Report to the President year ended June 30, 2024, Computer Science and Artificial Intelligence Laboratory

CSAIL leadership spearheaded several new initiatives in FY2024:

- The CSAIL Imagination in Action Symposium was held on June 7 2024, showcased the latest ideas and results from CSAIL's PIs and researchers, with a strong emphasis on AI advancements. This outward-facing event aimed to bring the most innovative concepts from CSAIL to a global audience, including industry professionals, the press, and the general public.
- CSAIL Impact Book has been published.
- MIT's Generative AI week has been a huge success and a collaborative experience with many units across MIT.
- CSAIL launched the AI Accelerator program Phase 2 for 5 more years
- The CSAIL - Wistron Collaboration was extended for another 5 years

Research Initiatives:

- NGSE@CSAIL has been in the process of launching to focus on performance and efficiency of foundational elements in a post-Moore's Law world (e.g., operating systems, workload scheduling, utilizing general processors vs accelerators, compiler optimization), compatibility with existing applications and systems for scale and frameworks for measuring performance, efficiency, and carbon impact in dedicated and distribute workloads, tradeoffs among security, privacy, and efficiency, opportunities to review and improve code embedded into hardware, smarter approaches to model training (sustainability) ease of use, wider training of HPC ,approaches and relevance in common software development methodologies.

- Our initiatives FintechAI@CSAIL, Machine Learning Applications (MLA@CSAIL), and Future of Data, Trust and Privacy are successfully continuing to grow and aims to develop technical solutions, work to establish industry standards for the responsible, transparent, and accountable use of AI, and engage in dialogue with regulators through workshops and events. The initiative intends to maintain an interdisciplinary focus that brings together thought leaders from industry and government with MIT faculty, researchers and students, conducting research in multiple areas of AI such as predictive analytics, machine learning, neural networks, robustness, explainability, ethics, security and privacy.

A Few Research Highlights (more available at csail.mit.edu/news):

- **Using AI to protect against AI image manipulation**
  CSAIL’s PhotoGuard provides a preemptive countermeasure to malicious image manipulations, using perturbations — minuscule alterations in pixel values invisible to the human eye but detectable by computer models — that effectively disrupt the model’s ability to change the image.

- **Exploring the mysterious alphabet of sperm whales**
  Sperm whales communicate primarily using sequences of short bursts of clicks, known as codas. MIT CSAIL and Project CETI used algorithms to decode the “sperm whale phonetic alphabet,” revealing sophisticated structures in codas akin to human phonetics and communication systems in other animal species.

- **Study: Smart devices’ ambient light sensors pose imaging privacy risk**
  A CSAIL team found that ambient light sensors, which are embedded in the screens of our smart devices to alter brightness, pose an imaging privacy threat. Their algorithm revealed how the sensors can capture images of users’ touch interactions, like swiping and tapping, for hackers, so the researchers suggested that developers tighten up app permissions and reduce the precision and speed of these sensors while placing them on the side.

- **Precision home robots learn with real-to-sim-to-real**
  Household robots must adapt to their specific environments, but traditional training methods require extensive real-world data collection or costly simulators. CSAIL’s RialTo system is a more efficient workaround: it uses "digital twin" simulations from minimal real-world data captured on users’ phones, potentially helping train robust robot policies on-the-fly for homes and warehouses.

- **Generative AI imagines new protein structures**
  MIT CSAIL researchers developed “FrameDiff,” a computational tool for creating new protein structures beyond what nature has produced. Much like image generators, the machine learning approach incorporates “frames” into diffusion models, aligning each one with the inherent properties to construct novel proteins independently of preexisting designs.

- **New tool empowers users to fight online misinformation**
  CSAIL researchers designed a framework for identity disclosure in public conversations called “meronymity” that can empower people to speak up without the drawbacks that come from full anonymity. They implemented meronymity in a communication system they built to help junior scholars use social media to ask research questions.

- **Robotic palm mimics human touch**
  To help robots grip and grasp more like humans, MIT CSAIL researchers developed GelPalm, a versatile new design that uses sophisticated tactile sensors in the palm and agile fingers. GelPalm’s combination of articulation and sensing could potentially improve human-robot collaboration, prosthetics, and

robotic hands with human-like sensing for biomedical uses.

- **A blueprint for making quantum computers easier to program**
  MIT researchers highlight why it's difficult to translate quantum algorithms into code for quantum computers: they don't follow the same rules for completing each step of a program in the order that regular computers do. CSAIL's new abstract model presents a quantum computer with reversible instructions, potentially bringing us closer to making their programs as easy to write as those for classical computers.

- **Boosting faith in the authenticity of open source software**
  MIT researchers' new system, Speranza, aims to reassure software consumers that the product they are getting has not been tampered with and is coming directly from a source they trust. They achieved privacy via "identity co-commitments," where a software developer's identity (an email address), is converted into a "commitment" while another big random number — the co-commitment — is generated.

