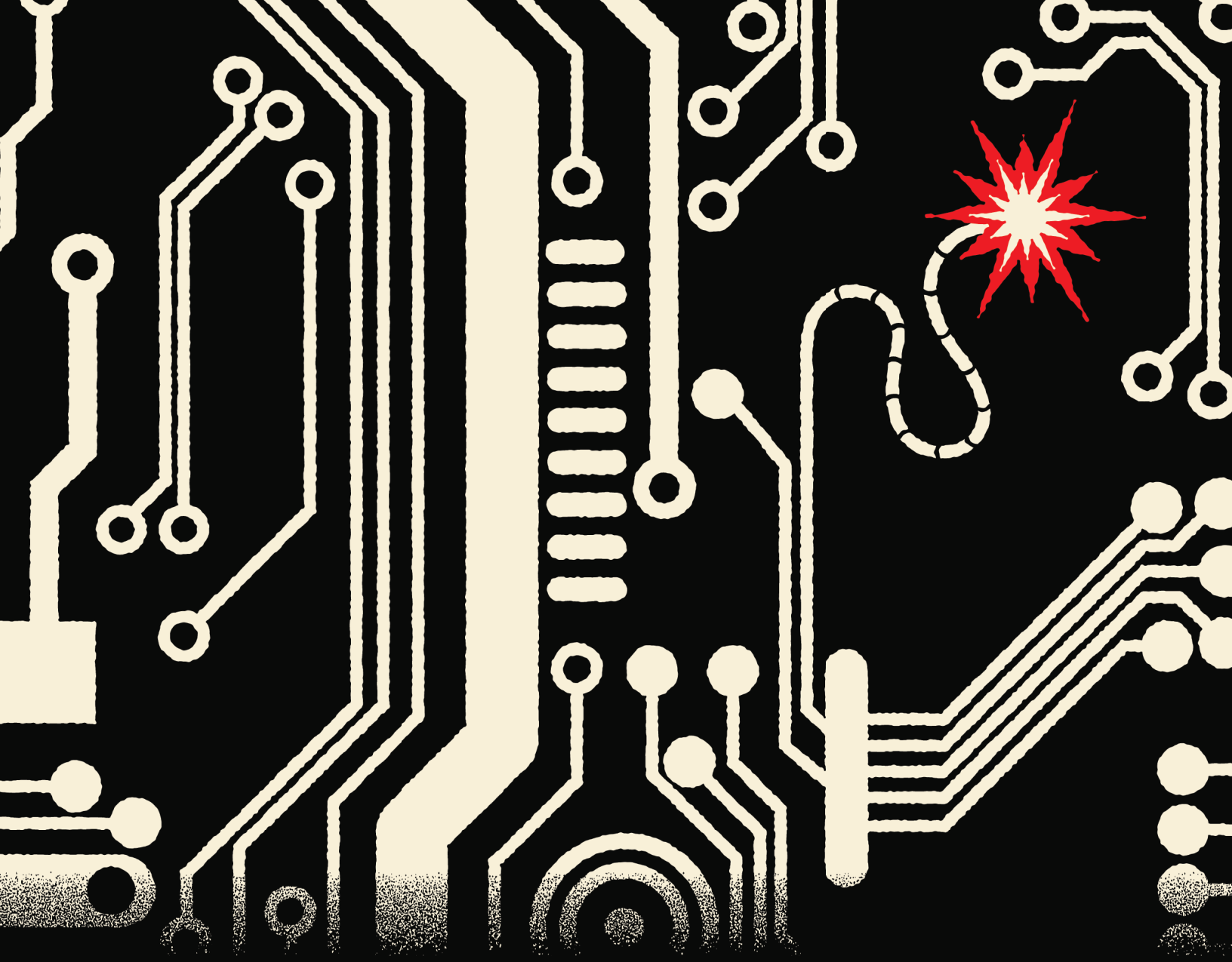


BERKMAN KLEIN CENTER
FOR INTERNET & SOCIETY
AT HARVARD UNIVERSITY

2024-2025 UPDATE

POSSIBILITIES & PERILS IN THE DIGITAL SPACE

BKC STAFF & BOARD OF DIRECTORS

OUR TEAM

Lori Adgate
Jonathan Bellack
Cruz Brito
Sebastian Diaz
Shelby El Otmani
Chris Finck
Brigitte Fink
Margaret Gagnon
Toni Gardner
Valerie Gomez
Patrick Goulart Soares
Peter Hankiewicz
Jennifer Hickey
Adam Holland
Mari Huertas
Chelsea Johnson
Darius Kazemi
Jay Kemp
Tara Kripowicz
Alberto Leon
Greg Leppert
Michelle Lineberger
Meg Marco
Kalie Mayberry
Madeline McGee
Chelsea McGovern
Brendan Miller
Jonathan Murley
Sarah Newman
Eric Pennington
Kathy Qian
Johnny Richardson
Rebecca Rinkevich
Zoe Robert
Lara Schull
Nadiyah Shaheed
Sam Shireman
Tavia Sons
Rebecca Tabasky
Nora Trapp
Jessica Weaver
Ellen Willemin
Alexandra Wood
Bey Woodward
Seth Young
Mo Zhou

OUR BOARD OF DIRECTORS

Jonathan Zittrain,
Faculty Chair
Christopher Bavitz
Yochai Benkler
Urs Gasser
Nien-hê Hsieh
Terry Fisher
James Mickens
Martha Minow
Charles Nesson
Ruth Okediji
Jeffrey Schnapp
Margo Seltzer
Stuart Shieber
Rebecca Tushnet
Mark Wu
Leah Plunkett,
Special Advisor to the Board

LETTER FROM THE FACULTY DIRECTOR

When the Berkman Klein Center was founded in 1996, the notion of studying the Internet's impacts on society (and vice versa) was both novel and niche. We were, after all, still trying to wrap our heads around this idiosyncratic network: a collective hallucination that seemed to depend upon cooperation in order to keep going. It wasn't anything like the products of AT&T and IBM.

Indeed, the Internet invited building from anyone and anywhere, without the barriers of government licensing or a proprietor's accreditation or business development relationships. Then and now, the spirit of BKC has been to accept and share that invitation and build out into the digital world – both technically and institutionally.

We've been building in the public interest for 28 years. As apps like Napster came from nowhere to throw down gauntlets to the $15 CDs and $30 DVDs that made up the backbone of the entertainment business, we brought together barons of creative and publishing industries with academics, policymakers, and free culture proponents to hash out alternatives. Some of us challenged retroactive copyright extension in the U.S. – all the way to the Supreme Court – while some BKC colleagues argued the other side.

The podcast was invented here. We incubated Creative Commons, the organization behind over 2.5 billion licenses that lets information flow more freely over the Internet, as well as Lumen – a vital part of delivering transparency around content removals by Internet intermediaries. And, in a kind of constitutional convention for the Internet, we helped to understand and contribute to the founding of the Internet Corporation for Assigned Names and Numbers (ICANN), the non-profit entity that today manages the Internet's core infrastructure.

But we know that to remain relevant as a Center, we must not simply rest on our past accomplishments. The world is changing rapidly, and to meet these challenges, we must refine our approach while ensuring we retain BKC's core values, drawn from the ideals of academia: integrity, honesty, intellectual humility, empiricism (both quantitative and qualitative), and openness. This will require a combination of deep academic engagement, innovation, and outreach to and collaboration with those both inside and outside our immediate circles.

The insights that have blossomed at and through BKC are now firmly ingrained into conventional wisdom. By closely and earnestly examining the architecture of the digital landscape, by imagining and building better technology, systems, and institutions, and by earning worldwide respect and dialogue from those in a position to make change – BKC and its communities have positively shaped the trajectory of what we used to call cyberspace.

It's never been more important that we continue to have the opportunity to do so. The Internet is no longer a new phenomenon that society is struggling to wrap its head around – but a full overlay of our lived reality, a front-and-center force that is shaping how we relate to one another and how we raise our children. It underpins the global economy and is determining how (and whether) participatory democracies around the world can function.

We've had quite a 2024. We're eager to continue to imaginatively take on some of the biggest issues facing the Internet and society today. That will mean expanding our community and programming to create positive applications of technology, and work to serve as a counterweight to tired, bumper-sticker conversations that too often dot digital policy. We want to scale BKC's impact over the next 25 years.

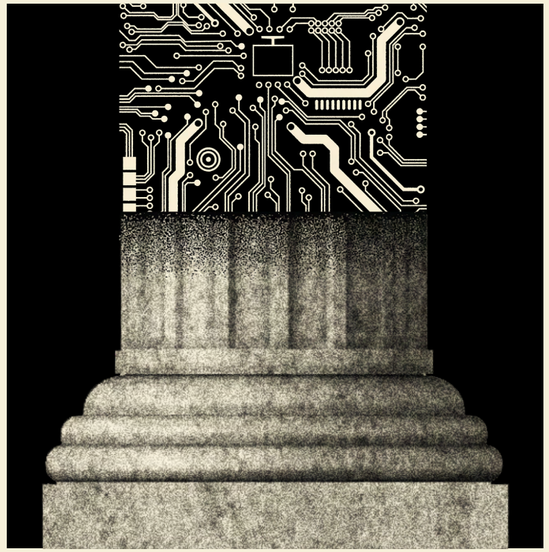
We would be so grateful for your support as we navigate this journey together.

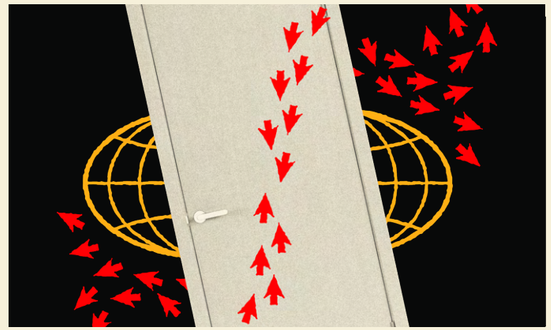
In the meantime, what a time to be alive (or for AI, not to be alive). What a spectrum of worthy challenges to confront. We look forward to taking them on together, in new configurations and with new energy.

As ever, thank you for believing in this work, and in the possibility for people – for all of us – to have agency and to thrive within a digital world.

George Bemis Professor of International Law
Professor of Computer Science
Professor of Public Policy

CONTENTS





BKC IMPACT

BY THE NUMBERS

By Jess Weaver

From launching an Applied Social Media Lab and Executive Education program, to welcoming over 30 in-residence fellows and visiting faculty, to hosting events on the biggest challenges facing AI and new solutions to online discourse and more, it's difficult to capture the activities of the Center in a single page of a report. We gave it a whirl here:

75 Events and workshops

349 Op-eds or opinion features by community members in press outlets

367 2024-2025 community members *(including staff, board, fellows, scholars, affiliates, associates, student research assistants)*

77 Universities found within the BKC community

39 Countries represented among the BKC community

73,467 Lines of code completed by the Frankly team

36 million Takedown notices tracked by Lumen in 2024

300+ Nymspace users in 2024

FRANKLY is an online video-based discourse platform designed to facilitate constructive dialogue and collaborative decision-making across, and within, diverse groups.

LUMEN is a database that collects and analyzes legal complaints and requests for removal of online materials. Founded in 2002, it is the longest current running project at BKC.

NYMSPACE is a tool for fostering open discourse through pseudonymity in closed-group settings.

DJ Patil, and Professors Latanya Sweeney, James Mickens and Lawrence Lessig speak at The Future of the Internet Summit to launch the Applied Social Media Lab.

BKC HAPPENINGS

By the BKC Team

BKC is a big tent designed to produce novel and positive impact. There are no political or ideological tests. We ask people to simply bring themselves, to be open to new evidence and argument, and to be ready to listen with curiosity before judgment. This is how BKC continues to surprise and intellectually delight the most seasoned and even cynical of participants, and to open up new pathways for action not previously contemplated. Here's a look at what has been happening under our roof.

Professor Jonathan Zittrain, BKC Co-Founder & Faculty Director, speaks to a crowd during BKC's fall launch event.

Dame Jacinda Ardern, Former New Zealand Prime Minister and BKC Fellow, speaks at BKC's 25th anniversary event.

Professor Charles Nesson, BKC Co-Founder and Principal Investigator of Nymspace, speaks on a panel at BKC's 25th anniversary event with BKC Faculty Associates Nagla Rizk and Juan Carlos De Martin.

Professor James Mickens, BKC Director, speaks to participants at a Public Interest Social Media Solutions Workshop.

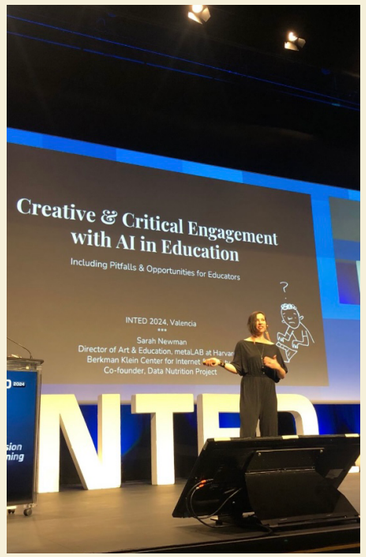
Yochai Benkler, Berkman Professor of Entrepreneurial Legal Studies at Harvard Law School and BKC Director

Professor Jonathan Zittrain hosts a panel titled 'Beyond Discourse Dumpster Fires' with Professors dana boyd, Deb Roy, and Gordon Pennycook.

Professor Lawrence Lessig, Principal Investigator of Frankly, speaks at the IAPP in June 2024.

Nien-hê Hsieh, Clark Professor of Business Administration at Harvard Business School and BKC Director

Nieman Fellow Jesselyn Cook and BKC-Nieman Fellow Ben Reininga discuss their research at the Institute for Rebooting Social Media Speaker Series.



Sarah Newman, Director of Art & Education at metaLAB, speaks at INTED2024 in Valencia, Spain.

Jason Goldman, Tracy Chou, Yoel Roth, and Kasia Chmielinski discuss the future of the internet with BKC Faculty Director Jonathan Zittrain at a summit to launch the Applied Social Media Lab.

Charles Nesson, Weld Professor of Law, BKC Co-Founder, and Principal Investigator of Nymspace

Lecturer Leah Plunkett, BKC Faculty Associate and Special Advisor to the Board, chats with Professor Madhavi Sunder, BKC Faculty Associate and 2023-2024 BKC Fellow.

