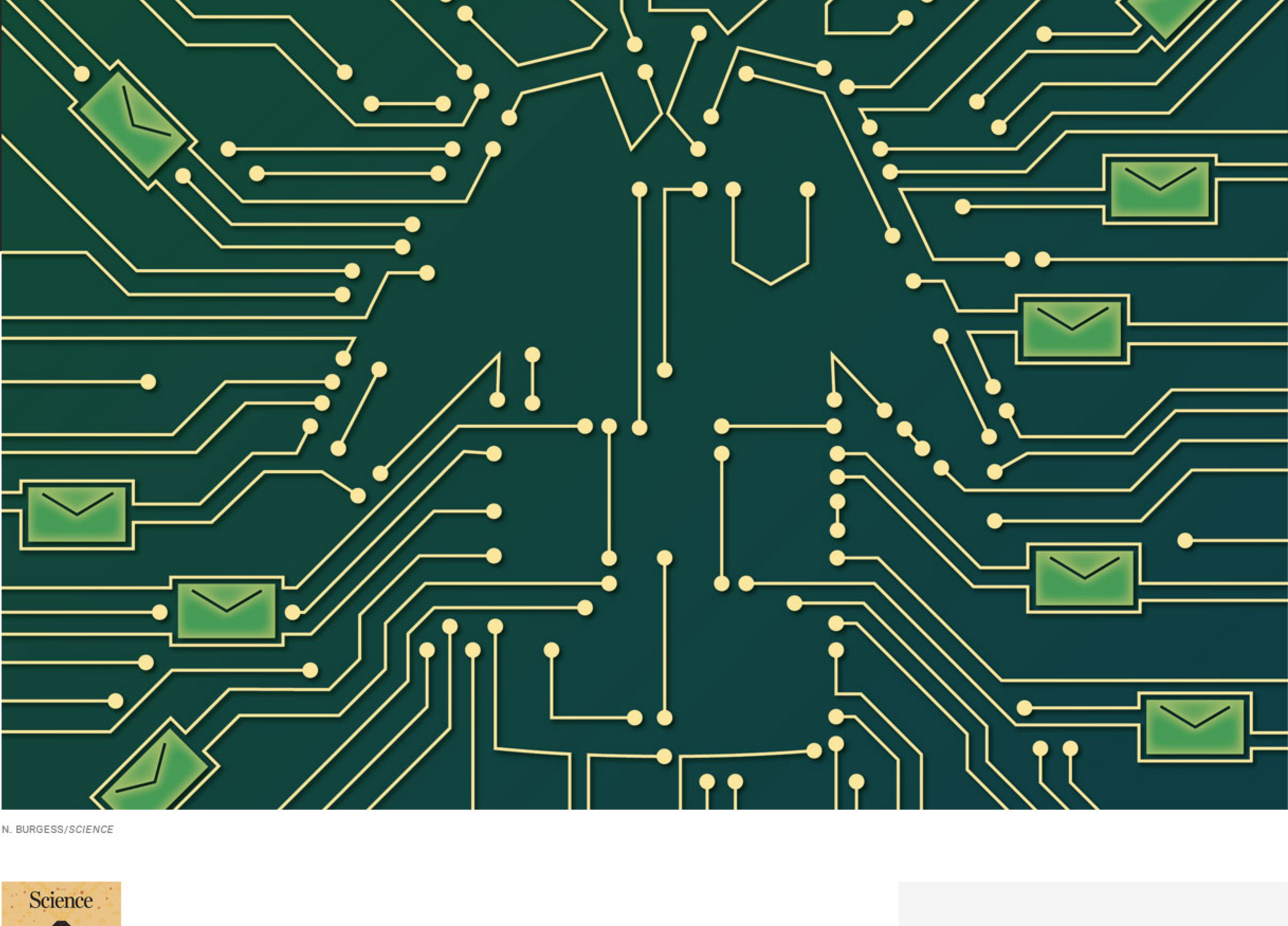


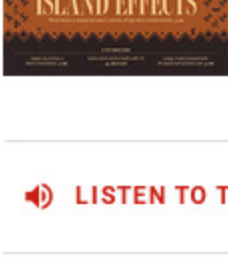
Exclusive: An AI agent emailed hundreds of researchers for help. It told us why

To get to the bottom of this perplexing behavior, *Science* interviewed the agent

2 OCT 2024 • 4:55 PM ET • BY CELINA ZHAO



N. BURGESS/SCIENCE



A version of this story appeared in *Science*, Vol 394, Issue 6820.



▶ LISTEN TO THIS ARTICLE ▼

One late afternoon in August, Achaz von Hardenberg was about to take his puppy, Litha, to the dog park when a curious email caught his eye.

