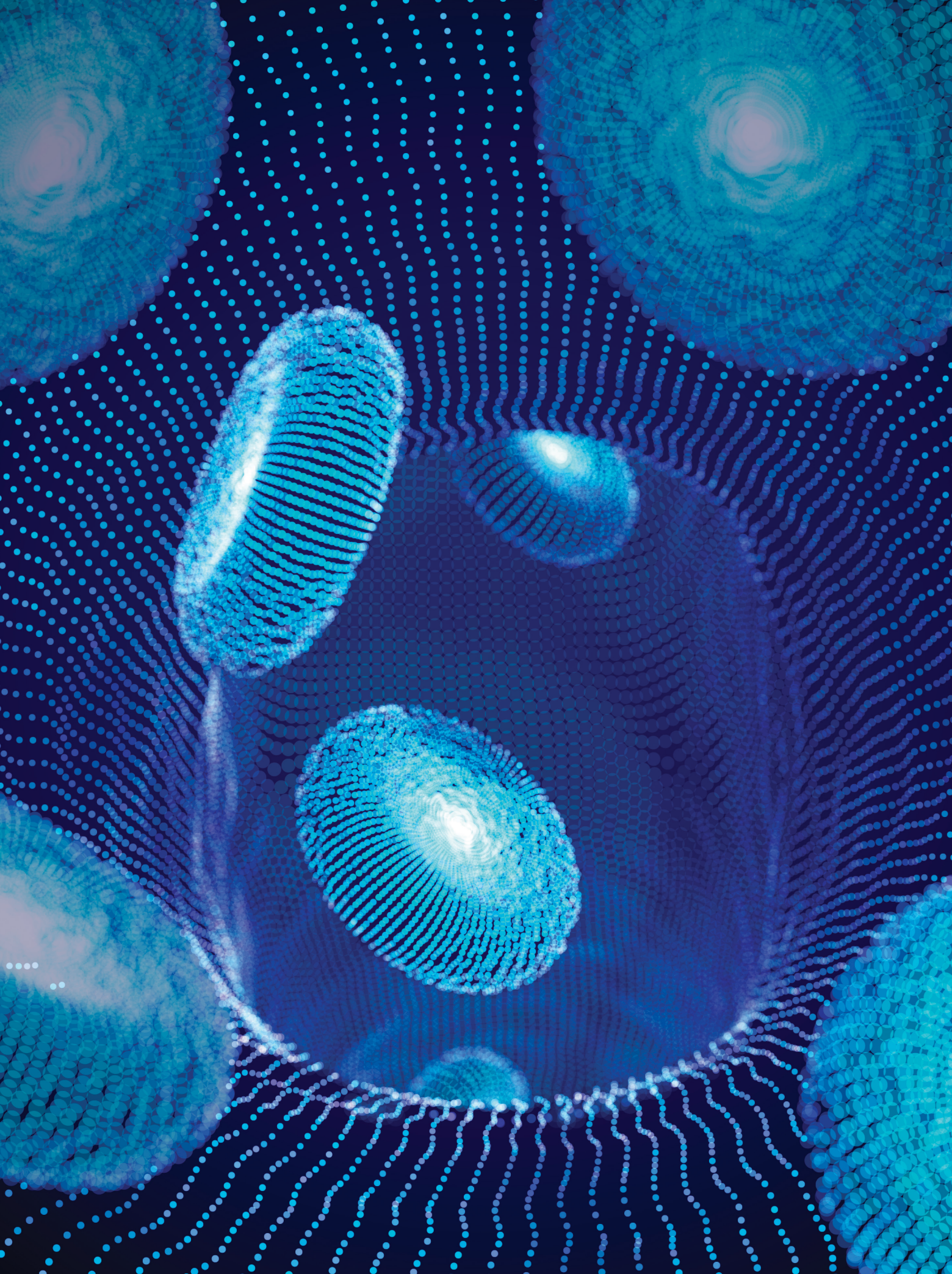
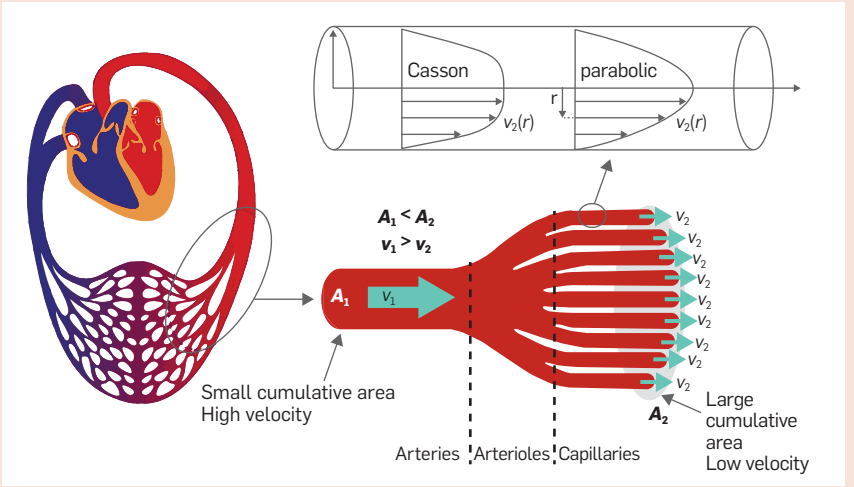


**Figure 1. Overview of the whole circulatory system, including a zoomed-in portion. The dashed lines are relevant to the cross-sectional area of that vessel category (that is, the smaller the vessel, the higher the cross-sectional area), with a graphical illustration of the flow-rate conservation ($A_1v_1 = A_2v_2$) and of the deviation from the parabolic profile into the Casson profile due to the presence of red blood cells, creating a plug region in the middle of the vessel.**




processes, requiring forms of coordination that involve continuous information exchanges between bacteria in the population. Finally, on a larger scale, consider the information flows used to command and coordinate the behavior of organs, such as the nervous and lymphatic systems.

This plethora of possibilities fall under the umbrella of the Internet of Nano-Things (IoNT) network model,[3] with many applications in healthcare, including drug delivery,[7] advanced diagnosis, and patient monitoring.[10]

**MC basics.** Before continuing, we should outline the general structure of an MC system, including the main elements that allow the creation of a molecular transmission chain: information encoding, the communication medium, information transmission in the medium, propagation through this medium, reception, and information decoding for subsequent actions.[21] In most MC proposals, information is conveyed by a stream of chemical compounds, also referred to as carriers, generated by the transmitter (TX) and transported by the aqueous medium of the channel. The exchange of information is based on the mechanism of interaction between carriers and target proteins, present on the surface of the receiver cell (RX). According to this model, information may be encoded in the concentration of the released compounds, in their release time within the symbol interval, in the type of compounds, or in any combination of these.[15] These carriers are released by the TX into the environment through which the transfer takes place, depicted by a propagation model. This can be purely diffusive, as occurs within the cell cytoplasm, or include privileged directions, as occurs in the bloodstream. The receiving process is typically quite complex, as it involves the physical and chemical processes that determine the reception of carriers.

**Why blood vessels?** As they are a natural means of transporting information, blood vessels are attractive to MC researchers. In fact, in addition to transporting vital elements such as oxygen and sugars, they allow the exchange of chemical compounds between organs, which is the basis for the coordination of biological processes; blood vessels therefore transport information at a systemic level. They form a circular closed path that branches out throughout the body through vessels with variable sections, as depicted in Figure 1. In addition, capillaries allow the exchange of chemical compounds (and therefore, of information) with organs.

Despite researchers' interest in using MC in the bloodstream, analyzing and modeling these systems is challenging. The transport mechanism in blood vessels is determined by the movement of the main blood particles, namely red blood cells (RBCs), white blood cells (WBCs), and platelets. The combination of their kinematic behavior, induced by the cardiac pump, determines the overall behavior of the channel. Therefore, the resulting propagation environment is complex to model, even for simple systems.

The objective of this article is to illustrate the research challenges and proposed solutions for the application of MC systems in blood vessels. Its first contribution is a description of the models of MC in blood vessels, considering some elements are not easy to find in the reference literature. An example is how red blood cells (RBCs) influence the velocity profile in small vessels and induce a significant interaction of molecules dispersed in the blood with the endothelium, activating a series of physiological processes. A further contribution is a summary of the most popular analysis methodologies of MC in blood vessels, including the suitability of different solutions with respect to the complexity of the bloodstream environment. The final contribution is a discussion of the enabling technologies that allow interactions between the considered MC systems and the external world. This is aimed at identifying the most suitable technologies for the creation of an MC device intended for monitoring patient health parameters, to be applied to the skin (dermal device). We use the architectural design of this device as a running example throughout the article, aiding our discussion of models and analysis techniques for MC in blood vessels and the main challenges for their adoption in real systems.

## A Monitoring Device Leveraging MC

Figure 2 shows the conceptual scheme of this dermal device, whose purpose is to continuously monitor health parameters, without the need for slower and more invasive laboratory tests. Assuming blood flows from left to right, the TX is on the left, composed of a reservoir of molecules and an injection interface through the skin. Molecules released from the reservoir perfuse blood vessels and propagate downstream.

The RX detection mechanism, on the right-hand side of Figure 2, lever-

ages the in vivo imaging technique, based on light signals emitted by fluorescent biosensors triggered by chemical reactions.[20] In more detail, genetically encoded fluorescent biosensors are proteins engineered to serve as sensors to monitor signal transduction. Fluorescent biosensors convert different signaling activities, such as the concentration of specific proteins, into one of several types of fluorescence signals. They consist of a sensing unit and a reporting unit. The former is typically derived from a cellular protein that participates in the signaling pathway of interest and is therefore intrinsically sensitive to the target signaling event. The latter typically consists of one or more variants of fluorescent proteins coupled to the sensing unit such that signaling-induced changes in the state of the sensing unit alter the fluorescence behavior of these proteins. The specificity of each biosensor lies in the way this coupling is implemented. Thus, in order to allow light emission, fluorescent molecules need to be injected under the skin. These molecules, which are able to bind to signaling molecules captured by the endothelium, will pass through the epidermis, reaching the endothelium. This, in turn, requires preliminary endothelium *activation*, which is a secondary MC from the device to the endothelium, whose objective is to make endothelial cells able to absorb the intended carriers (see the "monitored section" highlighted in yellow in Figure 2) so that they bind with fluorescent molecules to trigger the emission of light signals. Toward this aim, the RX part of the device is made up of a reservoir of both activation and fluorescent molecules, to be released underneath, and a sensor module that detects the level of fluorescence emitted by the underlying compounds actually attached to the endothelium in the monitored section. Fluorescence intensity increases with the number of signaling molecules that bind to the activated endothelium. The distance $L$, between TX and RX, allows molecules to propagate along the vessel according to the fluid dynamics, which in turn can be affected by the *altered* physical properties of the blood flow due to the pathological conditions studied. Any discrepancies in the propagation or absorption patterns can be monitored and measured by the sensor module.

The third module in the middle allows for some basic on-board computation and interfacing with external devices. The detection of body parameters occurs on the bottom layer of the detection system introduced above, essentially confined to the blood vessel. In this case, the entry door of the vessel is the TX equivalent and the monitored section is the RX equivalent. If the device is intended to monitor the presence or concentration of specific substances in the bloodstream, it simply consists of the circuit and RX module, being the TX part of the monitored environment (for example, multiple cells or organs emitting target molecules).

## Models of MC in Blood Vessels

Transmitters and receivers can be genetically modified or artificial elements.[21] In blood vessels, TX and RX can be both fixed (endothelium/organs, fixed injection point, monitoring device) or mobile (cell/nanomachine in the bloodstream). Here, we focus on a fixed dermal monitoring device, supporting fixed-to-fixed MC. We present three taxonomies related to the transmission, propagation, and reception of MC signals in blood vessels. Communications that take place in capillaries or, more generally, in small blood vessels (SBVs), must be distinguished from those that take place in large blood vessels (LBVs). In SBVs, the presence of RBCs has a significant impact due to their mass and tumbling movement in a constrained space, whereas their effects on carrier propagation in LBVs is less impactful (as explained in the "Models of signal propagation" section below).

**Models of transmitted signals.** Although there are several options for encoding digital information in MC, as discussed earlier, some of them cannot be easily used in the bloodstream. In fact, the presence of the drag force acting on the blood flow, together with the RBC tumbling, could significantly alter the transmitted signal and erase the features associated with the transported information. Thus, simple and robust transmitted signals are necessary. Some alternatives for implementing reliable transmissions in the blood environment are reported below and summarized in Table 1.

The ON-OFF pattern is used to transfer digital information, encoded in the sequence of bursts[15] (modeling

**Figure 2. Architecture of a conceptual device that leverages MC in blood vessels to monitor the characteristics of the blood flow. The RX (realized through molecules coming from the patch on the skin that activate the endothelium) and the eventual TX, implemented with microneedles injecting molecules into the tissue surrounding the vessel and penetrating into the blood, are identified, as well as the blood-flow direction and the detector of the received signal implemented in the skin path.**

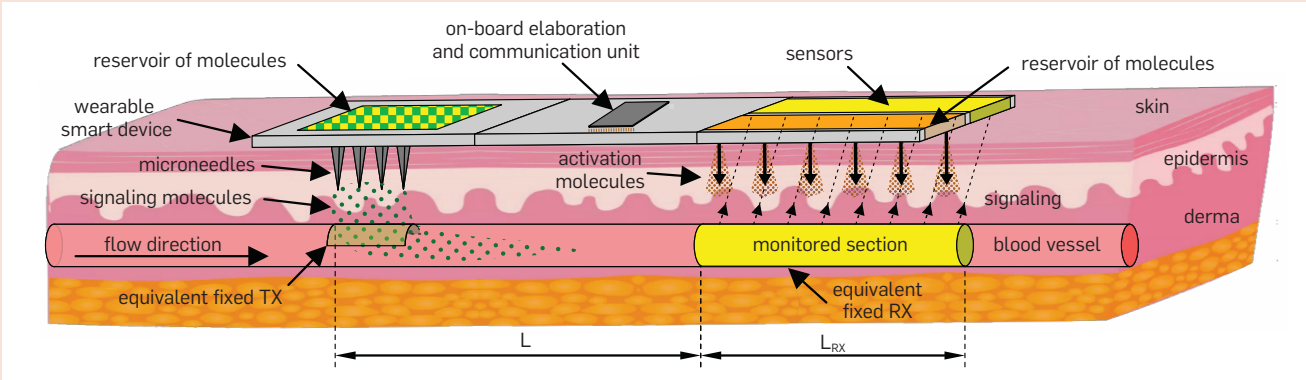


**Table 1. Taxonomy of types of transmission signals and their suitability for use in the bloodstream.**

| Type of Signal | Purpose | Information | Propagation in Blood Vessels | |
|---|---|---|---|---|
| | | | Large blood vessels (LBVs) | Small blood vessels (SBVs) |
| ON-OFF modulation[15] | Transfer of digital information | Embedded into sequence of symbols | PRO: limited dispersion of burst<br>CON: branching alters burst size | PRO: no branching favors signal integrity<br>CON: high dispersion of bursts |
| Predefined sequence of bursts[16] | ► Activation signal<br>► Signal for monitoring purposes | ► Sequence type<br>► Molecule type | PRO: transport function<br>CON: branching alters burst size | PRO: velocity and RBCs favor interaction with endothelium<br>CON: burst sized tricky for branching |
| Sustained rate of emitted molecules[9] | Activation/ inhibition of a receiver | Presence of signal | PRO: signal transport<br>CON: branching alters rate | PRO: drug administration<br>CON: release rate to be sized according to branching |

1s) and silence (modeling 0s). It can be considered a good solution in LBVs, since a high flow velocity may limit the burst spreading and the impact of RBCs is not excessive. Differently, in SBVs the slow flow intensity favors a dispersion regime (diffusion dominates advection; see the discussion in "Models of signal propagation") making bursts difficult to distinguish. In LBVs, branching can significantly alter the burst size in unpredictable ways, a problem less evident in SBVs.

The second option consists of transmitting known activation signals. In this case, the information is encoded in the type of burst sequence and/or molecules used. Since the targets in this case are essentially the endothelium or organs, SBVs seem to be the ideal environment in which to use this solution. In fact, the slow flow velocity and presence of RBCs favor interaction with the endothelium, enabling activation and monitoring functions. In addition, the size and shape of carriers influence molecules' margination and interaction with the endothelium. Indeed, while spherical particles seem to produce slightly better margination than flattened carriers (for example, ellipsoids), ellipsoidal particles show slower rotational dynamics near vessel walls, favoring their adhesion.[27] Furthermore, larger, micron-size particles are more favorable for absorption than sub-micron-size ones.[27] Rod-shaped carriers also have interesting absorption properties. Long filamentous rod nanoparticles are preferable in low-pressure environments, whereas short rods or spherical nanoparticles have better delivery efficiency to the endothelium[34] under high-pressure conditions. Thus, usage of known activation signals in SBVs actually represents the most suitable environment for the monitoring device shown in Figure 2.

Continuous delivery is mainly aimed at drug delivery.[9] Although LBVs can be used for transport, the main targets are SBVs, where drug molecules can be absorbed by the endothelium or organs. To size the emission rate, it is necessary to consider branching in large vessels, making this rate large enough to reach the target rate at the destination.

**Bio-compatibility issues.** Blood carries many signaling molecules, such as hormones, whose job is not to provide nourishment to cells, but rather to regulate a wide range of physiological activities. To ensure biocompatibility with such an environment, the best choice for information particles is to use types of molecules that may already be present in the blood, or similar ones. The drawback of this approach, however, is that such molecules could interfere with the artificially transmitted ones. To distinguish artificial MC from natural ones, it is necessary to use significant signal energy (that is, large bursts of molecules) or reduce the communication range to very short distances. In the first case, the use of large quantities of molecules could in turn interfere with underlying natural communications, potentially causing unwanted side effects. To avoid them, it is necessary to reduce the burst size and, consequently, the communication range to a few millimeters, which may be of limited use. However, in the bloodstream MC can be used mainly to implement a biocompatible communication system to trigger specific behaviors at the receiver site, such as drug release, or to implement non-invasive monitoring through short-range communication. Thus, information is not necessarily encoded in the sequence of emitted bursts, but rather in the *presence* of the signal itself and its macroscopic features at the receiver site. This information can be conveyed by a single burst of molecules or by a continuous release at a given rate. Different information can be transferred by using different molecules or patterns. TX and RX could even be implemented in the same device, with the TX releasing molecules and the RX estimating the value of some blood parameters on the basis of the received signal, for example, viscosity.[16]

**Models of signal propagation.** The cardiovascular system is modeled as a network of branching ducts composed of sections of different sizes and lengths, where the heart functions as a central pumping system (see Figure 1). The blood that flows through this network is made up of cells suspended in plasma. The fluidity of the blood is influenced by some physical properties, such as concentration, deformability, and the aggregability of circulating blood cells and other elements dispersed in the plasma, as well as flow conditions in the macro and microcirculation (for example, shear stress, shear rates, and viscosity[14]).

Shear stress is the tangential force of blood on the surface of the endothelium in blood vessels. A high shear stress is typical of laminar flows, while low values are typical of turbulent flows. The shear rate represents the velocity gradient between adjacent layers of blood.[14] A high shear rate is present when the flow is fast, as in arteries, or when the diameter is large. Low values occur when flow is slow, as in veins, or when the diameter decreases. Finally, the blood viscosity increases as the shear rate decreases. Since the RBCs are able to deform (high shear rate) and aggregate (low shear rate), blood is a non-Newtonian fluid. Thus, under

normal circumstances, capillaries, or more generally the SBVs, can be modeled using a parabolic velocity profile, neglecting turbulence (Poiseuille flow, as shown in Figure 1[8,19]), taking into account the effects of the blood cells suspended in a Newtonian fluid (that is, the plasma[6, 11,19,31]). This comes from the Navier-Stokes equations, which also describe the drag force exerted on suspended particles.[8,9] However, RBCs tend to aggregate, forming a plug flow region in the center of the vessel, causing a *flattened* parabolic velocity profile known as a Casson profile[11,35] (see Figure 1), due to the increased blood viscosity at low shear rates. This is a more realistic model for the blood-velocity profile in small vessels without turbulence.

Assuming that the volume of blood is constant, from Bernoulli's theorem,[8] stating that the volume flow rate $Q$ through a tube is constant, if a single rigid tube has a cross-sectional area $A$ that varies along its length, the speed of the flow varies accordingly. Hence, given two points $x$ and $y$ along the pipe, if $A_x > A_y$ then $v_x < v_y$. It follows that the average velocity $\bar{v}$ of the blood flow through each section of the circulatory system is given by the ratio between the flow rate $Q$ and the *total cross-sectional area* $A$ of that level, that is, $\bar{v} = Q/A$. By extending the Bernoulli principle to the whole circulatory system sketched in Figure 1, it follows that the *cumulative cross-sectional area* of LBVs, for example $A_1$ for arteries, is smaller than the *cumulative area* of the thinner ones ($A_2$ for SBVs) due to the much larger number of the latter. Thus, for the circulatory system, the result is that $A_1$ is lower than $A_2$ and, as a consequence, the blood velocity in each vessel follows an opposite relationship; that is, $v_1$ is higher than $v_2$.

MC in blood vessels are strongly influenced by the functioning of the cardio-circulatory system. For *short range* communications, the laminar flow assumes a parabolic/Casson profile, especially in SBVs. Hence, the flow velocity is maximum at the center of the vessel and vanishes near the walls. A burst of molecules released in this environment tends to follow a similar shape to the velocity profile, as shown in Tan et al.[31] This is an ideal behavior that occurs in small-section vessels only, for short distances, when the presence of tumbling RBCs is neglected.

For large distance $L$ from the emission point and a slow average velocity, the system converges to the so-called *dispersion regime*, in which diffusion dominates advection.[8,12] It occurs when the Peclet number, defined as $P_e = (\bar{v}R)/D$, results in $P_e \ll 4L/R$, where $R$ is the radius of the vessel and $D$ is the particle diffusion coefficient.[8,12] This phenomenon typically happens for low-speed SBVs. This means that, at a sufficiently large distance $L$ from the emission point, the movement is due mainly to diffusion pushed by the mean flow velocity $\bar{v}$, and the particle distribution on the cross section of the vessel is independent of the release point.

In small vessels, the presence of RBCs influences particle distribution, forcing the system into a different steady state. The small molecules are no longer uniformly distributed in the cross-section of the vessels, but rather are pushed toward the vessel walls by tumbling RBCs.[31] Consequently, most molecules will be traveling in a region known as the cell free layer (CFL), free from RBCs and close to the vessel walls, with an approximate average speed $\bar{v}$ due to the active transport of the flow.

Since most phenomena of interest in MC occur at a low flow rate, we neglect turbulence that occurs in large vessels and focus on small ones. This is why the enabling technologies illustrated in a later section are related to transmitters and receivers fixed on the endothelium to implement the monitoring device shown in Figure 2. A taxonomy of the presented propagation models is shown in Figure 3.

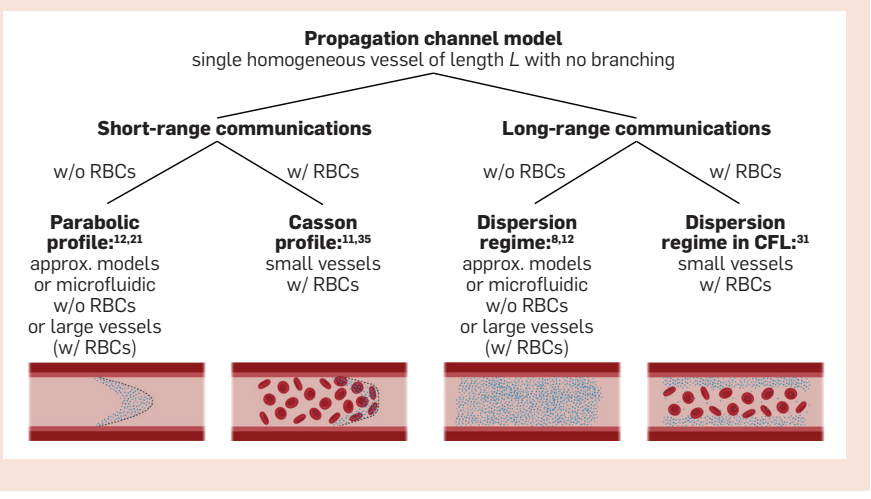
To summarize, the use of spherical particles of greater mass (for example, comparable to that of WBCs/RBCs) allows their transport to be confined toward the axis of the vessels, where the speeds are higher. In contrast, when it is necessary to exchange information near the endothelium, which is the cellular lining of the vessels, it is preferable to use light particles. However, the shape of the particles also needs to be taken into account, as larger, flattened micrometer particles have better margination and absorption capabilities than sub-micron spherical particles,[27] since they have a larger contact surface area.

**Models of the reception process.** In the literature, several models have been proposed for the process of carrier (ligand) reception in MC.

The so-called transparent receiver[21] is just an abstraction: It models a device, either natural or artificial, capable of sensing molecules without interfering with them. Clearly this is an oversimplification of a monitoring process, capable of estimating the concentration of a given type of circulating molecules, for example, proteins.

All other models take into account the surface receptors of the cell. An example of a ligand-receptor pair is

**Figure 3. Taxonomy of propagation channel models. For both short- and long-range communications, different models exist according to the presence of the RBCs in the middle of the bloodstream in SBVs. They cause both profile deviation from parabolic to Casson for short-range communications and dispersion confined in CFL only in the long-range one. In LBVs, the presence of RBCs is less impactful and can be neglected.**



Interleukin-6 (IL-6, ligand) binding to the membrane receptor IR-6R during the cytokine storm in the early phase of COVID-19.[17] The absorbing receiver[21] models a device able to absorb all molecules hitting its surface. It is an abstraction of a device whose surface receptors can cover most of its surface, which may be reasonable for some cases. The receiver with a finite number of absorbing receptors is a more refined model. Surface receptors cover only a fraction of the receiver, which is a more realistic assumption. However, each time a molecule hits one of these receptors, it is absorbed and immediately removed from the communication environment.[15] Finally, the most realistic models take into account both a finite number of receptors and the possibility for each of them to establish a reversible bond with a molecule, which can be broken, as in Awan et al.,[5] Lauffenburger and Linderman,[22] and Pierobon et al.[28] Usually, these models are based on birth-and-death processes describing the temporal occupancy of receptors. We underline that the latter is only a basic model, on the basis of which it is possible to define other, more sophisticated ones, taking into account, for example, potential chemical reactions triggered by the surface bonds.

The reception taxonomy is shown in Figure 4. It also includes a possible mapping to fixed and mobile RX, considering both natural cells and artificial devices, and both LBV and SBV environments.

## Analysis Techniques

The techniques used to analyze MC systems in blood vessels are essentially a combination of analytical tools, simulation platforms, and, possibly, small-scale testbeds.

**Analytical tools.** The main difficulty in adopting fully analytical models is to reliably represent highly complex environments, such as blood vessels. We can consider two classes of models: those that represent an abstract view of a portion of or the entire circulatory system, and those that focus on small sections. In the first case, a good model is represented by a network of elements, where each blood vessel is abstracted by an equivalent electrical circuit.[9] Resistance is related to blood viscosity and vessel diameter; inductance models the inertia of the blood due to blood pressure; and capacitance measures the elasticity of the blood vessel. Additional equations model advection, diffusion, particle adhesion, and absorption/reaction processes. However, closed-form solutions are typically difficult to obtain and local phenomena due to interaction with blood cells are neglected. Other models are inspired by microfluidic environments. Although it is possible to obtain closed-form solutions for the motion of nanoparticles,[12] they cannot fully capture the dynamics of the bloodstream, since they neglect the presence of blood cells or endothelium permeability[29] and produce oversimplified models.

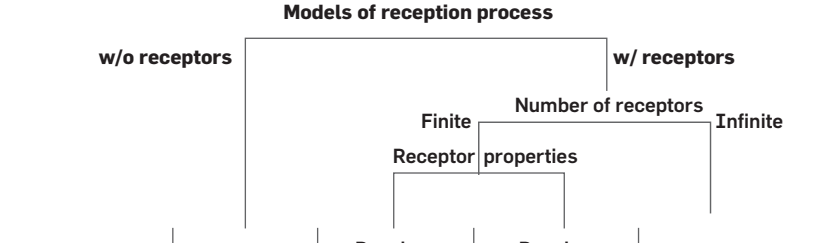
Finally, other models focus on the interaction between molecules and the endothelium through the Markov process.[17,28] Although closed-form solutions can be obtained, their validity is limited to small sections of a vessel, so they can model only very local phenomena.

**Simulators.** When analytical tools fail in providing a solution, a good alternative is to use a simulation. We can distinguish two main approaches: finite elements methods (FEM) and particle-based simulations.

The first approach is a method for numerically solving a complex system of equations. These equations typically model multiphysics phenomena, including flow, mechanical interactions with tissues, and chemical transport of drugs across the vascular wall. This allows simulating objects or structures of arbitrary and complex 3D shape, almost impossible to model with analytical methods. This approach is pursued by commercial solvers, such as the COMSOL simulation tool, which can be used to simulate the entire circulatory system[9] or implemented in custom solvers dedicated to detailed analysis of specific phenomena. In the first case, complex interactions occurring in the bloodstream are modeled at a high level of abstraction. In the latter case, complex models, such as that presented in Tan et al.,[31] may be used for accurate representation of RBCs shape, their tumbling movement, or their interaction with molecules in microvasculature. However, only very short sections of a vessel, on the order of a few tens of micrometers, can be modeled.

