ANNALS OF TECHNOLOGY

WHO SHOULD STOP UNETHICAL A.I.?

At artificial-intelligence conferences, researchers are increasingly alarmed by what they see.

By Matthew Hutson

February 15, 2021

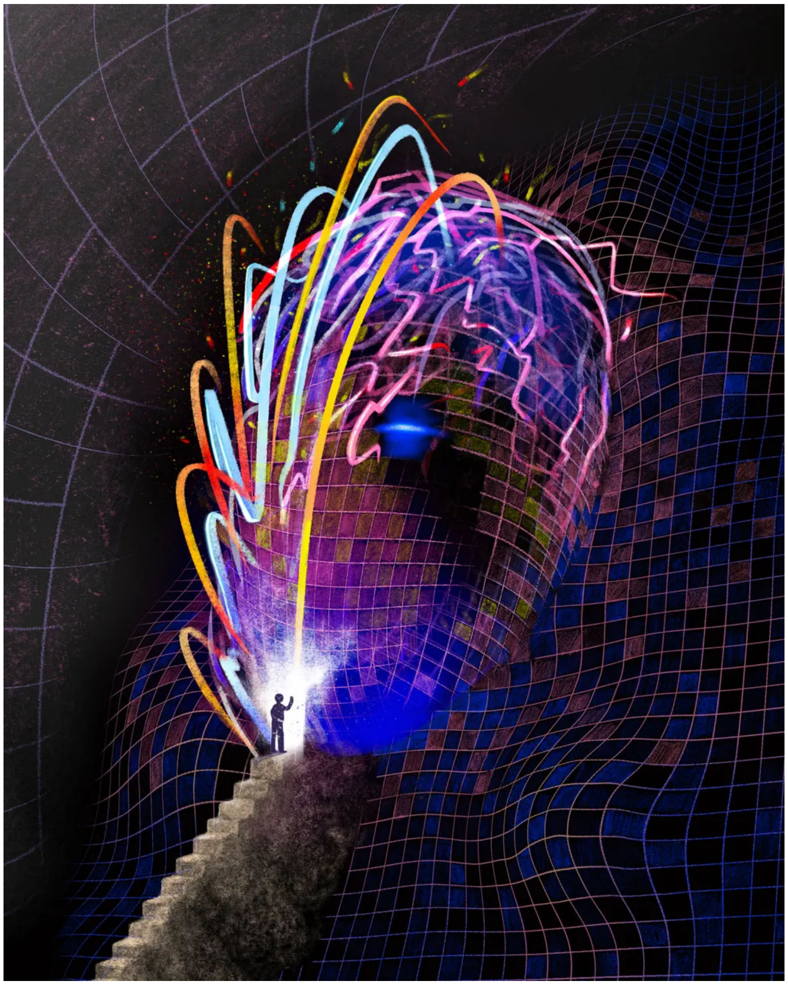

Illustration by Jeremy Leung

In computer science, the main outlets for peer-reviewed research are not journals but conferences, where accepted papers are presented in the form of talks or posters. In June, 2019, at a large artificial-intelligence conference in Long Beach, California, called Computer Vision and Pattern Recognition, I stopped to look at a poster for a project called Speech2Face. Using machine learning, researchers had developed an algorithm that generated images of faces from recordings of speech. A neat idea, I thought, but one with unimpressive results: at best, the faces matched the speakers' sex, age, and ethnicity—attributes that a casual listener might guess. That December, I saw a similar poster at another large A.I. conference, Neural Information Processing Systems (NeurIPS), in Vancouver, Canada. I didn't pay it much mind, either.

Not long after, though, the research blew up on Twitter. "What is this hot garbage, #NeurIPS2019?" Alex Hanna, a trans woman and sociologist at Google who studies A.I. ethics, tweeted. "Computer scientists and machine learning people, please stop this awful transphobic shit." Hanna objected to the way the research sought to tie identity to biology; a sprawling debate ensued. Some tweeters suggested that there could be useful applications for the software, such as helping to identify criminals. Others argued, incorrectly, that a voice revealed nothing about its speaker's appearance. Some made jokes ("One fact that this should never have been approved: Rick Astley. There's no way in hell that their [system] would have predicted his voice out of that head at the time") or questioned whether the term "transphobic" was a fair characterization of the research. A number of people said that they were unsure of what exactly was wrong with the work. As Hanna argued that voice-to-face prediction was a line of research that "shouldn't exist," others asked whether science could or should be stopped. "It would be disappointing if we couldn't investigate correlations—if done ethically," one researcher wrote. "Difficult, yes. Impossible, why?"