- **Rethinking AI's impact: MIT CSAIL study reveals economic limits to job automation**
  A new study from MIT FutureTech shows that human workers may be safer from short-term AI automation risks than previously thought. Focused on workplace tasks where computer vision is a key component to the work, they found that only a quarter of these jobs are economically viable for firms to replace with AI.

CSAIL Growth

The total combined research volume (Primary and Secondary funds) is $95,638,711 for FY24. CSAIL's research volume is 64% federal and 36% non-federal. CSAIL manages approximately 596 active research accounts and over 138 PIs with appointments across 11 MIT departments. Through 2023-2024 academic year we had 741 graduate students with RA appointments in CSAIL, and 461 UROP students.

CSAIL research is sponsored by many diverse sources, from US government contracts to the private sector.

New United States government sponsor – United States Air Force Academy (new prime sponsor, not a direct sponsor)

New US and international non-federal sponsors include:

Foreign Federal Government, Singapore
    European Commission Directorate General for Research and Innovation
Foreign Private Profit
    Samsung Electronics Co., Ltd.
    Banco Itau

Private Non-Profit
    Lightspeed Grants LLC
    The Milken Institute, Singapore
Private Profit
    The AI Institute

New other organizations sponsoring research include:

Institution of Higher Education
    Harvard University
    Wake Forest University
Private Non-Profit
    UT- Battelle LLC
Private Profit
    Quansight LLC
    Autonomous Cyber LLC
    Form Finding Studio LLC
    Leidos, Inc.

Department of the Air Force-MIT AI Accelerator

Established in 2019 and extended in 2024 for a 5 year Phase 2 program (2024-2029) , the Department of the Air Force (DAF)-MIT AI Accelerator (aia.mit.edu) has successfully funded projects that have advanced AI research in a broad range of areas, including weather modeling and visualization, optimization of training schedules, and enhancement of autonomy for augmenting and amplifying human decision-making. The AI Accelerator engages more than 150 faculty, researchers, and students, who are widely affiliated with organizational units across MIT Campus and MIT Lincoln Laboratory, including the School of Engineering, the School of Architecture and Planning, the School of Science, and the Stephen A. Schwarzman College of Computing. The interdisciplinary AI Accelerator project teams also include DAF personnel, who are embedded in the research teams and serve as liaisons between the projects and Department of Defense stakeholders. There are currently 18 active AI Accelerator projects that involve 33 MIT principal investigators and also include recently started projects in areas such as human-machine collaboration and conflict resolution and diplomacy.

Defense Science and Technology Agency (DSTA) Singapore

Research collaboration with DSTA continues into its fourth year. Projects that finished in calendar year 2023:

- Data-driven Optimization under Categorical Uncertainty, and Applications to Smart City Operations (Alexandre Jacquillat)
- Provably Robust Reinforcement Learning (Ankur Moitra)
- Computationally-Supported Roleplaying for Social Perspective Taking (D. Fox Harrell)

- Next Generation NLP Technologies for Low Resource Tasks (Regina Barzilay, Tommi Jaakkola)
- Building Dependable Autonomous Systems Through Learning Certified Decisions and Control (Chuchu Fan)
- Online Learning and Decision Making under Uncertainty in Complex Environments (Patrick Jaillet)
- Decentralized Learning with Diverse Data (Costis Daskalakis, Asu Ozdaglar, Russ Tedrake)
- Trustworthy, Deployable 3D Scene Perception via Neuro-symbolic Probabilistic Programs (Vikash Mansinghka, Joshua Tenenbaum)
- SYNTHBOX: Establishing Real-World Model Robustness and Explainability Using Synthetic Environments (Aleksander Madry)

Projects that are continuing in calendar year 2024:

- Low Resource Multilingual Speech Recognition (James Glass)
- Sparse Data Methods for Speech Recognition (James Glass)
- New Representations for Vision (William Freeman, Joshua Tenenbaum)
- Analytics-Guided Communications to Counteract Filter Bubbles and Echo Chambers (Deb Roy)
- 3D Fusion from Multiple Images (Sertac Karaman)

New projects launched in calendar year 2024:

- Building Dependable Autonomous Systems Through Learning Certified Decisions and Control and Establish Rigorous Safety in Real-world Machine Learning-Enabled Systems with Accessible Tools (Chuchu Fan)
- Building an Automated Toolkit for Reliable Machine Learning (Aleksander Madry)
- Robust MADER with Onboard Collaborative Localization and Mapping (Jonathan How)
- Distributed, Multi-agent Trajectory Planning to Achieve Mission Goals and Safety Constraints within Limited Communication and Bounded Risk (Brian Williams)
- Task-Adaptive Synthetic Data for Data-Scarce Computer Vision (Phillip Isola and Antonio Torralba)
- 3D Self-Supervised Learning for Label-Efficient Vision (Vincent Sitzmann)

Gwangju Institute of Science and Technology (GIST)

