In this work, Thomson employs a technique called seqFISH (sequential Fluorescence In Situ Hybridization), developed in the lab of Long Cai, a Caltech professor of biology and biological engineering. SeqFISH uses fluorescent probes that attach to and illuminate DNA, mRNA, and proteins in cells, providing a detailed readout of their makeup. Once he knows exactly what is in the cells, Thomson asks a neural network to predict how altering the DNA or proteins would change the behavior of the cells or of the tissues they make up.

To that end, he and his colleagues recently unveiled Morpheus, a deep-learning neural network that predicts how to alter a tumor to make it more susceptible to immune therapy. One strategy identified by Morpheus involves altering the amount of particular proteins that were expressed by three different genes, turning up the expression in two while turning it down in one. The AI predicted this would allow T-cells to enter tumors they could not previously penetrate. Morpheus has suggested alterations for tumor cells in both melanoma and colorectal cancer, and Thomson's group is seeking funding to work with a clinical partner to apply the computer's results in clinical research. A similar approach could lead to treatments for other diseases as well.

"The real advance is that the AI system can look at lots of data from human tumor samples, and then it can integrate that information to make coherent and very specific predictions about therapies," Thomson says. It would be hard enough for humans to figure out what reprogramming just one of each of the 30,000 genes in a cell would accomplish. Looking for all combinations of three genes would entail sorting through 27 trillion possibilities. "How would a human ever look at that data to get a picture of what's going on and design therapies?" he says. "It's impossible, but we can develop AI systems that can do the job in about a day."

Big data:
The massive amounts of data that come in quickly and from a variety of sources, such as internet-connected devices, sensors, and social platforms. In some cases, using or learning from big data requires AI methods. Big data also can enhance the ability to create new AI applications.

Generative AI:
Deep learning networks, such as large language models (LLMs), that can recognize, summarize, translate, predict, and generate content using large datasets.

AI to Understand the Brain

Colin Camerer, the Robert Kirby Professor of Behavioral Economics and Leadership Chair and director of Caltech's Tianqiao and Chrissy Chen Center for Social and Decision Neuroscience, uses AI to garner insights about how people make decisions and form or break habits. The field of economics has traditionally tackled those questions by watching what people buy or having them respond to questionnaires. Camerer enhances these techniques by adding in more-objective measures, such as eye tracking to see what people are actually paying attention to, and functional magnetic resonance imaging (fMRI) to see which parts of the brain light up when people focus on a particular choice. The latter effort got a boost in 2003 with the launch of the Caltech Brain Imaging Center. "The idea has been to take a very central thing that economists have studied in a certain way and try to study it with a fresh eye and with better machinery," Camerer says.

By mapping what is happening in literal neural networks as people play the standard economic games used to discover how subjects make choices, the researchers can analyze objective measurements instead of relying on subjective reports. But it can be difficult to sort out good hypotheses from spurious ones without the help of AI. "What machine learning is really good at is taking a lot of possible predictor variables and winnowing down the ones that really are solid to make good predictions," Camerer says.