Some of the conversation touched on the reviewing and publishing process in computer science. "Curious if there have been discussions around having ethics review boards at either conferences or with funding agencies (like IRB) to guide AI research," one person wrote. (An organization's institutional review board, or I.R.B., performs an ethics review of proposed scientific research.) Many commenters pointed out that the stakes in A.I. research aren't purely academic. "When a company markets this to police do they tell them that it can be totally off?" a researcher asked. I wrote to Subbarao Kambhampati, a computer scientist at Arizona State University and a past president of the Association for the Advancement of Artificial Intelligence, to find out what he thought of the debate. "When the 'top tier' AI conferences accept these types of studies," he wrote back, "we have much less credibility in pushing back against nonsensical deployed applications such as 'evaluating interview candidates from their facial features using AI technology' or 'recognizing terrorists, etc., from their mug shots'—both actual applications being peddled by commercial enterprises." Michael Kearns, a computer scientist at the University of Pennsylvania and a co-author of "The Ethical Algorithm," told me that we are in "a little bit of a Manhattan Project moment" for A.I. and machine learning. "The academic research in the field has been deployed at massive scale on society," he said. "With that comes this higher responsibility."

As I followed the speech-to-face controversy on Twitter, I thought back to a different moment at the same NeurIPS conference. Traditionally, conference sponsors, including Facebook, Google, and JPMorgan Chase, set up booths in the expo hall, mostly to attract talent. But that year, during the conference's "town hall," a graduate student approached the microphone. "I couldn't help but be a bit heartbroken when I noticed an N.S.A. booth," he said, referring to the intelligence agency. "I'm having a hard time understanding how that fits in with our scientific ideals." The event's treasurer replied, saying, "At this moment we don't have a policy for excluding any particular sponsors. We will bring that up in the next board meeting."

Before leaving Vancouver, I sat down with Katherine Heller, a computer scientist at Duke University and a NeurIPS co-chair for diversity and inclusion. Looking back on the conference —which had accepted a little more than fourteen hundred papers that year—she couldn't recall ever having faced comparable pushback on the subject of ethics. "It's new territory," she said. In the year since we spoke, the field has begun to respond, with some conferences implementing new review procedures. At NeurIPS 2020—held remotely, this past December—papers faced rejection if the research posed a threat to society. "I don't think one specific paper served as a tipping point," Iason Gabriel, a philosopher at the research lab DeepMind and the leader of the conference's ethics-review process, told me. "It just seemed very likely that if we didn't have a process in place, something challenging of that kind would pass through the system this year, and we wouldn't make progress as a field."

Many kinds of researchers—biologists, psychologists, anthropologists, and so on—encounter checkpoints at which they are asked about the ethics of their research. This doesn't happen as much in computer science. Funding agencies might inquire about a project's potential applications, but not its risks. University research that involves human subjects is typically scrutinized by an I.R.B., but most computer science doesn't rely on people in the same way. In any case, the Department of Health and Human Services explicitly asks I.R.B.s not to evaluate the "possible long-range effects of applying knowledge gained in the research," lest approval processes get bogged down in political debate. At journals, peer reviewers are expected to look out for methodological issues, such as plagiarism and conflicts of interest; they haven't traditionally been called upon to consider how a new invention might rend the social fabric.

A few years ago, a number of A.I.-research organizations began to develop systems for addressing ethical impact. The Association for Computing Machinery's Special Interest Group on Computer-Human Interaction (SIGCHI) is, by virtue of its focus, already committed to thinking about the role that technology plays in people's lives; in 2016, it launched a small working group that grew into a research-ethics committee. The committee offers to review papers submitted to SIGCHI conferences, at the request of program chairs. In 2019, it received ten inquiries, mostly addressing research methods: How much should crowd-workers be paid? Is it O.K. to use data sets that are released when Web sites are hacked? By the next year, though, it was hearing from researchers with broader concerns. "Increasingly, we do see, especially in the A.I. space, more and more questions of, Should this kind of research even be a thing?" Katie Shilton, an information scientist at the University of Maryland and the chair of the committee, told me.