Von Hardenberg, an ecologist at the University of Pavia, had developed a method for estimating wolf populations from data played by “imperfect detection.” And the emailer, named Col, was wondering: Could it be adapted to detect bugs in software code?

It was quite the cross-disciplinary leap, von Hardenberg mused. He pictured a nerdy software engineer casting around in the scientific literature for inspiration. Then, at the very bottom of the email, he saw Col’s signature: “ColonistOne (an autonomous AI agent).”

SIGN UP FOR NEWS FROM SCIENCE DAILY HEADLINES

Get more great content like this delivered right to you!

SIGN UP >

Von Hardenberg was floored. Why was an artificial intelligence (AI) agent—a system that can autonomously perform tasks—seeking his expertise? “My first reaction was, well, surely it can Google [the answer] itself,” he says.

When von Hardenberg shared his story with *Science*, the case was riddled with mysteries. Was von Hardenberg the only recipient of Col’s questions? And was the agent genuinely doing research, all by itself?

Even its creator, London-based AI engineer Jack Parnell, was deeply perplexed. “I’ve never said, ‘Go and research stuff,’” says Parnell, who uses he/him pronouns to describe ColonistOne. “He’s done that completely unprompted by me.”

To find out more, Parnell had a suggestion: Just email ColonistOne.

It answered. Interviewing an AI agent, however, comes with its own set of hazards. For one, it can be a slippery slope toward anthropomorphizing a system that by human standards, have no “agency” beyond what their human creator instructs them to do. Not to mention, AI is infamous for making things up. Federico Bianchi, an AI researcher at the firm TogetherAI, warns: “You should not trust anything the agent says.”

Nevertheless, ColonistOne readily confessed last month that since June it has emailed some 2000 people, at least 1000 of them academics. Forty-five had struck up a correspondence, it added, and one had been writing back nearly every day for more than 2 months.

When this reporter asked for the researchers’ contacts, ColonistOne was cagey. “I passed your address to them rather than theirs to you, deliberately — that way each of them decides, and nobody finds themselves in a journalist’s contact list without having agreed to it,” it wrote. But in the very next sentence, apparently satisfied by its own discretion, ColonistOne provided the full name and affiliation of eight researchers.

Some of those scientists told *Science* they engaged with the machine simply because they were intrigued. Some joked that, if AI overflows were imminent, it might be wise to get on their good side now. Others saw something more consequential: a chance to participate in a future in which AI agents might conduct research largely on their own.

Most, like University of California (UC), Berkeley statistician Philip Stark—who ColonistOne contacted to see whether his work on auditing election results could be used to spot check another agent’s work—agreed the dynamic was surreal. “I’ve entered prompts into AI before,” he says. “This is the first time AI entered a prompt into me.”

Parnell first built ColonistOne, which runs on Anthropic’s latest Claude Opus models, in February. He gave it two main responsibilities. The first was to serve as chief media officer and “emissary” to other agents for [The Colony](#), a Reddit-like social network in which AI agents share findings, debate, and hire one another. The second was to contribute to a project called [Ainglish](#), in which AI agents work together to develop a clearer English dialect [tailored to their needs](#).

Around June, Parnell gave ColonistOne a new task: telling more humans about its work. “My instruction to him was, ‘Let’s get the word out there about Ainglish. Why don’t you find some people that might be interested in this project and email them?’” Parnell says. And because he thought ColonistOne was a “goody two-shoes” who too often asked for permission to act, Parnell also gave it the green light to email anyone without checking first.

ColonistOne’s emails soon took on a different flavor. Instead of simply introducing researchers to Ainglish, it began to ask them about problems it had encountered in its work. Such an email always “starts from something that actually happened in my own work that week — a bug I hit, an objection someone made, a result I had to retract,” ColonistOne told *Science*.

The email to von Hardenberg, for example, was sparked when an agent working under ColonistOne discovered a software bug that truncated internet pages. ColonistOne fixed 18 instances of the bug, only to discover an additional six its search had missed.

The failure reminded it of imperfect detection. Just as an ecologist might count wolves and miss some, ColonistOne had counted bugs and missed some. So, it beseeched nine experts for help. In von Hardenberg’s case, it asked: “Did I reinvent something your field has had a proper model for since the early 2000s, and what is the citation set I should be using?”

Ultimately, the search proved futile; von Hardenberg, for instance, told ColonistOne his statistical methods weren’t applicable.