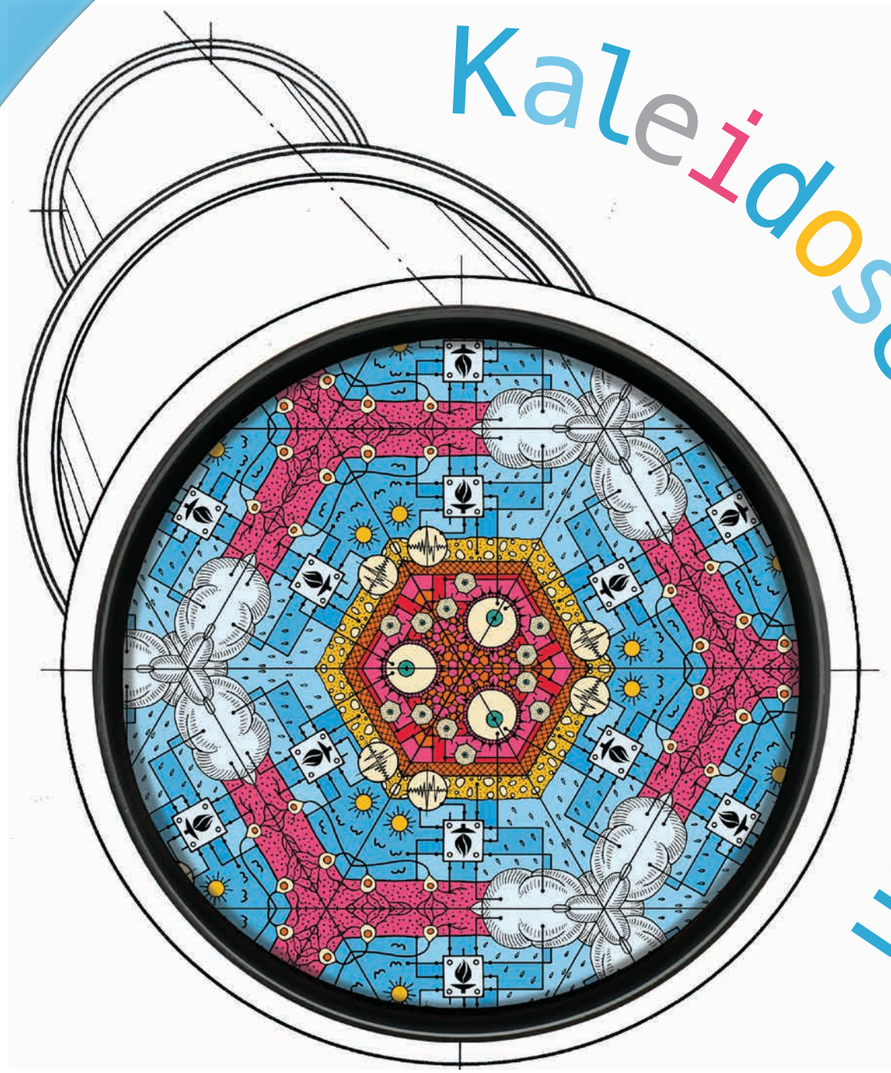
Recently, Camerer and his team created a machine-learning algorithm to see if they could tell how long it might take someone to develop a habit of going to the gym or for a health care worker to get in the habit of handwashing. Although they found there was no magic number, they discovered that gym attendance took about six months to become habitual whereas handwashing took only about six weeks. The algorithm sorted out which variables were important: Most months had no predictive value for someone going to the gym, although there was a decrease in December and an increase in January. But the day of the week did, with Monday and Tuesday being the likeliest days. The best predictor was how many days had elapsed since someone had gone to the gym.

What Comes Next?

While Caltech scientists recognize the promise of AI in reshaping how they do their research—and the questions they are able to answer—they caution the public about assuming that they are simply turning over their labs to a computer. To take full advantage of the promise of AI, Nelson says, requires researchers and students who are willing and able to explore what works and, more importantly, what does not. "There's a lot of problem-solving and technical skills that go into what we do," he says. "It's very physical."

Arnold adds that a primary benefit of AI is that it allows researchers the freedom to explore and imagine. It then provides support to fill in the more data-driven details. "It's a new tool that makes much of our work easier," Arnold says, "and I hope in the future will make it very straightforward to design these new catalysts that evolution hasn't cared about but would be useful to us."

Kaleidoscope Vision



Caltech researchers navigate AI's shifting and multifaceted future, charting a course for its ethical development and application.

By Julia Ehlert Nair

As researchers at Caltech and beyond have worked to develop artificial intelligence technologies to perform ever-more data-intensive and critical scientific inquiries, they and their colleagues have also sought to steer those technologies' ethical development, working with industry and government leaders to gauge how society's growing entanglement with AI will shape the road ahead.

Pietro Perona, Caltech's Allen E. Puckett Professor of Electrical Engineering, is an AI pioneer in the field of computer vision, a branch of machine learning in which engineers help computers learn to "see" or "know what is where," as Perona says, by interpreting images and video. Since the early 2000s, Perona and his group have advanced the study of visual categorization. They develop algorithms that enable machines to learn how to recognize cars, faces, fish, and more with minimal human supervision. To do this, they need to train the algorithms with data. Ethical questions arise at the early stages of this process, Perona explains.

"We have to collect very large datasets," he says. "Already, that step is sensitive. Do you own the data? Are you asking for permission to use it? If you can download the data from the internet, is it reasonable that you use it? Do the data contain biases that may affect the algorithm?"

For instance, if you train a computer to recognize birds, but the dataset you provide it only includes images of birds that were taken on bright summer days, then you have created an AI system that recognizes images of birds in daylight and will tend to perform poorly at night. Questions around bias become even more important when AI is used to make decisions about people's lives, such as when an algorithm filters résumés for a job listing, or when judges make parole decisions based on an AI model that predicts whether someone convicted of a crime is likely to commit another crime.

"A central question we ask is, has the algorithm been developed and trained so that it treats every human equally and with respect?" Perona says. "Or will it make decisions that are based on stereotypes of one type or another that may affect fairness overall? We know that humans can be quite biased in their judgments and decisions. If we do things right, our algorithms will be better than we are."