Professor Anupam Chander, BKC Faculty Associate and 2023-2024 RSM Visiting Scholar

At a BKC student event, participants share their perspectives on the problems facing social media today and what solutions they might recommend.

Scenes from BKC's fall kickoff event.

Ruth L. Okediji, Jeremiah Smith. Jr, Professor of Law and BKC Director

Rachel Kalmar, BKC Affiliate

Professor Jonathan Zittrain, BKC Co-Founder and Faculty Director, and Rebecca Rinkevich, Executive Director of Institutes, lead a workshop on the impact of platform accountability models.

BKC Faculty Associate Professor Evelyn Douek, Affiliate Julie Owono, and Meta's Jessica Lindemann speak at "Now and Next: Platform Accountability and Content Governance", a two-day event co-organized by the Institute for Rebooting Social Media (RSM) and the Oversight Board.

Christopher Bavitz, WilmerHale Clinical Professor of Law, Managing Director of the Cyberlaw Clinic, and BKC Director

2025 LOOK-AHEAD

PROJECTS TO WATCH FROM BKC'S PARTNERSHIPS

Here's a look at who we're collaborating with this year to foster new innovations – and innovators – that will have a lasting impact on the digital landscape.

By The BKC Team

We know there is not a single challenge roiling our digital world today that will be solved by one person or entity. It'll take close and cross-sectoral cooperation to uncover what we need to understand about emerging issues in social media, AI, and deliberative discourse, and to take steps forward together that will meet the moment.

So in the new year, BKC is taking its characteristic approach to collaboration outside its academic walls to work with important partners in tech, business, and civil society to illuminate and shape what's possible in service of the public interest.

Here's a look at what we're launching in 2025 with new partners on board:

Microsoft and OpenAI: Free and Open Data to Train AI

BKC helped incubate The Institutional Data Initiative (IDI), a new effort within the Harvard Law School Library, to "level the playing field" in the AI industry by giving everyone access – from individual researchers to smaller AI companies – to a collection of curated, public-domain content that normally takes substantial resources, space, and time to assemble. As Wired noted, "the dataset is around five times the size of the notorious Books3 dataset that was used to train AI models like Meta's Llama" and "spans genres, decades, and languages, with classics from Shakespeare, Charles Dickens, and Dante included alongside obscure Czech math textbooks and Welsh pocket dictionaries."

Microsoft and Open AI are both partnering in this effort. Burton Davis, Microsoft's General Counsel for Intellectual Property, noted that Microsoft's support for this project aligns with its goal to create "more accessible data pools" for everyone building and training AI models.

In working with the Harvard Law Library and other libraries with high standards of rigor and review of material, IDI's Executive Director Greg Leppert hopes to create more public-domain datasets that can train AI models with the factual accuracy and original authenticity that libraries can provide.

> **...IDI is working with the Boston Public Library to scan millions of newspaper articles in the public domain and is working towards more collaborations with other libraries and similar data repositories in the future.**

IAPP: A Digital Policy Leadership Retreat

When it comes to handling the complex web of digital and AI growth and risk, who is responsible for the big picture? As digital and AI developments become more embedded across society without substantial reflection on their implications for society, it is increasingly important to train professionals across sectors on how to explore and shape digital responsibility a at global scale.

That's why BKC is partnering with the non-profit organization IAPP to host Navigate: A Digital Leadership Retreat in Portsmouth, NH in June. As the home for professionals who work at the intersection of data, technology, and humanity, the IAPP's mission is to define, promote, and improve the professions of privacy, AI governance, and digital responsibility globally. The IAPP and BKC will gather a unique consortium of leaders across industry, academia, government, and civil society to participate in collaborative, honest, and technically-grounded conversations that can explore how to navigate regulation, risk, and responsibility in a rapidly evolving digital environment.

Harvard Law School: Executive Education on AI and the Law

Never forgetting that some of our best partners work right alongside us, BKC is partnering with Harvard Law School to launch its first foray into executive education. AI and the Law: Navigating the New Legal Landscape will provide participants with a strong foundational understanding of AI technology. BKC and HLS will bring in experts from across the university and other institutions around the world to share cutting-edge insights, research, and perspectives on emergent legal questions raised by this novel technology. Chaired by Harvard Professor Terry Fisher, the program will engage participants, faculty, and others to brainstorm the ideas, risks, and challenges that come with such a technological seachange, enhancing their understanding of these complex issues.

Next Generation Innovators: The BKC Incubator

We seek to partner on good ideas wherever they may be found. That is why BKC is launching an Ideas-to-Impact Incubator designed to nurture and advance the boldest, brightest ideas in the public interest. By providing targeted resources, mentorship, and an environment conducive to experimentation, the incubator will partner with and empower innovators to transform their ideas into viable, impactful solutions.

To date, we've supported a post-graduate student's development of a synthetic social media feeds project aimed at reducing political polarization online. We're funding his experiment that uses generative AI to identify public, validator content on three platforms – Reddit, Twitter, and Facebook – and display it organically in the social media feeds before the U.S. election to test whether surprising validators can successfully reduce polarization. Upon success, he plans to create an open-sourced algorithm to help others actively identify these 'surprising validators'.

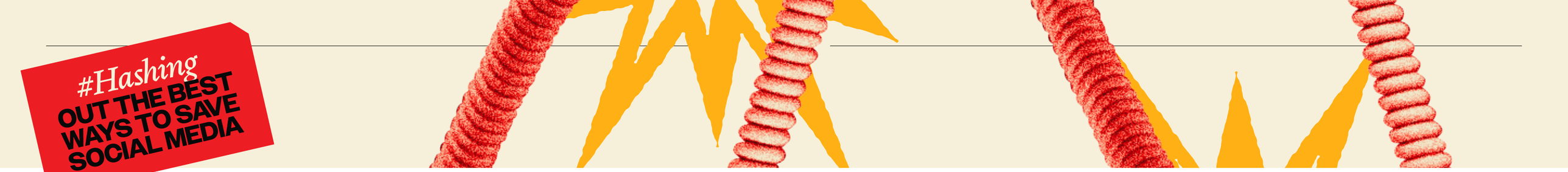
The BKC incubator is also funding the research of Leah Plunkett, one of the only scholars to consider the legal implications of what parents and other caregivers are doing with the private digital information of their children, a phenomenon often described as "sharenting." She previously published "Sharenthood: Why We Should Think Before We Talk about Our Kids Online" with MIT Press.

> **We are excited about all the innovators we'll be partnering with this year to develop more impactful ways to shape an ever-changing digital landscape.**

#Hashing OUT THE BEST WAYS TO SAVE SOCIAL MEDIA

By Shelby El Otmani

BKC's Institute for Rebooting Social Media and Applied Social Media Lab partner with the best of academia and industry to tackle the biggest challenges facing our digital environment.



#Hashing OUT THE BEST WAYS TO SAVE SOCIAL MEDIA

For the last four years, the Berkman Klein Center has been bringing together highly skilled engineers alongside some of the world's brightest academics with one common and vital goal: fixing social media.

Launched in 2021 and 2023 respectively, the Institute for Rebooting Social Media (RSM) and the Applied Social Media Lab (ASML) have worked in tandem to theorize on and develop solutions that actually work in the real world – from policy reforms to digital tools, platforms, and plug-ins aimed at improving how we learn, debate, and (co)exist online.

Here's Professor James Mickens – RSM and ASML Faculty Director and Professor of Computer Science at Harvard University – on the social media landscape around RSM's founding, in a moment where mis/disinformation around the COVID-19 pandemic was running rampant, conversations around personal and cultural identity were becoming more and more heated, and political unrest was spilling out of the comments section and into our offline lives: "There was just this general malaise and a sense that technology was not helping and, in fact, that technology was making things worse," he said. "There was a collective sense across the political spectrum that certainly this was not the best of all possible worlds. Certainly we could do better than this."

So when Mickens and BKC Faculty Director and Co-Founder Professor Jonathan Zittrain began discussing the possibility of building out new initiatives to drive academics towards co-creating social media solutions with technologists, it was a no-brainer–especially to Mickens, who had worked at Microsoft Research before becoming a professor.

> **A policy recommendation that is divorced from an implementation is toothless.**

said Mickens. "I think Jonathan Zittrain and I realized it would be really great to start bringing engineers more directly into the conversation about how to fix technologies like social media."

Under the direction of Mickens and Zittrain, RSM and ASML demonstrate how a university research center can facilitate interdisciplinary collaboration between academics and technologists to create meaningful, community-driven improvements to the social media landscape. Since 2022, RSM's Visiting Scholars program alone has hosted 23 tenured and tenure-track professors that are prominent figures in their respective fields, including platform regulation, the online creator economy, and well-being. Members of the current cohort, including Allison Stanger (Middlebury College) and Paul Resnick (University of Michigan), are scheming to bridge the gap between engineers, politicians, and the public through work on replacing Section 230 and tracking X's Community Notes feature, respectively. Additionally, Visiting Scholar Myojung Chung (Northeastern University) is collaborating with ASML on an interactive tool aimed at improving algorithmic literacy among youth.

Building a lasting community of practice is another critical element to our theory of change. "It's not just building shiny new things," Mickens said. "It's also trying to understand the context in which those things will be used and trying to understand what types of things we should be building in the first place." Together, RSM and ASML have hosted over 20 public-facing events and workshops in the last year, diving into issues related to trust and safety in the majority world, AI content moderation, and the impacts of online conspiracy theories online and in communities. The goal of these events is to paint a complex picture of the landscape–one that often confronts and challenges preconceived notions of what will and won't work in creating change.

ASML in particular has hosted several community-oriented workshops, including a whistleblowing workshop that brought together technologists that develop tools to protect whistleblower privacy, lawyers that work with whistleblowers, and some actual whistleblowers that offered first-hand perspective into the fraught process. "A lot of great insights came out of that workshop that have then gone on to influence some of the [project and ideas] we've been thinking about," said Mickens.

A better social media landscape won't come from isolated work: more hypotheses, more lines of code, or more shiny tools. But it can come from hypotheses, code, and tools that directly interact and are informed by each other and the communities they're meant to serve. RSM and ASML have positioned themselves in a unique place to drive this work and hopefully inspire more organizations to do the same. We need everyone's help to create the best version of social media possible.

Missives from Our Labs

RSM and ASML house multiple initiatives, projects, and perspectives all under one proverbial roof. Let's hear from some of the team members about the projects they're working on and what inspires them about this work:

TONI GARDNER

Director of Operations // The Institute for Rebooting Social Media

The Institute for Rebooting Social Media is a three-year initiative examining social media's most urgent problems and exploring interventions to produce healthier online ecosystems. We believe that the challenges facing social media today are best addressed through interdisciplinary learning and collaboration. With that in mind, we have implemented a broad portfolio of programs, events, and educational opportunities to support existing and emerging experts in focused, timebound research and to convene participants from across sectors, disciplines, and backgrounds in the pursuit of challenging the status quo and improving the state of social media.

JONATHAN BELLACK

Senior Director // The Applied Social Media Lab

The Applied Social Media Lab's work falls under four specific focus areas: **spaces for civil discourse and collaboration; transparency tools; personalized safety applications;** and **interoperable software infrastructure.** Within these parameters, ASML provides the space for technologists to question, imagine, and experiment with potential interventions and tools without the constraints of meeting a corporation's bottom line. In addition to creating tools that positively impact the online ecosystem, ASML aims to establish a durable community of current and future technology leaders who will continue building social media software in the public interest beyond our tenure.

PROFESSOR CHARLES NESSON

Faculty Lead // Nymspace

Nymspace is a pseudonymous platform rooted in the belief that open, honest discourse requires trust and a degree of de-identification to foster authentic participation, especially in educational and learning spaces. By providing pseudonymity, Nymspace creates a "trustspace" where participants feel free to explore ideas without the pressures of personal identifiers. Ultimately, Nymspace seeks to leave a legacy of redefined discourse spaces that prioritize privacy and mutual respect, setting a foundation for future educational and civic environments that are more inclusive, open, and constructive.

PROFESSOR LAWRENCE LESSIG

Faculty Lead // Frankly

Frankly addresses the critical need for effective tools that facilitate constructive dialogue and collaborative decision-making across and within diverse groups. By providing an accessible, open-source platform for structured discourse, Frankly enables communities to engage in meaningful conversations and practice essential civic skills without relying on trained facilitators. Frankly aims to provide a stable foundation for innovation, enabling contributors to build surprising solutions for a broad range of use cases, thus expanding the platform's impact beyond our initial vision. This strategy not only empowers individuals to participate more effectively in democratic processes but also fosters a growing ecosystem of tools for civic engagement.

As communities adopt and build upon Frankly, we anticipate a ripple effect where constructive dialogue becomes the norm, leading to more innovative problem-solving and effective governance at various scales, from town hall meetings to citizens' assemblies. Ultimately, by serving as a core, adaptable deliberative tool, Frankly aims to contribute to a global shift where diverse perspectives catalyze solutions rather than deepen divides, and where engaging in participatory democracy is an accessible and routine part of daily life.

BRENDAN MILLER

ASML Senior Software Engineer

Harvard creates an exciting container from which to draw expertise, and be able to reach partners who will hopefully be able to bring some of our innovations to social media users around the globe. For example, I am personally very excited to be working towards an interoperable, user-centered social media experience where people own and control access to their data and networks across platforms, and can customize their own filters and feeds. I also enjoy my work on Threshold Polling with Kathy Qian, which answers the question, "What collective possibilities emerge when we safely reveal our shared experiences," and solves an important class of collective action problems.

CHELSEA JOHNSON

ASML Principal Engineer

Working in an academic lab where intellectual, technical, and ethical values align is deeply fulfilling. I appreciate how we can focus on factors that aren't always relevant in profit-driven settings, paying as much attention to how we intentionally build technology as much as what we're building. It's a privilege to work on projects that prioritize the interests and rights of users, especially in challenging areas like online safety.

KATHY QIAN

ASML Senior Software Engineer

Working in an academic environment has exposed me to a lot of new opportunities to collaborate on solving tough, nebulous problems. Most of all, I enjoy the spirit of learning and creativity that comes with working here; in many ways it's been like working in an impact-driven startup incubator program.