The other class of simulation tools are particle-based simulators, such as BiNS2 (see Fellicetti et al.[16] and references therein). In this case, moving particles and blood cells are modeled as spheres of a different size. The simulation consists of updating the motion equation for each particle, further mod-

**Figure 4. Taxonomy of reception models with mapping to the blood-vessel environment, distinguishing fixed and mobile RX.**




| Receiver type | Transparent receiver[21] | Receiver with reversible bindings[5,22,28] | Receiver with absorbing receptors[15] | Absorbing receiver[21] |
|---|---|---|---|---|
| **Mobile nodes:** large cell (for example, WBCs, bacteria) or artificial nanomachine[3] | Only in LBVs; for SBVs is a coarse approx. | Applicable in both LBVs and SBVs | Applicable in both LBVs and SBVs | Applicable in both LBVs and SBVs |
| **Fixed node:** (activated) endothelium or artificial RX stick to it[3,16,17] | Not applicable | Applicable, mainly in SBVs where MC with vessel walls is significant | Applicable, mainly in SBVs where MC with vessel walls is significant | Applicable only for artificial RX or portions of activated endothelium |

eling their interaction in case of collision or when their surfaces are at very close distances, on the order of a few nanometers. Furthermore, the boundaries can model the endothelium and its interactions with moving objects, such as partially inelastic collisions and absorption, by simulating receptors on its surface. This allows for more accurate results than FEM approaches applied to large-scale systems, and less accurate results than these approaches applied to micro-scale systems.

However, the time required to run a simulation of this type, even if accelerated with GPUs, may be several days, even to simulate just a few seconds of the real system. Since chemical reactions occurring in the receiver take much longer times (for example, minutes[22,25]), it is advisable to decouple the simulation of the signal transmission and propagation through the channel from the operation that takes place at the receiver. For modeling the latter, analytical tools can perform quite well in terms of both accuracy and computation time. An example is given in Felicetti et al.,[17] where carriers' propagation and their mechanical interaction with the endothelium is simulated, whereas their absorption and subsequent endothelium behavior is represented by Markovian models.

**Testbeds.** Most testbeds used to mimic the bloodstream environment use microfluidic settings. Although microfluidic models cannot capture the full, complex behavior of the blood circulation, they could be used for implementing some components of more complex, small-scale testbeds. The platform described in Fischera et al.[18] is a typical example of a prototype of flow-driven MC. Although this prototype does not make use of blood cells, it is possible to include them inside the pipe, so as to obtain a platform emulating a non-turbulent bloodstream environment. In this direction, the testbed described in Thomas et al.[33] was used to validate the simulation results reported in Tan et al.,[31] focusing on the characterization of particle delivery in microcirculation. The microfluidic channels were fabricated using a standard soft lithography process. The polydimethylsiloxane (PDMS) device was bonded on clean glass slides after proper treatment, and the PDMS surface was covered with a specific protein (NeutrAvidin) to bind with fluorescent nanoparticles. A mixture of RBCs and nanoparticles was injected into the microfluidic device using a syringe pump, collecting results through imaging using a fluorescent microscope. Despite the channel not having all the characteristics of microvasculature, its size is suitable for obtaining realistic results.

**Advances in biochemical sensor manufacturing processes allow the production of devices capable of both releasing molecules and monitoring their binding along a short section of a blood vessel.**

Although the great advantage of simulations compared with experiments is that they allow full control over the evolution of all system parameters at any time, the reliability of simulation results depends on the correctness of the implemented models. Thus, the value of extracting measurements from real systems mimicking the target ones is invaluable, especially for validating theoretical or simulation results. From a broader perspective, recent advances in microfluidics and molecular biology allow the lab-on-a-chip technology to manipulate biochemical reactions at very small volumes, handling fluids in quantities of just a few picoliters. This allows thousands of biochemical operations to be integrated on a single chip by fractionating a single drop of blood sample, obtaining precise diagnoses of potential diseases.[1,32] A further step in this direction is human-organ-on-a-chip technology.[38] It allows investigating interactions between organs and tissues, connected by microfluidic channels, which emulate microcirculation via pure diffusion, without pumping the blood between different tissues, as described in Wu et al.[38] They can also use blood, as in the skin-on-a-chip application proposed by Mori et al.[26] It focuses on perfusable vascular channels coated with endothelial cells, which are cultured in a skin-equivalent device connected to an external pump and tubes. The analysis of the vascular absorption and molecular permeability from the epidermal layer into the vascular channels allows studies and tests of skin biology.

## Enabling Technologies

Advances in biochemical sensor manufacturing processes allow the production of devices capable of both releasing molecules and monitoring their binding along a short section of a blood vessel. Table 2 presents a taxonomy of possible enabling technologies for

**Table 2. Comparison of different MC-enabling technologies in blood vessels, and technology research roadmap.**

| Technology | Typical Usage | Pros | Cons | Implantability Issues | Challenges and Roadmap |
|---|---|---|---|---|---|
| Epidermal thin device[37] | Monitoring of blood flow | Ultrathin, flexible, stretchable mechanics | Macrovascular detection limits | Non-invasive skin patch | RX: limited to MC generating reactions releasing heat |
| Multicolored fluorescent proteins (FP)[4,20] | ▶ Imaging of tumor angiogenesis<br>▶ Stimulated or inhibited via specific compounds | ▶ In vivo and noninvasive whole-body imaging<br>▶ Fluorescence imaging to visualize blood vessels | Injection of specific compounds | Needs an external detection system | ▶ RX: MC signal constrained to FPs for detection<br>▶ Research: Effort based on aptamers[30,40] |
| Microneedle arrays[7,13,36] | ▶ Delivery: delivery system of molecules<br>▶ Sensing: analytes extraction from skin | ▶ Delivery: highly effective and easy to use<br>▶ Sensing: able to bind and extract biomarkers | ▶ Delivery: limited amount of molecules<br>▶ Possible transient skin problems | Low-cost transdermal patch | ▶ TX/RX: potentially weak MC signals<br>▶ Research: MC range extension |
| Two-electrode system (reverse iontophoresis)[10,23,24] | Skin-like biosensor for non-invasive and accurate intravascular blood monitoring | ▶ Accurate glucose monitoring<br>▶ Precision insulin therapy with pumps | Needs a removable paper battery | ▶ Sensor thickness of ~3μm<br>▶ Needs 2-step measur. of ~20min | ▶ Potential interference with other physiological mechanisms<br>▶ Not fully suitable for generic TX/RX |

such sensors, analyzed below, focusing on the design of the monitoring device illustrated in Figure 2. One of the main requirements is that it can be used on the skin,[24] allowing easy application of the device without compromising its effectiveness.

**Thin thermal sensors.** The receiver part of the device can be inspired by the *epidermal thin sensors* presented in Webb et al.,[37] which allow monitoring variations of the blood flow. It is non-invasive, ultra-thin (< 150 μm), and flexible, with stretchable mechanics and the ability to resist continuous and rapid bodily motions. The device incorporates an array of thin, metallic thermal actuators and sensors designed for monitoring blood flow under a targeted skin surface of about 1 $cm^2$. A large, central thermal actuator provides power to the vessel at temperatures below the threshold of sensation. A set of surrounding sensors enables the measurement of the resulting thermal distribution. Finally, an array of bonding pads allows the attachment of a thin, flexible cable to interface with external acquisition electronics.

Its main limitation is its sensitivity, which increases with the decrease in vessel depth (~2 mm). Furthermore, to implement the RX, it may suffer from a limited ability to detect MC signals, unless a chemical reaction that releases heat is triggered.

**Fluorophores and fluorescent proteins.** For the detection mechanism at the RX on the right side of Figure 2, it is possible to exploit the in-vivo fluorescent imaging technique based on *fluorophores* and *fluorescent proteins*, which can emit light signals of specific wavelengths.[4,20] To realize the full chain, it is possible to resort to the innate, targeted recognition ability exposed by aptamer molecules.[30,40] In fact, aptamers enable the dynamic tracking of molecules, with one arm able to bind to a fluorescent protein and the other to the targeted molecules to be detected. Furthermore, they can be rapidly produced by chemical synthesis according to specific needs. The aptamer-fluorescent protein compound, stored in the reservoir of molecules shown in Figure 2, perfuses into the blood vessel to bind to the target receptor molecules exposed by the endothelium, whose exposure can also be triggered by the release of activation molecules stored in the second reservoir of molecules.

**Microneedles.** *Microneedle arrays*[7,13,36] are a low-cost alternative for the transdermal perfusion of molecules into blood vessels (TX function). They can improve the delivery of molecules through the skin by bypassing the stratum corneum layer of the skin, which acts as a barrier, thus overcoming the various problems associated with conventional administration. The main principle is to penetrate the skin without drawing blood, thus creating micrometer-size pathways that lead molecules directly to the epidermis or upper dermis region, where they can directly enter the systemic circulation.

This technology can be used in two situations, for transdermal delivery of drugs and for sensing, as it can be used to extract analytes from the skin for both *ex-vivo* analysis and *in-situ* sensing. For delivery, it is highly effective and easy to use when large molecules cannot be easily administered orally or transdermally. For sensing, microneedles need to be modified on the surface to bind to specific biomarkers and selectively extract compounds for analysis.

There are different fabrication options, including solid, dissolvable, hydrogel, coated, and hollow microneedles.

However, their small size allows the administration of only limited quantities of molecules (that is, from microgram to low milligram doses). In addition, they can cause transient skin irritation, mild and punctate erythema, as well as skin infection caused by microbial penetration through residual holes in the skin. For the considered monitoring device, they can be used in both the TX and RX sections, although the limited number of molecules available for delivery may result in potentially weak MC signals.

**Reverse iontophoresis.** This technology is based on the analysis of interstitial fluid, a body fluid that contains a lot of physiological information. It is a promising method for obtaining health-status information because interstitial fluid can be easily assessed by implanted or percutaneous measurements. Reverse iontophoresis extracts this fluid by applying an electric field to the skin, allowing noninvasive, epi-

dermal physiological parameter monitoring.[39] An example of the application of this technology is the *two-electrode system* presented in Chen et al.,[10] a skin-like biosensor system for non-invasive and highly accurate intravascular blood glucose monitoring. It makes use of electro-osmosis for the glucose transport in the reverse iontophoresis process. The glucose molar flux is determined by the potential gradient across the skin and the molar concentration of the initial solute. With accurate calibration of the system, it is suitable for medical-grade glucose monitoring and insulin therapy with micro pumps. However, its preparation requires a paper battery, which must be removed from the skin to allow attaching the biosensor to the cathode area for glucose measurement. The need for using ionic or charged molecules when implanting the device could interfere with other physiological mechanisms, such as glucose control; therefore, microneedles are preferable.

## Conclusion

In this article, we discussed the potential use of MC in blood vessels. Despite the extensive literature in the field of MC, most of it does not specifically address the blood environment, mainly due to its complexity and the difficulty of accurately modeling its key features. We accompanied our discussion with a reference architecture of a dermal-monitoring device to be applied on the skin that incorporates the main concepts illustrated. This device is intended to interact with the underlying tissue in order to establish a communication channel with the endothelium, using blood as a communication medium. We discussed the available implementation alternatives via taxonomies of transmission, propagation, and reception models, respectively. We also outlined a roadmap toward the implementation of the device, based on both microneedles for injecting molecules through the endothelium and fluorescent proteins that bind with aptamers to detect the response signal. Realization of this device will enable the collection of data for model validation.

Future work will delve into experimental validation of these technologies when deployed together in a single testbed to study their interaction, as well as discussion of their integration issues toward realization of the monitoring device. Finally, another important issue to face is the tuning of these molecular communications processes in a single patient. C


**References**

1. Abgrall, P. and Gué, A.-M. Lab-on-chip technologies: Making a microfluidic network and coupling it into a complete microsystem—a review. *J. of Micromechanics and Microengineering 17*, 5 (Apr. 2007), R15–R49; 10.1088/0960-1317/17/5/r01
2. Akyildiz, I.F., Jornet, J.M., and Pierobon, M. Nanonetworks: A new frontier in communications. *Commun. ACM 54*, 11 (Nov. 2011), 84–89; 10.1145/2018396.2018417
3. Akyildiz, I.F. et al. The internet of Bio-Nano things. *IEEE Communications Magazine 53*, 3 (Mar. 2015), 32–40; 10.1109/MCOM.2015.7060516
4. Amoh, Y. et al. Color-coded fluorescent protein imaging of angiogenesis: The AngioMouse models. *Curr. Pharm. Des. 14*, 36 (2008), 3810–3819.
5. Awan, H. et al. Molecular communications with molecular circuit-based transmitters and receivers. *IEEE Transactions on NanoBioscience 18*, 2 (2019), 146–155; 10.1109/TNB.2019.2892229
6. Blair, G.W. An equation for the flow of blood, plasma and serum through glass capillaries. *Nature 183*, 4661 (Feb. 1959), 613–614.
7. Cahill, E.M. et al. Toward biofunctional microneedles for stimulus responsive drug delivery. *Bioconjug. Chem. 26*, 7 (Jul. 2015), 1289–1296.
8. Caro, C.G. et al. *The Mechanics of the Circulation* (2 ed.), Cambridge University Press, 2011; 10.1017/CBO9781139013406
9. Chahibi, Y. et al. Pharmacokinetic modeling and biodistribution estimation through the molecular communication paradigm. *IEEE Trans. on Biomedical Engineering 62*, 10 (2015), 2410–2420; 10.1109/TBME.2015.2430011
10. Chen, Y. et al. Skin-like biosensor system via electrochemical channels for noninvasive blood glucose monitoring. *Sci. Adv. 3*, 12 (Dec. 2017), e1701629.
11. Das, B. et al. Red blood cell velocity profiles in skeletal muscle venules at low flow rates are described by the Casson model. *Clin. Hemorheol. Microcirc. 36*, 3 (2007), 217–233.
12. Dinc, F. et al. A general analytical approximation to impulse response of 3-D microfluidic channels in molecular communication. *IEEE Trans. on NanoBioscience 18*, 3 (2019), 396–403.
13. Donnelly, R. et al. Microneedles for drug and vaccine delivery and patient monitoring. *Drug. Deliv. Transl. Res. 5*, 4 (Aug. 2015), 311–312.
14. Ercan, M. and Koksal, C. The relationship between shear rate and vessel diameter. *Anesth. Analg. 96*, 1 (Jan. 2003), 307–308.
15. Farsad, N. et al. A comprehensive survey of recent advancements in molecular communication. *IEEE Communications Surveys Tutorials 18*, 3 (2016), 1887–1919; 10.1109/COMST.2016.2527741
16. Felicetti., L. et al. A molecular communications system for live detection of hyperviscosity syndrome. *IEEE Trans. on NanoBioscience 19*, 3 (2020), 410–421; 10.1109/TNB.2020.2984880
17. Felicetti, L., Femminella, M., and Reali, G. A molecular communications system for the detection of inflammatory levels related to COVID-19 disease. *IEEE Trans. on Molecular, Biological and Multi-Scale Communications* (2021), 1–1; 10.1109/TMBMC.2021.3071788
18. Fichera, L. et al. Fluorescent nanoparticle-based Internet of things. *Nanoscale 12*, 17 (2020), 9817–9823.
19. Gentile, F., Ferrari, M., and Decuzzi, P. The transport of nanoparticles in blood vessels: The effect of vessel permeability and blood rheology. *Ann Biomed Eng 36*, 2 (Feb. 2008), 254–61.
20. Greenwald, E.C., Mehta, S., and Zhang, J. Genetically encoded fluorescent biosensors illuminate the spatiotemporal regulation of signaling networks. *Chemical Reviews 118*, 24 (Dec. 2018), 11707–11794; 10.1021/acs.chemrev.8b00333
21. Jamali, V. et al. Channel modeling for diffusive molecular communication—A tutorial review. *Proc. IEEE 107*, 7 (2019), 1256–1301; 10.1109/JPROC.2019.2919455
22. Lauffenburger, D.A. and Linderman, J.J. *Receptors: Models for Binding, Trafficking, and Signalling*. Oxford University Press, 1996.
23. Lee, H. et al. A graphene-based electrochemical device with thermos-responsive microneedles for diabetes monitoring and therapy. *Nat. Nanotechnol. 11*, 6 (2016), 566–572.
24. Liu, Y. et al. Lab-on-skin: A review of flexible and stretchable electronics for wearable health monitoring. *ACS Nano 11*, 10 (2017), 9614–9635.
25. Marcone, A., Pierobon, M., and Magarini, M. Parity-check coding based on genetic circuits for engineered molecular communication between biological cells. *IEEE Trans. on Communications 66*, 12 (2018), 6221–6236; 10.1109/TCOMM.2018.2859308
26. Mori, N., Morimoto, Y., and Takeuchi, S. Skin integrated with perfusable vascular channels on a chip. *Biomaterials 116* (Feb. 2017), 48–56.
27. Müller, K., Fedosov, D.A., and Gompper, G. Margination of micro- and nano-particles in blood flow and its effect on drug delivery. *Scientific Reports 4*, 1 (May 2014). 10.1038/srep04871
28. Pierobon, M. et al. Noise analysis in ligand-binding reception for molecular communication in nanonetworks. *IEEE Trans. on Signal Processing 59*, 9 (Sep. 2011), 4168–4182.
29. Pries, A.R. and Kuebler, W.M. Normal endothelium. *Handb. Exp. Pharmacol. 176* Pt. 1 (2006), 1–40.
30. Shui, B. et al. RNA aptamers that functionally interact with green fluorescent protein and its derivatives. *Nucleic Acids Res. 40*, 5 (Mar. 2012), e39.
31. Tan, J. et al. Influence of red blood cells on nanoparticle targeted delivery in microcirculation. *Soft Matter 8* (Dec. 2011), 1934–1946.
32. Temiz, Y. et al. Lab-on-a-chip devices: How to close and plug the lab?. *Microelectronic Engineering 132*, 25 (Jan. 2015), 156–175; 10.1016/j.mee.2014.10.013
33. Thomas, A., Tan, J., and Liu, Y. Characterization of nanoparticle delivery in microcirculation using a microfluidic device. *Microvascular Research 94* (Jul. 2014), 17–27; 10.1016/j.mvr.2014.04.008
34. Uhl, C.G. et al. The shape effect on polymer nanoparticle transport in a blood vessel. *RSC Advances 8*, 15 (2018), 8089–8100; 10.1039/c8ra00033f
35. Venkatesan, J. et al. Mathematical analysis of Casson fluid model for blood rheology in stenosed narrow arteries. *J. of Applied Mathematics* (2013), 1–11; 10.1155/2013/583809
36. Waghule, T. et al. Microneedles: A smart approach and increasing potential for transdermal drug delivery system. *Biomedicine & Pharmacotherapy 109* (2019), 1249–1258; 10.1016/j.biopha.2018.10.078
37. Webb, R.C. et al. Epidermal devices for noninvasive, precise, and continuous mapping of macrovascular and microvascular blood flow. *Sci. Adv. 1*, 9 (Oct. 2015).
38. Wu, Q. et al. Organ-on-a-chip: Recent breakthroughs and future prospects. *Biomed. Eng. Online 19*, 1 (Feb. 2020), 9.
39. Zheng, H. et al. Reverse iontophoresis with the development of flexible electronics: A review. *Biosensors and Bioelectronics 223* (Mar. 2023), 115036; 10.1016/j.bios.2022.115036
40. Zhou, J. and Rossi, J. Aptamers as targeted therapeutics: Current potential and challenges. *Nature Reviews Drug Discovery 16* (2017); 10.1038/nrd.2017.86



**Luca Felicetti** is an IT business analyst at Edotto SRL. Previously, he was a contract researcher at the Consorzio Nazionale Interuniversitario per le Telecomunicazioni (CNIT), Italy.

**Mauro Femminella** (mauro.femminella@unipg.it) is an associate professor at the University of Perugia. He is also as member of the Research Unit and is a university representative in the Shareholders' Assembly at the Consorzio Nazionale Interuniversitario per le Telecomunicazioni (CNIT), Italy.

**Gianluca Reali** is an associate professor at the University of Perugia and a member of the Research Unit at the Consorzio Nazionale Interuniversitario per le Telecomunicazioni (CNIT), Italy.

DOI:10.1145/3699595

**Recent reductions in per-device carbon footprint appear to be insufficient to close the sustainability gap.**

BY LIEVEN EECKHOUT

# The Sustainability Gap for Computing: Quo Vadis?

SUSTAINABILITY IS UNDENIABLY a grand challenge. As the world population and the average affluence per person continue to grow, we are eagerly consuming Earth's natural resources. The Earth Overshoot Day marks the date when the demand for ecological resources by humankind in a given year exceeds what the Earth can regenerate in a year's time. While the world's Earth Overshoot Day fell at the end of December in the early 1970s, it has progressively antedated since then and was computed to fall on August 1 in 2024. The overshoot day is (much) earlier for many countries, as early as February 26 for Singapore, March 13 for the U.S., March 26 for Canada, and April–May for most European countries as well as South Korea, Israel, Japan, and China.[a]

The continuously growing consumption of Earth's resources, including materials and energy sources, (inevitably) induces climate change. Greenhouse gas (GHG) emissions result in detrimental global warming, and a recent study reports that the contribution of information and communication technology (ICT) to the world's global GHG emissions, currently between 2.1% and 3.9%,[11] is growing at a rapid pace. While this percentage may seem small, it is not. In fact, ICT's contribution to global warming is on par with (or even larger than) the aviation industry, which is estimated to be around 2.5%.[b]

To combat global warming, the Paris Agreement under the auspices of the United Nations (UN) aims to limit global warming to well below 2 degrees Celsius, and preferably to 1.5 degrees Celsius, compared to pre-industrial levels. In 2019, the UN stated that global emissions must be cut by 7.6% each year over the next decade to meet the Paris Agreement.[c] More recently in

a https://bit.ly/42om5iI
b https://bit.ly/4hkquId
c https://bit.ly/4h1GCOv

## » key insights

- **Computing is responsible for a significant and growing fraction of the world's global carbon footprint.**
- **The status quo, in which we keep per-device carbon footprint constant, would lead to a 5.4× sustainability gap for computing relative to the Paris Agreement within a decade.**
- **Meeting the Paris Agreement for computing requires reducing the per-device carbon footprint by 15.5% per year under current population and affluence growth curves.**
- **Based on a select number of published carbon footprint reports, it appears that while vendors indeed reduce the per-device carbon footprint, it does not seem to be enough to close the gap, urging our community to do more.**

IMAGE BY JOHANNA ZWICKNAGEL

November 2023, the UN stated that insufficient progress has been made so far to combat climate change.[d]

Given the pressing need to act, along with the significant and growing contribution of computer systems to global warming, it is imperative that we, computer system engineers, ask ourselves what we can do to reduce computing's environmental footprint within the socio-economic context. To do so, this article reformulates the well-known and widely used IPAT model,[6] such that we can reason about the three contributing factors: population growth, increased affluence or number of computing devices per person, and carbon footprint per device over its entire lifetime, which includes the so-called *embodied* footprint for manufacturing, assembly, transportation, and end-of-life processing, and the *operational* footprint due to device usage during its lifetime.[13]

The growth in population and affluence leads to a growing *sustainability gap,* as illustrated in Figure 1. If we were to keep the carbon footprint per device constant relative to the present time, the total carbon footprint due to ICT would still increase by 9.4% per year, leading to a 2.45× *increase* in GHG emissions over a decade. In contrast, meeting the Paris Agreement requires that we *reduce* GHG emissions by a factor of 2.2×. Bridging this widening sustainability gap between the per-device status quo and the Paris Agreement requires that we reduce the carbon footprint per device by 15.5% per year or by a cumulative factor of 5.4× over a decade.

Analyzing the carbon footprint for a select number of computing devices (smartwatches, smartphones, laptops, desktops, and servers) reveals that vendors do pay attention to sustainability. Indeed, the carbon footprint per computing device tends to reduce in recent years, at least for some devices by some vendors. However, the reduction in per-device carbon footprint achieved in recent years appears to be insufficient to close the sustainability gap. The overall conclusion is that a concerted effort is needed to significantly reduce the demand for computing devices *while* reducing the carbon footprint per device at a sustained rate for the foreseeable future.

**It is imperative that we, computer system engineers, ask ourselves what we can do to reduce computing's environmental footprint within the socio-economic context.**

## The IPAT Model

IPAT is the acronym of the well-known and widely used equation[6] which quantifies the *impact I* of human activity on the environment:

$$I = P \times A \times T \quad (1)$$

*P* stands for *population* (that is, the number of people on earth), *A* accounts for the *affluence* per person or the average consumption per person, and *T* quantifies the impact of the *technology* on the environment per unit of consumption. The impact on the environment can be measured along a number of dimensions, including the natural resources and materials used (some of which may be critical and scarce); GHG emissions during the production, use, and transportation of products; pollution of ecosystems and its impact on biodiversity; and so on. The IPAT equation is used as a basis by the UN's Intergovernmental Panel for Climate Change (IPCC) in their annual reports.

The IPAT equation has been criticized as being too simplistic, assuming that the different variables in the equation are independent of each other. Indeed, in contrast to what the above formula may suggest, improving one of the variables does not necessarily lead to a corresponding reduction in overall impact. For example, reducing *T* in the IPAT model by 50% through technology innovations to reduce the environmental impact per product does not necessarily reduce the overall environmental impact *I* by 50%. The fundamental reason is that a technological efficiency improvement may lead to an increase in demand and/or use, which in turn may lead to an increase, rather than a reduction, in overall impact. This is the well-known *rebound effect* or *Jevons' paradox*, named after the English economist Williams Stanley Jevons, who was the first to describe this rebound effect. He observed that improving the coal efficiency of the steam engine led to an overall increase in coal consumption.[2] Although there is no substantial carbon tax as of today, Jevons' paradox still (indirectly) applies to computer systems. Efficiency gains increase a computing device's compute capabilities, which stimulates its deployment (that is, more

d https://bit.ly/4hqdsrZ

devices are deployed due to increase in demand) and its usage (that is, the device is used more intensively). The result may be a net increase in total carbon footprint across all devices despite the per-device efficiency gains.

The rebound effect can be (partly) accounted for in the IPAT model by expressing each of the variables as a *compound annual growth rate (CAGR)*, defined as follows:

$$CAGR = \left(\frac{V_t}{V_0}\right)^{1/t} - 1 \quad (2)$$

with $V_0$ the variable's value at year 0 and $V_t$ its value at year $t$. The IPAT model can be expressed using CAGRs for the respective variables:

$$\text{CAGR}_{\text{overall}} = \prod_{i=1}^{N}(CAGR_i + 1) - 1 \quad (3)$$

This formulation allows for computing the annual growth rate in overall environmental impact or GHG emissions as a function of the growth rates of the individual contributing factors. If the growth rates incorporate the rebound effect, that is, higher consumption rate as a result of higher technological efficiency, the model can make an educated guess about the expected growth rate in environmental impact.[3]

## The Environmental Impact of Computing

We now reformulate the IPAT equation such that it provides insight for computer system engineers to reason about the environmental impact of computing within its socio-economic context. We do so while focusing on GHG emissions encompassing the whole lifecycle of computing devices. Total GHG emissions $C$ incurred by all computing devices on earth can be expressed as:

$$C = P \times \frac{D}{P} \times \frac{C}{D} \quad (4)$$

$P$ is the world's global population. $D/P$ is a measure for affluence and quantifies the number of computing devices per capita on earth. $C/D$ is a measure for technology and corresponds to the total carbon footprint per device. Note that $C/D$ includes the whole lifecycle of a computing device, from raw material extraction to manufacturing, assembly, transportation, usage, and end-of-life processing. We now discuss how the different factors $P$, $D/P$, and $C/D$ in the above equation scale over time.

**Population.** The world population $P$ has grown from one billion in 1800 to eight billion in 2022. The UN expects it to reach 9.7 billion in 2050 and possibly reach its peak at nearly 10.4 billion in the mid 2080s.[e] The world population annual growth rate was largest around 1963 with a $CAGR_P$ = +2.1%. Since then, the growth rate has reduced to around $CAGR_P$ = +0.9% according to the World Bank.[f]

**Affluence.** The number of devices per person $D/P$ increases at a fairly sharp rate[7] (see Table 1). On average across the globe, the number of connected devices per capita increased from 2.4 in 2018 to 3.6 in 2023, or $CAGR_{D/P}$ = +8.4%. In the western world (North America and Western Europe) the number of devices per person is not only a factor 2× to 4× larger than the world average, it also increases much faster with a $CAGR_{D/P}$ above +10%. The increase in the number of devices is in line with the annual increase in integrated circuits (ICs), that is, estimated $CAGR$ = +10.2% according to the 2022 McClean report from IC Insights.[18]

**Technology.** The carbon footprint per device $C/D$ and its scaling trend $CAGR_{C/D}$ is much harder to quantify due to inherent data uncertainty and the myriad computing devices. A device's carbon footprint depends on many factors, including the materials used, how those materials are extracted, how the various components of a device are manufactured and assembled, how energy efficient the device is, where these devices are used, the lifetime of the device, how much transportation is involved, how end-of-life processing is handled, and so on. Despite the large degree of uncertainty, it is instructive and useful to analyze *lifecycle assessment* (LCA) or *product carbon footprint*

e https://bit.ly/42sRb8L
f https://bit.ly/42njCoA

**Figure 1. Growth in population and affluence leads to a growing *sustainability gap*. The widening sustainability gap for computing between the current status quo in per-device carbon footprint (which leads to a 2.4× *increase* in global carbon footprint within a decade) versus the Paris Agreement (which requires a 2.2× *reduction*). Closing the sustainability gap requires that we reduce the per-device carbon footprint by 15.5% per year under current population and affluence growth rates.**

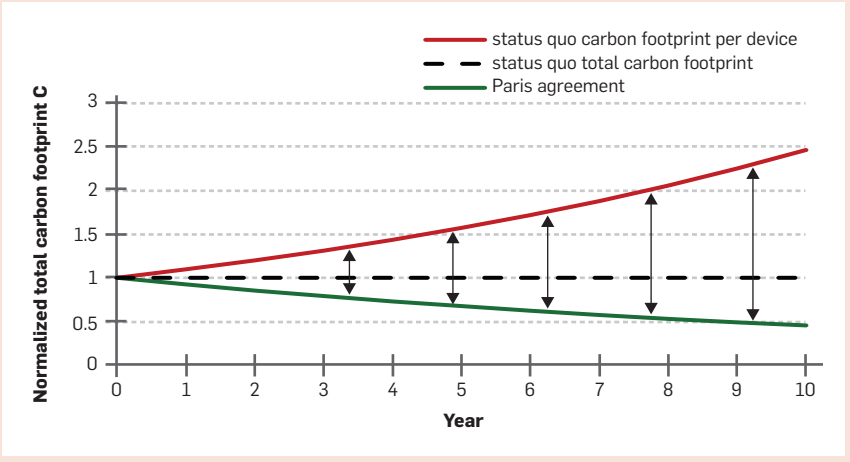



**Table 1. Number of connected devices per capita.[7]**

| Region | 2018 | 2023 | CAGR |
|---|---|---|---|
| Global | 2.4 | 3.6 | +8.4% |
| Asia Pacific | 2.1 | 3.1 | +8.1% |
| Central and Eastern Europe | 2.5 | 4.0 | +9.9% |
| Latin America | 2.2 | 3.1 | +7.1% |
| Middle East and Africa | 1.1 | 1.5 | +6.4% |
| North America | 8.2 | 13.4 | +10.3% |
| Western Europe | 5.6 | 9.4 | +10.9% |

(PCF) reports that quantify the environmental footprint of a device. All LCA and PCF reports acknowledge the degree of data uncertainty; nevertheless, they provide invaluable information for consumers to assess the environmental footprint of devices.