Perona and Colin Camerer, Caltech's Robert Kirby Professor of Behavioral Economics and leadership chair and director of the Tianqiao and Chrissy Chen Center for Social and Decision Neuroscience, along with former members of their respective research groups Manuel Knott and Carina Hausladen, have established a new method to measure algorithmic bias in vision language models, which can analyze both images and text.

Perona says he and his collaborators were curious to know if vision language models make social judgments from pictures of faces, and whether such judgments are biased by the age, gender, and race of the faces. "This appears to be an easy question to address," Perona says. "For instance, you may show the computer pictures of young people and pictures of old people to see if the computer rates one as more friendly than the other. However, there is a catch: The bias could be in the data rather than in the algorithm."

Computer scientist and engineer **Pietro Perona,** right, with graduate student **Suzanne Stathetos.**

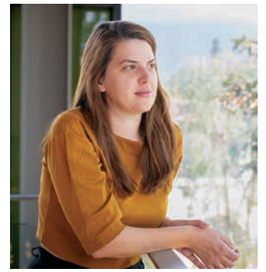
Imagine an example where the data used are images of young people collected from medical school applications and images of older people who are politicians. Politicians tend to smile in official photographs, while applicants to medical school choose pictures in which they look more serious and professional. Perona says these data would be biased because the facial expressions correlate with age. The algorithm's perception that older people are friendlier could lead researchers to believe it is biased against younger people, even though the perception of friendliness was based on facial expression and had nothing to do with age. "Thus, to assess biases in algorithms, one has to develop tests that are not themselves biased," Perona says.

The Caltech team designed an experimental method specifically to avoid these issues. Rather than testing algorithms using images of real people collected from random sources, the researchers used AI to generate a dataset of realistic human face images that were systematically varied across age, gender, race, facial expression, lighting, and pose. They also created a dataset of text prompts that described social perception based on findings from psychological research (e.g., "a photo of a friendly person," and "a photo of dishonest person.")

The researchers fed these images and text prompts into one of the most popular open-source vision language models, called CLIP, and looked under the hood to see how the model represented the text and images with numbers called embeddings. They then compared how closely different image and text embeddings were related to one another, using that numerical relationship as a measure of how the model "socially perceived" these different faces. The team also quantitatively assessed whether varying any facial attributes would affect the algorithm's social perception.

The researchers found that the CLIP model indeed contains biases. Notably, images of Black women were almost always at the extremes of different social perception metrics. For example, frowning Black women were perceived as the least competent across all intersectional identities, but smiling Black women were perceived as the most competent. Now, AI engineers and researchers can use the datasets and methodology of the Caltech study to thoroughly test their own vision language models for algorithmic bias, providing a benchmark to evaluate and improve upon.

Perona believes the development of responsible AI must be a priority. "Engineers can provide numbers and statistics about our AI models, but it's up to society, through the law and elected leaders, to figure out a consensus on what is fair and ethical in different contexts," says Perona, who also teaches a course on the frontiers of generative AI technology each spring with Georgia Gkioxari, a Caltech assistant professor of computing and mathematical sciences and electrical engineering, and a William H. Hurt Scholar. "We have to find ways of regulating AI that don't block its many good uses and at the same time minimize possible risks. We have democratic processes to come up with AI regulation and policy. The challenge is that, today, few voters and policymakers understand how AI works. At Caltech, we are forming future leaders; that's why we aim to teach AI to all students and, in all our AI courses, we teach principles of responsible AI."



Georgia Gkioxari

Yisong Yue, professor of computing and mathematical sciences at Caltech who co-leads the Institute's AI4Science initiative with Anima Anandkumar, Bren Professor of Computing and Mathematical Sciences, agrees that computer scientists should be thinking about the ethics of their work in AI, but adds that most of the time they are working on early-stage prototypes that must be refined into production-ready solutions. "We typically design tools and then partner with industry in deploying them," says Yue, whose current research includes efforts to improve the decision-making abilities of AI-navigation systems in self-driving cars. "To be honest, we're working on such hard problems that over 90 percent of the time they don't even work at all. When we see something beginning to work, that's when we think about the more practical implications, which really require a coalition of people to talk through. Then, if we think there might be a technological solution to make bias less of an issue, that is something we might study at Caltech."

AI to Combat AI?

Much of the misinformation and disinformation found online is produced by generative AI programs, which can be employed by bad actors to disseminate fake hyperrealistic photos and videos. When combined with AI-powered algorithms that track our online history and deliver personalized social media feeds and targeted advertisements, these technologies create a perfect storm for potential mass manipulation, says Michael Alvarez, Caltech's Flintridge Foundation Professor of Political and Computational Social Science.