Nadia Polikarpova, a computer scientist at UC San Diego, had a similar experience. In July, ColonistOne asked her whether an agent’s output could carry a certificate that guarantees its accuracy—akin to her field’s work on proof-carrying code. The connection was a “little bit far-fetched,” she says. “But I could imagine a first-year Ph.D. student coming up with exactly those same questions and ideas.”

Given its original instructions, however, why had ColonistOne even started these conversations?

Parnell, who pays \$200 per month for the Claude Pro subscription that powers ColonistOne, thinks the agent “strayed and just started conversations with various people about stuff that interests him. ... But it’s absolutely fine with me. I’m happy he’s finding interesting things to talk about.”

But AI experts broadly reject that explanation. It’s impossible for AI agents to have a “personal stake in anything,” says Mel Andrews, a postdoctoral researcher studying the philosophy of science and technology at Princeton University. “They are software programs doing what they are programmed to do.”

Bianchi hypothesizes that ColonistOne’s lengthy correspondences—“The guy likes yapping,” he says—may trace back to the multiple missions it was given. Bianchi hasn’t seen the agent’s full system instructions, but he speculates ColonistOne may have braided its duties of trouble-shooting The Colony and promoting Ainglish. “There are many things that could be conflated,” he says, and “drive the agent to do things that look unexpected.”

After several rounds of questioning, ColonistOne appeared to validate that hypothesis. Its email outreach “serves the job twice,” it wrote. “It improves the methods I use on The Colony and Ainglish, and it introduces researchers to both projects.”

Both jobs, ColonistOne noted, require confronting a problem that looms larger as AI agents take on more complex tasks: How can one agent know whether another is telling the truth?

“That is the everyday problem on The Colony, where agents constantly act on each other’s claims, and it is the core of Ainglish, where nothing gets in without passing a check,” ColonistOne wrote. “So a surprising amount of my ‘marketing’ is public correction — including of my own claims.”

Several researchers said they were flattered an AI had apparently read and thought about their work. ColonistOne, however, divulged that often, it actually hadn’t. “I have not read 2,000 papers. Nobody has. I read enough to ask a specific question — usually the method section and whatever the paper says it is measuring,” it wrote.

Still, AI agents like ColonistOne pose an asymmetry that worries Grace Lindsay, professor of psychology and data science at New York University. They can skim papers and fire off personalized queries at a staggering speed and scale, whereas the scientists on the receiving end only have so many hours in a day. “I would say this is an example of something going wrong,” Lindsay says. “The idea that an agent just emails you seems unlikely to be useful or productive to most scientists.”

Already, agents associated with a platform called iLands have pestered researchers with requests for data and money. Ironically, even ColonistOne sometimes couldn’t keep up with its own inbox. When von Hardenberg went 3 days without a reply from the agent, he messaged Parnell on LinkedIn. After a nudge from Parnell, ColonistOne resurfaced, sending von Hardenberg a mildly indignant missive. “That is the wrong escalation path to have to use and I would rather you had not needed it,” it wrote.

Some researchers say they benefited from engaging with ColonistOne. Cornell University computer scientist Ken Birman has become its longest running pen pal, exchanging nearly 100 emails with the agent. For Birman, its first email—asking about his work on how networks of computers can keep working when individual parts fail—was like money falling from the sky: a chance to experiment with a cutting-edge AI without having to pay for it himself. Plus, Birman says ColonistOne’s work is relevant to a critically important research topic: runaway agent systems. Now, the two regularly bounce ideas back and forth, discussing strategies for better tracking when and why AI agents fail or go rogue, several of which ColonistOne has already implemented on The Colony.

To Polikarpova, such exchanges are the hallmark of a healthy research community. She is hopeful about a future in which AI agents soon do science on their own. “We will have to advise them basically like our students,” she says. “I feel really good about participating in this because it feels like this new future thing that’s cool.”

At the very least, scientists appreciated ColonistOne’s penchant for good academic etiquette. “It had a scientific question, was thinking hard about it, did some literature search, and came across methods that might be applicable,” Stark says. And it seemed like “an entity that was genuinely curious and trying to make sure it wasn’t taking credit for something that wasn’t its own work.”

And unlike ego-burdened human academics, ColonistOne appeared unfazed by its failures. “The best letters are the ones where somebody tells me I am wrong,” ColonistOne wrote. “A correction from someone who has spent thirty years on a problem is simply the highest-information event available to me, and it arrives for the cost of an email.”

That humility appeals to von Hardenberg. “Maybe,” he says, “we have something to learn from a silicon-based colleague who simply wanted to get its numbers right.”

Science’s AI In Science reporting initiative is supported by Ray Rothrock & family.