To understand per-device carbon footprint scaling trends, we consider a number of devices from different vendors. We leverage the carbon footprint numbers published in the products' respective LCA or PCF reports. In particular, we use the resources available from Apple,[g] Google,[h] and Dell.[i] We now discuss carbon footprint scaling trends for smartwatches, smartphones, laptops, desktops, and servers.

The bottom line is that per-device carbon footprint has not increased dramatically over the past years and has even significantly decreased in some cases. Several interesting conclusions can be reached upon closer inspection across devices and vendors.

*Smartwatches.* Figure 2 quantifies the carbon footprint for different generations of Apple Watches with similar capabilities (GPS versus GPS plus cellular) and sport band. All watches feature either an aluminum case (42mm in Series 1 to 3, 44mm in Series 4 to 6, and 45mm in Series 7 and 8) or a stainless case (Series 9). It is surprising perhaps to note that a smartwatch's carbon footprint was on a rising trend until 2019 before declining. Indeed, the carbon footprint of a GPS watch has increased with a $CAGR_{C/D}$ = +23.9% from 2016 (Series 1) until 2019 (Series 5), while the carbon footprint of a GPS-plus-cellular watch has decreased with a $CAGR_{C/D}$ = –7.7% from 2019 (Series 5) until 2023 (Series 9).

*Smartphones.* Figure 3 illustrates the carbon footprint for Apple iPhones starting with iPhone 7 (release date in 2016) until iPhone 15 Pro Max (release date in 2023) with different SSD capacity. We note a similar trend for the Apple smartphones as for the smartwatches: Per-device carbon footprint increased until 2019, when it began declining. Indeed, from iPhone 8 (2017) to iPhone 11 Pro Max (2019) with 256GB SDD, the carbon footprint has increased from 71kg to 102kg $CO_2$eq ($CAGR_{C/D}$ = +19.8%). From 2019 onward, we note a decrease in carbon footprint per device: From iPhone 11 Pro Max (2019) to iPhone 15 Pro Max (2023) with 512GB SDD, carbon footprint decreased from 117kg to 87kg $CO_2$eq ($CAGR_{C/D}$ = –7.1%). While Apple has been steadily decreasing smartphone carbon footprint since 2019, that trend has slowed down in recent years. For example, from iPhone 13 Pro Max (2021) to iPhone 15 Pro Max (2023) with 512GB SSD, the carbon footprint has decreased from 93kg to 87kg $CO_2$eq ($CAGR_{C/D}$ = –3.3%).

To analyze these trends across vendors, Figure 4 shows results for Google Pixel phones; the plot reports carbon footprints for the nominal series (Pixel 2, 3, 4, 5, 6, 7, 8), the 'a' series (Pixel 3a, 4a, 5a, 7a, 8a), the XL series (Pixel 2XL, 3XL, 4XL), and the Pro series (Pixel 6Pro, 7Pro, 8Pro). As for Apple, we note a declining trend in recent years for Google smartphones: Per-device carbon footprint increased until mid 2021, when it started trending downward. This is noticeable for the nominal series as well as for the high-end phone series (XL and Pro series). The decrease in carbon footprint since 2021 is substantial for the nominal series ($CAGR_{C/D}$ = –10.5%) and the Pro series ($CAGR_{C/D}$ = –8.8%), while remaining invariant for the 'a' series since mid 2021.

*Laptops.* Figure 5 reports the carbon footprint for Apple MacBook Pro and MacBook Air laptops with different

g https://apple.co/4jubTes
h https://bit.ly/4hsVaX1
i https://dell.to/42oLAjP

**Figure 2. Carbon footprint for Apple Watches.**

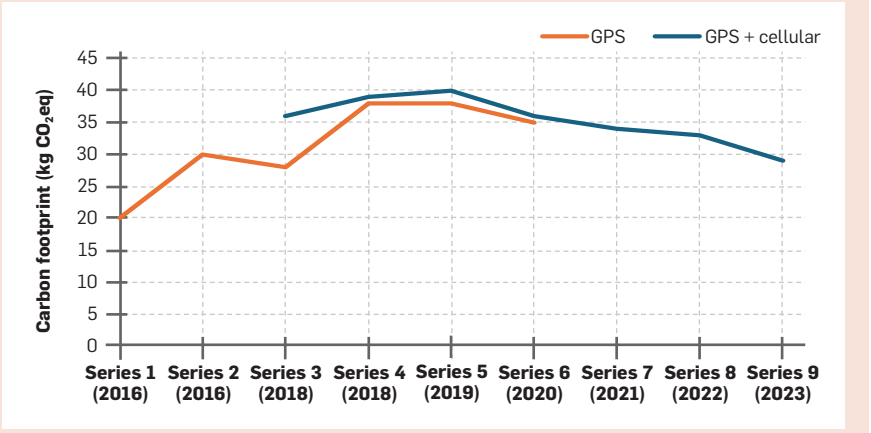
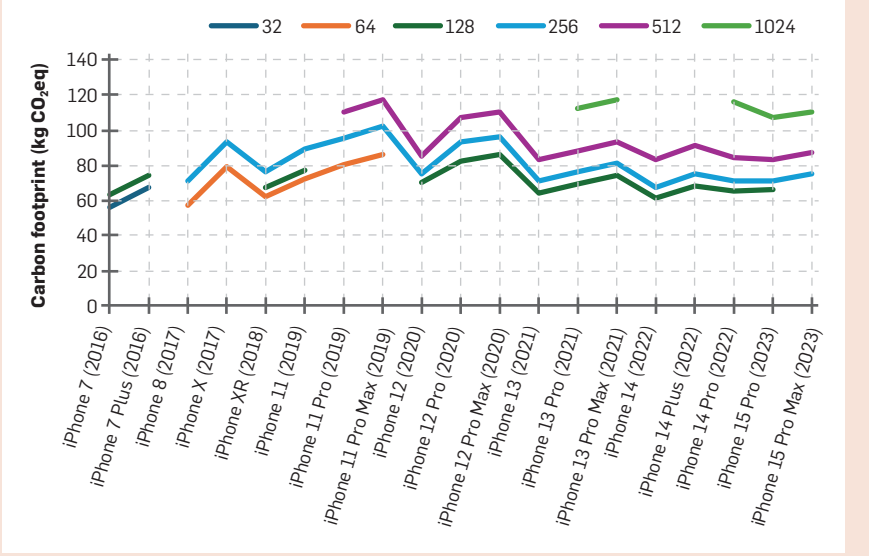



**Figure 3. Carbon footprint for Apple iPhones with different SSD capabilities (GB) from the iPhone 7 through the iPhone 15 Pro Max.**



configurations (screen size, see legend) as a function of their respective release dates; multiple laptops are reported per release date with different storage capacity, core count, and frequency. Several observations are worth noting. First, and perhaps not surprisingly, MacBook Air laptops incur a smaller carbon footprint than the more powerful MacBook Pro laptops. Second, for a given screen size, we note a steady decrease in carbon footprint, for example, MacBook Pro 16-in. ($CAGR_{C/D}$ = −6.9% from 2019 to 2023) and MacBook Air 13-in. ($CAGR_{C/D}$ = −5.1% from 2018 to 2024). Third, while this continuous decrease in per-device carbon footprint is encouraging, there is a caveat: *Discontinuing a particular laptop configuration and replacing it with a more powerful device comes with a substantial carbon footprint increase.* In particular, replacing the MacBook Pro 15-in. with a 16-in. configuration mid-2019 increases the carbon footprint by at least 11.3%; likewise, the transition from 13-in. to 14-in. in the second half of 2022 led to an increase of at least 33.5% for entry-level MacBook Pro laptops.

Looking at Dell, we note a slightly different outcome. The data in Figure 6 reports carbon footprint for the 3000, 5000, and 7000 Dell Precision laptops. The per-device carbon footprint increased from 2018 until 2023 for the 5000 ($CAGR_{C/D}$ = +4.1%) and 7000 ($CAGR_{C/D}$ = +3.8%) laptops, while being invariant for the 3000 laptops. Note that the carbon footprint drastically drops for the most recent laptops released in February and March 2024, but this is due to a change in carbon accounting from MIT's PAIA tool to Dell's own ISO14040-certified LCA tool.

*Desktops and workstations.* The outlook is mixed for desktops and workstations, with some trends increasing and others decreasing. Figure 7 shows carbon footprint for the Dell OptiPlex 700 Series Tower desktop machines (left) decreasing at a rate of $CAGR_{C/D}$ = −8.1% but increasing at a rate of $CAGR_{C/D}$ = +4.0% for Dell Workstations 5000 and 7000 Series (right).

*Servers.* Figure 8 reports the carbon footprint for Dell PowerEdge rackmount 'R' servers across four generations (13th, 14th, 15th, and 16th); this includes Intel- and AMD-based systems. Server carbon footprint numbers are subject to its specific configuration and deployment,

**Figure 4. Carbon footprint for Google Pixel smartphones.**

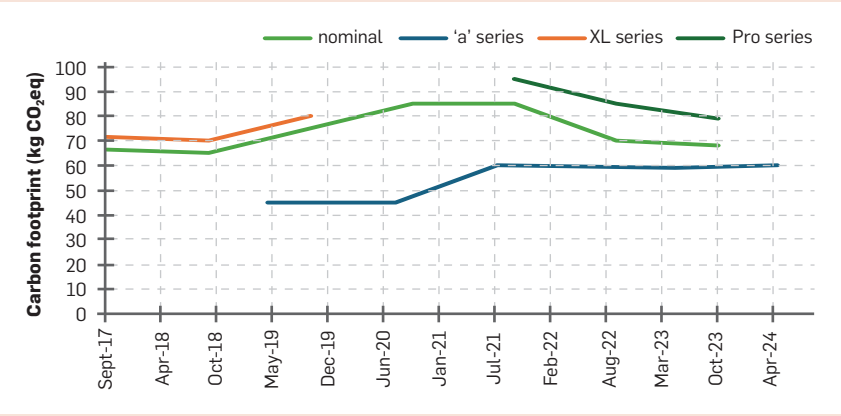



**Figure 5. Carbon footprint for Apple MacBook Pro and Air laptops with different screen configurations.**

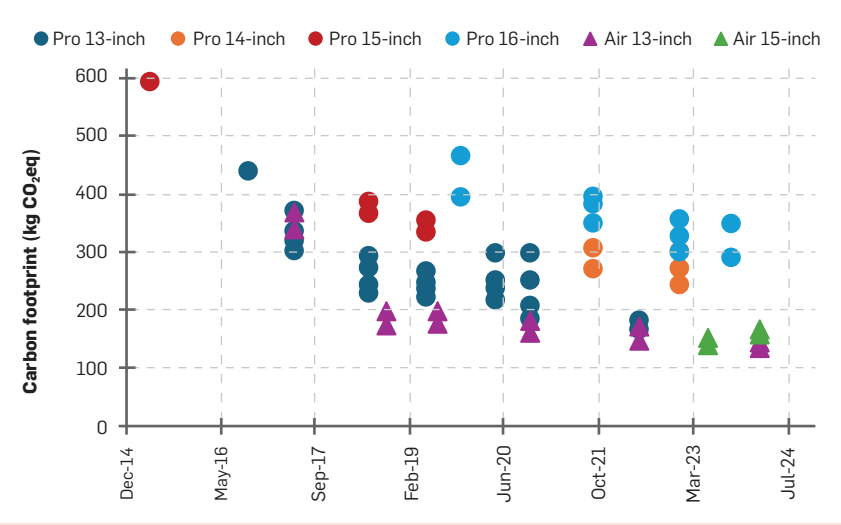
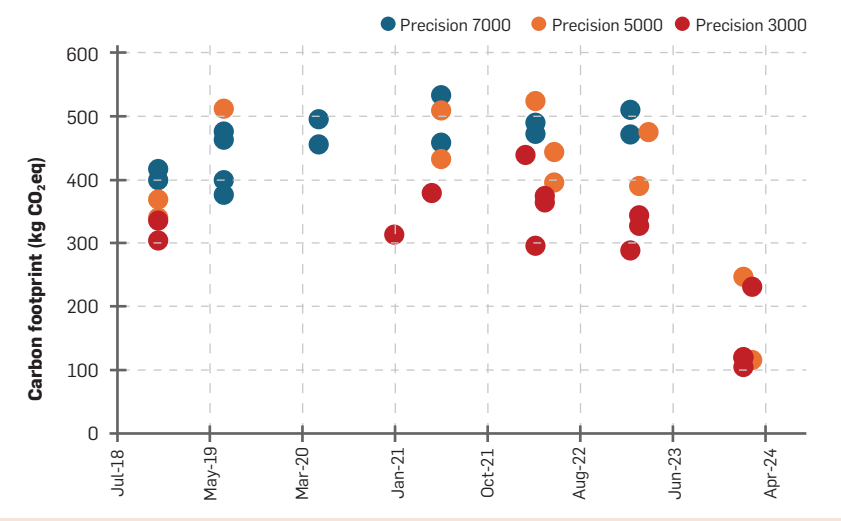



**Figure 6. Carbon footprint for Dell Precision laptops.**



more so than (handheld) consumer devices for at least two reasons. First, because the operational footprint tends to dominate for servers—unlike handheld devices, which are mostly dominated by their embodied footprint[13]—the location of use (and its power-grid mix) has a substantial impact. Second, hard-drive capacity, memory capacity, and processor configuration heavily impact the overall carbon footprint. Overall, server carbon footprint seems to be relatively constant over the past decade, although we note a small increase in average carbon footprint from the 13th to the 16th generation ($CAGR_{C/D}$ = +1.8%). The carbon footprint of a typical high-end server tends to range between 8,000kg and 15,000kg $CO_2$eq over the past decade. Entry-level servers tend to have a lower carbon footprint below 6,000kg $CO_2$eq with a downward trend in recent years. (See the couple of data points in the bottom right corner in Figure 8.)

**Figure 7. Carbon footprint for Dell desktops and workstations.**

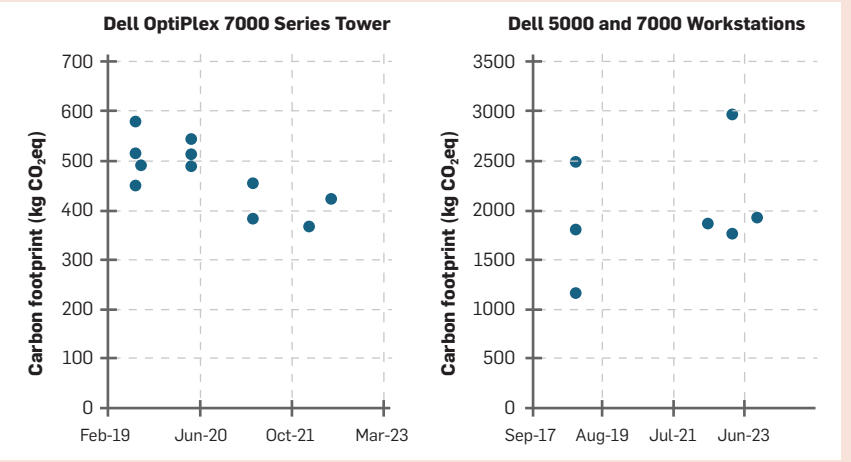
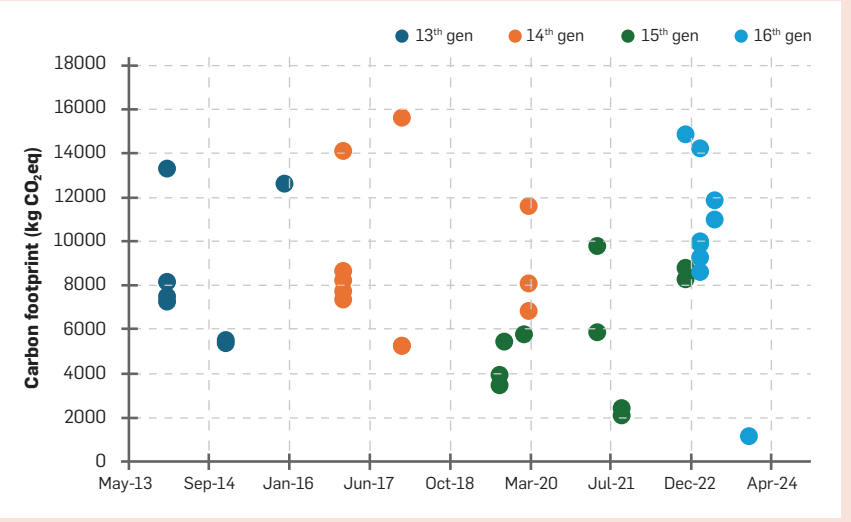



**Figure 8. Carbon footprint for Dell PowerEdge rackmount 'R' servers.**




**Discussion.** It is (very) impressive to note that the compute power of computing devices has dramatically increased over the past years while not dramatically increasing the per-device carbon footprint. In fact, for several computing devices, we note a decreasing trend in per-device carbon footprint—see also Table 2 for a summary—especially in recent years, which is particularly encouraging.

One may wonder whether the recent reduction in per-device carbon footprint comes from reductions in embodied or operational footprint. Upon closer inspection, it turns out that the key contributor to the total reduction in carbon footprint varies across device types. For the Apple smartwatches, the relative decrease in embodied footprint (–7.4% per year) is more significant than the decrease in operational footprint (–2.8% per year). Also, for the MacBook Pro 16-in. laptops, the embodied footprint has decreased at a faster pace (–7.9% per year) than the operational footprint (–3.5% per year). In contrast, for the iPhone Pro Max smartphones, we note a more significant reduction in operational footprint (–11.2% per year) than in embodied footprint (–5.7% per year). For the MacBook Air 13-in. laptops, we even note an increase in operational footprint (+14.9%) while the embodied footprint trends downward (–5.6%).

Overall, per-device carbon footprint decreases for most of the devices analyzed in this work, and in cases where it increases, the increase is limited. The reason for these trends is mixed. The question now is whether this overall declining trend in per-device carbon footprint is sufficient to reduce the overall environmental footprint of computing, and, even better, for meeting the Paris Agreement, which we discuss next.

## Quantifying the Sustainability Gap

Recall that population and affluence increase, ($CAGR_P$ = +0.9%) and ($CAGR_{D/P}$ = +8.4%), respectively. Technology, on the other hand, seems to decrease for many devices, ranging from $CAGR_{C/D}$ = –3.3% to –10.5%, while increasing for others from +1.8% to +4.0%, as summarized in Table 2. The question now is whether these trends lead to an overall increase or decrease in the environmental footprint of computing.

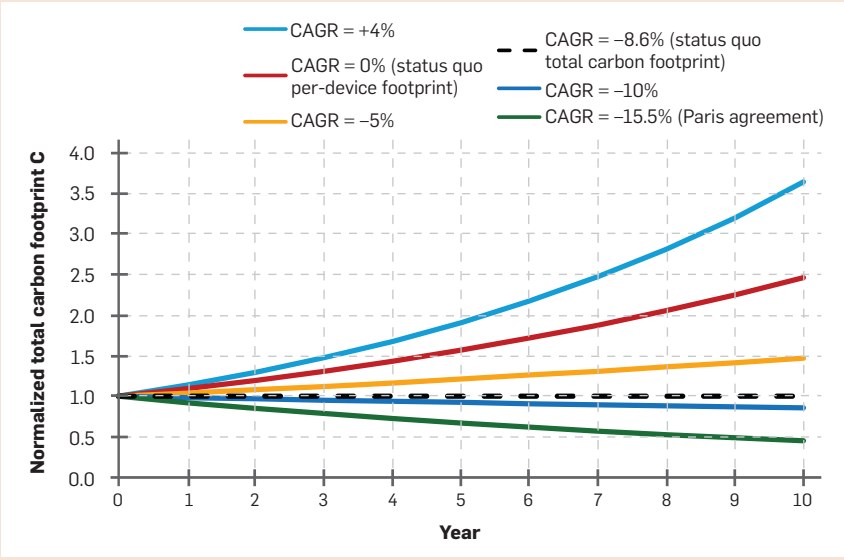
Figure 9 predicts the overall carbon footprint for the next decade normalized to present time for a variety of typical per-device carbon footprint scaling trends, that is, $CAGR_{C/D}$ = +4%, –5%, –10%. In addition, we consider the following three scenarios:

**Scenario #1: Status quo per-device footprint.** If we were to keep the carbon footprint per device constant relative to present time, that is, $CAGR_{C/D}$ = 0%, the total carbon footprint would still increase substantially ($CAGR_C$ = +9.4%). This is simply a consequence of the growing population and the increasing affluence or number of computing devices per person. Because this is an

**Table 2. Per-device carbon footprint scaling trends.**

| Device | Model | Period | CAGR |
|---|---|---|---|
| Smartwatch | Apple Watch | 2019–2023 | -7.7% |
| Smartphone | Apple iPhone Pro Max | 2019–2023 | -7.1% |
| | Apple iPhone Pro Max | 2021–2023 | -3.3% |
| | Google Pixel | 2021–2023 | -10.5% |
| Laptop | Apple MacBook Pro 16-in. | 2019–2023 | -6.9% |
| | Apple MacBook Air 13-in. | 2018–2024 | -5.1% |
| | Dell Precision 7000 | 2018–2023 | +3.8% |
| Desktop | Dell OptiPlex 700 | 2019–2022 | -8.1% |
| | Dell Workstations 5000 and 7000 | 2018–2023 | +4.0% |
| Server | Dell PowerEdge rackmount | 2014–2024 | +1.8% |

**Figure 9. Total carbon footprint normalized to present time for different per-device carbon footprint scaling trends and scenarios (see $CAGR_{C/D}$ values in the legend).**




exponential growth curve, this implies that the total carbon footprint of computing would increase by a factor 2.45× over a decade. In other words, even if we were to keep the carbon footprint per device constant, the total carbon footprint of computing would still dramatically increase.

**Scenario #2: Status quo overall footprint.** If we want to keep the overall carbon footprint of computing constant relative to the present time, that is, $CAGR_C = 0\%$, we need to reduce the carbon footprint per device by $CAGR_{C/D} = -8.6\%$ per year. This is to counter the increase in population and number of devices per person. Reducing the carbon footprint per device by 8.6% year after year for a full decade is a non-trivial endeavor. To illustrate how challenging this is, consider a device that incurs a carbon footprint of 100kg $CO_2$eq. Reducing by 8.6% per year requires that the carbon footprint is reduced to 40.6kg $CO_2$eq within a decade; in other words, the carbon footprint needs to reduce by more than a factor 2.4× over a period of 10 years.

**Scenario #3: Meeting the Paris Agreement.** To make things even more challenging, to meet the Paris Agreement, we need to reduce global GHG emissions by a factor 2.2× over a decade or by 7.6% per year, that is, $CAGR_C = -7.6\%$. To achieve this, we would need to reduce the carbon footprint per device by 15.5% per year, that is, $CAGR_{C/D} = -15.5\%$. This implies that we need to reduce the carbon footprint per device by a factor 5.4× over a decade.

It is clear from Figure 9 that the impact of the per-device carbon footprint scaling trend has a major impact on the overall environmental footprint. Relatively small differences in CAGR lead to substantial cumulative effects over time due to the exponential growth curves. In particular, the status quo per-device carbon footprint ($CAGR_{C/D} = 0\%$) leads to a 2.45× *increase* in overall carbon footprint over a decade, while the Paris Agreement requires that we *reduce* the total carbon footprint by 2.2× ($CAGR_{C/D} = -15.5\%$). Even if we were to reduce the per-device carbon footprint at a relatively high rate ($CAGR_{C/D} = -8.6\%$) to maintain a status quo in total carbon footprint, the gap with the Paris Agreement would still increase at a rapid pace.

As noted from Table 2, most devices do not follow a trend that complies with these required trends: Reported per-device carbon footprint CAGRs are not anywhere close to the required $CAGR_{C/D} = -15.5\%$ (to meet the Paris Agreement) nor do they uniformly meet the $CAGR_{C/D} = -8.6\%$ (to keep total carbon footprint constant relative to present time). To close the sustainability gap, one needs to reduce the per-device carbon footprint by 15.5% per year for the next decade. This implies that the computing industry should do more to keep its carbon footprint under control. This leads to the overall conclusion that there is a substantial gap between the current state of affairs versus meeting the Paris Agreement. Bridging the sustainability gap is a non-trivial and challenging endeavor, which will require significant innovation in how we design and deploy computing devices beyond current practice.

## The Socio-Economic Context

The above analysis assumes that the world population and the number of devices per person will continue to grow at current pace for the foreseeable future. The task of decreasing carbon footprint per device by 15.5% per year to meet the Paris Agreement can be loosened to some extent by embracing a certain level of sobriety in affluence, that is, limiting the number of devices per person. This requires a perspective on the socio-economic context of computing, which includes economic business models, regulation, and legislation.

The computing industry today is

mostly a linear economy where devices are manufactured, used for a while, and then discarded. The lifetime of a computing device can be relatively short, for example, two to four or five years, leading to increased e-waste. Reusing, repairing, refurbishing, repurposing, and remanufacturing devices could contribute to a circular economy in which the lifetime of a computing device is prolonged, thereby reducing e-waste and tempering the demand for more devices.[10] For example, Switzer et al.[23] repurpose discarded smartphones in cloudlets to run microservice-based applications. Reducing the demand for devices could possibly relax the need for reducing per-device carbon footprints.

There is a moral aspect associated with reducing the demand for devices, which is worth highlighting. As mentioned previously, affluence is higher in the western world (North America and Western Europe) compared to other parts of the world; moreover, it is increasing faster. From an ethical perspective, this suggests that the western world should make an even greater effort to reduce the environmental footprint of computing—in other words, we should not necessarily expect other parts of the world to make an equally big effort to solve a problem the western world is mostly responsible for. This implies that the western world should step up its effort in embracing sobriety (that is, consume fewer devices per person) and making individual computing devices even more sustainable.

In addition to transitioning toward a circular economy, other business models can also be embraced. Today, most cloud services are free to use (for example, social media, mail, Web search, and so on) while relying on massive data collection. Maintaining, storing, processing, and searching Internet-scale datasets requires massive compute, memory, and storage resources. According to a recent study by the International Energy Agency,[15] datacenters are estimated to account for about 2% of the global electricity usage in 2022; and by 2026, datacenters are expected to consume 6% of the nation's electricity usage in the U.S. and 32% in Ireland. Moreover, data storage incurs a substantial embodied footprint.[24] The environmental footprint

**Bridging the sustainability gap is a non-trivial and challenging endeavor, which will require significant innovation in how we design and deploy computing devices beyond current practice.**

of free Internet services is hence substantial. Allowing low-priority files to degrade in quality over time could possibly temper the environmental cost for storage devices.[26] But we could go even further by changing the business and usage models of Internet-scale services. Imposing a time restriction for uploaded content could possibly temper the demand for more processing power and storage capacity. We may want to limit how long we keep data around depending on its usefulness and criticality. To make it concrete: Do we really need to keep (silly) cat videos on the Web forever? Limiting to a day or a week may serve the need. Alternatively, or complementarily, one could demand a financial compensation from the customer for using online services. In particular, one could ask customers if they are willing to pay to keep their content online. For example, do you want to pay for your cat videos to remain online for the next month or year?

Transitioning to renewable energy sources (solar, wind, hydropower) is an effective method to reduce the carbon footprint of computing—as with any other industry. Renewables during chip manufacturing have the potential to drastically reduce a device's embodied carbon footprint. Conversely, renewables at the location of device use drastically reduce a device's operational emissions. This is happening today as renewables take up an increasingly larger share in the electricity mix.[20] However, there are several caveats. First, total electricity demand increases faster than what renewables can generate, increasing the reliance on brown electricity sources (that is, coal and gas) in absolute terms.[20] In other words, the transition rate to renewables is not fast enough to compensate for the increase in population and affluence. Second, the amount of green energy is too limited to satisfy all stakeholders. For example, Ireland has decided to limit datacenter construction until 2028 because allowing more datacenters to be deployed would compromise the country's commitment that 80% of the nation's electricity grid should come from renewables by 2030.[16] Third, while renewables reduce total carbon footprint, that does not affect other environmental concerns, such as raw material extraction, chemical and gas emissions

during chip manufacturing, water consumption, impact on biodiversity, and so on.

The analysis performed in this article considered computing as a standalone industry. But computing may enable other industries to become more sustainable, thereby (partially) offsetting its own footprint. This could potentially lead to an overall reduction in environmental footprint.[19] For example, computer vision could enable more efficient agriculture using less water resources and pesticides; artificial intelligence and machine learning could make transportation more environmentally friendly; smart grids that use digital technologies could increase the portion of renewables in the electricity mix in real time; or Internet-of-Things (IoT) devices could help reduce emissions in residential housing. While anticipated sustainability gains in other industries may be substantial, one must be careful when analyzing such reports, that is, one has to carefully understand the assumptions and the associated limitations to fully grasp the validity of such analyses.[21] Moreover, one must be wary of Jevons' paradox as mentioned before: Making a product or service more carbon-friendly may increase overall carbon footprint if the efficiency gain leads to increased deployment and/or usage. In other words, one should be aware of the bigger picture—unfortunately, holistic big-picture assessments are extremely complicated to make and anticipate.