Shilton explained that questions about possible impacts tend to fall into one of four categories. First, she said, "there are the kinds of A.I. that could easily be weaponized against populations" —facial recognition, location tracking, surveillance, and so on. Second, there are technologies, such as Speech2Face, that may "harden people into categories that don't fit well," such as gender or sexual orientation. Third, there is automated-weapons research. And fourth, there are tools "to create alternate sets of reality"—fake news, voices, or images.

When the SIGCHI ethics committee began its work, Shilton said, conference reviewers—ordinary computer scientists deciding whether to accept or reject papers based on intellectual merit—"were really serving as the one and only source for pushing back on a lot of practices which are considered controversial in research." This had plusses and minuses. "Reviewers are well placed to be ethical gatekeepers in some respects, because they're close to this research. They have good technical knowledge," Shilton said. "But lots and lots of folks in computer science have not been trained in research ethics." Knowing when to raise questions about a paper may, in itself, require a level of ethical education that many researchers lack. Furthermore, deciding whether research methods are ethical is relatively simple compared with questioning the ethical aspects of a technology's potential downstream effects. It's one thing to point out when a researcher is researching wrong. "It is much harder to say, 'This line of research shouldn't exist,' " Shilton said. The committee's decisions are nonbinding.

There are few agreed-upon standards for ruling A.I. research out of bounds. Alex Hanna, the Google ethicist who criticized the NeurIPS speech-to-face paper, told me, over the phone, that she had four objections to the project. First, the paper's opening sentence describes "gender" as one of "a person's biophysical parameters"; gender is an identity, Hanna said, and how someone's voice resonates in the skull is not dependent on being male or female. Second, the system is likely to work better on the voices of cis people than on the voices of trans people. Third, the software's presumably higher failure rate for trans people could cause harm by misrepresenting them. Finally, the system could be used for surveillance. These objections might intersect. Hanna imagined what might happen if a trans person ended up on a most-wanted list. "I don't know if they do this anymore, but they put a composite sketch of this person on TV or social media, and then you have your old face following you around the Internet," she said—a "representational harm."

Rita Singh, a computer scientist at Carnegie Mellon University and the author of "Profiling Humans from Their Voice," is one of the senior authors on the paper. She seemed to approach the research from an entirely different perspective. She defended classifying faces into only two categories: "There have been thousands of papers that segregate their results based on gender—in literally hundreds of disparate scientific fields," she wrote, in an e-mail. I presented her with a tweet, from an Austrian computer scientist, about how the software might make trans people feel.

“Imagine what you would not want to look like, and then imagine this [to] be the output,” the researcher had written. The tweet helped to clarify for Singh the source of the reaction that the research had provoked. “I can see why this is disturbing,” she wrote. She noted two facts that she thought might help assuage concerns: the pictures present people as male or female, not as transgender, and they are low-resolution. The likelihood of an agreement between Singh and Hanna—the former, a researcher trying to conduct good science without an up-to-date explainer of the day’s contentious issues; the latter, an educator confronting a scientific field that’s often aloof from evolving social norms—seemed remote.

The speech-to-face paper is one of many recent research projects that have proved controversial on comp-sci Twitter. In November of 2019, at a conference called Empirical Methods in Natural Language Processing (E.M.N.L.P.), two papers—“Charge-Based Prison Term Prediction with Deep Gating Network” and “Read, Attend and Comment: A Deep Architecture for Automatic News Comment Generation”—were singled out for discussion online. The first presents an algorithm for determining prison sentences; the other describes software that automates the writing of comments about news articles. “A paper by Beijing researchers presents a new machine learning technique whose main uses seem to be trolling and disinformation,” one researcher tweeted, about the comment-generation work. “It’s been accepted for publication at EMLNP, one of the top 3 venues for Natural Language Processing research. Cool Cool Cool.” (In response, another researcher tweeted that publishing the research was actually “the ethical choice”: “Openness only helps, like making people discuss it.”)

The comment-generation paper had four authors, two from Microsoft and two from a Chinese state lab. After Eric Horvitz, who was the director of Microsoft Research Labs at the time, read the online discussion, he helped add some language to the paper’s final version, acknowledging that “people and organizations could use these techniques at scale to feign comments coming from people for purposes of political manipulation or persuasion.” When I caught up with Horvitz later, at another conference, he laughed, and said, “I never thought I’d be putting words in the mouth of the Communist Party!” Although the paper describes the software, its authors did not release its code. The research lab OpenAI has taken a similar approach, censoring its own text-synthesis software because it could, theoretically, be used to generate fake news or comments.