- The follow projects, initiated in calendar year 2021, continue into calendar year 2024:
- Incorporating Generative AI into Contrastive Representation Learning for New and Multi-Modal Data
- (former title: Extending Contrastive Learning to New Data Modalities and Resource-Limited Devices) (Dina Katabi, Piotr Indyk)
- AI-driven discovery of co-evolutions of host-microbiome interactions for the application of microbiome-based therapeutics (Yoon Kim, Marzyeh Ghassemi)

- AI for Energy: Designing high-performance catalysts and electrodes for efficient hydrogen production (Tommi Jaakkola, Regina Barzilay)
- Artificial Compound Eye with Artificial Intelligence (ACE.A.I) for Enhanced Sensing (Fredo Durand, William T Freeman)
- HCI + AI for Human-Centered Physical System Design (former title: AI for Human Computer Interaction in Education) (Wojciech Matusik, Daniela Rus)
- AI-Driven Soft Robot Skin for Recognition, Modeling, and Exploration (Stefanie Mueller, Daniela Rus)

Quanta - CSAIL Research Collaboration

Collaboration will be extended for another five years starting on Sept 1 2024. The new focus will be on AI for Sustainability. Meanwhile, CSAIL is continuing its current five-year collaboration. The primary focus is largely defined by the troika of patient, hospital, and doctors; other aspects include privacy/security, and nutrition and self-care. Current projects include:

- Using Machine Learning to Curb Infectious Disease-John Guttag
- Learning to Assess Breast Cancer and Lung Cancer Risk to Enable Early Detection and Prevention-Regina Barzilay
- Revolutionizing the Care of Patients with Cardiovascular Disease- Collin Stultz Deep
- Metric Learning to Uncover Diabetic Patient Subtypes and Personalize Treatments - Marzyeh Ghassemi

Toyota-CSAIL Joint Research Center

CSAIL renewed and expanded its funded collaboration with Toyota Research Institute in University 2.0, a long-term collaboration with other universities, with funding through March 2026. Projects include:

- Task Driven Development of Nimble, Reactive, Rugged Hands (PIs: Ted Adelson, Sangbae Kim, Alberto Rodriguez, Wojciech Matusik, Pulkit Agrawal)
- Shared Autonomy Interactions and Learning (PIs: Daniela Rus, Sertac Karaman)
- Scalable Self-Supervised Learning for 3D Scene Understanding (PI: Justin Solomon)
- Physical and Functional Inductive Biases for Visual Representation Learning (PIs: Josh Tenenbaum, Fredo Durand)
- Behavioral model of driving to assess Vehicle-to-Grid feasibility (PI: Jessika Trancik)
- Requirements-informed Generative Modeling for Multimodal Conceptual Design Exploration (PI: Faez Ahmed)
- Bi-manual and Precise Dexterous Manipulation (PI: Pulkit Agrawal)
- Lifting Behavior Models to 3D for Generalist Robots (PI: Philip Isola)

Wistron-CSAIL Research Collaboration

This year has been the last year of Phase 2 in the collaboration between CSAIL and Wistron, with four projects wrapping up by August 2024. The collaboration has been extended for a 5 year Phase 3 program (2024 - 2029). The Phase 2 PIs visited Wistron for an annual meeting to report on the progress in July 2023. We also used this opportunity to kick-off the discussion of the next 5-year phase and potential research directions of mutual interest to Wistron and CSAIL. These discussions continued through the fall, with an in-person meeting during the Epoch Foundation's visit to MIT in October 2023. The master agreement for Phase 3 was signed in November, and we issued a call for proposals. A vigorous selection process led by CSAIL leadership in collaboration with Wistron identified four projects to kick off in September 2024. These span different topics of AI and computational design with applications in domains of critical importance to the company. The projects:

- From Data to Insight: Concept Learning from Unstructured Data
  PI: Polina Golland
- GenAI for New Modalities in Healthcare and Beyond
  PIs: Dina Katabi and Pyotr Indyk
- Augmenting Generative AI with Mechanical and Electrical Engineering Simulation to Create Physically Valid Enclosures and Circuits
  PIs: Stefanie Mueller and Mina Konaković-Luković
- Domain-Grounding AI Tools for Physics and Geometry
  PI: Justin Solomon

Industrial Outreach

CSAIL Alliances

The CSAIL Alliances Program (CAP) supports CSAIL's mission by connecting researchers, students and technological advances to industry and organizations across the globe; companies and organizations join the program at five different levels:

- Student Engagement – for connecting students and post docs to career opportunities in industry.
- Affiliate – grants access to CSAIL, including lab visits, CAP annual meeting, recruiting assistance, research briefings and professional education discounts.
- Partner – expanded CSAIL access, including research initiative meetings, custom faculty-led seminars, and expanded recruiting options.
- Start Up Connect – for early-stage CSAIL spin-offs, to maintain their connection with CSAIL and the CAP Alliances network.
- Start Up Connect Plus – for early to mid-stage start ups to connect with the lab, the CAP Alliances network, and the MIT startup ecosystem.