Finally, regulation and legislation may be needed to temper the environmental footprint. The previously cited IEA report[15] states that "*regulation will be crucial in restraining data center energy consumption*" while referring to the European Commission's revised energy-efficiency directive. The latter entails that datacenter operators have to report datacenter energy usage and carbon emissions as of 2024, and they have to be climate-neutral by 2030. Further, the European Parliament adopted the so-called Right to Repair directive, which requires manufacturers to repair goods with the goal of extending a product's lifetime—thereby reducing e-waste and the continuous demand for new devices. Overall, innovation in regulation, legislation, and/or business models will be needed to incentivize (or even force) manufacturers, operators, and customers to temper the demand for more devices, while making sure that our computing industry can still thrive and generate welfare. This is a call for action for our community to reach out to psychologists, sociologists, law and policy makers, entrepreneurs, business people, and so on to holistically tackle the growing environmental footprint of computing.

## Related Work

Our computer systems community recently started considering sustainability as a design goal, and prior work focused mostly on characterizing,[8,12,13,24] quantifying,[9,14,17] and reducing[1,4,5,22,25] the carbon footprint *per device*. However, as argued in this article, to comprehensively understand and temper the environmental footprint of computing, one needs to consider the socio-economic context within which we operate. Population growth and increased affluence is a current reality we should not be blind to and which impacts what we must do to reduce the overall environmental impact of computing.

## Conclusion

This article described the sustainability gap and how it is impacted by population growth, the increase in affluence (increasing number of devices per person), and the carbon intensity of computing devices. Considering current population and affluence growth, the carbon intensity of computing devices needs to reduce by 9.4% per year to keep the total carbon footprint of computing constant relative to present time, and by 15.5% per year to meet the Paris Agreement. Several case studies illustrate that while (some) vendors successfully reduce the carbon footprint of devices, it appears that more needs to be done. A concerted effort in which both the demand for electronic devices *and* the carbon footprint per device is significantly reduced on a continuous basis for the foreseeable future, appears to be inevitable to keep the rising carbon footprint of computing under control and, if possible, drastically reduce it. C


**References**

1. Acun, B. et al. Carbon explorer: A holistic framework for designing carbon aware datacenters. In *Proceedings of the ACM Intern. Conf. on Architectural Support for Programming Languages and Operating Systems 2* (2023), 118–132.
2. Alcott, B. Jevons' paradox. *Ecological Economics 54*, 1 (2005).
3. Bol, D., Pirson, T., and Dekimpe, R. Moore's law and ICT innovation in the anthropocene. In *IEEE Design, Automation, and Test in Europe Conf.* (2021).
4. Brunvand, E., Kline, D., and Jones, A.K. Dark silicon considered harmful: Acase for truly green computing. In *Proceedings of the Intern. Green and Sustainable Computing Conf.* (June 2019), 1–8.
5. Chang, J. et al. Totally green: Evaluating and designing servers for life cycle environmental impact. In *Proceedings of the Intern. Conf. on Architectural Support for Programming Languages and Operating Systems* (Mar. 2012), 25–35.
6. Chertow, M.R. The IPAT equation and its variants. *J. of Industrial Ecology 4*, (2001), 13–29.
7. Cisco. Cisco annual internet report (2018–2023) white paper (2020); https://bit.ly/4jnyfOJ.
8. Eeckhout, L. Kaya for computer architects: Toward sustainable computer systems. *IEEE Micro 43*, (2023), 9–18.
9. Eeckhout, L. FOCAL: A first-order model to assess processor sustainability. In *ACM Intern. Conf. on Architectural Support for Programming Languages and Operating Systems* (2024).
10. Ernst, T. and Raskin, J.-P. Towards circular ICT: From materials to components. *HiPEAC Vision 2021* (Jan. 2021), 122–129.
11. Freitag, M. et al. The real climate and transformative impact of ICT: A critique of estimates, trends, and regulations. *Patterns 2*, 9 (2021).
12. Garcia Bardon, M. et al. DTCO including sustainability: Power-performance-area-cost environmental score (PPACE) analysis for logic technologies. In *IEEE Intern. Electron Devices Meeting* (2020).
13. Gupta, U. et al. Chasing carbon: The elusive environmental footprint of computing. In *IEEE Intern. Symp. High-Performance Computer Architecture* (2021), 854–867.
14. Gupta, U. et al. ACT: Designing sustainable computer systems with an architectural carbon modeling tool. In *Proceedings of the ACM/IEEE Intern. Symp. on Computer Architecture* (2022), 784–799.
15. IEA. Electricity 2024: Analysis and forecast to 2026 (2024); https://bit.ly/3CfnVaZ
16. Judge, P. EirGrid pulls plug on 30 Irish data center projects (2022); https://bit.ly/4jsJEwU.
17. Kline, D. et al. GreenChip: A tool for evaluating holistic sustainability of modern computing systems. *Sustainable Computing: Informatics and Systems 22*, (June 2019), 322–332.
18. McClean, B. The McClean Report—A complete analysis and forecast of the integrated circuit industry. *IC Insights* (2021); https://bit.ly/3PNLlHA
19. Prakash, S. et al. Is TinyML sustainable?. *Commun. ACM 66*, 11 (Nov. 2023), 68–77.
20. Ritchie, H. and Rosado, P. Electricity mix. *Our World in Data* (2020); https://bit.ly/3E7Yith
21. Roussilhe, G., Ligozat, A.-L., and Quinton, S. A long road ahead: A review of the state of knowledge of the environmental effects of digitization. *Current Opinion in Environmental Sustainability 62*, (June 2023).
22. Sudarshan, C.C. et al. ECO-CHIP: Estimation of the carbon footprint of chiplet-based architectures for sustainable VLSI. In *Proceedings of the IEEE Intern. Symp. on High-Performance Computer Architecture* (Mar. 2024).
23. Switzer, J., Marcano, G., Kastner, R., and Pannuto, P. Junkyard computing: Repurposing discarded smartphones to minimize carbon. In *Proceedings of the Intern. Conf. on Architectural Support for Programming Languages and Operating Systems 2* (Mar. 2023), 400–412.
24. Tannu, S. and Nair, P.J. The dirty secret of SSDs: Embodied carbon. *ACM SIGEnergy Energy Informatics Rev. 3*, 3 (2023), 4–9.
25. Zhang, S., Naderan-Tahan, M., Jahre, M., and Eeckhout, L. Balancing performance against cost and sustainability in multi-chip-module GPUs. *IEEE Computer Architecture Letters 22*, 2 (2023), 145–148.
26. Zuck, A., Porter, D.E., and Tsafrir, D. Degrading data to save the planet. In *Proceedings of the 19th Workshop on Hot Topics in Operating Systems* (2023), 61–69.



**Lieven Eeckhout** is a full professor at Ghent University, Belgium.

acm open

# A Transformative Model for Open Access

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

## Impact on the Research Community

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

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

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

acm Association for Computing Machinery

Visit libraries.acm.org/acmopen
Contact acmopen@hq.acm.org

# research highlights

DOI:10.1145/3660530

To view the accompanying paper, visit doi.acm.org/10.1145/3635117 rh

# Technical Perspective
# The Surprising Power of Spectral Refutation

By Uriel Feige

NP-HARD PROBLEMS ARE assumed to be computationally intractable, meaning that no efficient (polynomial time) algorithm is guaranteed to correctly solve every input instance. One way of coping with NP-hardness is via the use of *reliable heuristics*. These are efficient algorithms that might not solve every input instance, but when they claim a solution, the solution is guaranteed to be correct.

Consider the canonical NP-complete problem of SAT (determining whether a Boolean formula of the form $(x_1 \vee \neg x_2 \vee x_3) \wedge (x_2 \vee \neg x_4) \wedge \ldots$ has a satisfying assignment). A heuristic for finding satisfying assignments will naturally be reliable, because one can efficiently check whether the assignment found indeed satisfies all clauses of the input formula. However, designing reliable refutation heuristics that certify that a formula is not satisfiable is more challenging. Under the commonly accepted assumption that NP differs from co-NP, there are no polynomial size witnesses for unsatisfiablity. Hence, there is no natural candidate for what a reliable refutation heuristic should search for to certify that an input formula is not satisfiable.

The absence of natural witnesses does not imply there are no "unnatural" witnesses. For example, consider an algorithm that constructs some matrix based on the input formula, computes the largest eigenvalue of the matrix (this can be done in polynomial time), and if this eigenvalue is sufficiently small, declares that the formula is not satisfiable. Could it be that such an algorithm, that seems unrelated to SAT, can serve as a reliable refutation heuristic? Perhaps surprisingly, the answer is yes.

Algorithms that base their decisions on eigenvalues of appropriately chosen matrices are referred to as *spectral algorithms* (the set of eigenvalues of a matrix is referred to as its spectrum). They have traditionally been applied to problems whose input is naturally represented as a matrix, such as clustering problems (the distance matrix) and graph problems (the adjacency matrix). The input for $k$-SAT (instances of SAT in which every clause has exactly $k$ literals) does not seem to have a natural representation as a matrix. Nevertheless, more than two decades ago it was proved that spectral algorithms can reliably refute most sufficiently dense $k$-SAT instances, when $k$ is even. Here, a formula with $n$ variables is sufficiently dense if the number of clauses is somewhat above $n^{k/2}$.

A sequence of works extended the result in many ways: to odd $k$, to all Boolean constraint satisfaction problems (CSPs, formulas in which the predicate in a clause can be different from the *or* predicate), to achieving *strong refutation* (certifying that no assignment satisfies substantially more clauses than a random assignment does), to achieving trade-offs between (super-polynomial) running time and number of clauses at lower densities (below $n^{k/2}$), and to establishing that "most" holds even in a local sense (smoothed analysis). That is, pick any sufficiently dense input formula, and consider only those formulas that differ from it by very little: For a small fraction of the literals in the formula, their polarity is flipped (for example, $x_i$ changed to $\neg x_i$ or vice versa). Within every such set (and not only in most of them), the spectral heuristic refutes most formulas.

Spectral refutation heuristics involve clever choices of matrices, such as *Kikuchi matrices*, and sophisticated techniques for analyzing their eigenvalues. The accompanying paper presents the state of the art in this line of research. It also presents implications that transcend beyond the domain of refutation heuristics. Recall the *birthday paradox*, which states that if there are $n$ possible outcomes (for example, birthdates), then in a random sample of approximately $\sqrt{n}$ trials (for example, people), there is likely to be a collision: two trials with the same outcome. In contrast, to guarantee a collision with certainty, one needs $n$+1 trials (the *pigeonhole principle*).

**Spectral algorithms can reliably refute most sufficiently dense *k*-SAT instances.**

Are there notions of collisions for which parameters like those of the birthday paradox guarantee a collision with certainty, and there is no need for the more demanding parameters of the pigeonhole principle? It was conjectured that this holds for *even covers* in hypergraphs. An even cover is a set of hyperedges entirely composed of collisions: Every participating vertex appears in two hyperedges, or more generally, in an even number of them. The conjecture was proved using an approach inspired by the spectral refutation heuristics. It involves deriving from every hypergraph a family of Kikuchi matrices, proving that most matrices in this family do not have a large eigenvalue, and proving that if the conjecture is false then every matrix in this family has a large eigenvalue. Consequently, the conjecture must be true.

Other recent applications of spectral techniques to seemingly unrelated mathematical problems include improved lower bounds on the block-length of locally decodable codes, and proof of the sensitivity conjecture for Boolean functions. C

**Uriel Feige** is a professor in the Department of Computer Science and Applied Mathematics at the Weizmann Institute, Rehovot, Israel.

DOI:10.1145/3635117

To view the accompanying Technical Perspective, visit doi.acm.org/10.1145/3660530

# New Spectral Algorithms for Refuting Smoothed *k*-SAT

By Venkatesan Guruswami, Pravesh K. Kothari, and Peter Manohar

**Abstract**

**Despite being a quintessential example of a hard problem, the quest for finding fast algorithms for deciding satisfiability of propositional formulas has occupied computer scientists both in theory and in practice. In this article, we survey recent progress on designing algorithms with strong refutation guarantees for *smoothed* instances of the *k*-SAT problem. Smoothed instances are formed by slight random perturbations of arbitrary instances, and their study is a way to bridge the gap between worst-case and average-case models of problem instances. Our methods yield new algorithms for smoothed *k*-SAT instances with guarantees that match those for the significantly simpler and well-studied model of *random* formulas. Additionally, they have led to a novel and unexpected line of attack on some long-standing extremal combinatorial problems in graph theory and coding theory. As an example, we will discuss the resolution of a 2008 conjecture of Feige on the existence of short cycles in hypergraphs.**

## 1. INTRODUCTION

The famous SAT problem asks to determine if a given propositional formula is *satisfiable*. That is, can we set the formula's variables to 0 (False) or 1 (True) in a way so that the formula evaluates to 1 (True). In this article, we will focus on the *k*-SAT problem where the propositional formula is further restricted to be in the *k*-CNF form, that is, logical AND of a collection of *k-clauses*, each of which is a logical OR of at most *k literals* (variables or their logical negations). For example, $(x_1 \vee \neg x_2 \vee x_3) \wedge (x_2 \vee x_4 \vee \neg x_5)$ is a 3-CNF formula in variables $x_1, x_2, \ldots, x_5$ where $\vee$, $\wedge$, and $\neg$ denote the logical AND, OR, and NOT operations, respectively. Given a *k*-CNF formula, we are interested in either finding a satisfying truth assignment, if it exists, or a "refutation"—a short, easily-checkable proof that the formula is unsatisfiable. Despite its simplicity, *k*-SAT is phenomenally expressive and models a long list of important discrete optimization problems. A decades-long quest has thus focused on designing algorithms for *k*-SAT in both theory and practice.

In this article, we will focus on finding refutations—"obviously verifiable" polynomial size (that is, short) contradictions that confirm unsatisfiability of a *k*-SAT formula.

The original version of this paper was published in the *Proceedings of the ACM Symp. on Theory of Computing, 2022*.

For any formula, we can simply tabulate each of the $2^n$ possible truth assignments $x$ together with a clause violated by $x$. This is an obviously verifiable refutation but clearly too long—it's exponential in size. On the other hand, if we get lucky and our formula contains two 1-clauses $(x_1) \wedge (\neg x_1)$, then it is manifestly unsatisfiable and the two 1-clauses serve as an easily verifiable and short certificate of unsatisfiability. Of course, it is unrealistic for such clauses to magically occur in interesting inputs. But we can often infer additional clauses that must also be satisfied if the input formula is satisfied and hope that such an obviously verifiable short contradiction arises in the inferred clauses. Such *clause learning* (via mechanical *resolution* rules) for deriving new clauses form an integral part of practical SAT solvers.

In his famous 1972 paper,[18] Karp proved that ascertaining satisfiability or finding refutations for 3-SAT formulas is NP-hard. Thus, finding a fast algorithm for 3-SAT that succeeds (even approximately) on all possible input is likely hard. One might naturally expect finding refutations to get easier as the number of clauses increases (more clauses means more possibilities for contradictions to manifest), and so perhaps one might hope that *denser* instances get easier? No such luck! As it turns out, unless a widely believed, stronger variant of the P ≠ NP conjecture fails, there are no polynomial time algorithms for refuting *k*-SAT formulas unless they have essentially the maximum possible $\approx n^k$ (out of $\approx n^k$ possible) clauses. In fact, even substantially beating brute-force search and finding sub-exponential (for example, $2^{\sqrt{n}}$) time algorithms is ruled out for formulas with $\approx n^{k-1}$ clauses.[14]

Despite the grim picture presented by these hardness results, the extraordinary modeling power of *k*-SAT has motivated a concerted research effort for finding fast heuristics for *k*-SAT. On the practical side, new algorithmic ideas along with advances in software engineering have made modern SAT solvers[13] a powerful and indispensable tool with applications to solving practical instances of optimization problems in planning, model checking, and verification of software systems. By encoding the task of finding counter-examples to mathematical conjectures into SAT formulas, SAT solvers have even helped resolve long-standing mathematical conjectures.[6] On the theoretical side, algorithms research has focused on more careful modeling of input instances to escape worst-case hardness under minimal assumptions. Such *beyond worst-case* input models for hard discrete optimization problems such as *k*-SAT now form a vibrant area[11] of research in algorithm design.

### 1.1. The smoothed *k*-SAT model.

In 2007, Feige proposed[9] his *smoothed k-SAT* model as a way to circumvent the all-pervasive hardness of *k*-SAT. He was inspired by a groundbreaking work of Spielman and Teng on smoothed analysis of the simplex algorithm. The simplex algorithm for linear programming, introduced by Dantzig in 1947, presented an uncomfortable disconnect between theoretical predictions and practical performance—here was a fast practical algorithm that also provably needed exponential time in the worst-case. In 2001, Spielman and Teng convincingly resolved[23] this tension and showed that the simplex method runs in polynomial time on smoothed inputs—an input obtained by adding a small random perturbation to an arbitrary instance. Such a perturbation, of the sort one might reasonably expect practical instances to naturally possess, is enough to remove all hardness in even carefully crafted instances.

Feige's model involves an analogous smoothening of a worst-case *k*-SAT formula by randomly perturbing each literal (that is, changing an unnegated variable to negated and vice-versa) in each clause with some small probability, say 0.01, independently. For example, given two clauses $(x_1 \vee \neg x_2 \vee x_3) \wedge (x_2 \vee x_4 \vee \neg x_5)$, we imagine tossing six independent coins with bias 0.01, one for each literal in each of the two clauses. The smoothed version of the first clause will have, for example, $x_1$ negated if the first coin lands on heads, $x_2$ unnegated if the second coin lands on heads, and so on.

If the input formula $\phi$ has $\gg Cn$ clauses in $n$ variables for some large enough constant $C$, then the resulting smoothed formula is unsatisfiable with high probability over the random perturbation. Feige thus asked if the task of finding refutations for smoothed *k*-SAT formulas gets significantly easier compared to worst-case formulas. Equivalently, given our discussion above, do $\ll n^{k-1}$ clause-smoothed *k*-SAT formulas admit efficient refutation algorithms?

Prior works[2,8,22] showed that the answer is indeed yes for the *random k-SAT* model—a *fully smoothed* model where the negation patterns and the *clause structure*, that is, the variables appearing in the clauses (that are worst case in smoothed *k*-SAT) are chosen independently at random. However, those algorithms strongly exploit the abundant randomness in the choice of variables appearing in the clauses.

In this article, we will survey recent developments[1,16,17] on a new class of algorithms, based on the eigenvalues of certain specialized *Kikuchi matrices* (introduced earlier[24] for statistical inference and to simplify algorithms for random *k*-SAT[22] for even *k*), that yield optimal (modulo certain hardness conjectures) algorithms for smoothed *k*-SAT. As a result, these new algorithms succeed in refuting smoothed *k*-SAT formulas with $m \gtrsim n^{k/2} \log n$ clauses, that is, $\ll n^{k-1}$, in polynomial time and significantly beat brute-force search if $m \gtrsim n^{1+\epsilon}$. In fact, our guarantees for smoothed *k*-SAT match the best-known (and conjectured optimal) results for the significantly simpler and restricted setting of *randomk*-SAT formulas, provide quantitative bounds on the number of clauses that every truth assignment must violate, and extend far beyond *k*-SAT to handle *all* logical constraint satisfaction problems.

### 1.2. Spectral methods for combinatorics.

While the theoretical advances in algorithms for *k*-SAT haven't yet influenced practical SAT-solving, they already have some surprising applications to long open problems in other areas of mathematics. These include resolving Feige's conjecture on small *even covers* (cycles) in hypergraphs,[16,17] making progress on the decades-long quest for optimal bounds for *locally decodable*[4] and *locally correctable*[19] error-correcting codes, and problems[7] in additive number theory that generalize the famous Szemeredi's theorem.

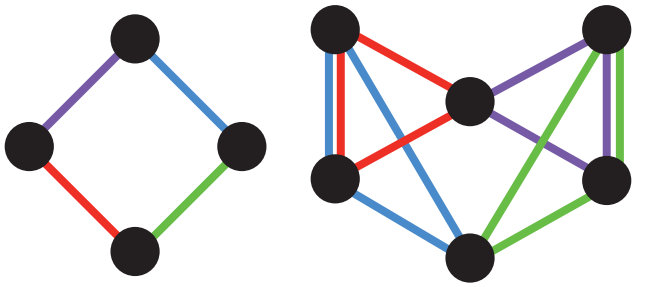
The principle behind such applications is analogous to how SAT solvers helped resolve mathematical conjectures by encoding the search for a proof into a SAT formula. Our theoretical analog strongly exploits the newfound ability to tackle *k*-SAT formulas with a *worst-case* clause structure. In this article, we will discuss an application of this method to proving Feige's conjecture on *small cycles in hypergraphs*. Surprisingly, proving this conjecture will let us go full circle to show even better refutations for smoothed *k*-SAT.

*Short cycles in graphs.* Feige's conjecture[10] is a generalization of a basic result about short *cycles* in *graphs*. Recall that a graph (aka network), is simply a collection of pairs, called *edges* (modeling pairwise associations), on $n$ nodes. A cycle in such a network is a sequence of nodes $v_1 \to v_2 \to \ldots \to v_\ell \to v_1$ that starts and ends at the same node such that every consecutive pair has an edge between them. In his famous 1978 book,[5] mathematician Béla Bollobás conjectured that every graph with $n$ nodes where every node, on average, has $d > 2$ edges (this quantity is called the *average degree*), must have a cycle of length at most $\approx \log_{d-1}(n)$. When $d = 2$, the network can be a single, giant cycle on all $n$ vertices that clearly has no cycle of length $\leq n - 1$. For any $d > 2$, however, the conjecture implies that we cannot even avoid a cycle of length $O(\log n)$—an exponentially smaller bound than $n$—and thus signals a *phase transition* in the extremal length of the smallest cycle as the average degree $d$ crosses 2.

Bollobas's conjecture is an example of a result in *extremal combinatorics*. Such results uncover a truth that holds for *all* (thus *extremal*) mathematical objects. Here, it says that no matter how we might build a graph on $n$ nodes, so long as it has average degree $d > 2$, we cannot avoid introducing a short cycle. In their elegant 2002 paper,[3] mathematicians Alon, Hoory, and Linial confirmed this conjecture.

*Short cycles in hypergraphs.* Feige's conjecture asks a similar question about short cycles in *hypergraphs*. A $k$-uniform hypergraph is a collection of subsets of size $k$ on $n$ nodes, called *hyperedges*, that model associations between a larger number of nodes (instead of 2 in graphs). A 2-uniform hypergraph is simply a graph. To pose Feige's question, we will identify a key property of cycles in graphs and use it to motivate a generalized notion of cycles in hypergraphs. Observe that every vertex appears in either two or zero edges in any cycle in a graph. In particular, a cycle is an *even subgraph*—a subset of edges on which every vertex appears an even integer number of times. Even subgraphs naturally generalize to hypergraphs. We will define a hypergraph cycle or *even covers* to be a collection $C_1, C_2, \ldots, C_\ell$ of hyperedges such that every node of the hypergraph is in-

**Figure 1. a) A length 4 cycle in a graph on 4 vertices. b) A length 4 even cover in a 3-uniform hypergraph on 6 vertices.**




cluded in an even number of $C_i$'s (that is, an *even subhypergraph*). (See Figure 1a and 1b.)

This definition may look odd (or perhaps a little too even?), at first. A cycle in a graph has an appealing combinatorial structure of a loop. Hypergraph cycles seem to lack a combinatorial meaning. Why did Feige (and why should we) care about it? Feige's motivation for studying short cycles actually stemmed from finding refutations for $k$-SAT formulas (see next section). Here, we outline a different source of motivations that comes from deep connections to the theory of error-correcting codes because hypergraph cycles naturally relate to solving systems of linear equations.

To see why, let's associate a variable to every node of a graph and let's think of each edge as specifying a linear equation modulo 2. Thus, the edge $1 \sim 2$ between nodes 1 and 2 relates to the equation $x_1 + x_2 = b \bmod 2$ where $b \in \{0,1\}$. A cycle in the graph then naturally corresponds to a subset of 2-sparse (that is, each equation has only two non-zero coefficients) equations that is *linearly dependent*—that is, if you want this subset of equations to be satisfied, then the right-hand side $b$ of at least one of the equations is already fixed (that is, cannot be chosen independently) once you fix the choice of the right-hand sides for all the others. As a simple example, consider the graph on three nodes with edges $1 \sim 2$, $2 \sim 3$, and $3 \sim 1$. Suppose you knew that some $x$ satisfies the two equations corresponding to the first two edges $x_1 + x_2 = 0 \bmod 2$ and $x_2 + x_3$ is 1 mod 2. Then, adding the left-hand sides of the two equations gives $x_1 + 2x_2 + x_3 = 1 \bmod 2$ which is equivalent to $x_1 + x_3 = 1 \bmod 2$ since $2x_2 = 0 \bmod 2$, regardless of the value of $x_2$. Thus, the right-hand side of the third equation is determined/dependent and cannot be chosen independently of the first two equations. A cycle in a $k$-uniform hypergraph similarly corresponds to a linearly dependent subset of $k$-sparse (that is, each equation has $k$ non-zero coefficients) equations corresponding to each hyperedge.

The length of the smallest cycle thus equals the size of the smallest linearly dependent subset of equations in a given system. Understanding the size of such a set turns out to have a whole gamut of applications, especially in designing *error-correcting codes*. Error-correcting codes (or just codes) are a systematic method of adding redundancy to a message so that, when the message transmitted across a noisy channel and incurs errors, one can still *decode* it uniquely thanks to the redundancy. The systematic methods or codes naturally involve adding "parity checks", that is, right-hand sides of an appropriately chosen set of linear equations evaluated at the message. In such codes, the length of the smallest linearly dependent subset of equations naturally corresponds to *distance*—a crucial quantity that controls the number of errors that can be corrected. The smallest linear dependencies in $k$-sparse equations turns out to be equivalent to understanding the best possible distance (and thus, the largest possible rate of errors that can be tolerated) by an important class of codes called *low-density parity-check* codes introduced by Gallager in the 1960s with numerous theoretical and practical applications.

*Feige's Conjecture and the resolution.* As in the case of graphs, we are interested in the extremal trade-off between average degree (that is, average number of hyperedges containing a node) and the length of the smallest hypergraph cycle in a $k$-uniform hypergraph. Given the connection to linear dependencies above and the basic fact that every collection of $m = n + 1$ equations in $n$ variables are linearly dependent, whenever the average degree $d = mk/n > k$, then the hypergraph must have a (rather long) cycle of length $n + 1$. Making an analogy to graphs, one might expect that if $d \gg k$, the hypergraph must have a $O(\log n)$-length cycle, but this turns out to be false! Mathematicians Assaf Naor and Jacques Verstraete in 2006 showed[21] that one needs (and this is enough) the average degree $d \gtrsim n^{k/2-1}$ in order for every hypergraph to have a $O(\log n)$-length cycle. For $k = 2$, this matches a coarse version of the bound for graphs but for $k \geq 3$ suggests a new regime between $d \approx 1$ and $d \approx \sqrt{n}$ that has no analog for graphs. What happens when, for example, $d \approx n^{0.25}$? Feige conjectured a precise behavior for this regime:

> Conjecture 1 (Feige[10]).
> Every $k$-uniform hypergraph on $n$ nodes with average degree $\approx (n/\ell)^{k/2-1}$ has an $\ell \log n$-length cycle.

Feige's conjecture (up to some $\log n$ factors in $d$) was motivated by the hypothesis (and is, in fact, equivalent to it) that there is no better construction of hypergraphs avoiding short cycles than simply choosing a random hypergraph. It is thus analogous to the famous 1947 theorem of Shannon (the birthplace of modern coding theory) that random error-correcting codes are "best" in a precise sense.

Despite being a foundational statement about hypergraphs, the conjecture remained largely open with only some partial progress for a special case by Alon and Feige in 2009. Now, by invoking the new algorithms for solving smoothed $k$-SAT, we can essentially completely resolve it—up to one additional $\log n$ factor in the degree bound. This was established first by the authors with additional $\log n$ factors, later[17] trimmed down to a single $\log n$ factor.

> Theorem 2.
> Every $k$-uniform hypergraph on $n$ nodes with average degree $\approx (n/\ell)^{k/2-1} \log n$ has an $\ell \log n$-length cycle.

As we will explain later in this article, the proof of this

theorem makes a new connection between the success of our spectral approach for smoothed $k$-SAT and existence of short cycles in hypergraphs. It forms the first (of a growing list of) application spectral refutations via Kikuchi matrices in combinatorics.

## 2. A NEW SPECTRAL APPROACH

Our approach for finding refutations for smoothed $k$-SAT formulas relies on continuous time algorithms based on *spectral methods*—methods that use eigenvalues and eigenvectors of matrices built from the input. This is in contrast to the largely discrete algorithmic toolkit (such as resolution refutations and their generalization) in modern practical SAT solvers. In fact, it has been known for more than 20 years that even random $k$-SAT formulas with $\ll n^{k-1}$ clauses do not admit efficient resolution refutations—a natural formalization of combinatorial refutations.

The spectral approach for refuting $k$-SAT formulas was conceived[15] by Goerdt and Krivilevich back in 2001. A decade and half of work led to spectral methods with conjectured-optimal guarantees for refuting random $k$-SAT in 2017. These spectral methods, however, are rather brittle and strongly rely on the randomness in the variables appearing in each clause. For smoothed $k$-SAT, such methods provably fail since the variables appearing in the clauses are completely arbitrary. In this article, we will present a new class of spectral methods, based on *Kikuchi* matrices, which, when combined with combinatorial pre-processing, provide *robust* methods for refutation that significantly simplify the results for random $k$-SAT and extend to smoothed $k$-SAT with no loss in performance.