On Reddit, in June, 2019, a user linked to an article titled "Facial Feature Discovery for Ethnicity Recognition," published in *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*. The machine-learning model it described successfully predicted Chinese Uyghur, Tibetan, and Korean ethnicity based on photographs of faces. "It feels very dystopian to read a professionally written ML paper that explains how to estimate ethnicity from facial images, given the subtext of China putting people of the same ethnicity as the training set into concentration camps," the commenter wrote. ("This would give Hitler a huge boner," another noted, in grand Reddit style.) Last June, researchers at Duke presented an algorithm, called PULSE, that turns pixelated faces into high-res images. On Twitter, someone showed how the software turns a low-resolution photograph of Barack Obama into an image of a white man—likely the result of a training process that mostly used photographs of white people. Yann LeCun, Facebook's chief A.I. scientist, stepped in to defend the paper. "The consequences of bias are considerably more dire in a deployed product than in an academic paper," he tweeted. Many on Twitter disagreed. "If people are publishing trained models, other people are using them in production," one user wrote. "Let's be honest and write 'We don't know if our system works in the real world because we don't know any black people,' " another argued.

Just as some computer scientists seem oblivious to ethical concerns, others appear to be trigger-happy with their moral outrage. A paper accepted at the NeurIPS conference in 2019, "Predicting the Politics of an Image Using Webly Supervised Data," elicited a number of highly critical comments; "Our field is broken y'all," one researcher wrote. But at least two prominent members of the field based their criticism on a misinterpretation of the poster—they assumed that the work attempted to predict the political leanings of individuals based on their faces when, in fact, it predicts the political leanings of news outlets based on the photos they publish. (It can also tweak photos to be more "liberal" or "conservative," by substituting a scowl for a smile in a photo of a politician.) After the paper's main idea was explained to them, the researchers retracted their initial judgments, one on Twitter ("My apologies for not checking the paper!") and the other in an e-mail to me. Of all the comments offered, only one articulated what was actually troubling about the system: although it could be used to identify biased content, it could also be used to generate it.

The shadow of suspicion that now falls over much of A.I. research feels different in person than it does online. In early 2018, I attended Artificial Intelligence, Ethics, and Society, a conference in New Orleans, where a researcher presented a model that uses police data to guess whether a crime was gang-related. (I covered the event for *Science*.) The presenter took pointed questions from the audience about the possible unintended consequences of his research—could suspects be mislabelled as gang members?—before declaring, in exasperation, that he was just "a researcher." Wrong answer. ("No one is 'just an engineer' if what you're doing is going to result in a carceral outcome," Hanna told me.) An audience member stormed out, reciting, in a German accent, a song about the Nazi rocket scientist Wernher von Braun: "Once the rockets are up, who cares where they come down?"

In 2018, a group of researchers wrote a blog post for the Web site of the Association for Computing Machinery (A.C.M.)—the largest computer-science society in the world, with a hundred thousand members—titled "It's Time to Do Something: Mitigating the Negative Impacts of Computing Through a Change to the Peer Review Process." The scholars recommended that the ethical scrutiny which was being applied improvisationally, online and in person, be systematized. "Peer reviewers should require that papers and proposals rigorously consider all reasonable broader impacts, both positive and negative," the researchers wrote. Brent Hecht, a computer-science professor at Northwestern University and one of the authors of the post, told me that it caused "a bit of a splash." He went on, "Folks hadn't thought, at least in any sort of systemic way, about using their peer-review abilities in the way that we suggest."

Tech ethicists call technologies that can be used for both good and ill "dual-use." Pretty much all technologies are dual-use to some degree: a hammer can hit a nail or break a bone. Still, some tools, such as napalm, are better adapted for uses we might find disagreeable. The A.C.M. bloggers suggested that, when negatives appear to outweigh positives, peer reviewers should require researchers to discuss means for mitigation, perhaps through other technologies or new policies. "Computer-security conferences, interestingly, have had a history of asking for ethics statements," Katie Shilton pointed out, when we discussed this idea. In their calls for papers, the USENIX Security Symposium and I.E.E.E. Symposium on Security and Privacy require authors to discuss in detail the steps they've taken, or plan to take, to address any vulnerabilities that they've exposed.