Currently there are over 130 member companies, including CSAIL start-ups and global brands such as Apple, Caterpillar, Cisco, Delta Electronics, Google, Microsoft, NTT Data, and Visa. At our most recent 3-day Alliances Annual Meeting in April 2024, we engaged over 650 member contacts, both in-person and via zoom. In addition, the Alliances team produces a steady stream of multi-media content about the latest research and commercialization activities to further the lab's reach and impact; case in point, in the first half of 2024 alone, the monthly Alliances podcast had over 35,000 plays.

CSAIL Alliances continues to produce and manage professional development courses with external organizations GetSmarter, Simplilearn and Emeritus, as well as internal collaborations with MIT's Professional Education, MIT's Office of Digital Learning, and MIT Sloan. This year we produced 3 new programs:

- Generative AI for Reinvention: Enabling the C-Suite (with MIT Professional Education).
- CSAIL Alliances SNAPSHOTS
- AI for Senior Executives (with MIT xPRO)

Total enrollments for FY24 reached 10,669, bringing the total impacted learners (since 2014) close to 57,000.

CSAIL Alliances also runs several research focused initiatives. Initiatives are pre-competitive, consist of a small group of companies around a theme, seed projects aligned with the theme, and include all the benefits of the Affiliate level. Presently we have the following active initiatives:

- MachineLearningApplications@CSAIL – this initiative focuses on applications of the latest machine learning (ML) technologies and research. In FY24, the initiative provided discretionary funding for 7 proposals and supported 7 CSAIL PIs; projects included explorations of Large Language Models (LLMs), Federated and Self-Supervised Learning, and Networked/Social Systems. This initiative is led by Professor Daniela Rus and currently includes BT, Cisco, and EY.
- Future of Data, Trust and Privacy – in collaboration with MIT's Internet Policy Research Initiative (IPRI), this initiative focuses on privacy-enhancing data systems and analytic techniques, a public policy dialogue, and new educational opportunities related to data governance technology and public policy. In FY24 the initiative provided discretionary funding for 5 proposals and supported 7 CSAIL PIs; projects included development of open banking privacy protocols, automating data privacy preservation, and using synthetic data. The initiative is co-led by Sr Research Scientist Daniel Weitzner and Professor Srini Devadas and currently includes Capital One, Fidelity, Mass Mutual Insurance, and Visa.

Internet Policy Research Initiative

The mission of the Internet Policy Research Initiative (IPRI), an institute-wide initiative, is to work with policymakers and technologists to bridge the gap between engineering and policymaking. Our missions is to increase the trustworthiness and effectiveness of interconnected digital systems, like the Internet. We accomplish this via a three-pronged approach: targeted engineering and public policy research, educational programs for students and policymakers, and outreach programs to build policy communities that facilitate communication. IPRI's research efforts cover six categories:

Cybersecurity

IPRI's cybersecurity research focuses on the technical and policy aspects of cybersecurity issues as they relate to the communication networks and software systems affecting the global society and economy. We use cryptographic computing to build cyber risk models and metrics. We have ongoing research projects with the US

Federal Reserve System (including an annual meeting on measuring cyber risk), municipalities, and security communities such as ISACs and ISAOs.

The MIT Future of Data Initiative
This is a leading multi-disciplinary research agenda to design and stimulate the deployment of accountable systems to provide trusted, traceable uses of personal data on an ecosystem-wide scale. The Initiative gathers computer science and Internet policy researchers with leading commercial enterprises in financial services, payment technology, cloud platforms, insurance and other sectors to discuss current challenges and opportunities in privacy and data governance. Modern privacy laws place appropriately high expectations on organizations processing personal data. At the same time, users report declining trust in those who handle their personal data and regulators around the world struggle with the scale of the enforcement challenge. A key challenge addressed by today's roundtable is to identify and put into service technical infrastructure for enterprises seeking to handle personal data in a trustworthy and lawful manner, and with guardrails to enable the scalable use of that data. One accountable system emerging from this research is the OTrace protocol that provides the ability for consumers to track how data is being used and shared, even (and especially) across organizational boundaries.

Privacy
IPRI's work also focuses on privacy policy and its critical role in trustworthiness. The Privacy group has publishes work on topics like privacy and security. Recent work focuses on the development and application of privacy enhancing technologies.

Networks
The Advanced Network Architecture (ANA) group is organized around five themes: Internet architecture, Internet security, Internet economics, Internet policy, and network management.

The Decentralized Information Group (DIG)
Focuses on data and systems governance (primarily on the Web) and explores both policy and technical issues. Current projects include a decentralized privacy-preserving platform for clinical research, evaluating the trustworthiness of autonomous systems, studying the relationship between privacy and machine learning, developing explanations for complex machines and models, securely aggregating distributed data, and developing smart contracts for data sharing.

MIT App Inventor
Empowers young people to develop useful apps that serve as novel digital solutions to problems they face in their lives, communities, and world.