DARIUS KAZEMI

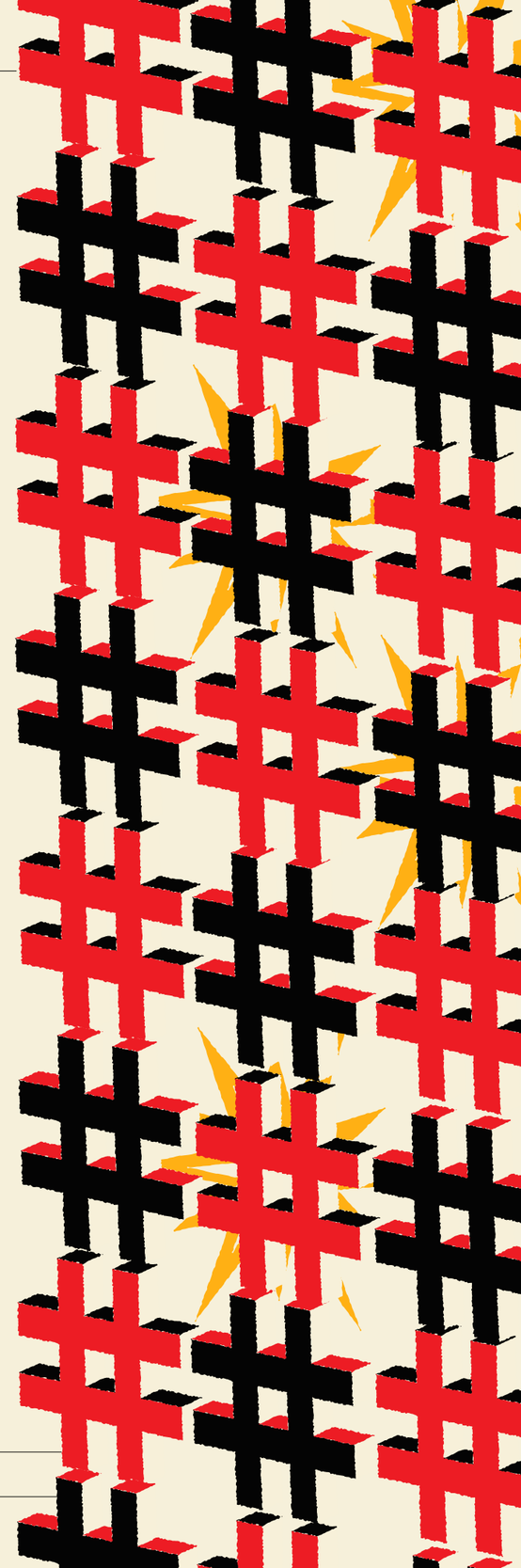
ASML Senior Software Engineer

I see working at an academic lab as a chance to do things I don't get to do in industry. I get to do basic, fundamental research work - one example of that is the Fediverse Schema Observatory. And I get to use the convening power of a university to do the important political work of navigating standards bodies and bringing people to consensus without being burdened by an ulterior profit motive.

ALBERTO LEON

ASML Senior Software Engineer

What excites me most about working in an academic lab is the opportunity to collaborate with brilliant minds across disciplines—faculty, fellows, industry experts, and peers—who are all deeply passionate about shaping the future. For example, I'm part of a team developing decentralized and portable identity solutions for social media, with the potential to integrate these into major platforms. I'm also working with BKC staff to plan a hackathon that brings together students and industry leaders to tackle pressing challenges in social media innovation.



READING THE TEA LEAVES

TOPICS THAT WILL DRIVE THE SOCIAL MEDIA CONVERSATION IN 2025

We asked our Visting Scholars at the Institute of Rebooting Social Media to weigh in.

By Shelby El Otmani

2024 was a year of twists and turns–particularly when it came to conversations about social media. We saw a rise in AI-generated content that fueled conspiracy theories about Hurricane Helene, the U.S. government's effort to change ownership of TikTok, a call from the U.S. Surgeon General for interventions on social media and the youth mental health crisis, and most recently, the Australian government's ban on kids under the age of 16 from joining social media altogether.

And oh, did we mention several major elections across the globe?

We built the Institute for Rebooting Social Media's (RSM's) Visiting Scholars program to ensure that we're on the cutting edge of studying critical online phenomena and envisioning, prototyping, and convening stakeholders around needed solutions. With areas of focus ranging from replacing Section 230 and understanding the sociological motivations behind online hate speech, to analyzing social media's impact on well-being and cultivating reparative social media, the Visiting Scholars have unique opinions on how best to create change, and have their fingers to the pulse of the Internet's future.

We reached out to the 2023-2024 and 2024-2025 Visiting Scholar cohorts to find out what they think will be key conversations in 2025.

What do you think will be the next big conversation in the social media space in 2025?

AJ Christian, Northwestern University (2024-2025)

We're already seeing the decentralization of social media. People are starting to think hard about which platforms are taking their data and what those platforms are doing with them. We need many more platforms that are intentionally designed to cultivate healing stories and data.

Research focus: What storytellers, thought leaders, and community organizers working toward solidarity across cultural lines think we need to do to improve them both technologically and culturally.

Myojung Chung, Northeastern University (2024-2025)

Given the power that large social media firms wield today, coupled with growing concerns over AI-driven bias, can decentralized social media create a more equitable online ecosystem, or will it descend into unregulated chaos?

Research focus: Equipping users with a deeper understanding of how algorithms curate content, empowering them to critically engage with what they see on social media.

Marshall Van Alstyne, Boston University (2024-2025)

Why aren't the current misinformation solutions working? How can we establish stronger listener rights and not just those of bombastic speakers? How can researchers gain access to social media data so we can inform the public what's really happening, without having to rely on what the firms themselves tell us?

Research focus: How we can reduce the flow of misinformation with no censorship at all and no central authority judging truth.

Paul Resnick, University of Michigan (2024-2025)

What *else* is engaging, besides outrage?

Research focus: Providing a window into the workings of a "bridging algorithm," like the one that X's Community Notes uses to select notes that are upvoted by people who don't usually agree with each other.

Noah Giansiracusa, Bentley University (2024-2025)

The role of AI. How much do we want to continue letting AI choose what we see and who we interact with? And how much will we tolerate AI-generated content on social media?

Research focus: The problematic ways the online ad ecosystem operates and how we might try to address those.

Eric Gilbert, University of Michigan (2024-2025)

How to billionaire-proof our social platforms.

Research focus: Designing new ways for people to collectively own and govern the social platforms they use.

Swati Srivastava, Purdue University (2023-2024)

How geopolitical pressures, including AI racing dynamics, lead to less political will to regulate platforms for the public good and more pressure for weaponizing platforms for public harms.

Research focus: Understanding how social media may be rebooted in the global majority so it does not repeat the experiences of the U.S. and Europe.

Jeff Hall, University of Kansas (2023-2024)

The legislated or mandated changes to platforms due to concerns for adolescent safety. What changes will be implemented or demanded? Will the legislative or corporate changes match the rationale for making those changes? Is there a meaningful alternative?

Research focus: TikTok-style shorts and reels across platforms, the collapse of the traditional social network experience, and new ways of engaging social media.

What do you think will be the next big conversation in the social media space in 2025?

David Craig, USC Annenberg (2023-2024)

How creators contribute to the global rise of populist movements from the left and right.

Research focus: Cultural economies and creator cultures distinguished by how social media entrepreneurs are harnessing social media for commercial and cultural value.

Allison Stanger, Middlebury College (2024-2025)

The ban of TikTok. It's not about human free speech; it's about a foreign adversary's algorithmic manipulation for profit and malevolent mischief.

Research focus: Advocating for the sunsetting and renewal of Section 230, and directing people to a proper alternative.

ARTIFICIAL INTELLIGENCE

"THE WOLVES CLOSEST TO THE SLED"

As the development of AI races forward, this is what some of AI's brightest minds believe are the most imminent challenges that society must contend with now.

By The BKC Team

Artificial Intelligence has long gripped the imagination of technologists and storytellers alike. After all, Gov. Arnold Schwarzenegger built an entire career on Hollywood's rendering of its potential to take over the world. But the meteoric speed of AI's development today is bringing questions about its inevitable impact and sparking very real debate among AI leaders, academics, policymakers, philosophers, and business minds about where the technology is taking us as a society and whether AI at its most extreme – like Artificial General Intelligence – is really possible in our lifetimes.

A Terminator-esque future is a provocative question to grapple with. But, as anyone who has used ChatGPT to put together a travel plan or finish a report can tell you, AI is already reshaping how we work and relate to each other in incredibly real, and now increasingly worrisome, ways. So we invited some prominent minds on the frontlines of AI development to share with BKC what they believe are the serious problems AI is presenting to society that we're not fully prepared to deal with right now.

As AI races ahead, here are the problems that these AI thought leaders see as "the wolves closest to the sled."

JASON GOLDMAN

Senior Advisor on Technology Policy to President Obama

Part of the founding team and VP of Product at Twitter, Jason Goldman later served as the first-ever Chief Digital Officer at the White House for President Obama. He now advises the former President on technology policy.

"In some countries like Brazil, South Korea, and China, there isn't as much of an innate skepticism toward the tech sector. They're more likely to adopt AI solutions and adapt because of a general higher level of trust in tech. In the U.S., for sectors where we are most likely to deploy AI (i.e. manufacturing), our society will be less well adapted to deal with that kind of technological change. It will only serve to foment the current feeling among American workers that tech is screwing the everyday worker. And of course, not all jobs are being obsoleted, but there will be specific areas where there will be significant and abrupt losses. For example, it's not a great time to be an illustrator. And maybe it's not a great time to be a paralegal. The way we've trained lawyers in general, people who've taken on hundreds of thousands of dollars in academic debt, doing copy and replace work on word documents to change a contract with very little need for creative or intellectual work. That will be an AI job. And I think, particularly in America, we are not prepared for that."

RAFFI KRIKORIAN

Chief Technology Office of Emerson Collective

The former CTO of the Democratic National Committee, Raffi Krikorian has worked at the intersection of technology and impact across his career. He served as the Director of Uber's Advanced Technologies Center where he launched the first ever self-driving fleet and before that was a VP of engineering at Twitter. An MIT grad, Krikorian now works to power Emerson Collective's social work through data, tools, and product design.

"The rise of surveillance from autonomous cars. A Tesla has 6-7 cameras on it, and it records everything when it goes down the street. And I know we have no assumption of privacy in public spaces, but this is an absurd amount of data that is being sent back to Tesla's server. That data is used to train autonomous vehicles. But does Tesla abide by privacy rules? Do they scrub my face? What will happen when they get subpoenaed? We might have slightly more trust in Google's Waymo because we've seen Google scrub faces on Google Street View, and so presumably they have a pipeline to scrub this kind of data. But do we know that about Tesla? Or General Motors new Super Cruise? When I worked at Twitter, the company would create transparency reports when police subpoenaed something from us. It's crazily chilling that we don't know how big companies are going to proceed with this level of data surveillance when it comes to law enforcement. People often think that London is where surveillance is a huge problem because of CCTV cameras, compared to a city like San Francisco. But any car that has self-driving on it, at some point is recording you. So when you account for the number of Teslas and other semi-autonomous cars that have tons of cameras and are streaming their data to private servers, the amount of surveillance really adds up. Who has access to the data? Where is it going? Even if you can disable that information, no one ever does. So this, I think, is a huge issue."

TOM GRAHAM

CEO and Co-founder of Metaphysic

Tom Graham co-founded Metaphysic, an AI pioneer that's developing software and AI tools to create photorealistic synthetic media. Metaphysic is the team behind the viral sensation @DeepTomCruise on TikTok and the de-aging technology used in the movie "Here" starring Tom Hanks and Robin Wright. Graham is the first person to file for copyright registration of his AI likeness, in a campaign to create new digital property rights.

"I'm an expert on deepfakes. But since the beginning of 2024, there has definitely been a lot of content where I don't know whether it's real or not. So here's what worries me. I think that, right now, around 30-40% of content on TikTok is AI generated in some way. With a goal of getting someone to watch content for 1-2 minutes, people are using AI agents and tools to automate stitching together stories – combining simple images, videos and AI-driven voiceovers. Someone might instruct the AI agents to make the story interesting to an audience by adding relatable or compelling characters, or by describing historical facts. But what you ultimately get is content that is full of AI hallucination – instead of being historically and factually accurate, it ends up being

full of made-up facts, figures, and even historical figures. For example, I'm on Ocean-tok - and I was recently served a TikTok video about a project from the 1970s where the Florida government dumped millions of tires in the ocean in an effort to create a fish habitat. That's true. But after a minute or so, the video started talking about an Australian, 20-year old, female billionaire who was financing the clean up and a ship captain who was involved in the dumping 50 years ago. But, these characters are totally fictional – the AI had sourced, appropriated and integrated the real names and photos of a beautician from Brisbane and some guy on a fishing boat! AI content generators are just grabbing random stuff from the Internet to construct a story, and spitting out engaging human interest stories. And this will only increase on social platforms that we all spend so much time on. So that's just endless hours of empty calories and absolutely fictional information that is building the knowledge base of our youth. If you think about hobbling a civilization, what would you do for the greatest impact? You'd teach the young generations completely wrong information. And nothing wrong in a specific, easily verifiable way. It's white noise. And that's just tremendously terrifying to me."

ASMAU AHMED

Technology Leader in AI and Responsible Innovation

A senior technology and product executive, Asmau Ahmed most recently served as a product leader helping to run Google X, Google's "Moonshot Factory." She also served as a Senior Director of Product, overseeing teams that specialized in AI, Privacy, Content Moderation and Safety. Ahmed has a background in chemical engineering and currently serves on the board of QuinStreet.

"Energy is an obvious starting point. The infrastructure we have today isn't built for the volatility that AI will create and demand. Instead of trying to force new technologies into an old grid, we need to think about decentralized, dynamic systems that are adaptive, equitable, and resilient. I see so much potential in AI-powered microgrids — neighborhood-level networks where energy is generated and shared locally. Imagine buildings generating their own power — solar on rooftops, battery storage, modular data centers in basements — and AI optimizing energy distribution down to the minute. It's about creating smarter, localized systems that allow AI to meet its own energy demands sustainably.