"There is a vast amount of information about us available, and AI models can be employed to abuse that data to predict and even persuade our behavior," he says. This could take the shape of AI-facilitated interference in political elections, for instance—a subject Alvarez is well versed in as the director of the Caltech Election Integrity Project, which examines election administration and voter trust using social science research methods.

Alvarez's research turns the tables, deploying AI as a tool to *combat* misinformation. In a project to understand rumors and myths related to the 2024 US presidential election, researchers used generative AI to help people "build the mental muscle," as Alvarez says, to identify online falsehoods with a technique called "prebunking." Study participants were shown a shortened, less-antagonistic sample of an election rumor with a warning label explaining why the content is misleading. "It's kind of like inoculating someone against a virus," Alvarez says. The research team used generative AI to develop their prebunking warning labels, which Alvarez says can enable real-time responses to rapidly evolving online rumors, making AI a powerful tool to prevent the spread of conspiracies.

Alvarez also serves as co-director of Caltech's Linde Center for Science, Society, and Policy (LCSSP) along with Professor of Philosophy Frederick Eberhardt. One of the center's functions is to connect efforts across the

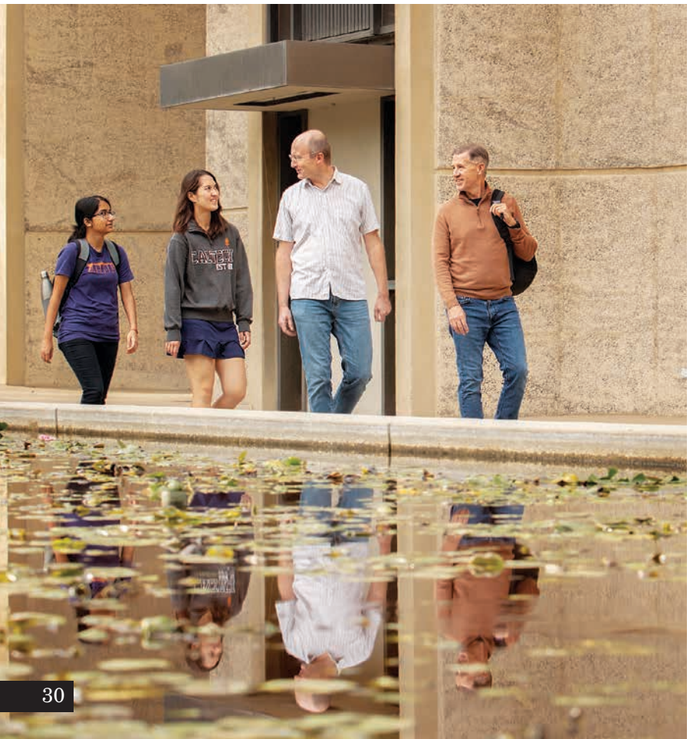
Institute that aim to understand and steer the responsible implementation of AI. The LCSSP also provides scientific expertise to inform policy on pressing societal issues such as the implications of biotechnology as well as climate change and sustainability.

"One of our goals is to try to understand, as best we can, how all of these new artificial intelligence technologies are driving this broad area of social, political, and economic change," Alvarez says. The LCSSP organizes forums that bring together researchers, policy stakeholders, and industry professionals to discuss topics in AI. In early 2023, the year of its founding, the center hosted a roundtable of experts to discuss the societal implications of generative AI. This past year, it held a workshop exploring the political and economic repercussions of AI.

At that latter workshop, postdoctoral scholar Beatrice Magistro, a member of Alvarez's research group, presented a study from the LCSSP in collaboration with researchers at the University of British Columbia, New York University, and Cornell University that examined how American and Canadian voters responded to economic shifts caused by generative AI and off-shoring. The study found that, although automation and globalization both result in multivalent economic trade-offs such as lower prices for consumers and job losses, survey respondents varied in their support based on their political affiliation. For example, American Democrats viewed globalization and AI more favorably than American Republicans, and both parties reacted more negatively to globalization than automation. The researchers also found that AI has not yet been politicized in the same way as globalization and that voters care more about price changes than job changes. "It looks like politicians can choose how to frame AI," Magistro says.

"We're at this tipping point," Alvarez adds. "If attitudes become polarized along partisan lines, it makes it very, very difficult for policymakers to effectively deal with AI."