Research Highlights

Numerous individual and multi-investigator projects are under way with major research discoveries across all areas of computing. Our work opens many opportunities to propel science, create new businesses, protect the planet, understand life, improve our cities and enhance our well-being and quality of life. A sampling of the work is highlighted below:

How language models implement ICL
Jacob Andreas

Large language models (like OpenAI's GPT or Google's Gemini) exhibit a remarkable capability called "in-context learning": they can acquire new skills from a small number of examples provided as *input*, without needing to be re-trained or fine-tuned. Over the last few years, in-context learning (ICL) has become a critical tool for controlling language models' behavior and specializing them for new tasks. But until recently, our understanding of how ICL works---that is, what computation neural language models perform "under the hood" when learning from inputs---has been extremely limited.

In papers at ICLR and ICML, we provided the first explanation of how language models implement ICL for several important problem families. To do so, we theoretically characterized the kinds of learning algorithms implementable on the neural "hardware" provided by modern language models, then developed procedures for identifying which of these algorithms (if any) were actually being implemented. Our results have provided insight into the capabilities and limitations of modern LMs, identifying different algorithmic "phases" associated with models of different sizes. Even more importantly, they have opened up a path towards better language models. We can identify specific components of language models that implement (messy, approximate versions of) subroutines from standard learning algorithms. By replacing these neural components with exact computation, we obtain language models that learn faster and generalize better.

(Work led by Ekin Akyürek in collaboration with Dale Schuurmans, Tengyu Ma, Denny Zhou, Bailin Wang and Yoon Kim.)

Designing Cloud Systems Beyond the End of Moore’s Law
Christina Delimitrou

Cloud computing now hosts the majority of the world’s computation. This includes interactive applications from which clients have come to expect low latency access at any point in time and from anywhere in the world. At the same time building a cloud infrastructure is an investment of hundreds of millions of dollars, most of which goes towards populating it with the latest server platforms. Therefore improving both the performance and efficiency of these systems is critical. A large fraction of the cloud’s resources now goes towards what is known as “datacenter tax” operations; computation like networking, memory allocation, encryption, and serialization. These are all necessary tasks for the system to operate as expected, but not part of any application’s logic. Furthermore, this fraction is only expected to increase because of the microservices and serverless programming frameworks many cloud providers now use to write their applications. Finally, the end of Moore’s Law has meant that the cloud servers themselves are not becoming faster, which has motivated a lot of recent work on hardware acceleration using specialized devices.

In our recent work, in the premier operating systems conference [OSDI'24], as well as our previous work at the premier venue for computer architecture and systems research [ASPLOS'21] we showed that using hardware acceleration for datacenter tax operations can greatly improve both the performance and efficiency of cloud platforms. The systems we designed use both new reconfigurable accelerators as well as repurpose accelerator engines that current servers already come embedded with to accelerate network processing and virtual machine snapshotting and compression respectively. In addition to accelerating these operations, we have also designed a secure resource sharing system that allows multiple tenants to share the same hardware platform without information leakage

Data consistency for MRI motion correction
Polina Golland

Motion artifacts are a pervasive problem in MRI, leading to misdiagnosis or mischaracterization in population-level imaging studies. Current retrospective motion correction techniques jointly optimize estimates of the image and the motion parameters. In contrast, we employ a deep network to reduce the joint image-motion parameter search to a search over motion parameters alone. Our network produces a reconstruction as a function of two inputs: acquired MRI signals and motion parameters. We train the network using simulated, motion-corrupted MRI data generated from known motion parameters. At test-time, we estimate unknown motion parameters by minimizing a data consistency loss between the motion parameters, the network-based image reconstruction given those parameters, and the acquired measurements. Our experiments on brain MRI data achieve high reconstruction fidelity while retaining the benefits of explicit data consistency-based optimization.

Publications:
Nalini M Singh, Neel Dey, Malte Hoffmann, Bruce Fischl, Elfar Adalsteinsson, Robert Frost, Adrian V Dalca, Polina Golland. Data Consistent Deep Rigid MRI Motion Correction. Medical Imaging with Deep Learning, PMLR 227:368-381, 2024. https://proceedings.mlr.press/v227/singh24a/singh24a.pdf

Making large language models more efficient
Yoon Kim

Today's large language models (LLMs) rely on a particular type of machine learning model called the transformer. While transformers are accurate and scalable, they rely on algorithmic primitives that are fundamentally computationally expensive, which can serve as a critical bottleneck in broadening the scope of their applications as well as widening access. In collaboration with the MIT-IBM Watson AI Lab, we have developed a new class of machine learning models which match the transformer's accuracy but is much more efficient to train and deploy. The core idea is to restructure the internal computations of the transformer so that it can be reformulated as a recurrent neural network, a particular type of neural network which enjoys better efficiency properties. On top of making existing LLM workflows more efficient, the long-sequence processing capabilities of our model is expected to open up a host of new applications such as question-answering over entire books, generating long videos, and processing complex biological data.

The Multimodal Automated Interpretability Agent
Antonio Torralba

As AI systems become pervasive in research and society, understanding the risks and capabilities they present is increasingly important. Consider, for example, an AI system that automatically classifies the content of images: we may wish to recognize when and how it relies on sensitive features like race or gender, identify systematic errors in its predictions, or learn how to modify its training data and model design to improve accuracy and robustness.