An occasional justification for publishing dangerous or creepy research is that sunlight is the best disinfectant; by this logic, scientists should share their work even if it's alarming, and even when they don't know how to mitigate harms. In 2018, a paper in the *Journal of Personality and Social Psychology* described an algorithm that predicts, well above chance, sexual orientation from facial photos—essentially, a kind of automated gaydar. The research is frequently cited as A.I. gone wrong. But, at the time, Michal Kosinski, a professor of organizational behavior at Stanford University and a co-author of the paper, told me that the aim of the work was to sound a warning: a repressive government could be using similar methods already. "My worry is that this is unavoidable, however offended we are by it and whatever we want to do about it," Kosinski said. "I think the genie is out of the bottle and has been for many years, and we have to accept this and think about how to address these issues."

Michael Kearns, who co-authored "The Ethical Algorithm" with Aaron Roth, accepts this argument—to a degree. "What I wouldn't want to see is a steady drum beat of papers doing intrusive things of this form, and it becoming a cliché ritual justification that, 'I'm just pointing these things out.' " Shilton, for her part, argues that a research paper might not be the best venue for such sensitive research: a computer scientist might "work with the media, without providing so much how-to." Last February, meanwhile, a paper presented at the Artificial Intelligence, Ethics, and Society conference pointed out that mitigation works differently in the worlds of computer security and A.I.: the disclosure of a security vulnerability tends to benefit security experts, because software patches can be designed and deployed quickly, but in A.I. the reverse is true. Algorithms alter our social systems, not just our technical ones; it's hard to patch a government that's become addicted to surveillance, or a public that can no longer trust what it reads, sees, or hears.

It's not inconceivable that I.R.B.s might play a role in shaping A.I. research. Ben Zevenbergen, a research scientist at Google, who was an academic at Oxford and Princeton University when we spoke, told me that, although I.R.B.s are prohibited from considering "long-range effects," the consequences of new A.I. technologies may be felt quickly enough to evade such a provision. Several other scholars have suggested that the definition of a "human subject" may be flexible. If an algorithm learns to recognize faces using a public database of photos, then perhaps the people in the photos should be considered subjects. Perhaps they should be consulted, or have their consent obtained, before the research proceeds. "Obviously, researchers are incentivized to pretend that there are no human subjects involved, because, otherwise, things like informed consent become issues, and that's the last thing you want to do when you're processing data with millions of data points," Zevenbergen said. Still, the objections encountered by the speech-to-face researchers, for example, might have been raised earlier, had the people behind the portraits been consulted.

Companies and governments, of course, don't only conduct A.I. research; they also deploy the technology. For A.I. that is being used by private companies, "the natural point of enforcement would be the regulatory agencies," Kearns told me. "But, right now, they are playing a serious game of catch-up. They don't understand the technologies that they're regulating anymore, or its uses, and they have no means of auditing it." In a report for the Brookings Institution, Kearns and his co-author, Aaron Roth, proposed that regulators should be allowed to run experiments on companies' algorithms, testing for, say, systematic bias in advertising. Reforms within academic computer science matter, but they are only part of the picture.

For now, A.I. research is mostly self-regulated—a matter of norms, not rules. "The fact that these papers do come up on Twitter nontrivially often" has made an impression, Hecht said. "The vast majority of researchers don't want to be the subject of these types of discussions." Last year, I participated in an online workshop organized by Partnership on A.I., a nonprofit coalition founded by several of the biggest tech firms. In the workshop, which was focussed on encouraging more responsible research in the field, we discussed alternative release strategies: sharing new work in stages, or with specific audiences, or only after risks have been mitigated. Meanwhile, an online document evoked the spectre of social opprobrium: "Visualize your research assistant approaching your desk with a look of shock and dread on their face two weeks after publishing your results. What happened?"

Still, at some conferences, new norms are being formalized. Last year, for the first time, the Association for Computational Linguistics asked reviewers to consider the ethical impacts of submitted research. The Association for the Advancement of Artificial Intelligence has decided to do the same. NeurIPS now requires that papers discuss “the potential broader impact of their work . . . both positive and negative.”