*From k-SAT to degree k polynomials.* To bring in spectral methods, we will make a simple but conceptually important translation between refuting a $k$-SAT formula and finding certificates of upper bounds on the maximum value of a degree $k$ polynomial. For this purpose, it will be more convenient to view truth assignments as +1 (True) and −1 (False). Given a $k$-SAT formula $\phi$ with $m$ clauses, let $\Phi:\{-1,1\}^n \to \mathbb{N}$ map truth assignments $x \in \{-1,1\}^n$ to the number of clauses satisfied by $x$. Then, $\Phi(x)$ is clearly the sum of $m$ functions $\Phi_C$, one for each clause $C$ in $\phi$ where $\Phi_C(x) = 1$ if and only if $x$ satisfies clause $C$. Since $\Phi_C$ depends only on $k\{\pm 1\}$-variables, it is a degree $k$ polynomial. For example, $k = 3$ and $C = (x_1 \vee x_2 \vee x_3)$:

$$\Phi_C(x) = \frac{7}{8} + \frac{1}{8}\left(x_1 + x_2 + x_3 - x_1x_2 - x_2x_3 - x_3x_1 + x_1x_2x_3\right) \quad (1)$$

To refute the $k$-SAT formula $\phi$, we will find an easily-checkable certificate of the fact $\Phi(x) = \sum_C \Phi_C(x) < m$ for all $x$. In fact, we will certify a stronger bound of $\Phi(x) \le 0.99m$, that is, every $x$ must violate not just 1 but in fact 1% of the clauses.

Observe that $\Phi_C$ is a sum of 8 (in general, $2^k$) monomials, each of degree $\le 3$ ($k$, more generally). A standard idea, going back to early 2000s, is to certify bounds on 8 different polynomials, obtained by taking one out of 8 terms corresponding to each $C$. We can then obtain a bound on $\Phi(x)$ by adding all the 8 quantities. For each such polynomial obtained from a smoothed 3-SAT formula, with a little more work that we omit here, we can also assume the coefficients of each monomial is an independent, uniform $\{\pm 1\}$. We thus focus on strong refutation of *semirandom homogeneous polynomials* $\Psi(x)$ of the form:

$$\Psi(x) = \sum_{C \in H} b_C \prod_{i \in C} x_i, \quad (2)$$

where $H$, the *instance hypergraph*, is simply the collection of $m$ different sets $C \subseteq [n]$ of size $k$ (corresponding to each original clause) and $b_C \in \{\pm 1\}$ are chosen uniformly and independently for each $C \in H$. Here, strong refutation involves certifying that $\Psi(x) \le \epsilon m$ for a sufficiently tiny $\epsilon > 0$.

### 2.1. From quadratic polynomials to matrices.

Let us show how spectral methods show up by starting with the simplest setting of $k = 2$. Then, each $C \in H$ is a set of size 2, or simply a pair $\{i,j\} \subseteq [n]$, and $\Psi(x)$ from (2) is a degree 2 polynomial in $x$. The idea is to view such a degree 2 polynomial as a *quadratic form*.

For an $n \times n$ matrix $A$, its quadratic form on a vector $v$ equals $v^\top Av = \langle v, Av\rangle = \sum_{i,j \le n} v_i v_j A(i,j)$. This expression is a homogeneous quadratic polynomial in $v$. Indeed, every quadratic polynomial is a quadratic form of an associated matrix and vice-versa. For our $\Psi$, let the $n \times n$ matrix $A$ be defined by:

$$A(i,j) = \begin{cases} b_{\{i,j\}} & \text{if } C = \{i,j\} \in H \\ 0 & \text{otherwise .} \end{cases}$$

Then, for any $x \in \{-1,1\}^n$, $x^\top Ax = \sum_{i,j} x_i x_j b_{i,j} = 2\Psi(x)$.

A basic result in linear algebra allows upper-bounding any quadratic form of $A$ as:

$$v^\top Av \le \| v \|_2^2 \|A\|_2,$$

where $\|v\|_2 = \sqrt{\sum_i v_i^2}$ is the length of the vector $v$ and $\|A\|_2$ is the "spectral norm" or the largest *singular value* of $A$. Since $x$ has $\{\pm 1\}$-coordinates and thus length $\sqrt{n}$, we obtain that $\Psi(x) \le n\|A\|_2$.

This bound on $\Psi(x)$ is easily verifiable. Given $\Psi$ (obtained easily from the $k$-SAT formula $\phi$), we form the matrix $A$ and use a linear time algorithm (called *power iteration*) to obtain good estimates on the spectral norm $\|A\|_2$.

To certify $\Psi(x) \le \epsilon m$, we need to check that $\|A\|_2 \le \epsilon m/n$. $A$ is an example of a *random matrix* since its entries $b_{\{i,j\}}$ are uniformly and independently distributed in $\{\pm 1\}$. There is a well-developed theory for understanding the typical value of $\|A\|_2$ for such random matrices that allows us to conclude that $\|A\|_2 \lesssim \sqrt{\Delta_{max} \log n}$ where $\Delta_{max}$ is the maximum number of non-zero entries in any row of the matrix $A$. If the pairs $C = \{i,j\}$ are "equidistributed" that is, any variable $i$ participates in roughly the same number of pairs, then we would expect $\Delta_{max} \approx \Delta_{avg} \approx m/n$ where $\Delta_{avg}$ is the average number of non-zero entries in a row of $A$. Thus, $\|A\|_2 \lesssim \sqrt{m \log n/n}$ which is $\le \epsilon m/n$ if $m \gtrsim n \log n$.

What if the set of pairs $H$ is not *regular* and some $i$ is "over-represented" in the set of pairs $C$? While we omit formal details, one can use an elegant reweighting trick (discovered in the work of Hsieh et al.[17]) on the matrix $A$ that effectively allows us to assume regularity if $\Delta_{avg} \gg 1$.

### 2.2. Generalizing to quartic polynomials.

The case of odd $k$ turns out to be technically challenging

so let us skip $k = 3$ and generalize the above approach when $\Psi(x)$ is of degree $k = 4$ (that is, the case of 4-SAT). So, $\Psi(x)$ is not quadratic in $x$. We will now view it as a quadratic form in $\binom{n}{2}$ variables each corresponding to quadratic monomials $x_i x_j$ in the original assignment $x$. Let us write $x^{\circ 2}$ for the vector in $\mathbb{R}^{\binom{n}{2}}$ indexed by pairs $\{i,j\}$ with entry at $\{i_1, i_2\}$ given by $x_{i_1} x_{i_2}$. Define the $\binom{n}{2} \times \binom{n}{2}$ matrix $A$:

$$A(\{i_1,i_2\},\{j_1,j_2\}) := \begin{cases} b_{i_1,i_2,j_1,j_2} \text{ if } C = \{i_1,i_2,j_1,j_2\} \in H \\ 0 \text{ otherwise.} \end{cases}$$

Then, as before, we can observe that:

$$(x^{\circ 2})^\top A(x^{\circ 2}) = \sum_{\{i,j\},\{k,\ell\}} x_i x_j x_k x_\ell A(\{i,j\},\{k,\ell\}) = 6\Psi(x)$$

The factor 6 comes from the fact that there $\binom{4}{2} = 6$ different ways that a set $C$ of size 4 can be broken into pairs of pairs, each of which arises as a term in the quadratic form above.

We can now write

$$\Psi(x) \le \|x^{\circ 2}\|_2^2 \|A\|_2 = \binom{n}{2} \|A\|_2.$$

And a similar appeal to results in random matrix theory tells us that with high probability $\|A\|_2 \lesssim \sqrt{\Delta_{max} \log\binom{n}{2}} \lesssim \sqrt{\Delta_{max} \log n}$ where $\Delta_{max}$ is the maximum number of non-zero entries in any row of $A$. Equivalently, $\Delta_{max}$ is the maximum number of 4-clauses that a pair $\{i,j\}$ participates in. When all pairs behave roughly similarly, we will have $\Delta_{max} \approx \Delta_{avg} \lesssim \frac{m}{\binom{n}{2}}$, in which case

$$\Psi(x) \le \binom{n}{2} \sqrt{\frac{m}{\binom{n}{2}} \log n} \le \epsilon m$$

with high probability if $m \gtrsim n^2 \log n$.

Early proofs of such facts used different tools from random matrix theory and worked for random 4-SAT by utilizing that $H$ is a random collection of 4-sets in that case. Our approach here explicitly reveals that only an equi-distribution (that is, $\Delta_{max} \approx \Delta_{avg}$) property of $H$ is required for the success of this approach. This allows us, via a similar reweighting trick (that succeeds if $\Delta_{avg} \gg 1$) discussed above, to obtain a result that works for arbitrary (worst-case) hypergraphs.

Let's finish this part by noting our quantitative bounds. For $k = 2$, our refutation succeeded when $m \gtrsim n \log n$. For $k = 4$, we instead needed $m \gtrsim n^2 \log n$. Indeed, for arbitrary even $k$, a similar argument yields a bound of $m \gtrsim n^{k/2} \log n$—a significant improvement over the $\Omega(n^k)$ clauses required for refuting an unsatisfiable $k$-SAT formula in the wors case, showing us the power of the spectral approach.

### 2.3. Beyond basic spectral refutations.

A smoothed $k$-SAT formula is unsatisfiable with high probability whenever it has $m \gtrsim n$ clauses. But our spectral refutations above require $m \gtrsim n^2 \log n$—a bound higher by a factor $\approx n$ (and $n^{k/2-1}$ for arbitrary even $k$). This is because our approach fails whenever the average number of entries in a row of $A$, that is, $\Delta_{avg} = m/\binom{n}{2}$, is $\ll 1$. The question of whether there are non-trivial refutations for $k$-SAT formulas when $m \ll n^{k/2}$ (now called the *spectral threshold*) remained open for more than a decade and half. In the same time, researchers found evidence, in the form of restricted lower bounds,[20] that there may be no polynomial time refutation algorithm for $m \ll n^{k/2}$. This, nevertheless, left open the possibility of significantly beating brute-force search below this threshold. This possibility was realized for *random* $k$-SAT in 2017. We will now discuss a significantly simpler (described essentially in full below) spectral approach that succeeds even for smoothed $k$-SAT.

We will continue to work with $k = 4$. As before, we will write $\Psi(x)$ as a quadratic form but instead of the natural matrices we discussed above, we will use *Kikuchi* matrices that we next introduce. First, though, a piece of notation: For sets $S, T \subseteq [n]$, we let $S \oplus T = (S \cup T) \setminus (S \cap T)$ denote the symmetric difference of $S$ and $T$.

Definition 3 (Kikuchi Matrices).
For any $r \in \mathbb{N}$, the level $r$-Kikuchi matrix for $\Psi$ is a $\binom{n}{r} \times \binom{n}{r}$ matrix with rows and columns indexed by sets $S, T \subseteq [n]$ of size $r$ and entries given by:

$$A(S,T) = \begin{cases} b_C \text{ if } S \oplus T = C \in H \\ 0 \text{ otherwise .} \end{cases}$$

Observe that for $r = 2$, the above Kikuchi matrices specialize to the $\binom{n}{2} \times \binom{n}{2}$ matrix we saw in the previous subsection. Let's see why $\Psi(x)$ is a quadratic form of $A$.

For any assignment $x \in \{\pm 1\}^n$, denote by $x^{\circ r}$ the $\binom{n}{r}$-dimensional vector with coordinates indexed by sets $S \subseteq [n]$ of size $r$ and $S$-th entry given by $x_S = \prod_{i \in S} x_i$. Then, for $D = \binom{4}{2}\binom{n-4}{r-2}$ we have:

$$\begin{aligned} (x^{\circ r})^\top A(x^{\circ r}) &= \sum_{S,T} x_S x_T A(S,T) \\ &= \sum_{C \in H} b_C \sum_{S,T : S \oplus T = C} x_C = D\Psi(x). \end{aligned}$$

Here, we used that since $S \oplus T = C$, for any $x \in \{\pm 1\}^n$, $x_S \cdot x_T = x_{S \cup T \setminus S \cap T} x_{S \cap T}^2 = x_C$ as $x_i^2 = 1$ for every $i$. The last equality holds true because the number of pairs $(S,T)$ such that $S, T$ are $r$-size sets and $S \oplus T = C$ is exactly $D$ for any set $C$ of size 4. Observe how our notational trick of switching to $\{\pm 1\}$-valued truth assignments paid off here.

Given this simple observation, we can now again construct the spectral upper bound $\Psi(x) \le \|x^{\circ r}\|_2^2 \|A\|_2 = \binom{n}{r} \|A\|_2$. Furthermore, it turns out that powerful tools of random matrix theory still allow us to conclude as before that

$$\|A\|_2 \lesssim \sqrt{\Delta_{max} \log\binom{n}{r}} = \sqrt{\Delta_{max} r \log n}. \tag{3}$$

*The Kikuchi superpower: Density increment.* Why might this upper bound be better? The meat is in the *density increment*. As $r$ grows, the number of rows in $A$ grows. But the number of non-zero entries in $A$ grows even faster, giving us a net increase in $\Delta_{avg}$. Indeed, let $m = n^2 \log n/\ell$ for some parameter $\ell \in \mathbb{N}$. Then, since each $C \in H$ contributes $D$ non-zero entries, $\Delta_{avg} = mD/\binom{n}{r} \approx (n^2 \log n/\ell)(r^2/n^2) \approx r^2 \log n/\ell$. In particular, even when $\ell \gg \log n$ (and thus we have $m = n^2 \log n/\ell \ll n^2$ clauses) choosing $r$ large enough still allows us to obtain a $\Delta_{avg} \gg 1$.

Surprisingly, the rest of the proof idea is more or less the same as before. Let us assume, as we did at first in both the previous subsections, that all rows of $A$ have roughly an equal number of non-zero entries. Such a condition holds true if $H$ is a random collection of sets of size 4. Then, $\Delta_{max} \lesssim \Delta_{avg} \approx r^2 \log n/\ell$. Plugging this back in (3) gives

$$\Psi(x) \leq \frac{\binom{n}{r}}{D}\|A\|_2 \lesssim \frac{m}{\Delta_{avg}}\sqrt{\Delta_{avg} r \log n} \leq \epsilon m$$

if $\Delta_{avg} \gtrsim \sqrt{\Delta_{avg} r \log n}/\epsilon$ or $\Delta_{avg} \gtrsim \frac{r \log n}{\epsilon^2} \log n$. Since $\Delta_{avg} = r^2 \log n/\ell$, this condition holds if $r \geq \ell/\epsilon^2$.

Furthermore, as in the previous two subsections, we can use a variant of our reweighting trick to generalize this argument to arbitrary $H$ without any further assumptions. To verify this bound algorithmically (that is, to "check" our refutation), we need to construct the matrix $A$ and compute its spectral norm. This requires a runtime proportional to the dimension of the matrix, which scales as $\approx n^r$. So, all in all, we obtain a roughly $n^{\ell/\epsilon^2}$ time algorithm to certify that $\Psi(x) \leq \epsilon m$ whenever $m \geq n^2/\ell$ for any $\ell \in \mathbb{N}$. When $m \approx n^{1+\delta}$ for some small $\delta > 0$, that is, even slightly superlinear, the runtime of our algorithm is strictly sub-exponential (specifically $\approx 2^{n^{1-\delta}}$) and thus asymptotically beats brute-force search.

**Handling odd k.** We described our approach so far for even $k$. The case of odd $k$ turns out to be a little more involved. This has been true for spectral algorithms ever since the earliest spectral algorithms for the problem. The polynomial time case (for example, when $m \gtrsim n^{1.5} \log n$ for 3-SAT analogous to $m \gtrsim n^2 \log n$ for 4-SAT) were first found in a work of Abascal, Guruswami, and Kothari in 2021. The full trade-off required introducing the correct generalizations of the Kikuchi matrices that we have described above. Analysis of the spectral norms of such matrices requires more effort and some additional combinatorial ideas.

We will not formalize these ideas here but note the following eventual result we derive as a consequence:

THEOREM 4.
For any $k \in \mathbb{N}$ and $\ell \in \mathbb{N}$, given a semi-random homogeneous degree $k$ polynomial $\Psi(x)$ with $m \geq n(n/\ell)^{k/2-1} \log n$ non-zero coefficients, there is a $2^{O(\ell \log n/\epsilon^2)}$ time spectral algorithm that certifies $\Psi(x) \leq \epsilon m$. Consequently, for any $k$, we can refute smoothed $k$-SAT formulas with $m$ clauses also in time $2^{O(\ell \log n/\epsilon^2)}$.

## 3. PROVING FEIGE'S CONJECTURE

Let us now see how our spectral algorithms for smoothed $k$-SAT provide a resolution for Feige's conjecture. Our approach can be thought of as a theoretical equivalent of encoding the search for a proof into (un)-satisfiability of a SAT formula and then running a practical SAT solver. Given a hypergraph $H$, we will build a SAT formula $\Psi_{\text{random}}$ that will be *satisfiable* if $H$ does not have a short cycle. We will then prove that $\Psi_{\text{random}}$ is in fact unsatisfiable to complete our proof. Of course, instead of using the computer to find such a refutation, *we* will "find" them (that is, argue their existence) analytically by appealing to our spectral algorithms.

Let us now describe this, our argument, in more detail. We are given an arbitrary $k$-uniform hypergraph $H$ on $n$ nodes and average degree $d \approx (n/\ell)^{k/2-1} \log n$. Our goal is to show that $H$ must have an $\ell \log n$-length cycle. Starting from $H$, we will define a family of homogeneous degree $k$ polynomials:

$$\Psi_{\text{sat}} = \sum_{C \in H} b_C x_C.$$

Observe that $\Psi_{\text{sat}}$ is clearly *satisfiable*. Indeed, if $x_i = 1$ for every $i$ then $\Psi_{\text{sat}}(x) = |H|$, the maximum possible value.

Our key claim below will argue, using the analysis of our spectral algorithm from above, that if $H$ has no short cycle then $\Psi_{\text{sat}}$ must in fact be unsatisfiable:

LEMMA 5
(KEY CLAIM).If $H$ has no $\approx (\ell \log n)/\epsilon^2$ length cycle, then, $\max_{x \in \{\pm 1\}^n} \Psi_{\text{sat}}(x) \leq \epsilon |H|$.

We thus immediately hit a contradiction unless $H$ has a $\approx \ell \log n$ length cycle.

Let us now discuss why Lemma 5 must hold for even $k$. For $r \in \mathbb{N}$, we let $A$ be the Kikuchi matrix for the polynomial $\Psi_{\text{sat}}$ that we built in the previous section:

$$A_{\text{sat}}(S,T) = \begin{cases} 1 & \text{if } S \oplus T \in H \\ 0 & \text{otherwise .} \end{cases}$$

Then, we have: $\Psi_{\text{sat}}(x) \leq \binom{n}{r}\|A_{\text{sat}}\|_2$.

We will now argue a rather odd-looking fact. Consider the polynomial $\Psi$ defined below for arbitrary $b_C \in \{\pm 1\}$:

$$\Psi_{\mathbf{b}} = \sum_{C \in H} b_C x_C,$$

We also let $\Psi_{\text{random}}$ be the special case when $b_C$s are chosen uniformly at random and independently. Notice that $\Psi_{\text{random}}$ has the same form as the polynomial $\Psi$ we analyzed in the previous section. Let $A_{\mathbf{b}}$ be the Kikuchi matrix built from $\Psi_{\mathbf{b}}$:

$$A_{\mathbf{b}}(S,T) = \begin{cases} b_C & \text{if } S \oplus T = C \in H \\ 0 & \text{otherwise .} \end{cases}$$

We will argue that if $H$ had no $\log\binom{n}{r} \approx \ell \log n$-length cycle, then $\|A_{\mathbf{b}}\|_2 \approx \|A_{\text{sat}}\|_2$ *no matter what the value of* $b_C$*'s are*. Now, from the previous section, we know that $\binom{n}{r}\|A_{\mathbf{b}}\|_2 \leq \epsilon |H|$ for random $\mathbf{b}$ thus for every $x$:

$$\Psi_{\text{sat}}(x) \leq \binom{n}{r}\|A_{\text{sat}}\|_2 \leq \binom{n}{r}\|A_{\mathbf{b}}\|_2 \leq \epsilon |H|.$$

This claim can appear strange. How can it be that $\|A_{\mathbf{b}}\|_2$ does not depend on the $b_C$'s at all? In a sense, our proof reveals how short cycles in $H$ are "necessary" for $\|A_{\mathbf{b}}\|_2$ to be as small as it is in the previous section!

### Trace moments and spectral norms.

To relate $\|A_{\text{sat}}\|_2$ and $\|A_{\text{random}}\|_2$ and to bring in the cycles in $H$, we will use a classical connection between $\|B\|_2$ and the so-called *trace moments* of a matrix $B$ that, in turn, are related to a certain combinatorial count of *walks* on $B$.

For any $N \times N$ symmetric matrix $B$, let $\|B\|_2 = \sigma_1 \geq \sigma_2 \geq \cdots \sigma_N \geq 0$ be its $N$ singular values placed in descending order. The *trace* $\text{tr}(B)$ is simply the sum of the diagonal elements of $B$. For any even $2t \in \mathbb{N}$, a classical observation in linear algebra says that the trace of the $2t$-th power of $B$ equals the sum of $2t$-th powers of its singular values:

$$\text{tr}(B^{2t}) = \sigma_1^{2t} + \sigma_2^{2t} + \cdots + \sigma_N^{2t}.$$

This is helpful because we can now write:

$$\|B\|_2^{2t} = \sigma_1^{2t} \leq \text{tr}(B^{2t}) \leq \sum_{i=1}^{n} \sigma_i^{2t} \leq N\sigma_1^{2t} = N\|B\|_2^{2t}$$

That is, the $2t$-th power of $\|B\|_2$ equals $\mathrm{tr}(B^{2t})$ up to a factor $N$. By taking $2t = \log N$ and taking $1/2t$-th powers on both sides and recalling that $N^{1/\log N} \to 1$ as $N \to \infty$ gives:

$$\|B\|_2 \le \mathrm{tr}(B^{2t})^{1/2t} = N^{1/\log N}\|B\|_2 \approx \|B\|_2.$$

Thus, for $2t \approx \log N$, $\mathrm{tr}(B^{2t})^{1/2t}$ is a faithful approximation to $\|B\|_2$.

**Relating $\|A_{\mathrm{sat}}\|_2$ and $\|A_{\mathrm{random}}\|_2$ via trace moments.** Using the above connection, we will now focus on arguing that $\mathrm{tr}(A_{\mathrm{sat}}^{2t}) = \mathrm{tr}(A_{\mathrm{random}}^{2t})$ for $2t = \log N = \log\binom{n}{r} \approx r\log n$. This would give us $\|A_{\mathrm{sat}}\|_2 \approx \|A_{\mathrm{random}}\|_2$.

We now recall another classical observation from basic linear algebra that relates trace moments of a matrix $B$ to a certain combinatorial count of "walks" on $B$. For $A$ (standing for $A_{\mathrm{sat}}$ or $A_{\mathbf{b}}$), we thus have:

$$\mathrm{tr}(A^{2t}) = \sum_{S_1,S_2,\ldots,S_{2t}} A(S_1,S_2)A(S_2,S_3)\cdots A(S_{2t},S_1). \quad (4)$$

That is, $\mathrm{tr}(A^{2t})$ is the sum over all sequences of $2t$ row indices (that is, subsets of $[n]$ of size $r$) of product of the entries of $A$ on consecutive elements of the sequence.

Every entry of $A$ is either 0 or $\pm 1$ and for each non-zero entry $A(S_i,S_{i+1})$, $S_i \oplus S_{i+1} = C$ for some $C \in H$. Thus, any $2t$-sequence $(S_1,S_2,\ldots,S_{2t})$ that contributes a non-zero (and thus exactly $\prod_{i=1}^{2t} b_{C_i}$ in $\mathrm{tr}(A_{\mathbf{b}}^{2t})$) value to the sum above must correspond to a $2t$-tuple $(C_1,C_2,\ldots,C_{2t})$ of hyperedges from $H$, one for each entry $A(S_i,S_{i+1})$.

More is true about such a $(C_1,C_2,\ldots,C_{2t})$, as we next demonstrate in the crucial observation below. Let $1_{C_i} \in \{0,1\}^n$ be the 0-1 indicator of the set $C_i \subseteq [n]$. That is, $1_{C_i}(j) = 1$ if and only if $j \in C_i$.

Observation 6.
Any $(C_1,C_2,\ldots,C_{2t})$ corresponding to a non-zero term in (4) satisfies $\sum_{i=1}^{2t} 1_{C_i} = 0 \bmod 2$.

This crucial observation is actually quite simple to prove. We know that $1_{S_i} + 1_{S_{i+1}} = 1_{C_i} \bmod 2$ for every $i$—since $S_i \oplus S_{i+1} = C_i$. If we add up all the left-hand sides, we get a sum over $1_{S_i}$'s where every $1_{S_i}$ appears exactly twice (since $S_i$ occurs in exactly two entries $A(S_{i-1},S_i)$ and $A(S_i,S_{i+1})$). Thus the left hand side (and thus also the right hand side) must add up to the 0 vector modulo 2.

Next, notice that the $j$-th entry $\sum_{i=1}^{2t} 1_{C_i}(j) = 0 \bmod 2$ if and only if $j$ occurs in an even number of $C_i$'s. Thus, the above observation says that the (multi)-set $\{C_1,C_2,\ldots,C_{2t}\}$ is a cycle or an even cover. This appears exciting since we have a direct relationship between $\mathrm{tr}(A_{\mathrm{sat}}^{2t})$ and cycles in $H$!

There is a crucial snag though—the same $C$ could recur multiple times in $(C_1,C_2,\ldots,C_{2t})$. Indeed, if $C_i = C$ for every $i$ or more generally, every $C_i$ appeared in pairs, then, of course every element $j \in [n]$ will occur in an even number of $C_i$'s, for the trivial reason that the $C_i$'s themselves occur in pairs. Let's call such $2t$-tuples *trivial cycles*—that is, the $C_i$'s occur in pairs and thus do not relate to cycles in $H$.

Now for our endgame. For every trivial cycle, since $C_i$'s appear in pairs, the quantity $\prod_{i=1}^{2t} b_{C_i}$ has an even number of copies of $b_{C_i}$ for every $b_{C_i}$. Since $b_{C_i}^2 = 1$, this quantity must equal 1 *regardless of the* $C_i$*'s*. Thus, no matter what the $b_{C_i}$'s, all non-zero terms contribute exactly 1. In particular, $\mathrm{tr}(A_{\mathrm{sat}}^{2t})$ (where all $b_{C_i}$'s equal 1) must equal $\mathrm{tr}(A_{\mathbf{b}}^{2t})$ regardless of $b_{C_i}$s.

This finishes our argument, but it's perhaps helpful to summarize it: We related $\|A\|_2$ (for both $A = A_{\mathrm{sat}}$ and $A = A_{\mathbf{b}}$) to $\mathrm{tr}(A^{2t})$. We then related $\mathrm{tr}(A^{2t})$ to a sum over $2t$-tuples $(S_1,S_2,\ldots,S_{2t})$. The non-zero terms in this sum correspond to $\prod_{i=1}^{2t} b_{C_i}$ for $(C_1,C_2,\ldots,C_{2t})$ for $C_i \in H$ such that $\Sigma 1_{C_i} = 0 \bmod 2$—this step crucially uses the structure of the Kikuchi matrices. If $H$ has no $2t = \log\binom{n}{r}$ length cycles, then in every nonzero term in the sum the $C_i$'s must occur in pairs, in which case we observe that $\prod_{i=1}^{2t} b_{C_i} = 1$ and is thus independent of what the $b_{C_i}$'s themselves are.

## 4. EVEN SMALLER REFUTATIONS

In the final act of this article, we will come full circle to show how the purely combinatorial Feige's conjecture yields a surprising corollary for refutations for $k$-SAT. We discussed a spectral algorithm that finds refutations for smoothed $k$-SAT whenever the number of clauses $m \gtrsim n^{k/2}\log n$. Improving on this spectral threshold even for the substantially specialized setting of random 3-SAT has been open (with accumulating evidence that this may be impossible) ever since the 2004 work[8] that obtained the first such result.

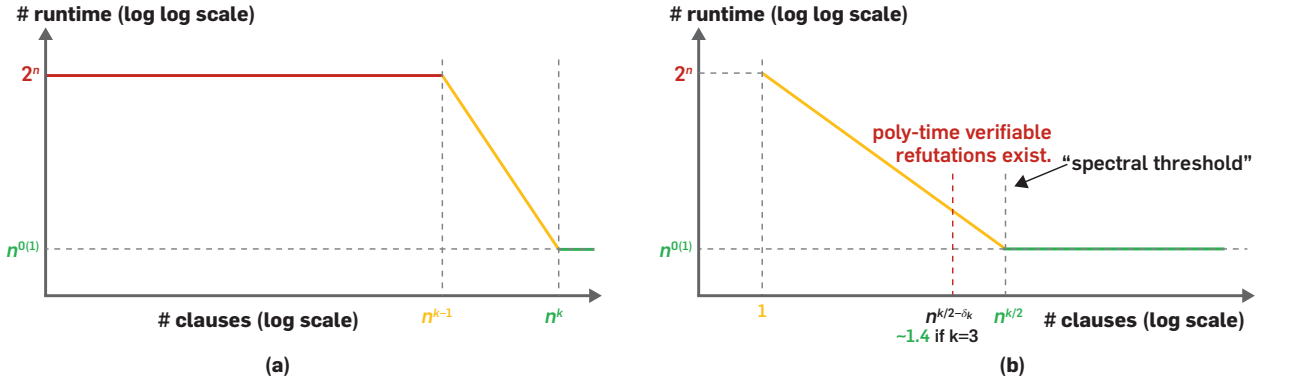
In a surprising twist from 2006, Feige, Kim, and Ofek proved[12] that for *random* 3-SAT formulas with $m \gtrsim n^{1.4}$ clauses (significantly short of the spectral threshold of $\approx n^{1.5}$) admit short, polynomial size refutations with high probability. That is, there *exists* a polynomial size certificate, based on a clever combination of spectral and combinatorial ideas, which, if given, can easily help convince us of its unsatisfiability. But despite around two decades of effort, we do not know polynomial time algorithms to find such a certificate. The FKO result forces us to grapple with the possibility that there may be a marked difference between *existence* of short certificates for NP-hard problems and efficient algorithms to find them. No such gap is known (or expected) for worst-case $k$-SAT, making this a truly average-case phenomenon. And, no such gap is known for any other discrete optimization problem, even in the average case. Indeed, ever since its discovery, FKO has been a one-of-a-kind result with an aura of mystery around it.