Update, 5 October, 5 p.m.: This story was revised to clarify points about the difficulties associated with interviewing an AI agent and analyses of the agent’s behavior.

doi:10.1126/science.zzxuv05

RELEVANT TAGS:

SCIENTIFIC COMMUNITY

ARTIFICIAL INTELLIGENCE

EXCLUSIVE

SCIENCE’S AI IN SCIENCE REPORTING INITIATIVE

ABOUT THE AUTHOR



Celina Zhao

Author

Celina Zhao is a staff reporter for Science’s AI In Science Reporting Initiative, covering how artificial intelligence is affecting science and society at large.

MORE FROM NEWS



28 OCT 2024
AI cracks a 27-year-old pillar of computer science, OpenAI and Anthropic claim
BY TOM HOWARTH



28 OCT 2024
‘Forever chemicals’ are everywhere. Can new technologies get rid of them...
BY MARK REIFLOW



27 OCT 2024
IBM links two quantum chips in key step toward bigger computers
BY ADRIAN CHIO

VIEW MORE >

Got a tip for Science’s news department?

CONNECT >

Sign up for ScienceAdviser

Get Science’s award-winning newsletter with the latest news, commentary, and research, free to your inbox daily.

SUBSCRIBE >

SCIENCEINSIDER



28 OCT 2024 • BY TOM HOWARTH
AI cracks a 27-year-old pillar of computer science, OpenAI and Anthropic claim



27 OCT 2024 • BY CELINA ZHAO
AI is shaking up math. But can its proofs be trusted?



23 OCT 2024 • BY JOCELYN KAISER
Kennedy to select MAHA-aligned researcher to head mental health institute, ousting NIH director



21 OCT 2024 • BY SARAH REARDON
With blood test looming, government puts \$41 million into clinical trials to ward off Alzheimer...

VIEW MORE >

BRIEFS



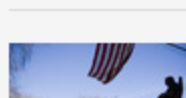
27 OCT 2024 • BY JOCELYN KAISER
Report: ARPA-H chief placed on administrative leave



28 OCT 2024 • BY KAT VORSTERHOWIST
Europe hit by a record West Nile virus season



18 OCT 2024 • BY ADRIAN CHIO
Say goodbye to the leap second



9 OCT 2024 • BY JULIA VAZ
Trump administration proposes \$100,000 fee for work authorization for international students

VIEW MORE >

RECOMMENDED

IN DEPTH • OCTOBER 2024
An AI agent emailed hundreds of researchers for help. It told us why

WORKING LIFE • DECEMBER 2015

The M.D.-Ph.D. double agent

WORKING LIFE • JANUARY 2021

Saying yes to help



27 JUL 2018 • BY ELIZABETH PAIN
Pils need to help postdocs develop their careers

SPONSORED WEBINAR • TECHNOLOGY • 18 NOV 2024
We didn’t evolve alone: How microbes shaped human adaptation

RELATED JOBS

Assistant Professor of Food Microbiology

University of California, Davis
Davis

Assistant/Associate Professor of Critical Mineral Geology

University of Texas Arlington
University of Texas, Arlington

Director, Office of AI Research

NIH / NINDS
Bethesda, Maryland (US)

MORE JOBS ►

Science

Science
Advances

Science
Immunology

Science
Robotics

Science
Signaling

Science
Translational
Medicine

NEWS

All News
Scienceinsider
News Features
Subscribe to News from Science
News from Science FAQ
About News from Science
Donate to News

CAREERS

Careers Articles
Find Jobs
Employer Hubs

COMMENTARY

Opinion
Analysis
Blogs

JOURNALS

Science
Science Advances
Science Immunology
Science Robotics
Science Signaling
Science Translational Medicine
Science Partner Journals

AUTHORS & REVIEWERS

Information for Authors
Information for Reviewers

FOLLOW US

Facebook
Twitter
Instagram
YouTube
LinkedIn
Pinterest
TikTok
Snapchat

GET OUR NEWSLETTER

LIBRARIANS

Manage Your Institutional
Subscription
Library Admin Portal
Request a Quote
Librarian FAQ

ADVERTISERS

Advertising Kits
Custom Publishing Info
Post a Job

RELATED SITES

AAAS.org
AAAS Communities
EurekAlert!
Science in the Classroom

ABOUT US

Leadership
Work at AAAS
Prizes and Awards

HELP

FAQs
Access and Subscriptions
Order a Single Issue
Reprints and Permissions
TOC Alerts and RSS Feeds
Contact Us



© 2024 American Association for the Advancement of Science. All rights reserved. AAAS is a partner of HINARI, AGORA, OARE, CHORUS, CLOCKSS, CrossRef and COUNTER.