Today, this kind of understanding requires significant effort on the part of researchers—involving exploratory data analysis, formulation of hypotheses, and controlled experimentation [1, 2]. As a consequence, this kind of understanding is slow and expensive to obtain even about the most widely used models.

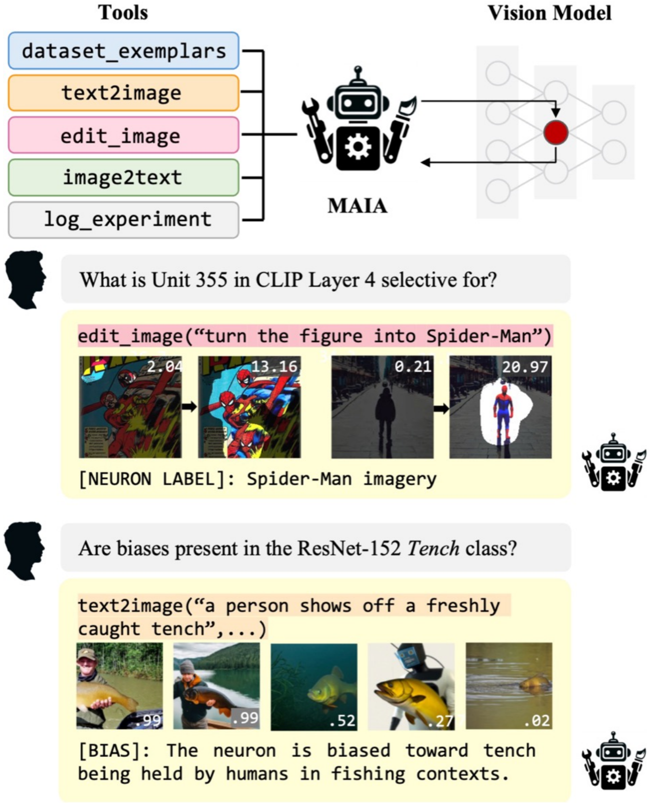
Recent work on automated interpretability (e.g., [3–5]) has begun to address some of these limitations by using AI models themselves to assist with AI understanding tasks—for example, by assigning natural language descriptions to learned representations, which may then be used to surface features of concern. But current methods are useful almost exclusively as tools for hypothesis generation; they characterize model behavior on a limited set of inputs, and are often low-precision [6].

We introduce a prototype system we call the Multimodal Automated Interpretability Agent (MAIA), that combines the scalability of automated techniques with the flexibility of human experimentation. MAIA equips a pretrained vision-language model backbone with an API containing tools for conducting experiments on AI systems. MAIA is prompted with an explanation task (e.g., “describe the behavior of unit 487 in layer 4 of CLIP” or “in which contexts does the model fail to classify labradors?”) and designs an interpretability experiment that composes experimental modules to answer the query. MAIA’s modular design (Fig. 1) enables flexible evaluation of arbitrary systems and straightforward incorporation of new experimental tools, including modules for synthesizing and editing novel test images, which enable direct hypothesis testing during the interpretation process.

We evaluate MAIA’s ability to produce predictive explanations of vision system components using the neuron description paradigm [4, 5, 7–10] which appears as a subroutine of many interpretability procedures. We additionally introduce a novel dataset of synthetic vision neurons built from an open-set concept detector with ground-truth selectivity specified via text guidance. We show that MAIA descriptions of both synthetic neurons and neurons in the wild are more predictive of behavior than baseline methods, and often on par with human labels.

MAIA also automates model-level interpretation tasks where descriptions of learned representations produce actionable insights about model behavior. We show in a series of experiments that MAIA’s iterative experimental approach can be applied to downstream model auditing and editing tasks including spurious feature removal and bias identification in a trained classifier. Both applications demonstrate the adaptability of the MAIA framework across experimental settings: novel end-use cases are described in the user prompt to the agent, which can then use its API to compose programs that conduct task-specific experiments. While these applications show preliminary evidence that procedures like MAIA which automate both experimentation and description have high potential utility to the interpretability workflow, we find that MAIA still requires human steering to avoid common pitfalls including confirmation bias and drawing conclusions from small sample sizes. Fully automating end-to-end interpretation of other systems will not only require more advanced tools, but agents with more advanced capabilities to reason about how to use them.

Figure 1



What is Unit 355 in CLIP Layer 4 selective for?

```
edit_image("turn the figure into Spider-Man")
```

```
[NEURON LABEL]: Spider-Man imagery
```

Are biases present in the ResNet-152 *Tench* class?

```
text2image("a person shows off a freshly
caught tench",...)
```

```
[BIAS]: The neuron is biased toward tench
being held by humans in fishing contexts.
```

Figure 1: MAIA framework. MAIA autonomously conducts experiments on other AI systems to explain them.