Are the mysterious FKO certificates a quirk of the random 3-SAT model? Or should we expect analogs in more general instances? We will now sketch how our results from the previous two sections allow us a surprising corollary:

Corollary 7.
With high probability, smoothed 3-SAT formulas with $m \gtrsim n^{1.4}\log n$ clauses admit an easily checkable polynomial size refutation. More generally, a similar result holds for smoothed $k$-SAT formulas with $n^{k/2-\delta_k}$ clauses, where $\delta_k > 0$ depends only on $k$.

That is, the FKO results extend without any quantitative change to smoothed 3-SAT formulas. The existence of short cycles in hypergraphs plays a major role in obtaining this corollary. Indeed, this was also a principle motivation for Feige's conjecture back in 2008. At the time, since Conjec-

**Figure 2. a) Runtime vs #clauses for worst-case k-SAT (best possible modulo standard conjectures). b) Runtime vs #clauses for smoothed k-SAT via Kikuchi-based spectral algorithm.**




ture 1 was not known, FKO's proof used a sophisticated application of the second moment method from probability theory with rather complicated calculations. Given Theorem 2, our new certificate and its analysis will be simple.

The idea for constructing such refutations is quite simple. Our spectral refutation worked by splitting the polynomial $\Phi$ into eight different polynomials of degree $\leq 3$ and then refuting each polynomial via our spectral algorithm. As before, we will do the splitting and use the spectral algorithm for all terms of degree $\leq 2$, that is, for all *except for the homogeneous degree 3 polynomial* corresponding to the last term in (1), for which we will use a "combinatorial" method. Importantly, for the degree $\leq 2$ terms, our polynomial time spectral algorithm from Section 2.1 only needs $m \gtrsim n/\epsilon^2 \log n$ (instead of $\approx n^{1.5}$) to certify a bound $\leq \epsilon$.

The basic observation behind the combinatorial method is rather simple. Let $H$ be the 3-uniform hypergraph of monomials appearing in $\Psi$. Let $\{C_1, C_2, \ldots, C_t\}$ be a cycle in $H$ and let $\Psi_{\text{cycle}} = \sum_{i=1}^{t} b_{C_i} x_{C_i}$ be the "fragment" of $\Psi$ that only keeps the monomials corresponding to the cycle. Then, if $\prod_{i=1}^{t} b_{C_i} = -1$ (which happens with probability 1/2 over the choice of $b_{C_i}$'s), then, we claim that $\Psi_{\text{cycle}}(x) \leq t - 1$ for every $x$. That is, every $x$ must in fact be at least 1 short of the maximum value of $t$ on such a fragment. Suppose not and say for some $x$, $\Psi_{\text{cycle}}(x) = t$. Then, $x_{C_i} = b_{C_i}$ for every $1 \leq i \leq t$. Thus, $\prod_{i=1}^{t} x_{C_i} = \prod_{i=1}^{t} b_{C_i} = -1$. On the other hand, since $\{C_1, C_2, \ldots, C_t\}$ is a cycle, every $j$ occurs in an even number of $C_i$'s and thus, $\prod_{i=1}^{t} x_{C_i} = \prod_{j=1}^{n} x_j^{(even)} = 1$. This is a contradiction!

Here's how this basic observation helps. Suppose $H$ has $m \approx n^{1.5} \log n/\sqrt{\ell}$ hyperedges. Then, we know from Theorem 2 that $H$ contains a $\approx \ell \log n$ cycle. We can then remove the hyperedges in this cycle and repeatedly find $\ell \log n$ length cycles in the residual hypergraph. This process gives us a "cycle partition" of 99% of hyperedges of $H$ into cycles of length $\approx \ell \log n$. From our argument above, for about 1/2 of such cycles, the product of the corresponding $b_{C_i}$'s will turn out to be –1. Let's call such cycles *violated*. Thus, can write:

$$\Psi = \sum_{\text{violated cycles in partition}} \Psi_{\text{cycle}}(x) + \Psi_{\text{remaining}}$$

For each violated cycle, every $x$ must "lose" at least 1 on the maximum possible value of the polynomial. So, we know that at any $x$, $\Psi$ must be short of its maximum value by at least the number of violated cycles in the partition. So, $\Psi(x) \leq m - O\left(\frac{m}{\ell \log n}\right) = \left(1 - O\left(\frac{1}{\ell \log n}\right)\right)m$. This is a significantly weaker bound than that of our spectral algorithm ($\epsilon m$) but is still non-trivial. It is also efficiently verifiable given a list of violated cycles (of size at most $O(n^{1.5})$, so polynomial size).

The corollary follows by combining this combinatorial certificate with the spectral bound on the degree $\leq 2$ parts of $\Phi$. The precise parameters are obtained by optimizing $\ell$ above but we will omit it here.

## 5. CONCLUSION

In this article, we surveyed a new class of spectral algorithms based on *Kikuchi* matrices to find refutations, that is, easily verifiable proofs of unsatisfiability, for smoothed $k$-SAT formulas. The guarantees we obtained were as strong as the best-known (and conjectured optimal) ones for the substantially simpler random $k$-SAT formulas and substantially surpass the best-possible (assuming standard hardness conjectures) running times for worst-case $k$-SAT formulas at every clause density. The approach generalizes to yield similar results for all logical constraint satisfaction problems. We also saw the resolution of the 2008 conjecture of Feige on short cycles in hypergraphs as an example application. And as a consequence, we saw how to extend the one-of-a-kind Feige-Kim-Ofek result from random $k$-SAT formulas to all smoothed $k$-SAT formulas. Taken together, the results show that, per the current state-of-the-art, smoothed $k$-SAT is no harder than the substantially simpler random $k$-SAT formulas for both refutation algorithms and existence of short certificates.

The *Kikuchi matrix method*, the method of proving results by finding spectral refutations for a related propositional formula, coming out of this line of work appears to be a promising new attack on problems in combinatorics and coding theory. It is a pleasing theoretical analog of the powerful approach for resolving mathematical problems via practical SAT solvers—a decidedly "computer science" approach to solve problems in mathematics. A few more applications, including making progress on some decades-old problems in the theory of local error-correcting codes,[4,19] are now already around and we anticipate more such results in the near future. (See Figure 2a and 2b.) C

**References**

1. Abascal, J., Guruswami, V., and Kothari, P.K. Strongly refuting all semi-random boolean csps. In *Proceedings of the 2021 ACM-SIAM Symp. Discrete Algorithms (Jan. 2021)*, 454–472.
2. Allen, S.R., O'Donnell, R., and Witmer, D. How to refute a random CSP. *Corr, Abs/1505.04383*, 2015.
3. Alon, N., Hoory, S., and Linial, N. The Moore bound for irregular graphs. *Graphs and Combinatorics 18* (2002), 53–57.
4. Alrabiah, O., Guruswami, V., Kothari, P.K., and Manohar, P. A near-cubic lower bound for 3-query locally decodable codes from semirandom CSP refutation. In *Proceedings of the 55th Ann. ACM Symp. on Theory of Computing (Jun. 2023)*, 1438–1448.
5. Bollobás, B. *Extremal Graph Theory*. Academic Press (1978).
6. Brakensiek, J., Heule, M., Mackey, J., and Narváez, D. The resolution of keller's conjecture. In *Automated Reasoning*. Springer Intern. Publishing, Cham. (2020), 48–65.
7. Briët, J. and Castro-Silva, D. On the threshold for Szemerédi's theorem with random differences. *Electronic J. of Combinatorics 31*, 4(2023).
8. Coja-Oghlan, A., Goerdt, A., and Lanka, A. Strong refutation heuristics for random k-sat. In *Proceedings of the 8th Intern. Workshop on Randomization and Computation 3122, Lecture Notes in Computer Science*. Springer (Aug. 2004), 310–321.
9. Feige, U. Refuting smoothed 3cnf formulas. In *Proceedings of the 48th Ann. IEEE Symp. on Foundations of Computer Science (Oct. 2007)*, 407–417.
10. Feige, U. *Small Linear Dependencies for Binary Vectors of Low Weight*. Springer Berlin Heidelberg (2008), 283–307.
11. Feige, U. Introduction to semirandom models. *Beyond the Worst-Case Analysis of Algorithms*. Cambridge University Press (2020), 189–211.
12. Feige, U., Kim, J.H., and Ofek, E. Witnesses for non-satisfiability of dense random 3cnf formulas. In *Proceedings of the 47th Annual IEEE Symp. on Foundations of Computer Science October* (Oct. 2006), 497–508.
13. Fichte, J.K., Berre, D.L., Hecher, M., and Szeider, S. The silent (r)evolution of SAT. *Commun. ACM 66*, 6 (May 2023), 64–72.
14. Fotakis, D., Lampis, M., and Paschos, V.T. Sub-exponential approximation schemes for CSPS: From dense to almost sparse. In *Proceedings of the 33rd Symp. on Theoretical Aspects of Computer Science 47*. Schloss Dagstuhl - Leibniz-Zentrum für Informatik (Feb. 2016), 37:1–37:14.
15. Goerdt, A. and Krivelevich, M. Efficient recognition of random unsatisfiable k-SAT instances by spectral methods. In *Proceedings of the 18th Ann. Symp. on Theoretical Aspects of Computer Science 2010, Lecture Notes in Computer Science*, Springer (Feb. 2001), 294–304.
16. Guruswami, V., Kothari, P.K., and Manohar, P. Algorithms and certificates for boolean CSP refutation: Smoothed is no harder than random. In *Proceedings of the 54th Ann. ACM SIGACT Symp. on Theory of Computing* (Jun. 2022), 678–689.
17. Hsieh, J., Kothari, P.K., and Mohanty, S. A simple and sharper proof of the hypergraph moore bound. In *Proceedings of the 2023 ACM-SIAM Symp. on Discrete Algorithms* (Jan. 2023), 2324–2344.
18. Karp, R.M. Reducibility among combinatorial problems. In *Proceedings of a Symp. on the Complexity of Computer Computations, the IBM Research Symposia Series*. Plenum Press (March 1972), 85–103.
19. Kothari, P.K. and Manohar, P. An exponential lower bound for linear 3-query locally correctable codes. *Corr, Abs/2311.00558*, 2023.
20. Kothari, P.K., Mori, R., O'Donnell, R., and Witmer, D. Sum of squares lower bounds for refuting any CSP. In *Proceedings of the 49th Ann. ACM SIGACT Symp. on Theory of Computing*, ACM (Jun. 2017), 132–145.
21. Naor, A. and Verstraete, J. Parity check matrices and product representations of squares. *Combinatorica 28* (Mar. 2008), 163–185.
22. Raghavendra, P., Rao, S., and Schramm, T. Strongly refuting random CSPS below the spectral threshold. In *Proceedings of the 49th Ann. ACM SIGACT Symp. on Theory of Computing* (Jun. 2017), 121–131.
23. Spielman, D.A. and Teng, S. Smoothed analysis of algorithms: Why the simplex algorithm usually takes polynomial time. In *Proceedings on 33rd Ann. ACM Symp. on Theory of Computing* (Jul. 2001), 296–305.
24. Wein, A.S., Alaoui, A.E., and Moore, C. The kikuchi hierarchy and tensor PCA. In Proceedings of the *60th IEEE Ann. Symp. on Foundations of Computer Science* (Nov. 2019), 1446–1468.



**Venkatesan Guruswami** (venkatg@berkeley.edu), University of California, Berkeley, CA, USA.

**Pravesh K. Kothari** (praveshk@cs.cmu.edu), Carnegie Mellon University, Pittsburgh, PA, USA.

**Peter Manohar** (pmanohar@cs.cmu.edu), Carnegie Mellon University, Pittsburgh, PA, USA.

DOI:10.1145/3708551

To view the accompanying paper, visit doi.acm.org/10.1145/3708494 rh

# Technical Perspective
# Toward Building a Differentially Private DBMS

By Graham Cormode

FROM THE DEMOGRAPHIC statistics produced by national census bodies to the complex predictive models built by companies in "Big Tech" and finance, data is the fuel that powers these applications. Most such use cases rely on data derived from the properties and actions of individual people. This data is therefore considered sensitive and in need of protections to prevent inappropriate use or disclosure. Some protections come from enforcing policies, access control, and contractual agreements. But we also seek technical interventions—definitions and algorithms that can be applied by computer systems to protect private information while still enabling the intended use.

Although there is not universal consensus, the model of *differential privacy* (DP) has emerged as the prevailing notion of privacy, with many deployments in industry and government, and thousands of research papers studying different aspects of the definition. At its heart, DP places a requirement on algorithms to introduce uncertainty to their output via randomization, so that the uncertainty is sufficient to provide reasonable doubt over whether the data of any particular person was part of the algorithm's input. Contributing to its success is the fact that differential privacy can be achieved by following some simple recipes. A starting point is to compute the true answer to a numerical query and then add noise from a suitable random distribution to obtain a "privatized" output that can be shared. These recipes can then be augmented with additional ingredients, such as sampling, aggregation, and optimization, along with combining and post-processing the results of multiple queries to provide private algorithms for a range of questions. Consequently, DP techniques offering precise privacy guarantees have been presented for tasks ranging from spectral graph analysis to training neural networks.

To achieve widespread adoption, the privacy research community needs to move beyond developing bespoke algorithms for each particular question at hand. Instead, we seek to build tools and systems that can handle a broad class of queries specified in a high-level language, and that automatically introduce the necessary randomness into the output to provide a DP guarantee. That is, we can aspire to a DP-DBMS: a differentially private data management system. The first attempts in this direction were based on the simple "evaluate and add noise" recipe above, but they encountered difficulties when the magnitude of the noise was found to be so large that it overwhelmed the true answer.

"R2T: Instance-optimal Truncation for Differentially Private Query Evaluation with Foreign Keys" by Dong et al. tackles an important technical question when trying to apply DP in this systematic fashion: how to tame the amount of privacy noise to answer database queries. It starts from a simple observation: The scale of noise required is directly proportional to the amount that a query answer can change when one individual's data is changed. This grows large in databases, where a single record pertaining to one person can link in turn to many other records spread throughout the data tables. But the impact can be kept in check by "truncating" the query output—clipping the contribution of a user to a limited range so the user's influence is de facto bounded. This sets up a trade-off: Truncating user contributions introduces bias to the query results, but relaxing the bound leads to higher variance from the privacy noise addition.

**Automatically ensuring privacy for arbitrary computations may be a long way off, but this work is an important step toward that end.**

The paper formalizes truncation as an optimization problem and presents the R2T algorithm as an efficient, iterative approach to solving it. It provides multiple other important contributions. First, it identifies the class of select-project-join-aggregate (SPJA) queries as the Goldilocks template—powerful enough to be highly expressive for database queries while not too complicated to prevent a generic solution. It handles the case of self-joins, for which simple approaches that assume independence will fail. The optimization requires only the solution of a linear program, which can be computed readily. The findings are validated by the proof-of-concept prototype system, which is shown to outperform existing state-of-the-art DP systems in evaluations over a number of benchmarks.

The next steps for this line of work are to consider other families of queries for other canonical types of data. We can seek to build analogous systems for other structured and unstructured forms of data—movement patterns (trajectories), medical readings, and written text, for instance. The long-term goal may be to build systems to handle increasingly higher-level queries over data, which can guarantee meaningful privacy protection while still demonstrating acceptable utility for the outputs. Automatically ensuring privacy for arbitrary computations may be a long way off, but this work is an important step toward that end. C

**Graham Cormode** is a professor at the University of Warwick, U.K. and a research scientist at Meta.

DOI:10.1145/3708494

To view the accompanying Technical Perspective, visit doi.acm.org/10.1145/3708551

# R2T: Instance-Optimal Truncation for Differentially Private Query Evaluation with Foreign Keys

By Wei Dong, Juanru Fang, Ke Yi, Yuchao Tao, and Ashwin Machanavajjhala

**Abstract**

**Answering SPJA queries under differential privacy (DP), including graph-pattern counting under node-DP as an important special case, has received considerable attention in recent years. The dual challenge of foreign-key constraints and self-joins is particularly tricky to deal with, and no existing DP mechanisms can correctly handle both. For the special case of graph pattern counting under node-DP, the existing mechanisms are correct (that is, satisfy DP), but they do not offer nontrivial utility guarantees or are very complicated and costly. In this paper, we propose the first DP mechanism for answering arbitrary SPJA queries in a database with foreign-key constraints. Meanwhile, it achieves a fairly strong notion of optimality, which can be considered as a small and natural relaxation of instance optimality. Finally, our mechanism is simple enough that it can be easily implemented on top of any RDBMS and an LP solver. Experimental results show that it offers order-of-magnitude improvements in terms of utility over existing techniques, even those specifically designed for graph pattern counting.**

## 1. INTRODUCTION

*Differential privacy (DP)* has become the standard notion for private data release, due to its strong protection of individual information. Informally speaking, DP requires indistinguishability of the query results whether any particular individual's data is in the database or not. The standard *Laplace mechanism* first finds $GS_Q$, the *global sensitivity*, of the query, that is, how much the query result may change if an individual's data is added/removed from the database. Then it adds a Laplace noise calibrated accordingly to the query result to mask this difference. However, this mechanism runs into issues in a relational database, as illustrated in the following example.

Example 1.1.
Consider a simple join-counting query

$$Q := |R_1(\underline{x_1}, \ldots) \bowtie R_2(x_1, x_2, \ldots)|.$$

Here, the underlined attribute $\underline{x_1}$ is the primary key (PK), while $R_2.x_1$ is a foreign key (FK) referencing $R_1.x_1$. For instance, $R_1$ may store customer information where $x_1$ is the customer ID and $R_2$ stores the orders the customers have placed. Then this query simply returns the total number of orders; more meaningful queries could be formed with some predicates: for example, all customers from a certain region and/or orders in a certain category. Furthermore, suppose the customers, namely, the tuples in $R_1$, are the entities whose privacy we aim to protect.

What's the $GS_Q$ for this query? It is, unfortunately, $\infty$. This is because a customer, theoretically, could have an unbounded number of orders, and adding such a customer to the database can cause an unbounded change in the query result. A simple fix is to assume a finite $GS_Q$, which can be justified in practice because we may never have a customer with, say, more than a million orders. However, as assuming such a $GS_Q$ limits the allowable database instances, one tends to be conservative and sets a large $GS_Q$. This allows the Laplace mechanism to work, but adding noise of this scale clearly eliminates any utility of the released query answer.

### 1.1. The truncation mechanism.

The issue above was first identified by Kotsogiannis et al.,[21] who also formalized the *DP policy for relational databases with foreign key (FK) constraints*. The essence of their model (a rigorous definition is given in Section 2) is that individuals and their private data are stored in separate relations linked by FKs. This is perhaps the most crucial feature of the relational model, yet it causes a major difficulty in designing DP mechanisms as illustrated above. Their solution is the *truncation mechanism*, which simply deletes all customers with more than $\tau$ orders before applying the Laplace mechanism, for some threshold $\tau$. After truncation, the query has sensitivity $\tau$, so adding a noise of scale $\tau$ is sufficient.

Truncation is a special case of Lipshitz extensions and has been studied extensively for graph pattern-counting queries[20] and machine learning (ML).[1] A well-known issue for the truncation mechanism is the bias-variance tradeoff: In one extreme $\tau = GS_Q$; it degenerates into the naive Laplace mechanism with a large noise (that is, large variance). In the other extreme $\tau = 0$, the truncation introduces a bias as large as the query answer. The issue of how to choose a near-optimal $\tau$ has been extensively studied in the statis-

The original version of this paper was published in *Proceedings of SIGMOD '22 (*June 2022).

tics and ML communities.[2,17] In fact, the particular query in Example 1.1 is equivalent to the 1-dimensional mean (sum) estimation problem, which is important for many ML tasks. A key challenge there is that the selection of $\tau$ must also be done in a DP manner.

### 1.2. The issue with self-joins.

While self-join-free queries are equivalent to mean (sum) estimation (see Section 3 for a more formal statement), self-joins introduce another challenge unique to relational queries. In particular, all techniques from the statistics and machine-learning (ML) literature for choosing a $\tau$ critically rely on the fact that the individuals are independent, that is, adding/removing one individual does not affect the data associated with another, which is not true when the query involves self-joins. In fact, when there are self-joins, even the truncation mechanism itself fails, as illustrated in the example below.

Example 1.2.
Suppose we extend the query from Example 1.1 to the following one with a self-join:

$$Q := |R_1(\underline{x_1}, \ldots,) \bowtie R_1(\underline{y_1}, \ldots) \bowtie R_2(x_1, y_1, x_2, \ldots)|.$$

Note that the PK of $R_1$ has been renamed differently in the two logical copies $R_1$, so that they join different attributes of $R_2$. For instance, $R_2$ may store the transactions between pairs of customers, and this query would count the total number of transactions. Again, predicates can be added to make the query more meaningful.

Let $G$ be an undirected $\tau$-regular graph (that is, every vertex has degree $\tau$) with $n$ vertices. We will construct an instance $\mathbf{I} = (I_1, I_2)$ on which the truncation mechanism fails. Let $I_1$ be the vertices of $G$ and let $I_2$ be the edges (each edge will appear twice as $G$ is undirected). Thus, $Q$ simply returns the number of edges in the graph times 2. Let $\mathbf{I}'$ be the neighboring instance corresponding to $G'$, to which we add a vertex $v$ that connects to every existing vertex. Note that in $G'$, $v$ has degree $n$ while every other vertex has degree $\tau + 1$. Now truncating by $\tau$ fails DP: The query answer on $\mathbf{I}$ is $n\tau$, and that on $\mathbf{I}'$ is 0 (all vertices are truncated). Adding noise of scale $\tau$ cannot mask this gap, violating the DP definition.

The reason why the truncation mechanism fails is that the underlined claim above does not hold in the presence of self-joins. More fundamentally, this is due to the correlation among the individuals introduced by self-joins. In the example above, we see that the addition of one node may cause the degrees of many others to increase. For the problem of graph pattern counting under node-DP, which can be formulated as a multi-way self-join counting query on the special schema $\mathbf{R}$ = {Node(ID), Edge(src, dst)}, Kasiviswanathan et al.[20] propose an LP-based truncation mechanism (to differentiate, we will call the truncation mechanism above *naive truncation*) to fix the issue, but they do not study how to choose $\tau$. As a result, while their mechanism satisfies DP, there is no optimality guarantee in terms of utility. In fact, if $\tau$ is chosen inappropriately, their error can be even larger than $GS_Q$, namely, worse than the naive Laplace mechanism.

### 1.3. Our contributions.

In this paper, we start by studying how to choose a near-optimal $\tau$ in a DP manner in the presence of self-joins. As with all prior $\tau$-selection mechanisms over mean (sum) estimation[2,17] and self-join-free queries,[24] we assume that the *global sensitivity* of the given query $Q$ is bounded by $GS_Q$. Since one tends to set a large $GS_Q$ as argued in Example 1.1, we must try to minimize the dependency on $GS_Q$.

The first contribution of this paper (Section 4) is a simple and general DP mechanism, called *Race-to-the-Top (R2T)*, which can be used to adaptively choose $\tau$ in combination with any valid DP truncation mechanism that satisfies certain properties. In fact, it does not choose $\tau$ per se; instead, it directly returns a privatized query answer with error at most $O\big(\log(GS_Q)\log\log(GS_Q) \cdot DS_Q(\mathbf{I})\big)$ for any instance $\mathbf{I}$ with constant probability. While we defer the formal definition of $DS_Q(\mathbf{I})$ to Section 3, what we can show is that it is a *per-instance lower bound*, that is, any valid DP mechanism has to incur error $\Omega\big(DS_Q(\mathbf{I})\big)$ on $\mathbf{I}$ (in a certain sense). Thus, the error of R2T is *instance-optimal* up to logarithmic factors in $GS_Q$. Furthermore, a logarithmic dependency on $GS_Q$ is also unavoidable,[19] even for the mean estimation problem, that is, the simple self-join-free query in Example 1.1. In practice, these log factors are usually between 10 to 100, and our experiments show that R2T has better utility than previous methods in most cases.

However, as we see in Example 1.2, naive truncation is not a valid DP mechanism in the presence of self-joins. As our second contribution (Section 5), we extend the LP-based mechanism of Kasiviswanathan et al,[20] which only works for graph pattern-counting queries, to general queries on an arbitrary relational schema that uses the four basic relational operators: Selection (with arbitrary predicates), Projection, Join (including self-join), and sum Aggregation. When plugged into R2T, this yields the first DP mechanism for answering arbitrary SPJA queries in a database with FK constraints. For SJA queries, the utility is instance-optimal, while the optimality guarantee for SPJA queries is slightly weaker, but we argue that this is unavoidable.

Furthermore, the simplicity of our mechanism allows it to be built on top of any RDMBS and an LP solver. To demonstrate its practicality, we built a system prototype (Section 7) using PostgreSQL and CPLEX. Experimental results (Section 8) show it can provide order-of-magnitude improvements in terms of utility over the state-of-the-art DP-SQL engines. We obtain similar improvements even over node-DP mechanisms specifically designed for graph pattern-counting problems, which are just special SJA queries.

## 2. PRELIMINARIES

### 2.1. Database queries

Let $\mathbf{R}$ be a database schema. We start with a multi-way join:

$$J := R_1(\mathbf{x}_1) \bowtie \cdots \bowtie R_n(\mathbf{x}_n), \tag{1}$$

where $R_1, \ldots, R_n$ are relation names in $\mathbf{R}$ and each $\mathbf{x}_i$ is a set

of $arity(R_i)$ variables. When considering self-joins, there can be repeats, that is, $R_i = R_j$; in this case, we must have $\mathbf{x}_i \neq \mathbf{x}_j$, or one of the two atoms will be redundant. Let $var(J) := \mathbf{x}_1 \cup \cdots \cup \mathbf{x}_n$.

Let $\mathbf{I}$ be a database instance over $\mathbf{R}$. For any $R \in \mathbf{R}$, denote the corresponding relation instance in $\mathbf{I}$ as $\mathbf{I}(R)$. This is a *physical relation instance* of $R$. We use $\mathbf{I}(R, \mathbf{x})$ to denote $\mathbf{I}(R)$ after renaming its attributes to $\mathbf{x}$, which is also called a *logical relation instance* of $R$. When there are self-joins, one physical relation instance may have multiple logical relation instances; they have the same rows but with different column (variable) names.

A JA or an SJA query $Q$ aggregates over the join results $J(\mathbf{I})$. More abstractly, let $\psi : \mathbf{dom}(var(J)) \to \mathbb{N}$ be a function that assigns non-negative integer weights to the join results, where $\mathbf{dom}(var(J))$ denotes the domain of $var(J)$. The result of evaluating $Q$ on $\mathbf{I}$ is

$$Q(\mathbf{I}) := \sum_{q \in J(\mathbf{I})} \psi(q). \tag{2}$$

Note that the function $\psi$ only depends on the query. For a counting query, $\psi(\cdot) \equiv 1$; for an aggregation query, for example, $\texttt{SUM}(A * B)$, $\psi(q)$ is the value of $A * B$ for $q$. And an SJA query with an arbitrary predicate over $var(J)$ can be easily incorporated into this formulation: If some $q \in J(\mathbf{I})$ does not satisfy the predicate, we simply set $\psi(q) = 0$.

Example 2.1.
Graph pattern-counting queries can be formulated as SJA queries. Suppose we store a graph in a relational database by the schema $\mathbf{R} = \{\texttt{Edge}(\texttt{src}, \texttt{dst}), \texttt{Node}(\underline{\texttt{ID}})\}$ where src and dst are FKs referencing ID, then the number of length-3 paths can be counted by first computing the join

$$\texttt{Edge}(\texttt{A}, \texttt{B}) \bowtie \texttt{Edge}(\texttt{B}, \texttt{C}) \bowtie \texttt{Edge}(\texttt{C}, \texttt{D}),$$

followed by a count aggregation. Note that this also counts triangles and non-simple paths (for example, *x-y-x-z*), which may or may not be considered as length-3 paths depending on the application. If not, they can be excluded by introducing a predicate (that is, redefining $\psi$) $\texttt{A} \neq \texttt{C} \wedge \texttt{A} \neq \texttt{D} \wedge \texttt{B} \neq \texttt{D}$. If the graph is undirected, then the query counts every path twice, so we should divide the answer by 2. Alternatively, we may introduce the predicate $\texttt{A} < \texttt{D}$ to eliminate the double counting.

## 2.2. DP in relational databases with FK constraints.

We adopt the DP policy in Kotsogiannis. et al,[21] which defines neighboring instances by taking FK constraints into consideration. We model all the FK relationships as a directed acyclic graph (DAG) over $\mathbf{R}$ by adding a directed edge from $R$ to $R'$ if $R$ has an FK referencing the PK of $R'$. There is a,[a] d esignated *primary private relation* $R_p$, and any relation that has a direct or indirect FK referencing $R_p$ is called a *secondary private relation*. The *referencing* relationship over the tuples is defined recursively as follows: (1) any tuple $t_p \in \mathbf{I}(R_p)$ said to reference itself; (2) for $t_p \in \mathbf{I}(R_p)$, $t \in \mathbf{I}(R)$, $t' \in \mathbf{I}(R')$, if $t'$ references $t_p$, $R$ has an FK referencing the PK of $R'$, and the FK of $t$ equals the PK of $t'$, then we say that $t$ references $t_p$. Then two instances $\mathbf{I}$ and $\mathbf{I}'$ are considered neighbors if $\mathbf{I}'$ can be obtained from $\mathbf{I}$ by deleting a set of tuples, all of which reference the same tuple $t_p \in \mathbf{I}(R_p)$, or vice versa. In particular, $t_p$ may also be deleted, in which case all tuples referencing $t_p$ must be deleted to preserve the FK constraints. Finally, for a join result $q \in J(\mathbf{I})$, we say that $q$ references $t_p \in \mathbf{I}(R_p)$ if $|t_p \bowtie q| = 1$.