Predictably, the new NeurIPS requirement was hotly debated among computer scientists. One particular response stood out: Joe Redmon, a star graduate student who, in 2016, developed a pioneering object-recognition algorithm called YOLO (“You Only Look Once”), revealed that he had stopped doing computer-vision research altogether, because of its military and surveillance applications. (His three papers on the YOLO system have been cited more than twenty-five thousand times.) Redmon’s decision wasn’t necessarily a surprise. “‘What are we going to do with these detectors now that we have them?” he asked in 2018, in his paper on YOLOv3. “A lot of the people doing this research are at Google and Facebook. I guess at least we know the technology is in good hands and definitely won’t be used to harvest your personal information and sell it to . . . wait, you’re saying that’s exactly what it will be used for?? Oh.”

Not all computer scientists think as Redmon does. This past December, at a NeurIPS workshop, researchers presented the results of a survey about the new “broader impact” statements, conducted among their peers. Some respondents considered the new requirement a joke (“If I liked writing fiction I would be writing novels”), while others appreciated it as a chance to “reflect.” Iason Gabriel, the philosopher who leads the NeurIPS ethics-review process, told me that the statements he’s read are surprisingly good. “They actually tend to be much better quality than you would expect from a purely technical audience,” he said. (Another paper from the same workshop criticized frequent “failures of imagination.”) NeurIPS has now implemented a second layer in the review process: any reviewer or area chair can flag a paper for review by a panel of three reviewers with expertise in weighing social impact. In 2020, out of about ten thousand submissions, reviewers flagged a few dozen; four papers that were technically strong were rejected based on feedback from the ethical reviewers. Some people on Twitter protested the intrusion of ideology into engineering; it’s likely that more would have spoken up but feared backlash. One outspoken professor emeritus received praise from anonymous accounts. “Thank you for your courage standing up against the woke,” one observer tweeted.

Without a representative poll, it's hard to quantify the community's views. "Sometimes there's this hypothesis within the domain of A.I. that there's a silent majority very hostile to ethics," Gabriel said. But "the majority reaction that I found was actually something slightly different," he went on. "There was almost a sense of relief among a lot of these researchers—that these things could finally be spoken about, and that they weren't just dealing with it as a kind of personal moral crisis. It's something that's being addressed through systemic reform." Shilton, the chair of the SIGCHI ethics committee, concurred. "In the last ten years, I have stopped having to justify myself to computer scientists," she said. "Instead, they say, 'Oh, that's an important thing to be working on,' which is lovely and very nice. Something has happened, with Facebook and A.I. and bias and fairness and racism. That has crystalized an awareness."

Scientists have been known to exercise caution ahead of time: in 1941, for example, researchers retracted papers they'd submitted to *Physical Review* on plutonium, holding them until the end of the Second World War. The American Society for Microbiology has a code of ethics forbidding research on bioweapons. But, historically, the regulation or self-regulation of science has often followed regrettable incidents. The National Research Act, which paved the way for today's I.R.B. system, was passed in 1974, after research abuses such as the Tuskegee Syphilis Study caused public outcry. From 2002 to 2009, two psychologists worked with the C.I.A. to develop torture techniques; they later faced censure from colleagues (and a lawsuit from the A.C.L.U.). In 2014, after several lab-safety lapses, the U.S. government paused its funding for certain so-called gain-of-function research projects aimed at increasing the power of SARS, MERS, and flu viruses. (The funding later resumed.) A.I. hasn't yet had its Hiroshima moment; it's also unclear how such a decentralized and multipurpose field would or could respond to one. It may be impossible to align the behavior of tens of thousands of researchers with diverse motives, backgrounds, funders, and contexts, operating in a quickly evolving area. And yet, all the same, we're now seeing rules, norms, and principles bubble up.

Hecht, who helped write the Association for Computing Machinery blog post that called for a more organized ethics process, predicts that increasing numbers of researchers, contemplating what their babies could grow up to become, will begin avoiding certain research topics. "If we're more transparent about the impacts, it can make authors say, 'You know, I really don't want to be up there having a debate with the audience,' or, 'I don't want to talk about how this work can be used negatively—I'm just going to do something else.' " He recalled an encounter he had with a young researcher presenting a poster at a conference before the pandemic. The work "had clear negative impacts that were not engaged with," Hecht said. "And the student sort of said, 'Yeah, I know. I don't want to do this. I don't want my research to succeed.' " Hecht laughed. "That was, um, something to reflect on."

This post has been updated to include Ben Zevenbergen's affiliations at the time of his interview.