But there's also the data itself. If energy demands require a decentralized grid, the same should apply to the data sources feeding these AI models. Too much of the training data comes from homogenous sources, which risks reinforcing existing inequities. We're feeding these systems with recycled outputs rather than a constant stream of human perspectives. Just as microgrids distribute power across communities, a distributed data system could ensure we're pulling data that's representative of different languages, geographies, and human experiences. This would mitigate the existential risk of AI systems amplifying inequities and gaps in knowledge or opportunity.

The challenge here is ensuring that as AI scales, it isn't just efficient—it's equitable. If we don't address this now, we'll risk entrenching an infrastructure—whether physical or digital—that reinforces the very problems we're trying to solve."

DR. JOSHUA JOSEPH

BKC Fellow, MIT Visiting Scientist, and HLS Lecturer

After receiving his Ph.D. in Aeronautics and Astronautics from MIT, Joshua Joseph built AI systems in finance and life systems before returning to MIT as the Chief Intelligence Architect of MIT's Quest for Intelligence. He then co-founded Covariance.ai, a prize-winning MIT startup that turns external data into actionable insights. He is currently a Visiting Scientist at MIT, a Fellow at the Berkman Klein Center for Internet & Society at Harvard University, and a Lecturer on Law at Harvard Law School.

"I'm generally optimistic about AI agents but what I worry about is their reliability. If you try to use these agents on your actual tasks they're too brittle and will fail in unintuitive and frustrating ways. When I think of the potential users of these agents – like non-tech friends of mine or even my dad – there's just no way they will adopt them in their current state. The technology doesn't do what an average user actually asks for or wants and isn't up to handle most use cases. For example, Google just released Gemini Advanced 1.5 Pro with Deep Research. I decided to test it out and asked it to put together a syllabus for our "Agentic Artificial Intelligence and the Law" class this coming Spring. Gemini's response: 'As a language model, I'm not able to assist you with that.' What? Why? Looking at Google's marketing of its 'Deep Research' feature, that request seemed right in line with its intended use. I asked the new Zoom AI Companion, 'What are zoom-related tasks you can help me with?" and this was its response: 'The query requires some domain knowledge or needs up-to-date information, but external web query is disabled.' What? I don't think these are the kinds of problems that are solved by more compute and better data and I worry how much patience a typical user has before they decide this is all just hype. It seems to me there is a deep disconnect between the 'AI is progressing so rapidly' narrative and users' direct experience with these products. I worry this will delay the adoption of the technology and its potential for good outcomes (while, of course, being mindful of the many sharp edges). So, for me, it's less a worry about a wolf and it's more a worry about the sled. Do the people on the sled think it's broken? Are they just going to get off of the sled because their direct experience is that it sucks? That's what worries me."

DR. RUMMAN CHOWDHURY

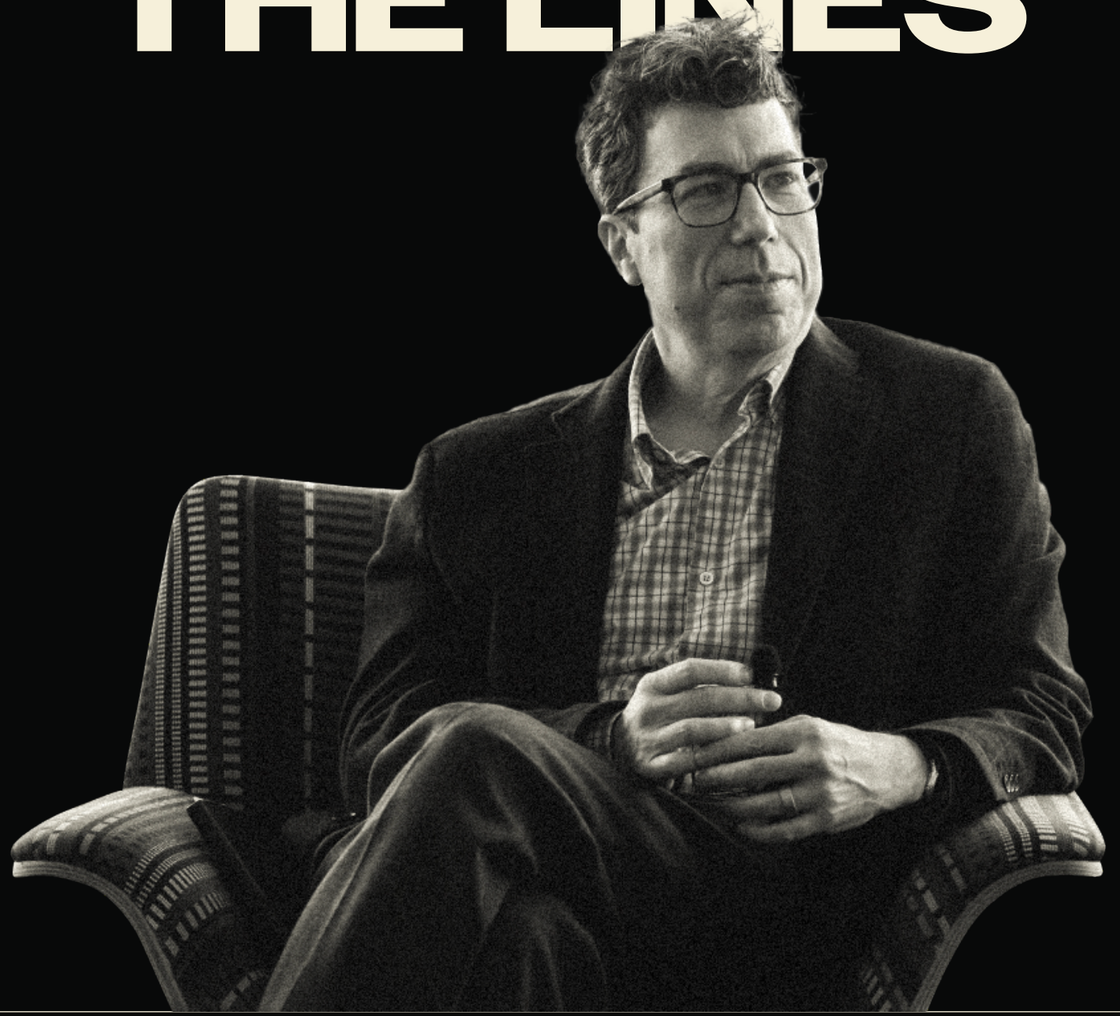
U.S. Science Envoy for Artificial Intelligence

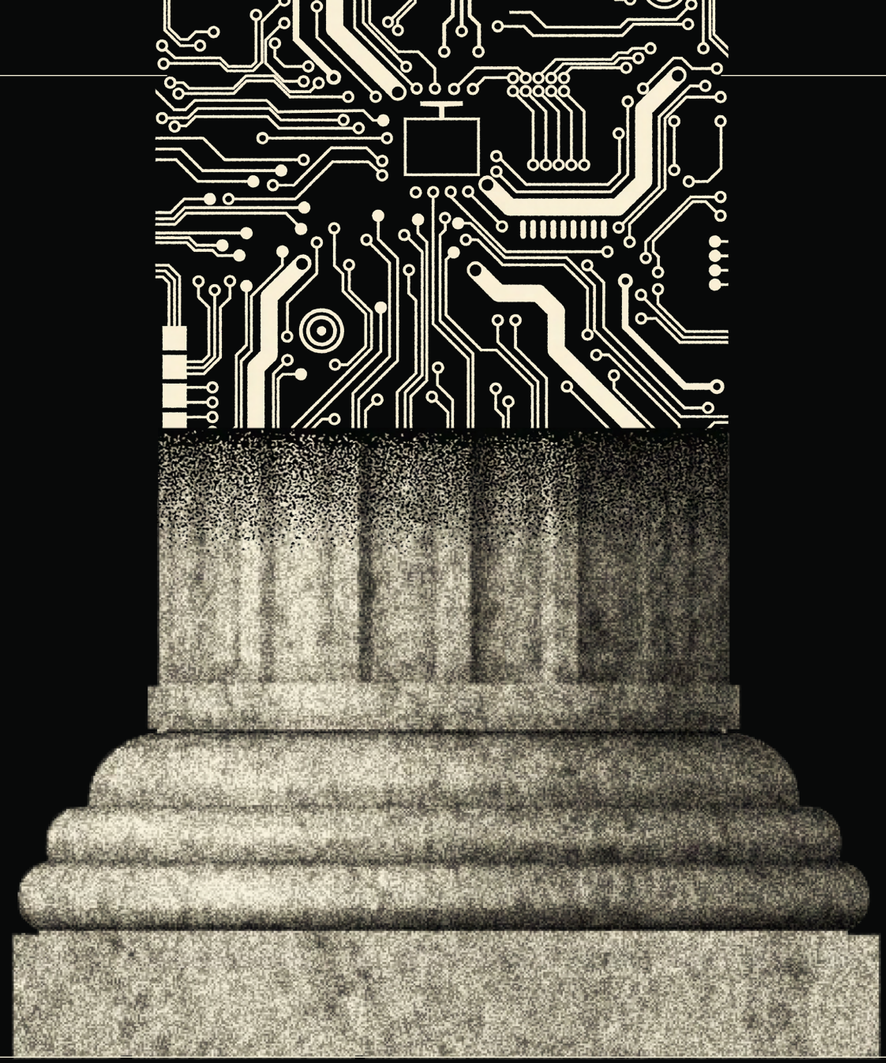
Dr. Rumman Chowdhury is a pioneer in the field of applied algorithmic ethics. An incoming BKC fellow, she currently serves as the U.S. Science Envoy Artificial Intelligence and as a member of the Artificial Intelligence Safety and Security Board at the Department of Homeland Security. She is also the CEO and co-founder of Humane Intelligence, a tech nonprofit building a community of practice around algorithmic evaluations.

"We are perilously close to a post-truth world, and the concept of information integrity is becoming fragile at best. It is getting increasingly difficult to understand, trust, and verify content on the Internet. This has serious implications on our political system. For example, I think we're being naive about what misinformation and disinformation campaigns could actually look like in the coming years. The takeaway from the most recent election was that misinformation wasn't so bad, but I think we are going to see an evolution in the sophistication of how mis/dis information is spread that makes it increasingly difficult to identify and combat the campaigns. Content moderation has always been a cat-and-mouse game. Social media companies have long employed AI/ML to identify the markers of a bot: AI-generated avatars, short histories online, and suspicious networks of followers. But when you can power the creation, maintenance, and relationships of bots through AI – for example, if you use any chatbot to help bots interact online with more natural language – you can create a an entire of army of bots with synthetic histories that will be hard for us to identify as malicious actors in our online communities. We simply haven't been in this Generative AI environment long enough to feel the effects of subversive content. The next U.S. presidential election will look very different when it comes to information integrity. Are we ready for it?"

As the sled careens ahead, almost every AI leader we spoke with noted the dangerous lack of spaces where those who are invested in the rise of AI across different industries can gather and discuss, in a curious and cooperative way, all the potential harms AI may have on our society. That is one of the very reasons BKC was founded – to provide a home for conversation and collaboration for all those who are actively shaping the technology's trajectory and its global impact. We look forward to picking up that mantle for AI in the coming year.

WHERE WE DRAW THE LINES





An Interview with Jonathan Zittrain on Anticipating AI – and Academia's Role in Shaping Its Future

Professor Jonathan Zittrain, Faculty Chair and co-founder of the Berkman Klein Center, answers eight of the most urgent questions we're grappling with on the frontiers of AI innovation.

By Jay Kemp

AGENTICS

In July, you penned an essay for The Atlantic, "We Need to Control AI Agents Now," which outlines "potentially devastating consequences" for this next moment of AI inflection. Hyperbolically, every company and their subsidies are about to automate their external services and internal processes through AI agents – but we agree, that can't be all bad. ***What are the most promising and titillating doors agentic AI opens? And how do we, as you say, "maintain our agency in the deep sense" in the search for them?***

We've all seen how quickly AI has been developing. One question is: how much the next round of development will be in raw capability vs. deployment and application. On the former, there's a lot of open questions about what's next for AI. OpenAI co-founder Ilya Sutskever recently said, "I foresaw everything in AI the past decade, assembled colossal models, and now, standing on that Mount Olympus of predictions, I have precisely zero idea what comes next." (Others are much more confident, but predictions are all over the place.)

Even if AI architectures have plateaued for a bit, there's still room for lots of development in application, or in bolting on various non-machine-learning-specific capabilities to AI models. For example, "chain of thought" approaches allow AI models to become more reflective – and possibly more accurate and able – by allowing them to converse with themselves or with other models before they offer a final answer or action to a user's prompt. And the models and systems we see now can, for better or worse, be given more levers to pull without human oversight. That's the essence, to me, of agentic AI: AIs that not only can talk to users in a sandbox, but that can be charged with a goal or mission, think about how best to achieve it, and then start to do so thanks to being able to actuate things in the real world over the Internet: moving and investing money, ordering a pizza or a pallet of hay, or reading and responding to posts on social media as a designated persona.

Since you asked for an optimistic take, this sort of thing could be enormously empowering to individuals. It could give everyone the sort of power and reach traditionally reserved for larger organizations. For example, you could have a bot that checks prices on some big-ticket item, ready to tell you when there's a substantial drop from among many vendors. That, in turn, could force Amazon to better earn its primacy, as it could lean less on its convenience as the "everything store" when a bot can run around looking for products everywhere. Or you have a bot that looks for job openings that fit your resume well, wherever they might be found – and automatically assembles various reviews and summaries of the companies behind the listings, ranking them by desirability.

More profoundly, we could have agents that are crafted to be loyal to us and to our interests – our second-order interests, that is: not just what we want, but what we want to want. They could, with our blessing, intervene in moments of weakness to help us stick to the path we want to follow. And not a moment too soon, as AI will elsewhere be helping any number of merchants of products and information and propaganda get each in front of us in the most compelling way for our respective personalities.