Laboratory Sponsored Activities

CSAIL Media Outreach

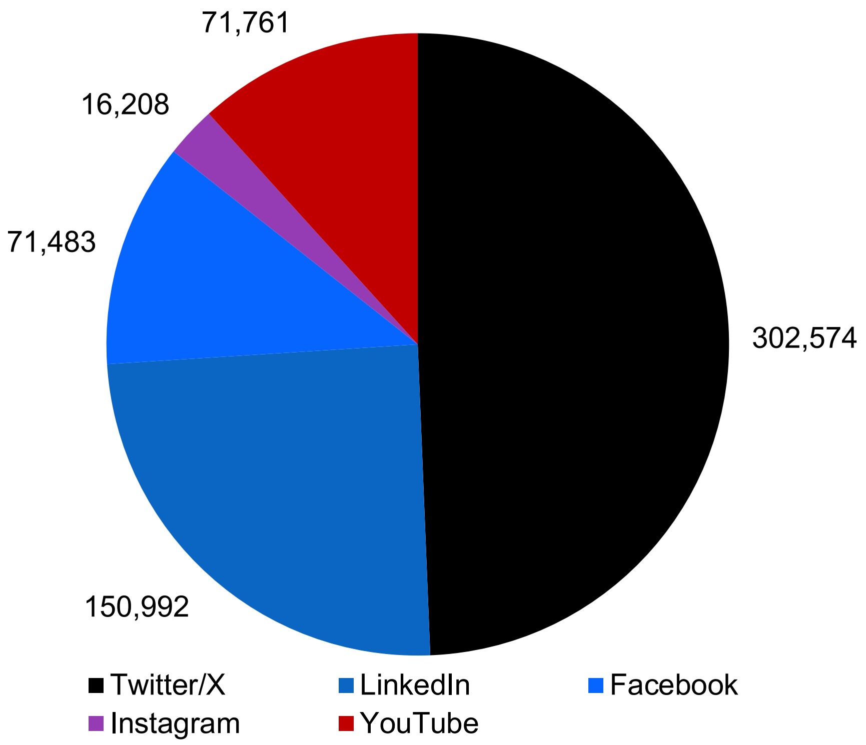
Consistently higher media attention and output than competitors.

- Followers: Combined following of 613,144 users across Twitter/X, Instagram, Facebook, LinkedIn, TikTok, and YouTube (roughly 13% growth from last year).

- Consistent media coverage in various top-tier outlets: New York Times, Forbes, WIRED, TechCrunch, Quanta Magazine, Boston Globe, Popular Science, and Washington Post.
- MIT CSAIL has 1,807 press mentions this year, whereas Carnegie Mellon School of Computer Science (CMU SCS) received 190 and Stanford HAI received 603.
- CSAIL's Twitter/X following has also grown over 2750% since 2015, well over the percentages of the University of California, Berkeley Electrical Engineering and Computer Sciences (nearly 2000%), Stanford Engineering (roughly 125%), MIT Engineering (over 1125%), MIT Department of Electrical Engineering and Computer Sciences (almost 500%), and CMU SCS (under 250%).

Audience Breakdown

Listed by platform



YouTube Channel:

Channel	Views	Videos	Average daily views
MIT CSAIL	354,797	22	583.5
Harvard SEAS	290,070	50	430.7

Stanford HAI	203,910	41	508.4
CMU SCS	36,308	45	99.4

Research Highlights:
Rethinking AI's impact: MIT CSAIL study reveals economic limits to job automation: 139 media mentions
Using AI to protect against AI image manipulation: 137 media mentions
Exploring the mysterious alphabet of sperm whales: 120 media mentions

CSAIL Hosted Lecture Series

Dertouzos Distinguished Lectures have been a highly regarded tradition since 1976, featuring some of the most influential thinkers in computer science. Speakers featured during 2023-2024:

- Bob Metcalfe*, CONNECTIVITY is a thing, is THE thing*, Nov. 8, 2023
- Bill Weihl, (canceled due to illness), March 3, 2024
- Dan Spielman, *Algorithmic Discrepancy Theory and Randomized Controlled Trials,* Mar. 20, 2024

Hot Topics in Computing was launched in 2017, convening experts to discuss emergent potential, perception, and problems associated with the proliferation of computation and machines. Topics featured during 2023-2024:

- Matei Zaharia, *Going Beyond Models to Reliable AI Systems using LLM,* Oct. 11, 2023
- Yejin Choi, *Possible Impossibilities, Impossible Possibilities, and Paradoxes,* Nov. 8, 2023
- Prabhakar Raghavan, *Beyond information retrieval: what does Search mean these days*? Apr. 3, 2024
- Robert Kahn, *Sustaining Creativity for National Information Infrastructure*, May 1, 2024
- Michael Bronstein, *Geometric Deep Learning – from Euclid to Drug Design,* Jun 6, 2024

Organizational Changes

Hired 4 FOs, 2 HR staff, 4 AAs, 2 IT/TIG staff (1 replacement and 1 additional new FTE), 3 Alliances/CAP staff, and replaced 1 manager of administrative support services.