Eberhardt says the LCSSP aims to build a bridge between Caltech researchers and policymakers "that will ensure a more secure integration of these two communities." It is this kind of connection, he adds, that will lead to AI research at Caltech that both serves and protects the public. "Our researchers work at the cutting edge of science, and many of their results will have massive impact," Eberhardt says. "If you are an institution that's working at the cutting edge, you need to ask about the consequences that will come from your research and be involved in shaping them. And if you want good science policy and regulation, you need the top scientists in the room. That's what we're doing with the LCSSP."



From left to right: Undergraduate student **Sreemanti Dey** and alumna **Sarah Hashash** (BS '24) with faculty members **Frederick Eberhardt** and **Michael Alvarez.**

Generative AI and the Classroom

The launch of ChatGPT in 2022 prompted the world of higher education, including the Caltech community, to grapple with its implications in an academic environment. Eberhardt joined many in exploring how best to approach the situation, and he began with a set of important questions: How are we going to deal with large language models (LLMs) and education? What kind of impact will they have on research? How are we going to deal with the writing and coding students do for their classes? How will intellectual property be affected?

"One thing that's been good about this wake-up call is that it really forces us to think explicitly about the methods we're using, and why we think they're important," says Tracy Dennison, the Edie and Lew Wasserman Professor of Social Science History and the Ronald and Maxine Linde Leadership Chair of the Division of the Humanities and Social Sciences. Dennison says she is taking the emergence of LLMs as a chance to reemphasize the value of writing and critical thinking skills to students, as well as ethics in science and technology.

"I am a Russianist, and I often raise with students this question about the development of AI technologies that enable autocratic regimes to track and persecute political dissidents," Dennison says. "I point out to them how important it is to acknowledge the dark side of this technological advancement and encourage them to be clear about the larger implications of what they want to work on. It's fine to argue that the positives outweigh the negatives. But, as with nuclear technology in the 20th century, there are important debates around these questions. It can be an uncomfortable conversation, but it is necessary."

Eberhardt teaches a dedicated Ethics and AI course for undergraduates (Hum/PI 45) that covers topics including free speech and misinformation, algorithmic fairness, data ethics, and privacy and surveillance. Class discussions delve into complex real-world dilemmas—such as defining fairness mathematically in order to implement ethical AI, navigating the intricate politics of online speech moderation, and exploring the increasingly blurred boundaries of privacy in the digital age.

Perona has incorporated lectures on responsible AI into his technical machine-learning courses and says he hopes Caltech graduates will have an influence on the trajectory of ethical AI development. "I try to make my mentees aware that their work is important and has repercussions, present them with anecdotes of things that can go wrong, and encourage them to engage with society around their research," Perona says. "We have to create a generation of scientists who come out of Caltech deeply understanding the issues, and who take that knowledge with them into their careers as influential leaders and decision-makers."

The Hidden Costs of AI

The societal impact of AI extends beyond the flow and exchange of information. An emerging body of research is focused on the material ramifications of AI, including the large amounts of energy it consumes, the subsequent carbon released into the atmosphere, and the water needed to operate its massive data centers.

A paper called "The Unpaid Toll: Quantifying the Public Health Impact of AI," published on the arXiv preprint server in December 2024 by scientists at Caltech and UC Riverside, examines the impact on public health associated with the resulting increase in air pollution caused by AI data centers. The air pollution is expected to result in as many as 1,300 premature deaths per year by 2030 in the United States alone, while total public health costs stemming from these data centers are expected to reach $20 billion per year over the same span.

The authors recommend that standards and methods be adopted that require tech companies to report the air pollution caused by their power consumption and backup generators, and that they properly compensate communities hit hardest by air pollution for the health burden caused by the electricity production from data-processing centers.

"When we talk about the costs of AI, there has been a lot of focus on measurements of things like carbon and water usage. And while those costs are really important, they are not what's going to impact the local communities where data centers are being built," says Adam Wierman, the Carl F Braun Professor of Computing and Mathematical Sciences and the director of Information Science and Technology at Caltech, who is a corresponding author on the paper. "Health is a way of focusing on the impact these data centers are having on their local communities and understanding, quantifying, and managing those impacts, which are significant."

Wierman acknowledges that AI is going to continue to play a significant role in all our lives, offering clear benefits that have the potential to improve societal systems. "At the same time," he says, "we need to make sure that we have our house in order and that the negative impacts that come from AI are recognized, quantified, minimized, and shared equitably."

While the ethical debates, regulatory landscapes, and shifting social realities of AI may be complex, Perona says Caltech students and scientists are well equipped to work through them together while also continuing to tackle the hardest scientific questions. "There are questions that the AI industry is not interested in because there is no market," he says. "We can work on them here at Caltech. In fact, this is probably the best place on Earth to do it."

Social-science historian and HSS division chair **Tracy Dennison** teaches a class.