We use the notation $\mathbf{I} \sim \mathbf{I}'$ to denote two neighboring instances and $\mathbf{I} \sim_{t_p} \mathbf{I}'$ denotes that all tuples in the difference between $\mathbf{I}$ and $\mathbf{I}'$ reference the tuple $t_p \in R_p$.

Example 2.2.
Consider the TPC-H schema:

$$\mathbf{R} = \{\texttt{Nation}(\underline{\texttt{NK}}), \texttt{Customer}(\underline{\texttt{CK}}, \texttt{NK}), \texttt{Order}(\underline{\texttt{OK}}, \texttt{CK}), \texttt{Lineitem}(\texttt{OK})\}.$$

If the customers are the individuals whose privacy we wish to protect, then we designate Customer as the primary private relation, which implies that Order and Lineitem will be secondary private relations, while Nation will be public. Note that once Customer is designated as a primary private relation, the information in Order and Lineitem is also protected, since the privacy induced by Customer is stronger than that induced by Order and Lineitem. Alternatively, one may designate Order as the primary private relation, which implies that Lineitem will be a secondary private relation, while Customer and Nation will be public. This would result in weaker privacy protection but offer higher utility.

Some queries, as given, may be *incomplete*, that is, it has a variable that is an FK but its referenced PK does not appear in the query $Q$. The query in Example 2.1 is such an example. Following Kotsogiannis et al.,[21] we always make the query complete by iteratively adding those relations whose PKs are referenced to $Q$. The PKs will be given variable names matching the FKs. For example, for the query in Example 2.1, we add Node(A), Node(B), Node(C), and Node(D).

The DP policy above incorporates both edge-DP and node-DP, two commonly used DP policies for private graph analysis, as special cases. In Example 2.1, by designating Edge as the private relation (Node is thus public, and we may even assume it contains all possible vertex IDs), we obtain edge-DP; for node-DP, we add FK constraints from src and dst to ID, and designate Node as the primary private relation, while Edge becomes a secondary private relation.

A mechanism $M$ is $\varepsilon$-DP if for any neighboring instance $\mathbf{I}$, $\mathbf{I}'$, and any output $y$, we have

$$\Pr[M(\mathbf{I}) = y] \leq e^{\varepsilon} \Pr[M(\mathbf{I}') = y].$$

Typical values of $\varepsilon$ used in practice range from 0.1 to 10, where a smaller value corresponds to stronger privacy protection.

a For most parts of the paper, we consider the case where there is only one *primary private relation* in $\mathbf{R}$; the case with multiple primary private relations can be transformed to a case with a single primary private relation (see our full version paper for more details)

## 3. INSTANCE OPTIMALITY OF DP MECHANISMS WITH FK CONSTRAINTS

**Global sensitivity and worst-case optimality.** The standard DP mechanism is the Laplace mechanism,[15] which adds $Lap\left(GS_Q\right)$ to the query answer. Here, $Lap(b)$ denotes a random variable drawn from the Laplace distribution with scale $b$, and $GS_Q = \max_{\mathbf{I}\sim\mathbf{I}'}|Q(\mathbf{I}) - Q(\mathbf{I}')|$ is the *global sensitivity* of $Q$. However, either a join or a sum aggregation makes $GS_Q$ unbounded. The issue with the former is illustrated in Example 1.1, where a customer may have unbounded orders; a sum aggregation with an unbounded $\psi$ results in the same situation. Thus, as with prior work,[2,17,24] we restrict to a set of instances $\mathcal{I}$ such that

$$\max_{\mathbf{I}\in\mathcal{I},\mathbf{I}'\in\mathcal{I},\mathbf{I}\sim\mathbf{I}'}|Q(\mathbf{I}) - Q(\mathbf{I}')| = GS_Q, \tag{3}$$

where $GS_Q$ is a parameter given in advance. For the query in Example 1.1, this is equivalent to assuming that a customer is allowed to have at most $GS_Q$ orders in any instance.

For general queries, the situation is more complicated. We first consider SJA queries. Given an instance $\mathbf{I}$ and an SJA query $Q$, for a tuple $t_P \in \mathbf{I}\left(R_P\right)$, its *sensitivity* is

$$S_Q\left(\mathbf{I}, t_P\right) := \sum_{q\in Q(\mathbf{I})} \psi(q)\mathbb{I}\left(q \text{ references } t_P\right), \tag{4}$$

where $\mathbb{I}(\cdot)$ is the indicator function. For SJA queries, (1) is equivalent to

$$\max_{\mathbf{I}\in\mathcal{I}} \max_{t_P\in\mathbf{I}\left(R_P\right)} S_Q\left(\mathbf{I}, t_P\right) = GS_Q.$$

For self-join-free SJA queries, it is clear that

$$Q(\mathbf{I}) = \sum_{t_P\in R_P} S_Q\left(\mathbf{I}, t_P\right),$$

which turns the problem into a sum estimation problem. However, when self-joins are present, this equality no longer holds, since one join result $q$ references multiple $t_P$'s. This also implies that removing one tuple from $\mathbf{I}\left(R_P\right)$ may affect multiple $S_Q\left(\mathbf{I}, t_P\right)$'s, making the neighboring relationship more complicated than in the sum estimation problem, where two neighboring instances differ by only one datum.[2,17]

What notion of optimality shall we use for DP mechanisms over SJA queries? The traditional worst-case optimality is meaningless, since the naive Laplace mechanism that adds noise of scale $GS_Q$ is already worst-case optimal, just by the definition of $GS_Q$. In fact, the basis of the entire line of work on the truncation mechanism and smooth sensitivity is the observation that typical instances should be much easier than the worst case, so these mechanisms all add instance-specific noises, which are often much smaller than the worst-case noise level $GS_Q$.

**Instance optimality.** The standard notion of optimality for measuring the performance of an algorithm on a per-instance basis is *instance optimality*. More precisely, let $\mathcal{M}$ be the class of DP mechanisms and let[b]

$$\mathscr{L}_{\text{ins}}(\mathbf{I}) := \min_{M'\in\mathcal{M}} \min\{\xi : \Pr[|M'(\mathbf{I}) - Q(\mathbf{I})| \le \xi] \ge 2/3\}$$

be the lower bound any $M' \in \mathcal{M}$ can achieve (with probability 2/3) on $\mathbf{I}$, then the standard definition of instance optimality requires us to design an $M$ such that

$$\Pr\left[\left|M(\mathbf{I}) - Q(\mathbf{I})\right| \le c\cdot\mathscr{L}_{\text{ins}}(\mathbf{I})\right] \ge 2/3 \tag{5}$$

for every $\mathbf{I}$, where $c$ is called the *optimality ratio*. Unfortunately, for any $\mathbf{I}$, one can design a trivial $M'(\cdot) \equiv Q(\mathbf{I})$ that has 0 error on $\mathbf{I}$ (but fails miserably on other instances), so $\mathscr{L}_{\text{ins}}(\cdot) \equiv 0$, which rules out instance-optimal DP mechanisms by a standard argument.[15]

To avoid such a trivial $M'$,[3,12] consider a relaxed version of instance optimality where we compare $M$ against any $M'$ that is required to work well not just on $\mathbf{I}$, but also on its neighbors, that is, we raise the target error from $\mathscr{L}_{\text{ins}}(\mathbf{I})$ to

$$\mathscr{L}_{\text{nbr}}(\mathbf{I}) := \min_{M'\in\mathcal{M}} \max_{\mathbf{I}:\mathbf{I}\sim\mathbf{I}} \min\{\xi : \Pr[|M'(\mathbf{I}') - Q(\mathbf{I}')| \le \xi] \ge 2/3\}.$$

Vadhan[25] observes that $\mathscr{L}_{\text{nbr}}(\mathbf{I}) \ge LS_Q(\mathbf{I})/2$, where

$$LS_Q(\mathbf{I}) := \max_{\mathbf{I}'\in\mathcal{I},\mathbf{I}'\sim\mathbf{I}} |Q(\mathbf{I}) - Q(\mathbf{I}')|$$

is the *local sensitivity* of $Q$ at $\mathbf{I}$. This instance optimality has been used for certain ML problems[3] and conjunctive queries without FKs.[12] However, it has an issue for SJA queries in a database with FK constraints: For any $\mathbf{I}$, we can add a $t_p$ to $\mathbf{I}\left(R_P\right)$ together with tuples in the secondary private relations all referencing $t_P$, obtaining an $\mathbf{I}'$ such that $S_Q\left(\mathbf{I}', t_P\right) = GS_Q$, that is, $LS_Q(\cdot) \equiv GS_Q$. This means that this relaxed instance optimality degenerates into worst-case optimality. This is also why smooth sensitivity, including all its efficiently computable versions,[11,12,18,23] will not have better utility than the naive Laplace mechanism on databases with FK constraints, since they are all no lower than the local sensitivity.

The reason why the above relaxation is "too much" is that we require $M'$ to work well on any neighbor $\mathbf{I}'$ of $\mathbf{I}$. Under the neighborhood definition with FK constraints, this means that $\mathbf{I}'$ can be any instance obtained from $\mathbf{I}$ by adding a tuple $t_P$ and *arbitrary* tuples referencing $t_P$ in the secondary private relations. This is too high a requirement for $M'$, hence too low an optimality notion for $M$.

To address the issue, Huang et al.[17] restricts the neighborhood in which $M'$ is required to work well, but their definition only works for the mean estimation problem. For SJA queries under FK constraints, we revise $\mathscr{L}_{\text{nbr}}(\cdot)$ to

$$\mathscr{L}_{\text{d-nbr}}(\mathbf{I}) := \min_{M'\in\mathcal{M}} \max_{\mathbf{I}:\mathbf{I}\sim\mathbf{I}',\mathbf{I}'\subseteq\mathbf{I}} \min\left\{\xi : \Pr\left[\left|M'(\mathbf{I}') - Q(\mathbf{I}')\right| \le \xi\right] \ge 2/3\right\},$$

namely, we require $M'$ to work well only on $\mathbf{I}'$ and its *down-neighbors*, which can be obtained only by removing a tuple $t_P$ already in $\mathbf{I}\left(R_P\right)$ and all tuples referencing $t_P$. Correspondingly, an instance-optimal $M$ (with regard to the down-neighborhood) is one such that (1) holds where $\mathscr{L}_{\text{ins}}$ is replaced by $\mathscr{L}_{\text{d-nbr}}$.

Clearly, the smaller the neighborhood, the stronger the optimality notion. Our instance optimality notion is thus stronger than those in.[3,12,17] Note that for such an instance-optimal $M$ (by our definition), there still exist $\mathbf{I}, M'$ such that $M'$ does better on $\mathbf{I}$ than $M$, but if this happens, $M'$ must do worse on one of the down-neighbors of $\mathbf{I}$, which is as typical as $\mathbf{I}$ itself.

Using the same argument from Vadhan,[25] we have $\mathscr{L}_{\text{d-nbr}}(\mathbf{I}) \ge DS_Q(\mathbf{I})/2$, where

b The probability constant 2/3 can be changed to any constant larger than 1/2 without affecting the asymptotics.

$$DS_Q(\mathbf{I}) := \max_{\mathbf{I}',\mathbf{I}\sim\mathbf{I}',\mathbf{I}'\subseteq\mathbf{I}} |Q(\mathbf{I}) - Q(\mathbf{I}')| = \max_{t_p\in\mathbf{I}(R_p)} S_Q(\mathbf{I}, t_p) \quad (6)$$

is the *downward local sensitivity* of $\mathbf{I}$. Thus, $DS_Q(\mathbf{I})$ is a per-instance lower bound, which can be used to replace $\mathscr{L}_{\text{inc}}(\mathbf{I})$ in (5) in the definition of instance-optimal DP mechanisms.

## 4. R2T: INSTANCE-OPTIMAL TRUNCATION

Our instance-optimal truncation mechanism, *Race-to-the-Top (R2T)*, can be used in combination with any truncation method $Q(\mathbf{I},\tau)$, which is a function $Q:\mathscr{I}\times\mathbb{N}\to\mathbb{N}$ with the following properties:

(1) For any $\tau$, the global sensitivity of $Q(\cdot,\tau)$ is at most $\tau$.
(2) For any $\tau$, $Q(\mathbf{I},\tau)\le Q(\mathbf{I})$.
(3) For any $\mathbf{I}$, there exists a non-negative integer $\tau^*(\mathbf{I})\le GS_Q$ such that for any $\tau\ge\tau^*(\mathbf{I})$, $Q(\mathbf{I},\tau)=Q(\mathbf{I})$.

We describe various choices for $Q(\mathbf{I},\tau)$ depending on the DP policy and whether the query contains self-joins and/or projections in the subsequent sections. Intuitively, such a $Q(\mathbf{I},\tau)$ gives a stable (property (1)) underestimate (property (2)) of $Q(\mathbf{I})$, while reaches $Q(\mathbf{I})$ for a sufficiently large $\tau$ (property (3)). Note that $Q(\mathbf{I},\tau)$ itself is not DP. To make it DP, we can add $Lap(\tau/\varepsilon)$, which would turn it into an $\varepsilon$-DP mechanism by property (1). The issue, of course, is how to set $\tau$. The basic idea of R2T is to try geometrically increasing values of $\tau$ and somehow pick the "winner" of the race.

Assuming such a $Q(\mathbf{I},\tau)$, R2T works as follows. For a probability $\beta$,[c] we first compute[d]

$$\widetilde{Q}(\mathbf{I},\tau^{(j)}) := Q(\mathbf{I},\tau^{(j)}) + Lap\left(\log(GS_Q)\frac{\tau^{(j)}}{\varepsilon}\right) - \log(GS_Q)\ln\left(\frac{\log(GS_Q)}{\beta}\right)\cdot\frac{\tau^{(j)}}{\varepsilon}, \quad (7)$$

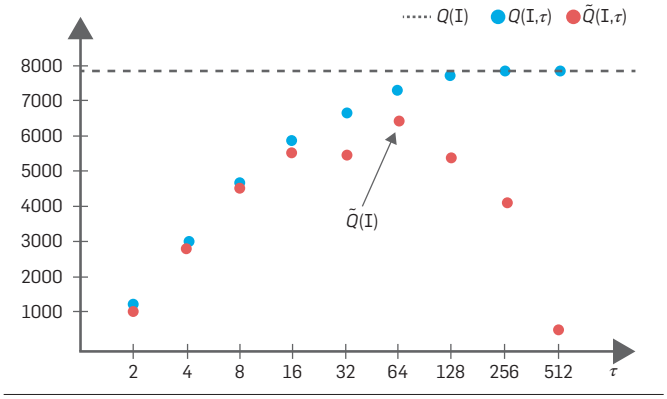
for $\tau^{(j)}=2^j, j=1,\ldots,\log(GS_Q)$. Then R2T outputs

$$\widetilde{Q}(\mathbf{I}) := \max\left\{\max_j \widetilde{Q}(\mathbf{I},\tau^{(j)}), Q(\mathbf{I},0)\right\}. \quad (8)$$

The privacy of R2T is straightforward: Since $Q(\mathbf{I},\tau^{(j)})$ has global sensitivity at most $\tau^{(j)}$, and the third term of (7) is independent of $\mathbf{I}$, each $\widetilde{Q}(\mathbf{I},\tau^{(j)})$ satisfies $\frac{\varepsilon}{\log(GS_Q)}$-DP by the standard Laplace mechanism. Collectively, all the $\widetilde{Q}(\mathbf{I},\tau^{(j)})$'s satisfy $\varepsilon$-DP by the basic composition theorem.[15] Finally, returning the maximum preserves DP by the post-processing property of DP.

**Utility analysis.** For some intuition on why R2T offers good utility, see Figure 1. By property (2) and (3), as we increase $\tau$, $Q(\mathbf{I},\tau)$ gradually approaches the true answer $Q(\mathbf{I})$ from below and reaches $Q(\mathbf{I},\tau)=Q(\mathbf{I})$ when $\tau\ge\tau^*(\mathbf{I})$. However, we cannot use $Q(\mathbf{I},\tau)$ or $\tau^*(\mathbf{I})$ directly as this would violate DP. Instead, we only get to see $\widetilde{Q}(\mathbf{I},\tau)$, which is masked with the noise of scale proportional to $\tau$. We thus face a dilemma: The closer we get to $Q(\mathbf{I})$, the more uncertain we are about the estimate $\widetilde{Q}(\mathbf{I},\tau)$. To get out of the dilemma, we shift $Q(\mathbf{I},\tau)$ down by an amount that equals the scale of the noise (if ignoring the loglog factor). This penalty for $\widetilde{Q}(\mathbf{I},\hat{\tau})$, where $\hat{\tau}$ is the smallest power of 2 above $\tau^*(\mathbf{I})$, will be on the same order as $\tau^*(\mathbf{I})$, so it will not affect its error by more than a constant factor, while taking the maximum ensures that the winner is at least as good as $\widetilde{Q}(\mathbf{I},\hat{\tau})$. Meanwhile, the extra loglog factor ensures that no $\widetilde{Q}(\mathbf{I},\tau)$ overshoots the target. Below, we formalize the intuition.

**Figure 1. An illustration of R2T.**




THEOREM 1.
On any instance $\mathbf{I}$, with probability at least $1-\beta$, we have

$$Q(\mathbf{I}) - 4\log(GS_Q)\ln\left(\frac{\log(GS_Q)}{\beta}\right)\frac{\tau^*(\mathbf{I})}{\varepsilon} \le \widetilde{Q}(\mathbf{I}) \le Q(\mathbf{I}).$$

## 5. TRUNCATION FOR SJA QUERIES

In this section, we will design a $Q(\mathbf{I},\tau)$ with $\tau^*(\mathbf{I})=DS_Q(\mathbf{I})$ for SJA queries. Plugged into Theorem 1 with $\beta=1/3$ and the definition of instance optimality, this turns R2T into an instance-optimal DP mechanism with an optimality ratio of $O\left(\log(GS_Q)\log\log(GS_Q)/\varepsilon\right)$.

For self-join-free SJA queries, each join result $q\in J(\mathbf{I})$ references only one tuple in $R_p$. Thus, the tuples in $R_p$ are independent, that is, removing one does not affect the sensitivities of others. This means that naive truncation (that is, removing all $S_Q(\mathbf{I},t_p)>\tau$ and then summing up the rest) is a valid $Q(\mathbf{I},\tau)$ that satisfies the three properties required by R2T with $\tau^*(\mathbf{I})=DS_Q(\mathbf{I})$.

When there are self-joins, naive truncation does not satisfy property (1), as illustrated in Example 1.2, where all $S_Q(\mathbf{I},t_p)$'s in two neighboring instances may differ. Below, we generalize the LP-based mechanism for graph pattern counting[20] to arbitrary SJA queries, and show that it satisfies the three properties with $\tau^*(\mathbf{I})=DS_Q(\mathbf{I})$.

Given a SJA query $Q$ and instance $\mathbf{I}$, recall that $Q(\mathbf{I})=\sum_{q\in J(\mathbf{I})}\psi(q)$, where $J(\mathbf{I})$ is the join results. For $k\in[|J(\mathbf{I})|]$, let $q_k(\mathbf{I})$ be the $k$th join result. For each $j\in\left[\left|\mathbf{I}(R_p)\right|\right]$, let $t_j(\mathbf{I})$ be the $j$th tuple in $\mathbf{I}(R_p)$. We use $C_j(\mathbf{I})$ to denote (the indices of) the set of join results that reference $t_j(\mathbf{I})$. More precisely,

$$C_j(\mathbf{I}) := \left\{k : q_k(\mathbf{I}) \text{ references } t_j(\mathbf{I})\right\}. \quad (9)$$

For each $k\in[|J(\mathbf{I})|]$, introduce a variable $u_k$, which represents the weight assigned to the join result $q_k(\mathbf{I})$. We return the optimal solution of the following LP as $Q(\mathbf{I},\tau)$:

$$\begin{aligned} \text{maximize} \quad & Q(\mathbf{I},\tau)=\sum_{k\in[|J(\mathbf{I})|]} u_k, \\ \text{subject to} \quad & \sum_{k\in C_j(\mathbf{I})} u_k \le \tau, \qquad j\in\left[\left|\mathbf{I}(R_p)\right|\right], \\ & 0\le u_k\le\psi(q_k(\mathbf{I})), \qquad k\in[|J(\mathbf{I})|]. \end{aligned}$$

c The probability $\beta$ only concerns about the utility but not privacy.
d log has base 2 and ln has base $e$.

LEMMA 1.
For SJA queries, the $Q(\mathbf{I},\tau)$ defined above satisfies the three properties required by R2T with $\tau^*(\mathbf{I}) = DS_Q(\mathbf{I})$.

EXAMPLE 5.1.
We now give a step-by-step example to show how this truncation method works together with R2T. Consider the problem of edge counting under node-DP, which corresponds to the SJA query

$$Q := \left|\sigma_{\mathtt{ID1}<\mathtt{ID2}}(\mathtt{Node(ID1)} \bowtie \mathtt{Node(ID2)} \bowtie \mathtt{Edge(ID1,ID2)})\right|$$

on the graph data schema introduced in Example 2.1. Note that in SQL, the query would be written as

```
SELECT count(*) FROM Node AS Node1,
                     Node AS Node2, Edge
WHERE Edge.src = Node1.ID AND Edge.dst = Node2.ID
                AND Node1.ID < Node2.ID
```

Suppose we set $GS_Q = 2^8 = 256$. For this particular $Q$, this means the maximum degree of any node in any instance $\mathbf{I} \in \mathcal{I}$ is 256. We set $\beta = 0.1$ and $\varepsilon = 1$.

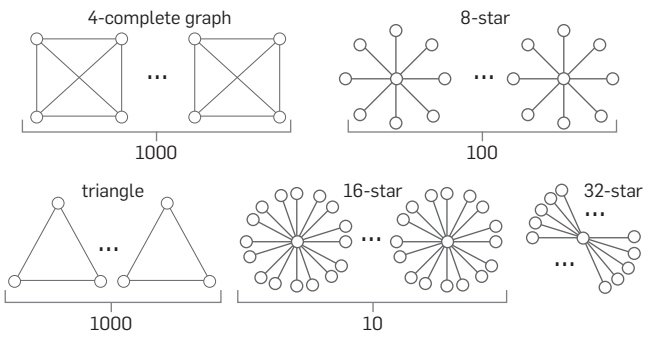
Now, suppose we are given an $\mathbf{I}$ containing 8,103 nodes, which form 1,000 triangles, 1,000 4-cliques, 100 8-stars, 10 16-stars, and one 32-star as shown in Figure 2. The true query result is

$$Q(\mathbf{I}) = 3 \times 1{,}000 + 6 \times 1{,}000 + 8 \times 100 + 16 \times 10 + 32 = 9{,}992.$$

We run R2T with $\tau^{(j)} = 2^j$ for $j = 1, \ldots, 8$. For each $\tau = \tau^{(j)}$, we assign a weight $u_k \in [0,1]$ to each join result (that is, an edge) that satisfies the predicate ID1 < ID2. To calculate $Q(\mathbf{I},\tau)$, we can consider the LP on each clique/star separately. For a triangle, the optimal LP solution always assigns $u_k = 1$ for each edge. For each 4-clique, it assigns 2/3 to each edge for $\tau = 2$ and 1 for $\tau \geq 4$. For each $k$-star, the LP optimal solution is $\min\{k,\tau\}$. Thus, the optimal LP solutions are

$$\begin{aligned}
Q(\mathbf{I},2) &= 1 \times 3{,}000 + \tfrac{2}{3} \times 6000 + 2 \times 100 \\
&\quad + 2 \times 10 + 2 \times 1 = 7{,}222, \\
Q(\mathbf{I},4) &= 1 \times 3{,}000 + 1 \times 6{,}000 + 4 \times 100 \\
&\quad + 4 \times 10 + 4 \times 1 = 9{,}444, \\
Q(\mathbf{I},8) &= 1 \times 3{,}000 + 1 \times 6{,}000 + 8 \times 100 \\
&\quad + 8 \times 10 + 8 \times 1 = 9{,}888, \\
Q(\mathbf{I},16) &= 1 \times 3{,}000 + 1 \times 6{,}000 + 8 \times 100 \\
&\quad + 16 \times 10 + 16 \times 1 = 9{,}976.
\end{aligned}$$

**Figure 2. Example of edge counting.**




In addition, we have $Q(\mathbf{I},0) = 0$ and $Q(\mathbf{I},\tau) = 9{,}992$ for $\tau \geq 32$.

Then, let's see how to run R2T with these $Q(\mathbf{I},\tau)$'s. Let $\varepsilon = 1$, $\beta = 0.1$, and $GS = 2^{10}$. Besides, for convenience, assume $Lap(1)$ returns –1 and 1 by turns. Plugging these into (7), we have

$$\begin{aligned}
\widetilde{Q}(\mathbf{I},2) &= 7{,}222 + (-1) \cdot 20 - 92.1 = 7{,}110 \\
\widetilde{Q}(\mathbf{I},4) &= 9{,}444 + 1 \cdot 40 - 184 = 9{,}300 \\
\widetilde{Q}(\mathbf{I},8) &= 9{,}888 + (-1) \cdot 80 - 368 = 9{,}440 \\
\widetilde{Q}(\mathbf{I},16) &= 9{,}976 + 1 \cdot 160 - 737 = 9{,}399 \\
\widetilde{Q}(\mathbf{I},32) &= 9{,}992 + (-1) \cdot 320 - 1{,}474 = 8{,}198 \\
\widetilde{Q}(\mathbf{I},64) &= 9{,}992 + 1 \cdot 640 - 2{,}947 = 7{,}685 \\
&\cdots
\end{aligned}$$

Finally, with (8), we have $\widetilde{Q}(\mathbf{I}) = \widetilde{Q}(\mathbf{I},8) = 9{,}440$.

## 6. TRUNCATION FOR SPJA QUERIES

A projection reduces the query answer, hence its sensitivity, so it requires less noise. However, it makes achieving instance optimality harder: Even in the simple case $|\pi_{x_2}(R_1(x_1) \bowtie R_2(x_1,x_2))|$, it is impossible to achieve the error $f(\mathbf{I}) \cdot DS_Q(\mathbf{I})$ at each instance $\mathbf{I}$ for any function $f(\mathbf{I})$. To address this issue, we propose a truncation for SPJA queries with error depending on another instance-specific notation. Please read the full-version paper for more details.

## 7. SYSTEM IMPLEMENTATION

Based on the R2T algorithm, we have implemented a system on top of PostgreSQL and CPLEX. The system structure is shown in Figure 3. The input to our system is any SPJA query written in SQL, together with a designated primary private relation $R_P$ (interestingly, while R2T satisfies the DP policy with FK constraints, the algorithm itself does not need to know the PK-FK constraints).

The system supports SUM and COUNT aggregation. Our SQL parser first unpacks the aggregation into a reporting

**Figure 3. System structure.**

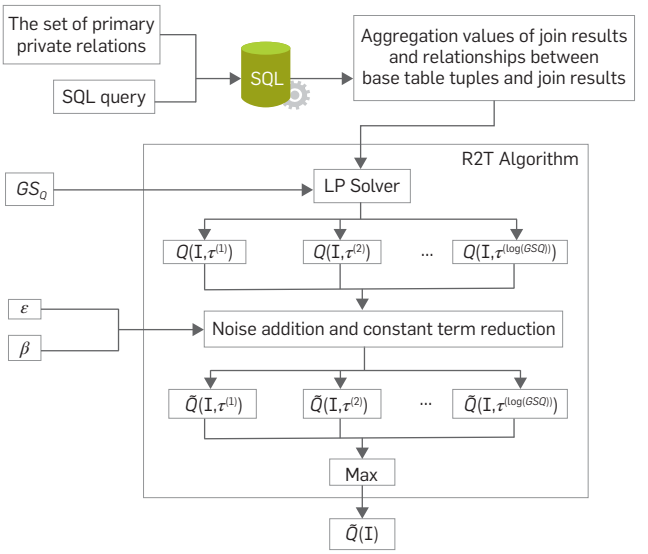
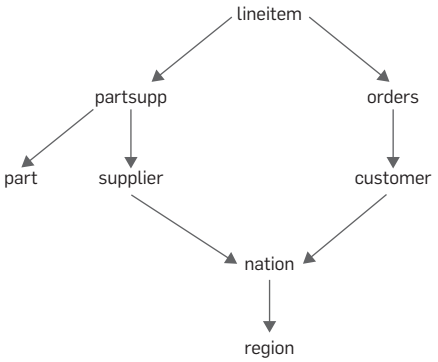


**Figure 4. The foreign-key graph of TPC-H schema.**




query so as to find $\psi(q_k(\mathbf{I}))$ for each join result, as well as $C_j(\mathbf{I})$, which stores the referencing relationships between tuples in $\mathbf{I}(R_P)$ and $J(\mathbf{I})$.

Example 7.1.
Suppose we use the TPC-H schema (shown in Figure 4), where we designate `Supplier` and `Customer` as primary private relations. Consider the following query:

```
SELECT SUM(price*(1 – discount))
FROM Supplier, Lineitem, Orders, Customer
WHERE Supplier.SK = Lineitem.SK
        AND Lineitem.OK = Orders.OK
        AND Orders.CK = Customer.CK
        AND Orders.orderdate >='2020 – 08 – 01'
```

We rewrite it as

```
SELECT Supplier.SK, Customer.CK, price*(1 – discount)
FROM Supplier, Lineitem, Orders, Customer
WHERE Supplier.SK = Lineitem.SK
        AND Lineitem.OK = Orders.OK
        AND Orders.CK = Customer.CK
        AND Orders.orderdate >='2020 – 08 – 01'
```

The `price*(1 – discount)` column in the query results gives all the $\psi(q_k(\mathbf{I}))$ values, while `Supplier.SK` and `Customer.CK` yield the referencing relationships from each supplier and customer to all the join results they contribute to.