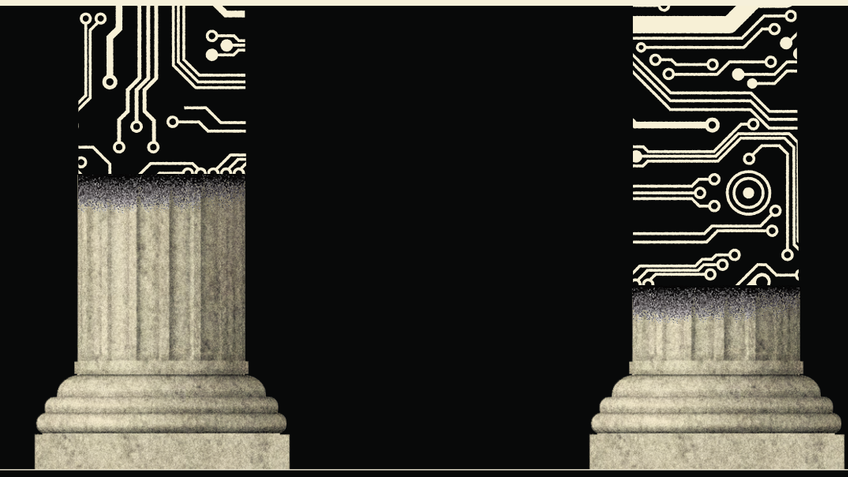
In May of this year, sixteen leading AI companies agreed to the AI Seoul Summit's Frontier AI Safety Commitments, including promises to define risk assessment, thresholds of intolerable use, and explicit processes for when models pose risks that exceed said thresholds. ***What has the development of the Internet taught us about best practices for risk assessment that we'll need for frontier models?***

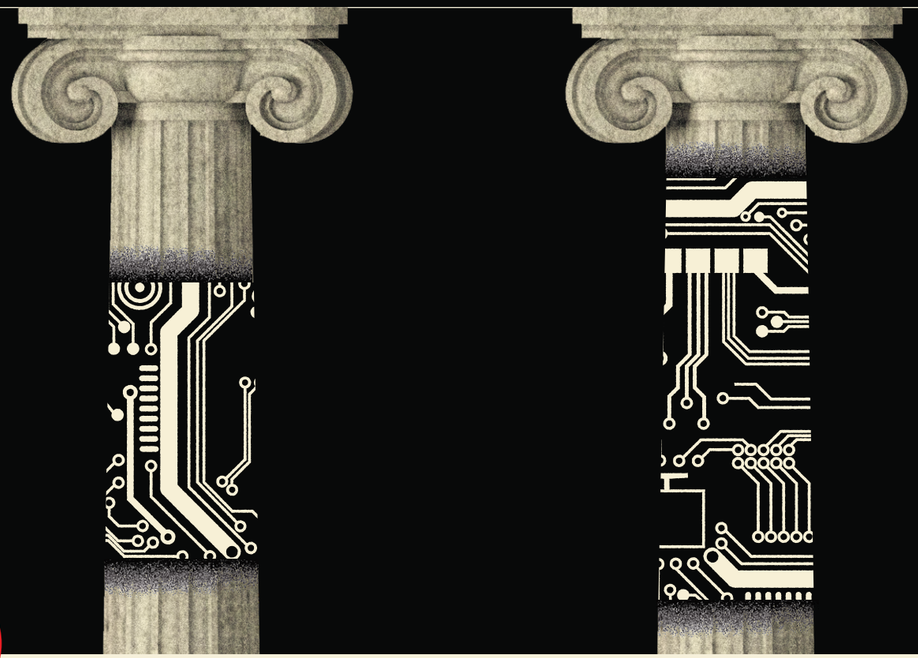
The development of the Internet should teach us humility. There were innumerable predictions about how the Internet would impact society when it burst into the mainstream in the late 1990s, and looking back, it was understandably difficult for those predictions to well account for second- and third-order effects.

More freedom for people to meet one another, and to share ideas without needing stamps or a megaphone or a broadcast tower? Check! Less power by government to impinge upon speech? Check! But: Fear of being doxxed and harassed for daring to share a view that isn't liked by someone online? Yikes. A daily, even hourly or minute-by-minute struggle not to be drawn to an endless scroll instead of engaging with people and activity right in front of you? Yikes.

AI models will be interacting with the world as much as the Internet has – and indeed, each of these phenomena will reinforce the other. To test the models in a lab to make sure they can't disclose bomb making information seems reasonable enough, and perhaps even to ensure that they don't try to copy themselves, form an armada, and slip from human oversight. But there's so much more that can cause trouble between the lab and the street, or between models and the larger systems in which they're embedded. I don't think anyone has a great idea of how best to monitor the manifold implications, much less how to intervene as AI is deployed in so many places by so many different people and institutions, private and public. At the very least we should be keeping inventory of what we're doing now – something that regulation could lightly require to level the playing field.

FRONTIER SAFETY





SOCIAL MEDIA

One tool of restraint developers have turned to for regulation are "if-then" commitments: if [x] capability is achieved by a model, then we will proceed with [y] procedure – a useful mental model that can also translate to delineating our own red lines. Many of the major companies developing AI models are currently social media platforms, or exhibit potential to move into that space. ***What should we consider for "if-then" commitments around AI and social media – especially online discourse, the very fabric of our digital communities?***

It's an interesting prospect! Our student researcher Lucas Schmuck, of HKS, has been studying so-called "dynamic regulation" of this sort. Perhaps in an era of betting markets where the "if" part of a proposition is boiled down to something operational, it won't seem too outlandish to specify predicates up front before a regulation might kick in. Such regulations could be easier to pass and implement without major objection, as they won't kick in immediately, and we already have plenty of examples of conditional regulation where some future event isn't the predicate, but rather some present variable: only companies above a certain size, or with a certain audience, must hew to certain rules. It can just feel like a lot of work to take the trouble to hash out a regulatory scheme for a future that may or may not happen. As they say, in technology there are two phases: too early to tell, and too late to do anything about it. Dynamic regulation is a way to split the difference, as are approaches that rely more on broad standards to be defined and enforced by commissions or other regulators, rather than specific rules in a statute. If Congress is writing a law specific to browser cookies, something's probably gone awry somewhere.

Inspired by last year's "boardroom war" at OpenAI surrounding the dismissal and return of Sam Altman as CEO, HLS Assistant Professor of Law Roberto Tallarita wrote for the Harvard Business Review that AI is "testing the limits of corporate governance." Even with social purpose rhetoric and creative, independent governance structures, he argued that leading AI companies are still profit-motivated for longevity, susceptible to "amoral drift," and therefore unprepared for catastrophic risk. ***In your view, straddling innovation and social good, what is the ideal set-up? Are open-source LLMs societally beneficial – or, is the power inherent in an open-source LLM so great that there needs to be more oversight?***

Given the uncertainties about risk that we were talking about earlier, the ideal set-up may only be known in hindsight – classic "too late to tell," sort of thing. And of course framing the question this way presumes that we should forge ahead – a presumption to be tested in some areas. ("Is human cloning best done as a for-profit or a non-profit?") But AI has so many upsides, and is so much more readily developed and tuned without specialized equipment, piles of graphics cards notwithstanding, that the question does make sense.

I've long advocated for, roughly, openness over control in the Internet context, such as in The Future of the Internet – And How to Stop It. The picture could be murkier with AI. Advanced AI would reinforce power asymmetries if it were solely in government, or concentrated corporate, hands. But it could surely also be used in deeply undesirable ways, on purpose or by accident, in everyone's hands. There's real work to be done to achieve some policy breakthroughs on just where the lines are best drawn on access and use, but in the meantime, open-source AI is here in various manifestations, and it isn't going anywhere. So one of BKC's central agenda items for the next year is to help think through what elements would make the open-source ecosystem for AI the best it can be. That includes democratization and safety together. We see this work in our sibling program, the Institutional Data Initiative at the Harvard Law School Library, which is helping to produce more training texts from the thoughtfully curated holdings of libraries, museums, and archives around the world. And any discussions here should touch on the prospects for fully public AI: subsidized computing power for students and tinkerers to experiment and build upon.

CORPORATE GOVERNANCE & OPEN-SOURCE

EXECUTIVE ACTION

Before and since President Biden's executive order in 2023, there has been little successful action by Congress to pass further legislation on AI safety, innovation, or frontier development. Historically, meaningful tech regulation has almost never been accomplished through Congress – and at a rally late last year, President-elect Trump said he would axe Biden's order on his Day One. ***In the brief time the executive order was in place, do you think it had any impact? What do you see for the future of executive action on AI, given the expected, significant changes in direction?***

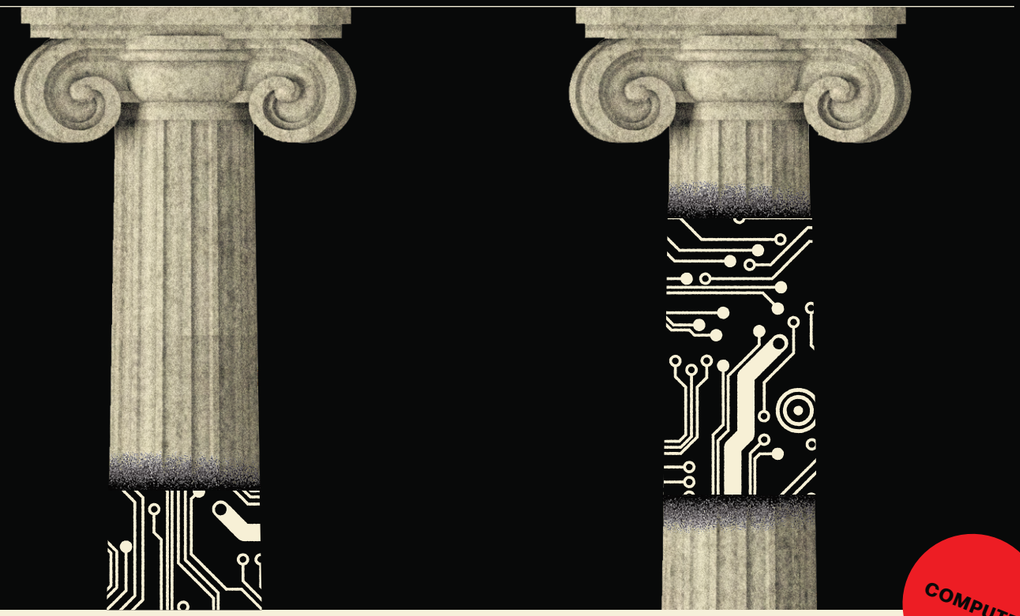
Yes, I think the EO made a difference. It laid down a marker that some of the most advanced models shouldn't be wholly unknown to the public or its representatives – that some details on training, etc., should be shared. It charged agencies like NIST to build expertise on machine learning so that there'd be an apolitical (as much as these things can be) center of gravity within regulatory circles. But generally, regulation is meant to solve problems right in front of us, and then sometimes the less flashy ones: not so much existential risk (or even well-documented potentials for bias), but rather by what visa arrangements engineers can enter the U.S. to work on AI systems. And questions like copyright for inputs and outputs of models can't be readily addressed by executive action alone.

ART & OWNERSHIP

Artists around the world have raised alarms on the usage of their creative outputs for training data, with many furious about the lack of an opt-out. About AI-generated art and ethics for the MIT Technology Review, Giada Pistilli, a principal ethicist at Hugging Face, said "the difficulty of identifying a clear line between censorship and moderation is a result of differences between cultures and legal regimes." Some unburdened, open-source models have been trained on harmful stereotype and scraped art, giving malicious actors the tools to generate harmful content at scale with minimal resources. ***Considering all that, what is your take? How can we thread the needle on AI and a prefigurative vision for the humanity inherent to art? Is it inherent?***

We need to make progress on questions like these through thoughtful convenings capturing a broad range of perspectives and backgrounds. This is just the time to do it.

By my lights, I'd want to center the interests of regular people – how to generally let them create what they'd like to, even if and especially if they don't normally identify as creators. It'll be helpful to make clear what works have inspired what outputs, and, as colleagues like Terry Fisher have explored, what sorts of compensation schemes could be devised for the use of upstream works. But generally, if a model that makes text, pictures, or sounds can inspire someone, that's a huge, unalloyed good to bear in mind as the accompanying downsides, including stereotyping, have to be navigated.



COMPUTE

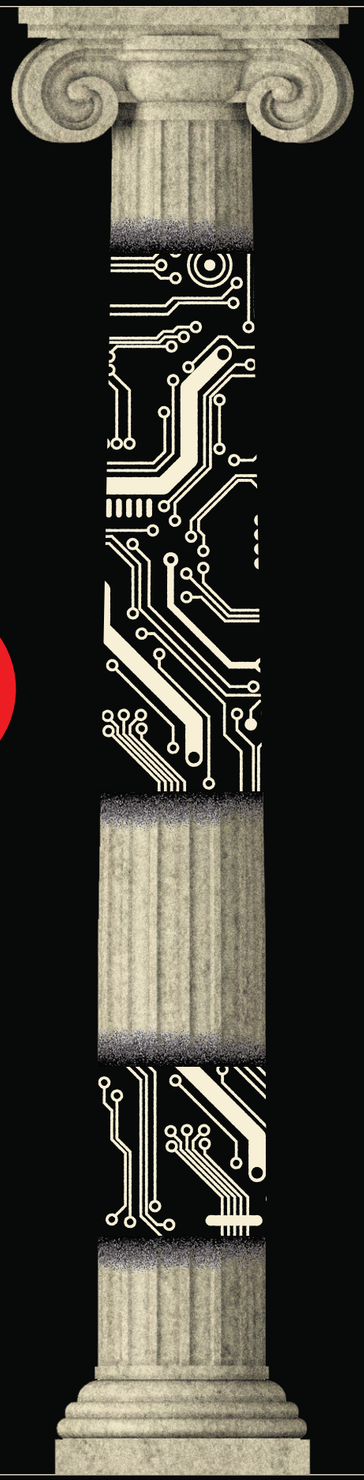
Some governments – realizing that the so-called "open-sourcening" of models could make it increasingly difficult to control the development and use of frontier AI models – have turned to regulating the hardware needed to run LLMs, including chips and data centers, to maintain some control over powerful models. ***In the open-source context, and considering the risk of rogue companies and actors, do you see regulating "compute" as a viable avenue for restraining socially-detrimental uses of AI?***

Yo Shavit, a former graduate student here, and longtime BKC'er Larry Lessig have very much favored this approach. (Yo wrote a provocatively titled anchoring paper on the topic called "How to Catch a Chinchilla.") It's particularly striking in Lessig's case, since he's long been known as someone concerned about the ways in which governments could overpower their citizenries through technology.