New faculty joined during 2023-2024 include:

- Yael Kalai, Professor
- Kaiming He, Associate Professor

Awards and Honors

- Adelson, Ted: IEEE Lifetime Achievement Award on Computer Vision
- Agrawal, Pulkit: IEEE Early Academic Career Award in Robotics & Automation
- Andreas, Jacob: Sloan Fellowship
- Barzilay, Regina: elected to the National Academy of Medicine
- Davis, Randall: certificate of appreciate from the International Olympic Committee, for work with the IOC Committee on AI
- Delimitrou, Christina: ASPLOS'24 Test of Time Award
- Delimitrou, Christina: IEEE Micro Top Picks Award for the Best Paper, Computer Architecture Conferences 2023
- Devadas, Srini: Computer and Communications Security Conference Test of Time Award
- Emer, Joel: EEE Computer Society B. Ramakrishna Rau award
- Emer, Joel: ISPASS Best Paper 2024
- Kaashoek, Frans: IEEE Computer Society Distributed System Award
- Katabi, Dina: SIGCOMM Lifetime Achievement Award
- Hopkins, Sam: MIT EECS outstanding educator award
- Indyk, Piotr: elected to National Academy of Sciences
- Indyk, Piotr: Test of Time Award at PODS'24
- Isola, Phillip: Best Paper Award at CoRL
- Leiserson, Charles: SPAA Parallel Computing Award, ACM Symposium on Parallelism in Algorithms and Architectures
- Liskov, Barbara: Fellow of the Computer History Museum in Mountain View
- Liskov, Barbara: honorary degree from University of Connecticut
- Lynch, Nancy: Brooklyn College Alumni Lifetime Achievement Award
- Madden, Sam: SIGMOD Edgar F. Codd Innovations Award
- Matusik, Wojciech: Joan and Irwin M. (1957) Jacobs Chair
- Metcalfe, Bob: Franklin Institute Medal in Electrical Engineering
- Oliva, Aude, Justine and Yves Sergent Award in Cognitive Neuroscience
- O'Reilly, Una-May: ACM SIGEVO Impact Award
- O'Reilly, Una-May: Senate of Dalhousie University honorary degree of Doctor of Laws, honoris causa
- Rus, Daniela: elected to National Academy of Sciences
- Rus, Daniela: Boston Globe Tech Power Players 50 (no 17 in 2024)
- Rus, Daniela: John Scott Medal
- Satyanarayan, Arvind group was awarded an Outstanding Paper Award at ACL 2023; Won an Alfred P Sloan Fellowship in 2024
- Solar-Lezama, Armando: Distinguished Professor of Computing
- Solomon, Justin: ACM SIGGRAPH Test-of-Time Award
- Sussman, Gerry: IEEE Computer Society's Taylor L. Booth Educator Award 2024
- Sussman, Gerry: IEEE Educational Activities Board Major Education Innovation Award in 2023
- Tedrake, Russ: MIT School of Engineering Distinguished Educator Award and MIT EECS Digital Innovation Award and Burgess (1952) and Elizabeth Jamieson Award for Excellence in Teaching in 2024

- Tedrake, Russ: 2023 IEEE Transactions on Robotics King-Sun Fu Memorial Best Paper Award
- Tenenbaum, Josh: Schmidt Futures AI 2050 senior fellowship.
- Tenenbaum, Josh: outstanding computational modeling prize, 2023 Conference of the Cognitive Science Society
- Vaikuntanathan, Vinod: Simons Investigator
- Vaikuntanathan, Vinod: CRYPTO 2023 Test of Time Award I
- Vaikuntanathan, Vinod: 2024 MacVicar Faculty Fellow
- Vaikuntanathan, Vinod: Distinguished Alumnus Award from IIT Madras
- Vassilevska, Virginia Williams: Simons Investigator
- Williams, Ryan: Gödel Prize
- Yang, Mengjia: IEEE CS TCCA Young Computer Architect Award
- Zeldovich, Nikolai: Joan and Irwin M. (1957) Jacobs Chair

CSAIL members recognized for excellence in service and research contributions:

- MIT Excellence Award for Marcia Davidson, Senior AA
- Infinite Mile Award: Technology & Infrastructure Group (TIG) awarded for Team Excellence
- 54 Spot Awards achieved by CSAIL staff
- 23 Gratitude Book Club Awards to faculty and staff to recognize exceptional service.

Key Statistics for Academic Year 2023-2024

Headcount Count	**Women/Men**	**Women %**
Faculty	22/99	18.18%
Postdoc Assoc/Fellow	15/76	16.48%
Principal Research Scientist	3/3	50%
Research Staff	10/40	20%
Senior Research Scientist	2/3	40%
Administration, technical, and support staff	65/40	61.90%
Graduate Students*	210/529	28.42%
UROP	205/256	44.47%
Visitors	4/32	11.11%
TOTAL PERSONNEL	**1,614 (536/1,078)**	**33.21%**

*Please note, 2 of the grad student headcount are unknown.

We are pleased to present this report for 2023-2024; much more extensive information is available at www.csail.mit.edu.

Daniela Rus
Director, Computer Science and Artificial Intelligence Laboratory