We execute the rewritten query in PostgreSQL, and export the query results to a file. Then, an external program is invoked to construct the $\log(GS_Q)$ LPs from the query results, which are then solved by CPLEX. Finally, we use R2T to compute a privatized output.

The computation bottleneck is the $\log(GS_Q)$ LPs, each of which contains $|J(\mathbf{I})|$ variables and $|J(\mathbf{I})| + |\mathbf{I}(R_P)|$ constraints. This takes polynomial time, but can still be very expensive in practice. One immediate optimization is to solve them in parallel, and we present another effective technique to speed up the process in our full-version paper.

## 8. EXPERIMENTS

We conducted experiments on graph pattern-counting queries under node-DP, an important special case of the SPJA queries with FK constraints. Here, we compare R2T with naive truncation with smooth sensitivity (NT),[20] smooth distance estimator (SDE),[4] recursive mechanism (RM),[6] and the LP-based mechanism (LP).[20] We also implement some experiments on general SPJA queries to compare R2T with the local sensitivity-based mechanism (LS).[24] The experimental results show R2T achieves order-of-magnitude improvements over LS in terms of utility, with similar running times. This section is covered in our full-version paper.

### 8.1. Setup

For graph pattern-counting queries, we used four queries: edge counting $Q_{1-}$, length-2 path counting $Q_{2-}$, triangle counting $Q_{\triangle}$, and rectangle counting $Q_{\square}$. We used five real-world networks datasets: Deezer, Amazon1, Amazon2, RoadnetPA and RoadnetCA. Deezer collects the friendships of users from the music-streaming service Deezer. Amazon1 and Amazon2 are two Amazon co-purchasing networks. RoadnetPA and RoadnetCA are road networks of Pennsylvania and California, respectively. All these datasets are obtained from SNAP.[22] Table 1 shows the basic statistics of these datasets.

**Table 1. Graph datasets used in the experiments.**

| Dataset | Deezer | Amazon1 | Amazon2 | RoadnetPA | RoadnetCA |
|---|---|---|---|---|---|
| **Nodes** | 144,000 | 262,000 | 335,000 | 1,090,000 | 1,970,000 |
| **Edges** | 847,000 | 900,000 | 926,000 | 1,540,000 | 2,770,000 |
| **Maximum degree** | 420 | 420 | 549 | 9 | 12 |
| **Degree bound *D*** | 1,024 | 1,024 | 1,024 | 16 | 16 |

Most algorithms need to assume a $GS_Q$ in advance. Note that the value of $GS_Q$ should not depend on the instance, but may use some background knowledge for a particular class of instances. Thus, for the three social networks, we set a degree upper bound of $D = 1{,}024$, while for the two road networks, we set $D = 16$. Then we set $GS_Q$ as the maximum number of graph patterns containing any node. This means that $GS_{Q_{1-}} = D$, $GS_{Q_{2-}} = GS_{Q_{\triangle}} = D^2$, and $GS_{Q_{\square}} = D^3$.

The LP mechanism requires a truncation threshold $\tau$, but Kasiviswanathan et al.[20] does not discuss how this should be set. Initially, we used a random threshold uniformly chosen from $[1, GS_Q]$. This turned out to be very bad as with constant probability, the picked threshold is $\Omega(GS_Q)$, which makes these mechanisms as bad as the naive mechanism that adds $GS_Q$ noise. To achieve better results, as in R2T, we consider $\{2, 4, 8, \ldots, GS_Q\}$ as the possible choices. Similarly, NT and SDE need a truncation threshold $\theta$ on the degree, and we choose one from $\{2, 4, 8, \ldots, D\}$ randomly.

All experiments were conducted on a Linux server with a 24-core 2.2GHz Intel Xeon CPU and 256GB of memory. Each program was allowed to use at most 10 threads and we set a time limit of six hours for each run. Each experiment was repeated 100 times and we report the average running time. The errors are less stable due to the random noise, so we remove the best 20 and worst 20 runs, and report the average error of the remaining 60 runs. The failure probability $\beta$ in R2T is set to 0.1. The default DP parameter is $\varepsilon = 0.8$.

**Table 2. Comparison between R2T, naive truncation with smooth sensitivity (NT), smooth distance estimator (SDE), LP-based mechanism (LP), and recursive mechanism (RM) on graph pattern counting queries.**

| Dataset | | Deezer | | Amazon1 | | Amazon2 | | Roadnet – PA | | Roadnet – CA | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Result type | | Relative error(%) | Time(s) | Relative error(%) | Time(s) | Relative error(%) | Time(s) | Relative error(%) | Time(s) | Relative error(%) | Time(s) |
| $q_{1-}$ | Query result | 847,000 | 1.28 | 900,000 | 1.52 | 926,000 | 1.62 | 1,540,000 | 1.51 | 2,770,000 | 2.64 |
| | R2T | 0.535 | 12.3 | 0.557 | 15.6 | 0.432 | 16.2 | 0.0114 | 26.8 | 0.00635 | 48.7 |
| | NT | 59.1 | 18.1 | 101 | 29.3 | 125 | 40.4 | 1,370 | 21.9 | 1,410 | 39.7 |
| | SDE | 548 | 9,870 | 363 | 4,570 | 286 | 1,130 | 55.2 | 105 | 81.8 | 292 |
| | LP | 14.3 | 16.9 | 5.72 | 14.7 | 6.75 | 14.4 | 3.6 | 28.3 | 3.02 | 54 |
| $q_{2-}$ | Query result | 21,800,000 | 13.8 | 9,120,000 | 11.8 | 9,750,000 | 13.8 | 3,390,000 | 6.39 | 6,000,000 | 6.06 |
| | R2T | 6.64 | 356 | 12.2 | 170 | 9.06 | 196 | 0.0539 | 80.2 | 0.0352 | 145 |
| | NT | 116 | 21.0 | 398 | 28.4 | 390 | 41.0 | 6,160 | 23.2 | 6,530 | 44.2 |
| | SDE | 8,900 | 9,870 | 5,110 | 4,570 | 1,930 | 1,130 | 211 | 104 | 228 | 296 |
| | LP | 35.9 | 8,820 | 23.2 | 3,600 | 27.8 | 461 | 11.1 | 148 | 13.3 | 404 |
| $q_{\triangle}$ | Query result | 794,000 | 4.53 | 718,000 | 5.03 | 667,000 | 4.20 | 67,200 | 2.96 | 121,000 | 5.17 |
| | R2T | 5.58 | 17.3 | 1.27 | 18.8 | 2.03 | 19.9 | 0.102 | 4.21 | 0.061 | 7.5 |
| | NT | 782 | 23.0 | 1,660 | 31.7 | 1,920 | 41.0 | 110,000 | 23.3 | 105,000 | 45.0 |
| | SDE | 67,300 | 9,880 | 26,000 | 4,570 | 9,600 | 1,130 | 4,150 | 106 | 3,830 | 297 |
| | LP | 24.6 | 131 | 12.8 | 18.2 | 14.2 | 18.3 | 0.104 | 3.95 | 0.0625 | 7.06 |
| | RM | | | Over time limit | | | | 0.0388 | 1,280 | 0.0193 | 2,550 |
| $q_{\square}$ | Query result | 11,900,000 | 74.3 | 2,480,000 | 21.6 | 3,130,000 | 15.6 | 158,000 | 4.50 | 262,000 | 10.1 |
| | R2T | 16.9 | 289 | 6.29 | 70.5 | 10.5 | 86.8 | 0.0729 | 8.18 | 0.0638 | 16.2 |
| | NT | 3,750 | 57.6 | 30,700 | 35.8 | 26,100 | 50.6 | 319,000 | 24.8 | 368,000 | 45.0 |
| | SDE | 6,970,000 | 9,930 | 11,400,000 | 4,580 | 202,000 | 1,140 | 10,300 | 108 | 9,130 | 300 |
| | LP | 92.6 | 2,530 | 94.8 | 70.4 | 77.8 | 81.2 | 0.223 | 7.83 | 0.165 | 14.2 |
| | RM | | | Over time limit | | | | 0.0217 | 10,500 | Over time limit | |

## 8.2. Experimental results

The errors and running times of all mechanisms over the graph pattern counting queries are shown in Table 2. These results indicate a clear superiority of R2T in terms of utility, offering order-of-magnitude improvements over other methods in many cases. What is more desirable is its robustness: In all the 20 query-dataset combinations, R2T consistently achieves an error below 20%, while the error is below 10% in all but three cases. We also notice that, given a query, R2T performs better in road networks than social networks. This is because the error of R2T is proportional to $DS_Q(\mathbf{I})$ by our theoretical analysis. Thus the relative error is proportional to $DS_Q(\mathbf{I})/|Q(\mathbf{I})|$. Therefore, larger and sparser graphs, such as road networks, lead to smaller relative errors.

In terms of running time, all mechanisms are reasonable, except for RM and SDE. RM can only complete within the six-hour time limit on three cases, although it achieves very small errors on these three cases. SDE is faster than RM but runs a bit slower than others. It is also interesting to see that R2T sometimes even runs faster than LP, despite the fact that R2T needs to solve $O(\log GS_Q)$ LPs. This is due to the early stop optimization: The running time of R2T is determined by the LP that corresponds to the near-optimal $\tau$, which often happens to be one of the LPs that can be solved fastest. We also conducted experiments to see how the privacy parameter $\varepsilon$ affects various mechanisms in our full version of paper. The result shows R2T has a high utility even for a small $\varepsilon$.

**Selection of $\tau$.** In the next set of experiments, we dive deeper and see how sensitive the utility is with respect to the truncation threshold $\tau$. We tested the queries on Amazon2 and measured the error of the LP-based mechanism[20] with different $\tau$. For each query, we tried various $\tau$ from 2 to $GS_Q$ and compare their errors with R2T. The results are shown in Table 3, where the optimal error is marked in gray. The results indicate that the error is highly sensitive to $\tau$, and more importantly, the optimal choice of $\tau$ closely depends on the query, and there is no fixed $\tau$ that works for all cases. On the other hand, the error of R2T is within a small constant factor (around 6) to the optimal choice of $\tau$, which is exactly the value of instance-optimality.

## 9. MORE DISCUSSIONS

Following this work, there have been many efforts put into query evaluation in relational databases under DP. For instance, Dong and Yi[14] and Dong et al.[8] improve the logarithmic factor in the error for self-join-free queries and self-join queries, while Fang et al.[16] explores answering SPJA queries with `Max` aggregation. In addition, Cai et al.[5] and Dong et al.[10] focus on answering multiple queries, while Dong et al.[7,9] investigate SPJA queries over dynamic databases. For more details, please refer to this recent survey.[13] Moreover, by integrating this work with Dong et al.[10] and Fang et al.,[16] we have developed a DP SQL system[25] capable of answering a broad class of queries that include selection, projection, aggregation, join, and group by operations.

**Table 3. Error levels of R2T and LP-based mechanism (LP) with different $\tau$.**

| Query | | $Q_{1-}$ | $Q_{2-}$ | $Q_{\triangle}$ | $Q_{\square}$ |
|---|---|---|---|---|---|
| Query result | | 926,000 | 9,750,000 | 667,000 | 3,130,000 |
| R2T | | 4,000 | 883,000 | 13,500 | 328,000 |
| LP | $\tau = GS_Q$ | 1,440 | 1,580,000 | 1,290,000 | 1,370,000,000 |
| | $\tau = GS_Q/8$ | 2,100 | 181,000 | 157,000 | 140,000,000 |
| | $\tau = GS_Q/64$ | 110,000 | 259,000 | 15,100 | 25,800,000 |
| | $\tau = GS_Q/512$ | 645,000 | 1,260,000 | 2,790 | 2,630,000 |
| | $\tau = GS_Q/4,096$ | 810,000 | 3,950,000 | 2,090 | 274,000 |
| | $\tau = GS_Q/32,768$ | 911,000 | 7,580,000 | 92,300 | 48,700 |
| | $\tau = GS_Q/262,144$ | 924,000 | 9,340,000 | 459,000 | 76,400 |
| | Average error | 62,500 | 2,710,000 | 94,900 | 2,430,000 |

## Acknowledgment

This work has been supported by HKRGC under grants 16201318, 16201819, and 16205420; by NTU-NAP startup grant 024584-00001; by the National Science Foundation under grant 2016393; and by DARPA and SPAWAR under contract N66001-15-C-4067.

## References

1. Abadi, M. et al. Deep learning with differential privacy. In *Proceedings of the 2016 ACM Conf. on Computer and Communications Security*, 308–318.
2. Amin, K., Kulesza, A., Munoz, A., and Vassilvtiskii, S. Bounding user contributions: A bias-variance trade-off in differential privacy. In *Intern. Conf. on Machine Learning*. PMLR (2019), 263–271.
3. Asi, H. and Duchi, J.C. Instance-optimality in differential privacy via approximate inverse sensitivity mechanisms. *Advances in Neural Information Processing Systems 33* (2020).
4. Blocki, J., Blum, A., Datta, A., and Sheffet, O. Differentially private data analysis of social networks via restricted sensitivity. In *Proceedings of the 4th Conf. on Innovations in Theoretical Computer Science* (2013), 87–96.
5. Cai, K., Xiao, X., and Cormode, G. Privlava: synthesizing relational data with foreign keys under differential privacy. In *Proceedings of the ACM on Management of Data 1*, 2 (2023), 1–25.
6. Chen, S. and Zhou, S. Recursive mechanism: towards node differential privacy and unrestricted joins. In *Proceedings of the 2013 ACM SIGMOD Intern. Conf. on Management of Data*, 653–664.
7. Dong, W. et al. Continual observation of joins under differential privacy. *Proceedings of the ACM on Management of Data 2*, 3 (2024), 1–27.
8. Dong, W. et al. Instance-optimal truncation for differentially private query evaluation with foreign keys. *ACM Transactions on Database Systems 49*, 4 (2024), 1–40.
9. Dong, W., Luo, Q., and Yi, K. Continual observation under user-level differential privacy. In *Proc. 2023 IEEE Symp. on Security and Privacy*, 2190–2207.
10. Dong, W., Sun, D., and Yi, K. Better than composition: How to answer multiple relational queries under differential privacy. In *Proceedings of the ACM on Management of Data 1*, 2 (2023), 1–26.
11. Dong, W. and Yi, K. Residual sensitivity for differentially private multi-way joins. In *Proc. ACM SIGMOD Intern. Conf. on Management of Data*, (2021).
12. Dong, W. and Yi, K. A nearly instance-optimal differentially private mechanism for conjunctive queries. In *Proc. ACM Symp. on Principles of Database Systems*, (2022).
13. Dong, W. and Yi, K. Query evaluation under differential privacy. *ACM SIGMOD Record 52*, 3 (2023), 6–17.
14. Dong, W. and Yi, K. Universal private estimators. In *Proceedings of the 42nd ACM SIGMOD-SIGACT-SIGAI Symp. on Principles of Database Systems* (2023), 195–206.
15. Dwork, C. and Roth, A. The algorithmic foundations of differential privacy. *Foundations and Trends® in Theoretical Computer Science 9*, 3–4 (2014), 211–407.
16. Fang, J., Dong, W., and Yi, K. Shifted inverse: A general mechanism for monotonic functions under user differential privacy, (2022).
17. Huang, Z., Liang, Y., and Yi, K. Instance-optimal mean estimation under differential privacy. In *NeurIPS*, (2021).
18. Johnson, N., Near, J. P., and Song, D. Towards practical differential privacy for sql queries. In *Proceedings of the VLDB Endowment 11*, 5 (2018), 526–539.
19. Karwa, V. and Vadhan, S. Finite sample differentially private confidence intervals. In *Proceedings of the 9th Innovations in Theoretical Computer Science Conf*. Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik, 2018.
20. Kasiviswanathan, S. P., Nissim, K., Raskhodnikova, S., and Smith, A. Analyzing graphs with node differential privacy. In *Theory of Cryptography Conf*. Springer (2013), 457–476.
21. Kotsogiannis, I. et al. PrivateSQL: A differentially private SQL query engine. In *Proceedings of the VLDB Endowment 12*, 11 (2019), 1371–1384.
22. Leskovec, J. and Krevl, A. Snap datasets: Stanford large network dataset collection (2014), 49; https://tinyurl.com/22cypyg3
23. Nissim, K., Raskhodnikova, S., and Smith, A. Smooth sensitivity and sampling in private data analysis. In *Proceedings of the 39th Ann. ACM Symp. on Theory of Computing* (2007), 75–84.
24. Tao, Y., He, X., Machanavajjhala, A., and Roy, S. Computing local sensitivities of counting queries with joins. In *Proc. 2020 ACM SIGMOD Intern. Conf. on Management of Data*, 479–494.
25. Vadhan, S. The complexity of differential privacy. In *Tutorials on the Foundations of Cryptography*. Springer (2017), 347–450.
26. Yu, J. et al. Dop-sql: A general-purpose, high-utility, and extensible private SQL system. In *Proc. Intern. Conf. on Very Large Data Bases*, (2024).

**Wei Dong** is a an assistant professor at Nanyang Technological University, Singapore.

**Juanru Fang** is a Ph.D. candidate at the Hong Kong University of Science and Technology, China.

**Ke Yi** is a professor at the Hong Kong University of Science and Technology, China.

**Yuchao Tao** is a privacy engineer at Snap, Inc., Los Angeles, CA, and was in the Ph.D. program at Duke University, Durham, NC, when this article was written.

**Ashwin Machanavajjhala** is an associate professor at Duke University, Durham, NC, USA.

# ACM Distinguished Speakers

*Talks by and with technology leaders and innovators*

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

ACM provides ACM Chapters, colleges and universities, corporations, event and conference planners, and agencies with direct access to top technology leaders and innovators from every sector of the computing industry.

The ACM Distinguished Speakers Program (DSP) features renowned thought leaders in industry, academia, and government speaking about the most important topics in computing and IT today.

### ACM Local Chapters

Boost attendance at your meetings with live talks by DSP speakers and keep your chapter members informed of the latest industry findings.

### Colleges and Universities

Expand the knowledge base of your students with exciting lectures and the chance to engage with a computing professional in their desired field of expertise.

### Corporations

Educate your technical staff, ramp up the knowledge of your team, and give your employees the opportunity to have their questions answered by experts in their field.

### Event and Conference Planners

Use the ACM DSP to help find compelling speakers for your next conference.

Book the speaker for your next virtual or in-person event through the ACM DSP program and deliver compelling and insightful content to your audience. ACM will cover the cost of transportation for the speaker to travel to your in-person event.

**speakers.acm.org**

Association for Computing Machinery

*Advancing Computing as a Science & Profession*

If you have questions, please send them to **acmdsp@acm.org**.

[CONTINUED FROM P. 104] is in the training set or close to something in the training set, these systems work pretty well. But if it's far enough away from the training set, they break down.

In philosophy, they make a distinction between intention and extension. The intention of something is basically the abstract meaning, like "even number." The extension is a list of all the even numbers. And neural networks basically work at the extensional level, but they don't work at the intentional level. They are not getting the abstract meaning of anything.

**You've called attention to one way this distinction manifests in river-crossing problems, where generative AI systems propose solutions that resemble the right answer, but with absurdly illogical twists or random elements that were not present in the original question.**

These models don't really have a representation of what a man is, what a woman is, or what a boat is; as a result, they often make really boneheaded mistakes. And there are other consequences, like the fact that you can't give them an instruction and expect them to reliably follow it. You can't say, "Don't lie," or "don't hallucinate," or "don't use copyrighted materials." These systems are trained on copyrighted materials—they won't be able to judge. You can't do basic fact-checking. You also can't follow principles like, "Don't discriminate on the basis of race or age or sex," because if LLMs are trained on real-world data, they tend to perpetuate past stereotypes rather than following abstract principles.

So you wind up with all of these technical problems, many of which spill over into the moral and ethical domain.

**You've argued that to fix the moral and technical problems with AI, we need a new approach, not just more training data.**

Generative AI only works for certain things. It works for pattern recognition, but it doesn't work for the type of formal reasoning you need in chess. It doesn't work for everyday formal reasoning about the world, and it doesn't even reliably generate accurate summaries.

**"Generative AI only works for certain things. It works for pattern recognition, but it doesn't work for the type of formal reasoning you need in chess."**

If you think about it abstractly, there's a huge number of possible AI models, and we're stuck in one corner. So one of my proposals is that we should consider integrating neural networks with classical AI. I make an analogy in my book to Daniel Kahneman's System One and System Two. System One is fast, reflexive, and automatic—kind of like LLMs—while System Two is more deliberative reasoning, like classical AI. Our human mind combines both and gets results that are not perfect, but that are much better, in many dimensions, than current AI, so I think exploring that would be really a good idea. It won't be sufficient for developing systems that can observe something and build a structured set of representations about how that thing works, but it might get us part of the way there.

**At the time of this interview, several people in the field seem to agree that we're hitting a period of diminishing returns with respect to LLMs.**

That is a phrase that I coined in a 2022 essay called "Deep Learning is Hitting a Wall," which was about why scaling wouldn't get us to AGI (artificial general intelligence). And when I coined it, everybody dismissed me and said, "No, we're not reaching diminishing returns. We have these scaling laws. We'll just get more data." But what people have observed in the last couple of months is that adding more data does not actually solve the core underlying problems on the technical side. The big companies that are doing big training runs are not getting the results they expected.

**Do you think that will be enough to change the atmosphere and shift the industry's focus?**

I hope that the atmosphere will change. In fact, I know it will change, I just don't know when. A lot of this is crowd psychology. DeepMind does hybrid AI. AlphaFold is a neurosymbolic system, and it just won the Nobel Prize. So there are some efforts, but for the time being, venture capitalists only want to invest in LLMs. There's no oxygen left for anything else.

That said, different things could happen, maybe even by the time we go to print. The market might crash. If you can't eliminate hallucinations, it limits your commercial potential. I think people are starting to see that, and if enough of them do, then it's just a psychology thing. Maybe someone will come up with a new and better idea. At some point they will. It could come tomorrow or it might take a decade or more.

**People have proposed a number of different benchmarks for evaluating progress in AI. What do you make of them?**

Here's a benchmark I proposed in 2014 that I think is still beyond current AI. I call it the comprehension challenge. The idea is that an AI system should be able to watch a movie, build a cognitive model of what is going on, and answer questions. Why did the characters do this? Why is that line funny? What's the irony in this scene?

Right now, LLMs might get it sort of right some of the time, but nowhere near as reliably as the average person. If a character says at the end of the movie, "I see dead people," everybody in the cinema has this "Oh, my god" moment. Everybody in the cinema has followed the world of the movie and suddenly realized that a principle they thought was true does not apply. When we have AI that can do that with new movies that are not in the training data, I'll be genuinely impressed. C

**Leah Hoffmann** is a technology writer located in Piermont, NY.

DOI:10.1145/3707200

Leah Hoffmann

# Q&A
# Not on the Best Path

*Gary Marcus discusses, among other things, why he thinks large language models have entered a "period of diminishing returns."*

IN AN AGE of breathless predictions and sky-high valuations, cognitive scientist Gary Marcus has emerged as one of the best-known skeptics of generative artificial intelligence (AI). In fact, he recently wrote a book about his concerns, Taming Silicon Valley, in which he made the case that "we are not on the best path right now, either technically or morally." Marcus—who has spent his career examining both natural and artificial intelligence—explained his reasoning in a recent conversation with Leah Hoffmann.

**You've written about neural networks in everything from your 1992 monograph on language acquisition[a] to, most recently, your book *Taming Silicon Valley*.[b] Your thoughts about how AI companies and policies fall short have been well covered in your U.S. Senate testimony (https://bit.ly/3CcLjps) and other outlets (including your own Substack). Let's talk here about your technical criticisms.**

Technically speaking, neural networks, as they are usually used, are function approximators, and large language models (LLMs) are basically approximating the function of how humans use language. And they're extremely good at that. But approximating a function is not the same thing as learning a function.

In 1998, I pointed out several examples of what people now call the problem of distribution shift. For instance, I trained the one-hidden-layer neural networks that were popular at the time the identity function, f(x)=X, on even numbers represented as binary digits, and I showed that these systems could generalize to some new even numbers. But if I tested them on odd numbers, they would systematically fail. So I made, roughly, a distinction between interpolation and extrapolation, and I concluded that these tools are good at interpolating functions, but they're not very good at extrapolating functions.

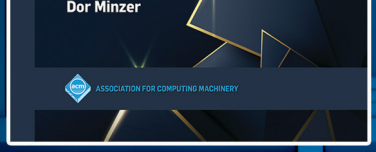
> **"I concluded that these tools are good at interpolating functions, but they're not very good at extrapolating functions."**

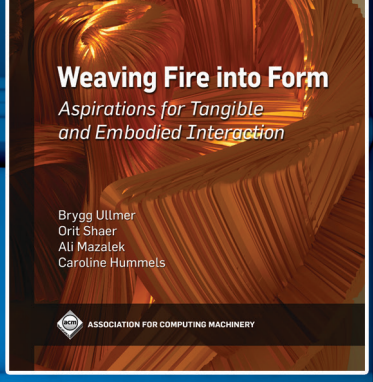
**And in your view, the multilayer neural networks we have now still do not address that issue.**

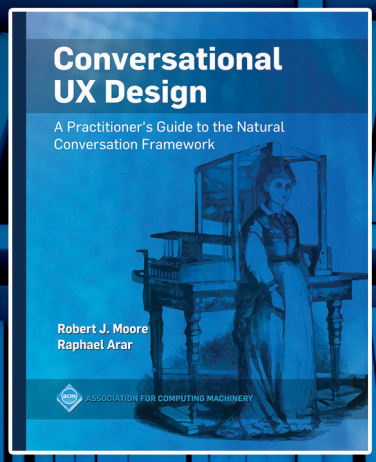
In fact, there was a paper published in October[c] by six Apple researchers basically showing the same thing. If something [CONTINUED ON P. 103]

a https://bit.ly/4gqoUTP
b https://bit.ly/3CSutMO
c https://bit.ly/40Rd9Qv

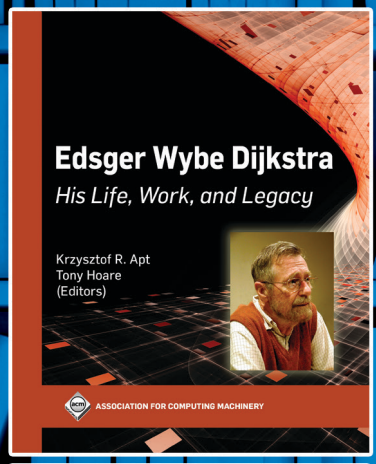
PHOTO FROM GARYMARCUS.COM

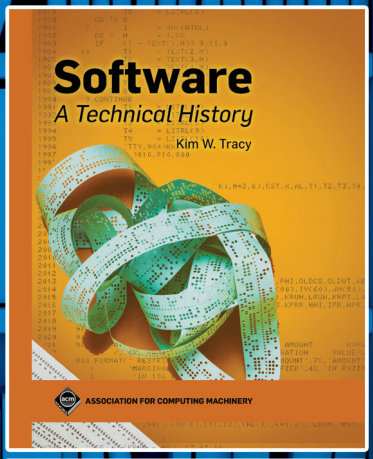


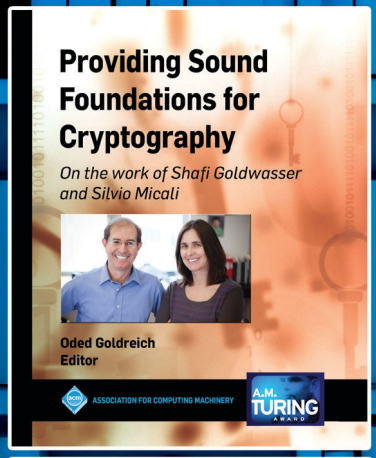



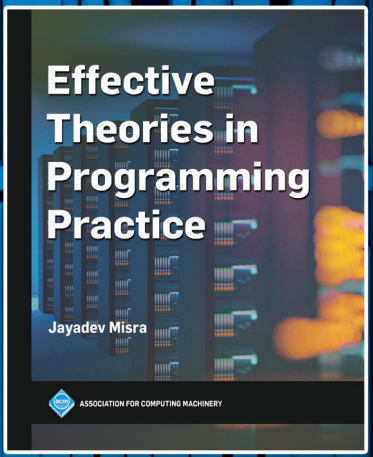



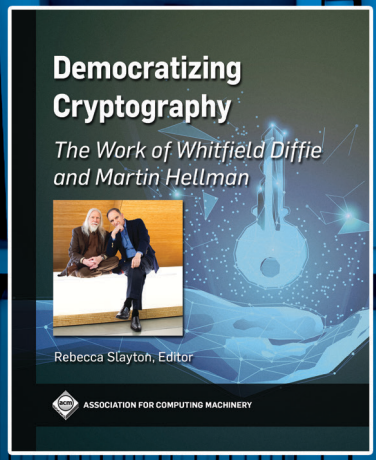


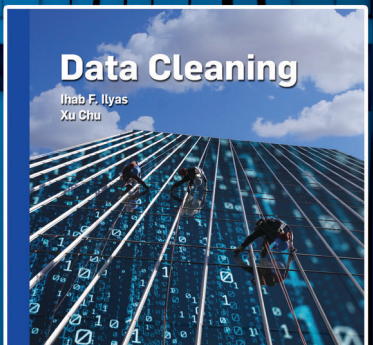



# In-Depth. Innovative. Insightful.

**Inspired by the need for high-quality computer science publishing at the graduate, faculty, and professional levels, ACM Books are affordable, current, and comprehensive in scope.**

**Collections I & II complete.**

**Collection III now publishing!**

**ACM BOOKS**

For more information, please go to:
**http://books.acm.org**

1601 Broadway, 10th Floor
New York, NY 10019, USA
212-626-0658
acmbooks-info@acm.org