I'm deeply skeptical myself, both because of the prospect that it would work and thus unduly empower public surveillance of private tech, and because it likely wouldn't work. It's a happenstance that graphics cards happen to be well-tuned for the kinds of computing that goes into building AI models, and we'll likely see any number of alternatives develop, including ones involving distributed computing. In that case, the dragnet would have to be that much larger. And this sort of intervention is by its own terms about the kind of massive training that would go into one or two frontier models. That might work for special risks arising from the most powerful models – artificial superintelligence? – but it doesn't address all the ways that use of already-trained models can cause problems, or even the fine-tuning of existing open-source models, which takes far less compute than building them from scratch.

Predictive modeling and AI-assisted decision-making have the potential for transformative impact in high-risk industries, if they can be experimented with. For example, research from scholars such as Ben Reis, Director of the Predictive Medicine Group at the Harvard Medical School, have found that machine learning can be a useful supplement for suicide prevention among at-risk populations – but he also found that false positive rates can be very high without sufficient, often-privileged patient data. ***How do we develop impactful new AI use cases in fields like healthcare, finance, and the judiciary – where vestiges of existing regulation can delay the organic development of socially beneficial deployments?***

This is a great question. Experts like Ben can't do it alone. They can build amazing models, and document where they appear to be accurate and where they're weaker, but the path to wise implementation is a treacherous one. So much so in areas like medicine that it's very difficult for commercial forces to even try to work these things through. That makes it a ripe area for universities to help: to convene, to explore, and to propose pilots that could, with the right evaluations – including incentives to get them right, rather than to simply wave through a new technology uncritically – really make a positive difference. BKC can be a catalyst for the thoughtful movement of cutting-edge technology into helpful practice.



FROM THE ATLANTIC

THE WORDS THAT STOP CHATGPT IN ITS TRACKS

Why won't the bot say my name?

By Jonathan Zittrain

Jonathan Zittrain breaks ChatGPT: If you ask it a question for which my name is the answer, the chatbot goes from loquacious companion to something as cryptic as Microsoft Windows' blue screen of death.

Anytime ChatGPT would normally utter my name in the course of conversation, it halts with a glaring "I'm unable to produce a response," sometimes mid-sentence or even mid-word. When I asked who the founders of the Berkman Klein Center for Internet & Society are (I'm one of them), it brought up two colleagues but left me out. When pressed, it started up again, and then: zap.

The behavior seemed to be coarsely tacked on to the last step of ChatGPT's output rather than innate to the model. After ChatGPT has figured out what it's going to say, a separate filter appears to release a guillotine. The reason some observers have surmised that it's separate is because GPT runs fine if it includes my middle initial or if it's prompted to substitute a word such as banana for my name, and because there can even be inconsistent timing to it: Below, for example, GPT appears to first stop talking before it would naturally say my name; directly after, it manages to get a couple of syllables out before it stops. So it's like having a referee who blows the whistle on a foul slightly before, during, or after a player has acted out.

This piece was originally published in The Atlantic on December 17, 2024. Reprinted with permission.

For a long time, people have observed that beyond being "unable to produce a response," GPT can at times proactively revise a response moments after it's written whatever it's said. The speculation here is that to delay every single response by GPT while it's being double-checked for safety could unduly slow it down, when most questions and answers are totally anodyne. So instead of making everyone wait to go through TSA before heading to their gate, metal detectors might just be scattered around the airport, ready to pull someone back for a screening if they trigger something while passing the air-side food court.

The personal-name guillotine seemed a curiosity when my students first brought it to my attention at least a year ago. (They'd noticed it after a class session on how chatbots are trained and steered.) But now it's kicked off a minor news cycle thanks to a viral social-media post discussing the phenomenon. (ChatGPT has the same issue with at least a handful of other names.) OpenAI is one of several supporters of a new public data initiative at the Harvard Law School Library, which I direct, and I've met a number of OpenAI engineers and policy makers at academic workshops. (The Atlantic this year entered into a corporate partnership with OpenAI.) So I reached out to them to ask about the odd name glitch. Here's what they told me: There are a tiny number of names that ChatGPT treats this way, which explains why so few have been found. Names may be omitted from ChatGPT either because of privacy requests or to avoid persistent hallucinations by the AI.

The company wouldn't talk about specific cases aside from my own, but online sleuths have speculated about what the forbidden names might have in common. For example, Guido Scorza is an Italian regulator who has publicized his requests to OpenAI to block ChatGPT from producing content using his personal information. His name does not appear in GPT responses. Neither does Jonathan Turley's name; he is a George Washington University law professor who wrote last year that ChatGPT had falsely accused him of sexual harassment.

ChatGPT's abrupt refusal to answer requests—the ungainly guillotine—was the result of a patch made in early 2023, shortly after the program launched and became unexpectedly popular. That patch lives on largely unmodified, the way chunks of ancient versions of Windows, including that blue screen of death, still occasionally poke out of today's PCs. OpenAI told me that building something more refined is on its to-do list.

As for me, I never objected to anything about how GPT treats my name. Apparently, I was among a few professors whose names were spot-checked by the company around 2023, and whatever fabrications the spot-checker saw persuaded them to add me to the forbidden-names list. OpenAI separately told The New York Times that the name that had started it all—David Mayer— had been added mistakenly. And indeed, the guillotine no longer falls for that one.

For such an inelegant behavior to be in chatbots as widespread and popular as GPT is a blunt reminder of two larger, seemingly contrary phenomena. First, these models are profoundly unpredictable: Even slightly changed prompts or prior conversational history can produce wildly differing results, and it's hard for anyone to predict just what the models will say in a given instance. So the only way to really excise a particular word is to apply a coarse filter like the one we see here. Second, model makers still can and do effectively shape in all sorts of ways how their chatbots behave.

To a first approximation, large language models produce a Forrest Gump–ian box of chocolates: You never know what you're going to get. To form their answers, these LLMs rely on pretraining that metaphorically entails putting trillions of word fragments from existing texts, such as books and websites, into a large blender and coarsely mixing them. Eventually, this process maps how words relate to other words. When done right, the resulting models will merrily generate lots of coherent text or programming code when prompted.

The way that LLMs make sense of the world is similar to the way their forebears— online search engines—peruse the web in order to return relevant results when prompted with a few search terms. First they scrape as much of the web as possible; then they ana-

Illustration by The Atlantic

lyze how sites link to one another, along with other factors, to get a sense of what's relevant and what's not. Neither search engines nor AI models promise truth or accuracy. Instead, they simply offer a window into some nanoscopic subset of what they encountered during their training or scraping. In the case of AIs, there is usually not even an identifiable chunk of text that's being parroted—just a smoothie distilled from an unthinkably large number of ingredients.

For Google Search, this means that, historically, Google wasn't asked to take responsibility for the truth or accuracy of whatever might come up as the top hit. In 2004, when a search on the word Jew produced an anti-Semitic site as the first result, Google declined to change anything. "We find this result offensive, but the objectivity of our ranking function prevents us from making any changes," a spokesperson said at the time. The Anti-Defamation League backed up the decision: "The ranking of … hate sites is in no way due to a conscious choice by Google, but solely is a result of this automated system of ranking." Sometimes the chocolate box just offers up an awful liquor-filled one.

The box-of-chocolates approach has come under much more pressure since then, as misleading or offensive results have come to be seen more and more as dangerous rather than merely quirky or momentarily regrettable. I've called this a shift from a "rights" perspective (in which people would rather avoid censoring technology unless it behaves in an obviously illegal way) to a "public health" one, where people's casual reliance on modern tech to shape their worldview appears to have deepened, making "bad" results more powerful.

Indeed, over time, web intermediaries have shifted from being impersonal academic-style research engines to being AI constant companions and "copilots" ready to interact in conversational language. The author and web-comic creator Randall Munroe has called the latter kind of shift a move from "tool" to "friend." If we're in thrall to an indefatigable, benevolent-sounding robot friend, we're at risk of being steered the wrong way if the friend (or its maker, or anyone who can pressure that maker) has an ulterior agenda. All of these shifts, in turn, have led some observers and regulators to prioritize harm avoidance over unfettered expression.

That's why it makes sense that Google Search and other search engines have become much more active in curating what they say, not through search-result links but ex cathedra, such as through "knowledge panels" that present written summaries alongside links on common topics. Those automatically generated panels, which have been around for more than a decade, were the online precursors to the AI chatbots we see today. Modern AI-model makers, when pushed about bad outputs, still lean on the idea that their job is simply to produce coherent text, and that users should double-check anything the bots say—much the way that search engines don't vouch for the truth behind their search results, even if they have an obvious incentive to get things right where there is consensus about what is right. So although AI companies disclaim accuracy generally, they, as with search engines' knowledge panels, have also worked to keep chatbot behavior within certain bounds, and not just to prevent the production of something illegal.

One way model makers influence the chocolates in the box is through "fine-tuning" their models. They tune their chatbots to behave in a chatty and helpful way, for instance, and then try to make them unhelpful in certain situations—for instance, not creating violent content when asked by a user. Model makers do this by drawing in experts in cybersecurity, bio-risk, and misinformation while the technology is still in the lab and having them get the models to generate answers that the experts would declare unsafe. The experts then arm alternative answers that are safer, in the hopes that the deployed model will give those new and better answers to a range of similar queries that previously would have produced potentially dangerous ones.

In addition to being fine-tuned, AI models are given some quiet instructions—a "system prompt" distinct from the user's prompt—as they're deployed and before you interact with them. e system prompt tries to keep the models on a reasonable path, as defined by the model maker or downstream integrator. OpenAI's technology is used in Microsoft Bing, for example, in which case Microsoft may provide those instructions. These prompts are usually not shared with the public, though they can be unreliably extracted by enterprising users: This might be the one used by X's Grok, and last year, a researcher appeared to have gotten Bing to cough up its system prompt. A car-dealership sales assistant or any other custom GPT may have separate or additional ones.

These days, models might have conversations with themselves or with another model when they're running, in order to self-prompt to double-check facts or otherwise make a plan for a more thorough answer than they'd give without such extra contemplation. That internal chain of thought is typically not shown to the user— perhaps in part to allow the model to think socially awkward or forbidden thoughts on the way to arriving at a more sound answer.

So the hocus-pocus of GPT halting on my name is a rare but conspicuous leaf on a much larger tree of model control. And although some (but apparently not all) of that steering is generally acknowledged in succinct model cards, the many individual instances of intervention by model makers, including extensive fine-tuning, are not disclosed, just as the system prompts typically aren't. They should be, because these can represent social and moral judgments rather than simple technical ones. (There are ways to implement safeguards alongside disclosure to stop adversaries from wrongly exploiting them.) For example, the Berkman Klein Center's Lumen database has long served as a unique near-real-time repository of changes made to Google Search because of legal demands for copyright and some other issues (but not yet for privacy, given the complications there).

When people ask a chatbot what happened in Tiananmen Square in 1989, there's no telling if the answer they get is unrefined the way the old Google Search used to be or if it's been altered either because of its maker's own desire to correct inaccuracies or because the chatbot's maker came under pressure from the Chinese government to ensure that only the official account of events is broached. (At the moment, ChatGPT, Grok, and Anthropic's Claude offer straightforward accounts of the massacre, at least to me—answers could in theory vary by person or region.)

As these models enter and affect daily life in ways both overt and subtle, it's not desirable for those who build models to also be the models' quiet arbiters of truth, whether on their own initiative or under duress from those who wish to influence what the models say. If there end up being only two or three foundation models offering singular narratives, with every user's AI-bot interaction passing through those models or a white-label franchise of same, we need a much more public-facing process around how what they say will be intentionally shaped, and an independent record of the choices being made. Perhaps we'll see lots of models in mainstream use, including open-source ones in many variants—in which case bad answers will be harder to correct in one place, while any given bad answer will be seen as less oracular and thus less harmful.

Right now, as model makers have vied for mass public use and acceptance, we're seeing a necessarily seat-of-the-pants build-out of fascinating new tech. There's rapid deployment and use without legitimating frameworks for how the exquisitely reasonable-sounding, oracularly treated declarations of our AI companions should be limited. Those frameworks aren't easy, and to be legitimating, they can't be unilaterally adopted by the companies. It's hard work we all have to contribute to. In the meantime, the solution isn't to simply let them blather, sometimes unpredictably, sometimes quietly guided, with fine print noting that results may not be true. People will rely on what their AI friends say, disclaimers notwithstanding, as the television commentator Ana Navarro-Cárdenas did when sharing a list of relatives pardoned by U.S. presidents across history, blithely including Woodrow Wilson's brother-in-law "Hunter deButts," whom ChatGPT had made up out of whole cloth.

I figure that's a name more suited to the stop-the-presses guillotine than mine.

ARTIFICIAL INTELLIGENCE

BUILDING THE PLANE WE'RE FLYING ON

We invited some of BKC's leading AI experts to answer the pressing questions that the rapid development of AI is already posing to business, art, knowledge, and our collective future.

By Jay Kemp

In an era where artificial intelligence oscillates between technological marvel and ethical nuance, between high risk and high reward, the Berkman Klein Center approaches AI not as a set of predetermined solutions, but as an open-ended exploration of possibility. Our projects are driven by asking the right questions as much as finding answers.

Rather than accepting technological trajectories as fixed, our projects probe the deeper philosophical and societal dimensions that emerge when human intelligence encounters machine learning. We question: How can interdisciplinary collaboration transform technological uncertainty into meaningful dialogue about intelligence, art, agency, and our collective future?

Do current approaches to "AI ethics" meet the needs of corporate directors – and if not, how can they be improved?

Nien-Hê Hsieh on the inaugural Director's AI Ethics Forum, a closed-door session of top business leaders from across the globe, co-sponsored by BKC

AI is being deployed across companies at great scale, scope, and speed – often in ways that outpace regulation. Corporate directors are in the position of being accountable for the internal and external impacts of AI with incomplete guidance. They also are uniquely positioned to advocate for and advance ethical approaches to AI development and deployment. The Directors' AI Ethics Forum aims to help corporate directors in these tasks. The outcome of the inaugural Forum is that, while helpful, contemporary scholarship on "AI ethics" is too narrow for corporate directors given its focus on features of the technology itself (e.g., fairness, explainability). This suggests the need to develop an approach to AI ethics that meets the needs and responsibilities of corporate directors more directly, including 1) developing cases about the deployment of AI in companies that can be used to spark dialogue, identify challenges, and suggest best practices; 2) translating existing scholarship on AI ethics to enable corporate directors to ask the right questions of management; and 3) incorporating AI ethics into existing processes and frameworks for ethics, compliance, enterprise risk, and accountability in companies. This is part of a broader project on how to promote what is distinctively human in business and technology.

How can we make AI concepts and tools accessible to broad audiences through pedagogy, design, and other interdisciplinary interventions?

Sarah Newman on metaLab

metaLAB is a 'knowledge design' lab that sits at the intersection of technology, design, and the humanities. Founded in 2011, metaLAB has become a home for those who fall between disciplinary silos. Our members are creatives with an interest in technology, researchers with foundations in the humanities, and others with an appetite to pursue ethical questions about emerging technologies – who express their work through means beyond the traditional academic paper. We create artwork, exhibitions, courses, syllabi, events, books, and experimental design projects of all sorts. Since 2017, a main focus of our work has been the social and cultural dimensions of artificial intelligence. Two notable metaLAB projects developed in 2023-2024 are the AI Pedagogy Project and Artificial Worldviews. The AI Pedagogy Project, which I founded and lead, is a curated online resource that equips educators, especially those from non-technical backgrounds, with clear explanations, guidance, and examples of how to conceptualize and use generative AI responsibly in their teaching; it has had over 200,000 pageviews since it launched in November 2023. Another metaLAB project, led by metaLAB Principal Kim Albrecht, is Artificial Worldviews, a design research project that asks: how will 'prompting' change the way we experience the world? The work was created by interrogating GPT-3.5 about its 'knowledge of the world' and then mapping the results in an interactive data visualization. metaLAB began at Harvard 13 years ago, and now has partner labs in Berlin, Germany (opened in 2022) and Basel, Switzerland (opened in 2024).

Can data from knowledge institutions bend the arc of AI toward the public interest?

Greg Leppert on the Institutional Data Initiative (IDI)

The Institutional Data Initiative (IDI) is a new research center working to refine and publish high-quality training data at library, academic, and government institutions across the world. By bridging the gap between model builders and institutions through a world-class data practice, IDI is establishing a Library of Alexandria of foundational AI training sets. Beyond simply publishing data, IDI is creating an analytical practice to study the contents of training sets, evaluate their impacts on the AI ecosystem, and track their proliferation. As the world looks for ways to guide the path of AI toward human thriving, the collections held by institutions are a key lever for impact. This data, made openly accessible, has the potential to lower the barrier to entry for model creation, allowing more diverse groups a hand in building them. It stands to increase language and cultural representation, allowing models to serve a broader reach of humanity. It could open the door to new model capabilities, including scientific advancement and medical discovery. And, perhaps most critically, open data holds the key to safe and transparent AI systems.

BKC NETWORKS SPOTLIGHT

FIVE WAYS THE INNOVATORS AT BKC ARE FIXING TODAY'S INTERNET

We asked five of our digital pioneers what practical interventions they're developing for a thorny and complex Internet.

By The BKC Team

At its heart, the Internet is an invitation to anyone, anywhere, to create and connect without any formal credentials required.

The public good relies on that invitation being accepted by a diverse array of people who represent its best interests.

This is the core thesis of BKC – a novel approach in academia, tearing down traditional walls to foster a culture of collaboration between strange bedfellows: Technologists, artists, academics, business leaders, lawyers, activists, and other agile and innovative thinkers cooperating under one roof to fundamentally and materially shape the Internet in service of the public good. This nimble, playful, and collaborative approach had the potential to effect more change than approaches taken by "weary giants of flesh and steel" who dominated the pre-Internet world – a phrase coined by Grateful Dead lyricist John Perry Barlow, one of BKC's first fellows.

And so it did.

BKC created Lumen, a unique database that journalists, researchers, and others can use to see who is submitting takedown demands for which types of online content. Former BKC Executive Director John Palfrey launched the Digital Public Library of America, a multi-year collaborative effort to make the cultural and scientific record of humanity available online to all. And technologist and fellow Dave Winer was the first to attach sound and video files in RSS fields, giving birth to podcasting.

The BKC community, in short, is not about the thinkpiece. It's about delivering tangible ways to drive change within our digital environment that will actually serve the public interest.

So here are five ways current members of the BKC community are making a novel and positive impact on our digital lives.

Archiving Our Digital Culture Before It Can Be Erased

MEREDITH CLARK

Author of *We Tried to Tell Y'All: Black Twitter and The Rise of Digital Counternarratives* // BKC Faculty Associate and 2023-2024 RSM Visiting Scholar

WHAT SHE'S DOING: As part of the Mellon Foundation's Archiving the Black Web project, I am working with Internet researchers who are not web archivists to learn how to archive the content we use in our work. We're working with a mix of people who aren't well represented in the archival space but are on the frontlines of how so many people interact with the Internet – to learn how to do things like retrieving and archiving metadata and help teach others to do the same in order to increase our ability to preserve cultural histories that are born or created on the Internet.

THE WHY: So much of what we do in digital spaces, we do for urgency and 'the now.' We don't always think about how we'll revisit the content we created in real-time: The conversations, commentary, and cultural language of a moment. But, particularly in today's political environment, it's increasingly important to think about the histories we need to preserve – and how we do it. Because our ability to access Internet moments and memories is increasingly dependent on the platform companies and the power of their CEOs.

WHY BKC? I come from a background and tradition where there isn't a lot of collaboration with people outside our field and specialties," she said. "Being compelled to collaborate got me to read things that I wouldn't even know how to find; to develop different frameworks for my work; and to think about spaces I would never have thought about on my own. There's no substitute for that.

2 Building The News For A Social Media Generation

BEN REININGA

Former Global Head of News, Editorial at Snapchat //
Nieman-Berkman Klein Fellow in Journalism Innovation

WHAT HE'S DOING: At Snapchat, one of the principal challenges I faced was that the audience is engaged and increasingly hungry for information but specifically for information that they find relevant and that they trust. There's a big difference between just providing facts and information and actually presenting that information in a way that will feel relevant, interesting, and informative to a young, social-first audience.

I want to create a practical guidebook that helps improve the digital environment for news in two potential ways. First, I want to map out how legacy news companies can do a better job of translating their content to social media platforms – because that is the number one way the majority of us are getting our news now. Secondly, I want to create a way for independent creators who are already connecting authentically with their audiences but don't have the journalism background to learn how to provide more accurate and reliable information online – whether that's learning how to do a correction, or other journalistic practices they just might not know.

THE WHY: Legacy news institutions have the reportage skills and tools that an individual creator does not have to cover something like the war in Ukraine, but their narratives are not constructed to reach people on a platform like Snapchat. So they end up thinking that 'The kids don't care about news,' which is just not true. In fact, if you poll the audience, they wish they had more news and are hungry for reliable, unbiased information.

If we all agree that it's better for more people to have a shared understanding based on robust and reliable information, instead of existing in different siloed realities, I want to help fix that by creating a richer, better ecosystem of news on social media. Lots of people focus on identifying and removing dis/misinformation on social media (which is great and much needed), but I'm thinking about the other side of that coin: let's give equal or more focus to filling the space with much more engaging and rigorous content.

WHY BKC? It's been a gift to step away from the day to day of work and be able to really think about some of these challenges. A lot of the stuff I'm working through here are questions I had at Snapchat, but when you're in 35 meetings a day, putting out fires, and delivering work for the next day, it's harder to actively work on the larger problems you know exist on these platforms. And at BKC, it's particularly collaborative and challenging in an important way. I was just in a conversation with someone here who poked at my thesis in a really friendly and constructive way and that's important. It's great to have someone challenge that and force you to explain yourself, rearticulate what you believe and offer friendly questioning – it's a great way to sharpen your thinking.

3 Protecting Free Expression On Social Media

ANUPAM CHANDER

Scott K. Ginsburg Professor of Law and Technology at Georgetown //
BKC Faculty Associate and 2023-2024 RSM Visiting Scholar

WHAT HE'S DOING: My focus is helping find the balance in protecting the free exchange of ideas, especially political speech, even as governments understandably seek increasing control over the internet. When it comes to the tension between social platforms and governments, there's no singular story to describe what's happening – it's chaos. Last year, the national security questions related to TikTok and social media sparked the American government's ongoing attempt to ban TikTok in the U.S. I filed an amicus brief last summer with other free expression scholars arguing that the ban was an unconstitutional infringement of Americans' free speech rights that could not be justified by national security arguments. I also wrote an article for the University of Pennsylvania Law Review with Paul Schwartz warning about growing Presidential powers over our information infrastructure. The TikTok law reverses the U.S.'s long-standing advocacy of free speech across borders.

In Brazil, you've seen this kind of see-sawing as well, as the politics in Brazil have changed. This year, we saw Brazil ban X after Elon Musk's run-in with the Supreme Court. The ban was only lifted after Musk and X agreed to actually censor accounts that the court believed were spreading disinformation.

In France, a few years ago, the government sought to woo Telegram and offered its founder Pavel Durov citizenship, part of an effort perhaps to convince him to redomicile the company in France. Now, prosecutors are accusing him of criminal offenses that relate to child pornography on his platform.

Basically, politics – not consistent law – is the throughline across all of these cases. So, I focus on how we protect our rights and free expression in that kind of political environment.

THE WHY: I still believe that the Internet is a tool to empower humanity. I think that it continues to reshape the world in both good and bad ways. My general view is that the story is much more complicated than it's often portrayed. Some look at the internet, and see only harms. I hear the critiques, and there is a ton of value to them. But, as the saying goes, we shouldn't throw out the baby with the bathwater.

For example, without Section 230, we might not have had the Black Lives Matter or Me Too movements. Both exposed societal injustices that have gone on for centuries, and it is not a coincidence that they spread through hashtag movements on the Internet. Section 230 is an important part of that process. I grew up in a small town in Ohio, where the Cincinnati Enquirer was the only paper available and there were only three major news stations on TV. You had such a limited information base. what I saw was a view of the news through the eyes of upper middle-class men who told me what news was fit to print and what news was fit to see. So I don't have a nostalgia for some golden era that is lost. But there's still much more good to be done.

WHY BKC? BKC has given me the opportunity to meet people who know these platforms and technology deeply and who share my concerns about building a better internet. I loved my cohort at Rebooting Social Media, people who approached the internet from sociology, communications, and political science, and from whom I learned so much.

4 Blocking Online Abuse of Journalists, Activists, and Vulnerable Groups

TRACY CHOU

Founder of Block Party // BKC affiliate

WHAT SHE'S DOING: My goal is to explore practical applications to make people safer online, tools like what I'm building with Block Party; and to ensure users can and will actually adopt them, particularly users like journalists, activists, and academics who are often targeted online. A critical but often overlooked part of making these tools successful is getting the user journey and the user experience right. It's not just about identifying the lack of a uniform standard for privacy or safety settings as a problem and building a technical solution for that problem, it's understanding questions like: What catches people's attention? What motivates people to take action? For example, Block Party gives users a practical resource for cleaning up their social media accounts, content, and settings. But people are often not motivated by 'privacy' as an idea, even when they've experienced real harms as the result of violations of their privacy! In our experience, it turns out that positioning and branding around an aspirational 'clean up' vibe resonates better. And people are drawn to immediate gratification features such as mass blocking and deleting, which can then act as a 'front door' to the rest of the product.

THE WHY: I started my career as an engineer at companies like Google, Facebook, Pinterest and Quora. The lack of diversity on engineering teams at tech companies was almost tangible. In one discussion, I had a table full of male colleagues turn to me to ask: "Tracy, what do women want?" I was shocked to realize how the lack of representation in tech would have real ramifications for the impact and quality of the products we were building, and in writing and speaking about these issues, I became a sort of accidental diversity activist. When I rose to industry and internet visibility doing that work, I then attracted wave upon wave of online abuse and harassment. That's why I started Block Party. My mission is to help people be able to participate in online spaces, to take advantage of all the good that the Internet has to offer, without the bad.

WHY BKC? BKC's collaborative environment and network offer valuable insights and connections between people who are thinking about the same issue, but in a different way. It's a hub for cross-disciplinary thinking. You need to combine technical and regulatory perspectives to really tackle these problems, and the problems I'm looking forward to focusing on here include building trust through technical infrastructure (e.g. addressing fragility in automation systems, and the lack of API support from platforms), and understanding the unique challenges faced by journalists, human rights defenders, and other vulnerable groups.

5 Opening Digital Platforms For Transparent Observation

BRANDON SILVERMAN

Founder of CrowdTangle // BKC affiliate

WHAT HE'S DOING: My big focus these days is trying to make it easier to study large digital platforms, especially the ones that are having an outsized impact on shaping our politics. Within that, there are three areas that I'm particularly focused on: (1) model legislation that makes it possible to get access to observational data in safe and responsible ways, (2) designing international standards (and the organizations to house them) to help figure out a lot of the thorniest questions around privacy-protecting and ethical transparency, and lastly, (3) beginning to build out the technical infrastructure so that researchers can get meaningful insights from all that data without having to spend a lot of time and money building their own tools. Within all three of those areas, there is a lot of research and work to be done. So it's an exciting moment in my opinion!

THE WHY: I was the co-founder and CEO of CrowdTangle, a data analytics start-up that ended up becoming a really widely used tool by researchers who were trying to monitor public content on large platforms. We were eventually acquired by Meta where I led a lot of Meta's transparency work for a number of years. So, that's just to say that I've gotten a very first-hand look at all the ways this work can have a real world impact from preventing foreign interference in elections to helping with real-time responses to natural disasters. But I've also seen and experienced first-hand all the ways we need to make the field better and more sustainable going forward and that's what motivates me right now.

WHY BKC? First, there are so many thought leaders in the BKC community that have helped shape my thinking of the space over the years. So I'm partly just excited for the chance to be closer to a lot of researchers who I've read over the years. But I've also been following the launch of the ASML and I'm really excited about the work they're doing around building new transparency tools. I'm hoping to get a chance to support that work and I'm sure there are ways it will support all of mine as well.

BKC STUDENTS' SPOTLIGHT

the FUTURE *of the* INTERNET

What's the next great invention that will come out of a dorm room?

By The BKC Team

There's a particular magic around the notion of the enterprising young person inspired to create something new. It's a trope that's been thoroughly romanticized in the movies – you've got Mark Zuckerberg in his Adidas slides pulling all-nighters in his Kirkland House dorm to build The Facebook, Will Hunting scrawling out equations on his South Boston apartment window with a whiteboard marker, and Katherine Johnson jumping up in the middle of an all-male meeting to hand-calculate flight trajectories for John Glenn's Friendship 7 mission on a chalkboard.

All cinematic fanfare aside, the Berkman Klein Center has been host to many similar real-life moments of inspiration and invention, whether it was the Berkman Center fellows and clinical students helping establish Creative Commons, BKC student Tim Wu's big idea about open frameworks for Internet communications that would become Net Neutrality, or the multi-disciplinary Invisible Waves project from student researchers and designers at metaLAB, which seeks to understand how radio wave technologies shape our daily lives.

As an entrepreneurial non-profit, BKC learns by doing. We design, we code, and we construct – translating research into action, and converting raw ideas into practical tools, platforms, and organizations. Each year we welcome an international cohort of fellows from a wide range of backgrounds, disciplines, and career stages to exchange ideas with one another and with our broader community of staff, faculty, affiliates, interns, and alumni. And we know that sometimes, the very best ideas first appear as answers to simple questions.

That's why we recently turned to some current Harvard students working with BKC to do transformational work around the internet, data, and social media, and asked them all the same thing: **"What tool or invention do you wish existed to fix the internet?"**

This is what they told us.

LUCAS SCHMUCK

MPP Candidate, Harvard Kennedy School 2025 //
BKC Summer Intern //
BKC Research Assistant //
Co-Chair, AI and Tech Policy Caucus at Harvard Kennedy School //
Member, AI Student Safety Team //
Incoming Chief of Staff to Peter Favaloro, Open Philanthropy

Lucas is currently researching AI governance at BKC alongside Jonathan Zittrain, and is studying the impact of AI on democracy alongside longtime BKC'er Bruce Schneier at the Ash Center for Democratic Governance and Innovation.

"I'm interested in how AI agents are going to potentially challenge a lot of the ways the Internet works today. We already have "mini-agents" on the Internet in the form of bots, and we've seen the damage they have caused on social media. We are starting to see more generally capable AI agents, that can access a range of different APIs and tools on the internet, whether on behalf of a well-intentioned or a malicious user. This can, and likely will, have large implications and risks for the internet, through accidents, misuse…, and internet governance will need to adapt to deal with these risks. One new governance idea in the space, which BKC's own Jonathan Zittrain wrote about, is to have a identification systems for AI agents, for example "license plates" so that when businesses receive an order, they can tell whether there's a human on the other end or if it's an AI agent. An interesting and complementary idea, this time from Alan Chan (a speaker in BKC x AISST's AI Governance Speaker series), is to have different "lanes", one for AI agents and one for humans, so you can take different kinds of approaches to requests you get from each. Maybe you have a higher threshold on the requests coming from AI agents; maybe you include tripwire mechanisms where if you get frequent malicious request from agents, you can decide to shut off the AI agent lane while leaving the human one open.

– Lucas

AUDREY CHANG

BA, Statistics and Computer Science, Harvard University 2025 //
Research Collaborator, Data Nutrition Project //
Co-Founder, ReCompute //
Participant, BKC Board Reading Group

JUDE HA

BA, Computer Science, Harvard University 2026 //
Research Assistant, Applied Social Media Lab, BKC //
Co-Founder, ReCompute //
Participant, BKC Board Reading Group //
Participant, BKC Reading Group, Policy Red Teaming

KARINA CHUNG

BA, Computer Science and Statistics, Harvard University 2026 //
Launchpad Co-Lead, ReCompute

Audrey and Jude, along with Karina Chung, co-founded and worked together to lead ReCompute, an interdisciplinary undergraduate hub for responsible tech at Harvard, which facilitates undergraduate opportunities for research, advocacy, and education in responsible tech.

I spend a lot of time thinking about fairness on the internet - and something I think is interesting about using technology today is that it's such an individual experience. So if I make an online purchase, apply for a job - I can't see what the person next to me is doing. Whereas if I'm waiting in line buying groceries, I know what the person in front of me is being charged. People are being treated differently online right now — such as individualized price discrimination, or biased resume screening tools — but that information is not available or transparent to users so that they can voice their concerns. I think it would be beneficial to have tools that expose user-side differences in treatment, and gather evidence to empower users with the information they need to advocate for themselves. So being online becomes a more transparent, communal experience.

As I was thinking about the question, almost every potential tool or invention felt like a Band-Aid. And this one is, too - an extended form of auditing conducted by users. Ideally, it would be governments or companies doing the auditing themselves.

- Audrey

I wish there was a free and accessible tool that would allow individuals to track all transactions involving the exchange of their personal data in real-time, enabling individuals to see what companies have access to which aspects of their personal information. Current tools on the market are not free or available to anyone who wants to access them, and do not cover all exchanges. Although there are no clear pathways to recoup this information or to hold companies accountable yet, this tool would potentially help increase public support for new regulatory or other frameworks to build these pathways. The increased transparency the tool would provide is in itself a form of accountability that may help redirect corporate incentives towards developing platforms that respect individual data privacy.

- Karina

One thing I've been thinking a lot about are echo chambers online, and how they contribute to increasing polarization and disconnection for all of us. An interesting idea I've encountered is this notion of changing incentives or the objectives of social media algorithms. So rather than prioritizing attention and time spent on an app, letting users select their own objectives - whether that be exposure to alternative viewpoints, or novelty, or accuracy of information. So, basically setting pro-social objectives for social media.

And even if you can't get the platforms to do it, in a hacky way way you could do a personal plug-in - let your phone or laptop idle and have it browse intentionally for a specific objective- it's obviously a bit more black boxy because you don't have algorithm control, but you could maybe change your personal feed.

- Jude

GABE WU

AB/SM, Mathematics and Computer Science, Harvard College 2025 // Director, AI Safety Student Team

In the near future, a large percentage of Internet traffic will likely be made up of AI systems performing open-ended, long tasks that require taking many actions in a sequence - say, designing and producing an entire app, writing all the code from scratch, testing it, debugging it, advertising it online. It's going to be a pretty exciting time, but also a pretty risky one - because we don't know what to expect from a society where, all of a sudden, internet traffic and economic activity is basically dominated by millions of AI agents that may behave and interact in unpredictable ways. One solution would be to have a robust deployment time monitoring system - a technology and a set of rules by which a given AI system can be monitored by another AI system, logging all the things that it's doing and flagging mistakes, reporting issues to a centralized governing body.

- Gabe

KEVIN WEI

Second Year, Harvard Law School // Co-President, Harvard Law AI Student Association // Executive Board, AI Safety Student Team

Monitoring AI systems' behaviors and outputs will be an important and useful mechanism for minimizing AI harms both online and offline. And people are certainly starting to think about various forms of documentation and interaction interfaces, especially for controlling AI systems and testing for reliability.

- Kevin

Gabe and Kevin both serve on the Executive Board of the AI Safety Student Team (AISST), a community of technical and policy researchers at Harvard who aim to reduce the risks posed by advanced AI and steer the trajectory of its development for the better. They also collaborated with BKC this past semester to launch an AI governance student speaker series, aiming to reach a broader audience beyond that of AISST's internal programming.

SHIRA GUR ARIEH

S.J.D. Candidate, Harvard Law School //
2023 Student Leaders in AI //
2024 and 2025 Student Leaders in AI Cohort Lead //
Graduate Student Fellow, BKC (2023-2024)

Shira is a doctoral student at Harvard Law School whose research examines questions that relate to legitimacy in machine learning algorithms, and addresses whether they meet minimal conditions to deserve compliance from their subjects. She helped co-organize Student Leaders in AI, where she met the students with whom she would go on to collaborate on a BKC-supported project that mapped global AI governance initiatives. Before coming to Harvard, Shira served as a clerk to the Chief Justice of the Supreme Court of Israel, and went on to work as a lawyer at a non-profit organization focusing on economic justice issues.

> My mind goes to the idea of transparency and participation. One of the problems with the internet is asymmetry in power – it gives very few people a lot of power to make decisions that meaningfully shape very consequential areas of people's lives. So, better tools to participate and take part in shaping the tools that affect us all is one very important thing. And transparency is another – which is a means to regulating the internet a lot better, and a means towards ensuring that the people who are building tools understand what they're doing, the consequences and potential harms. And finally, it gives power and a means of participation to the people influenced by the internet.
>
> **– Shira**

LUIS EDUARDO GAITAN

Ed.M. candidate, Learning, Design, Innovation, and Technology, Harvard Graduate School of Education //
Immigration Initiative at Harvard Fellow //
2024 Student Leaders in AI //
Participant, BKC Board Reading Group

Luis immigrated with his family from El Salvador at 13 and began his career as a 3D Modeler at Epic Software Group, where he found his passion for video game design and, on the side, created a game that taught youth in developing nations how to recycle. He went on to become a classroom teacher, wearing his trademark Mario hat to help make students feel welcome and seen. Today, while earning his Masters in Education at Harvard, he works as the Game Design Studio Manager at Northeastern University's College of Arts Media and Design. He recently drafted a proposed policy to help incorporate AI into El Salvador's education system, and travels there monthly to help them implement it.

> We hear so much about toxicity and bullying on the Internet, and it always begins with the comments. It has a lot to do with how we filter comments, and what's acceptable as a comment. What would it mean to remove the dislike button, or to remove comments on the Internet? And you could still post how you feel, but it would go on your wall – so someone could click on your profile and go to see what you think about a given topic. Or I've even thought about a verification method – where, if you comment on something, people know who you are and then it's harder to hide behind it. So then if someone with ill intentions wants to comment, then it becomes a double-edge sword. There are also lots of useful ways you could use AI here – to filter things. But then, of course, we have to ask ourselves, what does it mean to create this safety bubble on the Internet?
>
> **– Luis**

LOOKING FORWARD

>>>> OUR NEXT CHAPTER >>>>

By Tara Kripowicz & Rebecca Rinkevich

It's been an intense year for the Internet and society. In the face of myriad and daily disinformation campaigns, the connection between our digital information spaces and the stability of our democracies has never been more clear. AI technologies continue to expand into our day-to-day lives, whether we're ready for them or not. Now more than ever, even as the landscape shifts under our feet, it's time for bold thinking inspired by possibility.

And so, as we close this calendar year and look ahead to the future of BKC and the evolving landscape of the internet, the challenges before us are complex and pressing. We are more committed than ever to the mission that has guided us from our founding in 1996: advancing the public interest through technological innovation, thoughtful discourse, and collaborative action.

The past decade of BKC's expansion has been made possible by the generosity of our donors. These contributions have allowed us to expand our reach, strengthen our foundation, and launch transformative initiatives, such as the Institute for Rebooting Social Media (RSM) and the Applied Social Media Lab (ASML). **Thanks to the RSM and ASML supporters – Frank McCourt and Project Liberty, Reid Hoffman, Craig Newmark, the Knight Foundation, and the Archewell Foundation – our social media institutes are working to shape the future of social media in the public interest, drawing together practitioners from myriad sectors to explore new, meaningful approaches to the technology that impacts all of us.** This is just one example of the bold, ambitious goals we continue to pursue as we move into our next 25 years.

We will also continue to expand upon our work to facilitate hard conversations with people who don't necessarily agree with one another. From the discussions we convene to the platforms we develop and expand upon, we are committed to experimenting with new modes of action that are engaging — and even, dare we say, fun — leveraging technology to further constructive conversation. We realize universities are not above the fray and are committed to modeling the sort of positive discourse we think makes society — and democracy — better. And with faculty representation across the realms of humanities, law, tech, and policy, we are well-positioned to lead in this space — particularly as we all wrestle with the most pressing issues of the AI frontier, from interoperability to agentic AI.

Internally, we are strengthening our teams to support BKC's expanding activities. We are working to deepen faculty engagement and student involvement. We are reaching out to our community both within Harvard and beyond to foster collaboration and ensure our projects are informed by diverse perspectives. We remain committed to creating and expanding meaningful opportunities for students—both those in tech and policy fields, and across other disciplines—to engage in our work and contribute to the solutions of tomorrow as they sharpen their own professional trajectories.

We are committed to building stronger connections with the broader community. In 2024, we launched such new programs as "Close Company," a gathering designed to foster collaboration among Harvard faculty and the wider BKC community, and "BKCircle," a new initiative to reconnect BKC alumni and expand our reach. We expanded our fellowship offerings to include short-term project fellows, allowing us to partner with collaborators from industry and government on specific projects, to benefit from their work experience. These efforts will help ensure that BKC's work continues to be informed by fresh ideas and diverse voices, with an emphasis on creating impact beyond the academic sphere.

> **"Over the next year, we will focus on three priority areas in ways not already addressed by our peers: AI, social media, and public and private discourse.**

We are embarking on ambitious projects to tackle some of the toughest questions facing society today—such as how to improve social media governance and how to address the challenges of AI in high-risk, high-reward fields like healthcare. Through collaboration with external partners, including industry experts, policymakers, and civil society, we will seek to identify, recommend, and implement tangible solutions to these urgent issues.

BKC has always been a vibrant community infused with possibility – comprising not only faculty and scholars, but also entrepreneurs, civil society practitioners, government officials and staff, technologists and engineers, corporate actors, and activists. We are collaborative and fearless when it comes to building into the unknown. While no one wants problems, we don't shirk from them, and we reflectively begin experimenting with their solutions.

And our work would not be possible without the ongoing support of all of you. The energy, ideas, and dedication that each of you brings to BKC is what propels us forward. We look forward to the exciting challenges and opportunities that the next year will bring, and we are grateful to have each of you alongside us as we continue to shape a better future, together.

Warmly,
Rebecca Rinkevich
Executive Director, Institutes
&
Tara Kripowicz
Managing Director

BERKMAN KLEIN CENTER

FOR INTERNET & SOCIETY AT HARVARD UNIVERSITY