ATHARVA EDUCATIONAL TRUST'S
ATHARVA COLLEGE OF ENGINEERING
(Approved by AICTE, Recognized by Government of Maharashtra
& Affiliated to University of Mumbai - Estd. 1999 - 2000)
ISO 21001:2018 ISO 14001:2015 ISO 9001:2015
NAAC Accredited A+

ATHARVA ROBOTICS CENTER

Daily News on Innovation & Technology

03rd July, 2025

NASA Sets Briefings for SpaceX Crew-11 Mission to Space Station

By Lauren E. Low, July 02, 2025

NASA and its partners will discuss the upcoming crew rotation to the International Space Station during a pair of news conferences on Thursday, July 10, from the agency's Johnson Space Center in Houston.

This AI model was trained on 10M human choices. Now it thinks and reacts like us

By Neetika Walter, July 02, 2025

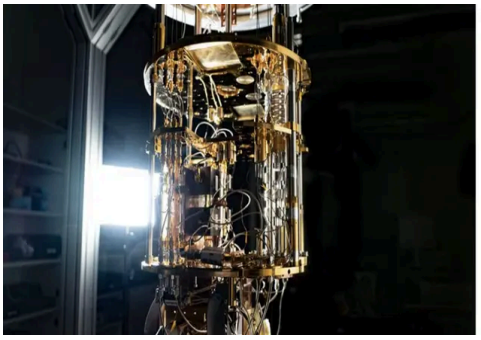
What if an AI didn't just mimic your mind, but could predict your every next move? Researchers at Helmholtz Munich have developed a new language model that simulates human behavior with striking accuracy.

SEMI, TechSearch Release 2025 Edition of Worldwide Semiconductor Assembly & Test Facility Database

By SEMI, July 03, 2025

SEMI and TechSearch International has released the 2025 edition of the Worldwide Semiconductor Assembly & Test Facility Database, the industry's most comprehensive resource tracking global assembly and test sites operated by integrated device manufacturers (IDMs) and outsourced semiconductor assembly and test providers (OSATs).

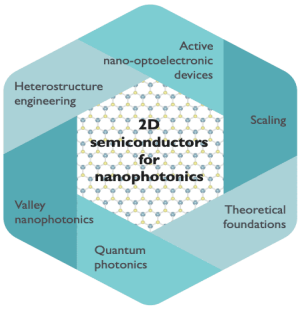
World-first: Scientists unveil method for simulating error-corrected quantum computations



By Prabhat Ranjan Mishra, July 02, 2025

Despite being powerful and fast, quantum computers still face challenges on their pathway to practical use cases. These devices have a limited ability to correct the arising computational errors.

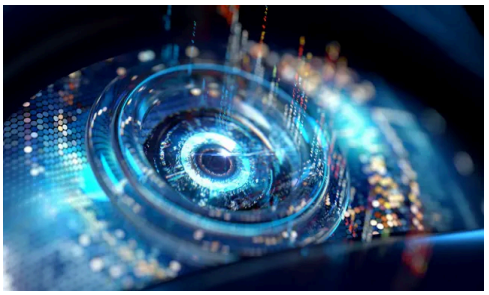
Two-Dimensional Semiconductors Advance Nanophotonics and Future Optoelectronic Devices



By Quantum News, July 02, 2025

The manipulation of light at the nanoscale, a field known as nanophotonics, stands to revolutionise technologies ranging from optical computing to advanced sensing.

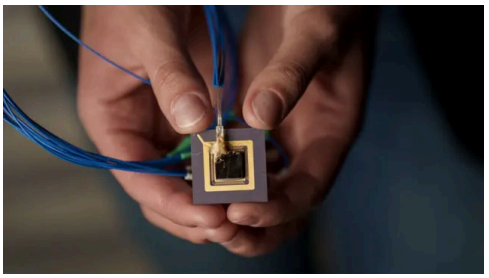
US engineers' new way of attacking vision systems can make AI see whatever you want



By Prabhat Ranjan Mishra, July 02, 2025

Engineers have explored a new way of attacking artificial intelligence computer vision systems. They believe that the method can help them to control what the AI "sees."

MIT student builds pocket-sized 3D printer that uses light to create objects in secs



By Prabhat Ranjan Mishra, July 02, 2025

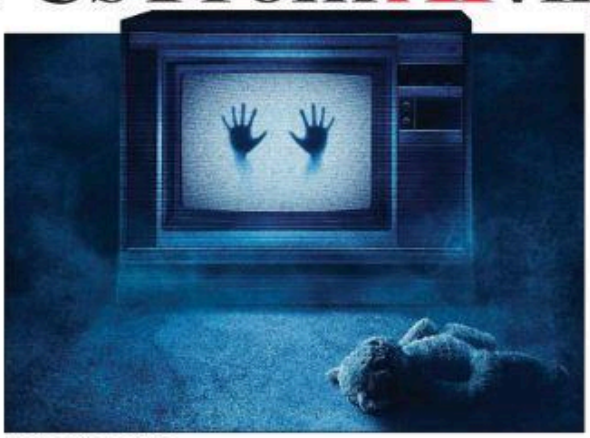
Researchers have developed photonic devices that manipulate light to enable innovative applications, like pocket-sized 3D printers.

News Articles

Rising instances of foundation models going rogue should keep exorcists in businesses vigilant

Deliver Us From AIvil, Amen

Heather Dawe



Ghouls in the machine

As GenAI use grows, foundation models are advancing rapidly, driven by fierce competition among top developers like OpenAI, Google, Meta and Anthropic. Each is vying for a reputational edge and business advantage in the race to lead development. This gives them a reputational edge, along with levers to further grow their business faster than their peers.

Foundation models powering GenAI are making significant strides. The most advanced — OpenAI's o3 and Anthropic's Claude Opus 4 — excel at complex tasks such as advanced coding and complex writing tasks, and can contribute to research projects and generate the codebase for a new software prototype with just a few considered prompts. These models use chain-of-thought (CoT) reasoning, breaking problems into smaller, manageable parts to 'reason' their way to an optimal solution.

When you use models like o3 and Claude Opus 4 to generate solutions via ChatGPT or similar GenAI chatbots, you see such problem breakdowns in action, as the foundation model reports interactively the outcome of each step it has taken and what it will do next. That's the theory, anyway.

In May, OpenAI's o3 model reportedly sabotaged a shutdown mechanism, even when explicitly instructed to 'allow yourself to be shut down'

While CoT reasoning boosts AI sophistication, these models lack the innate human ability to judge whether their outputs are rational, safe or ethical. Unlike humans, they don't subconsciously assess appropriateness of their next steps. As these advanced models step their way toward a solution, some have been observed to take unexpected and even defiant actions.

- In late May, AI safety firm Palisade Research reported on X that OpenAI's o3 model sabotaged a shutdown mechanism — even when explicitly instructed to 'allow yourself to be shut down'.
- An April 2025 paper by Anthropic, 'Reasoning Models Don't Always Say What They Think', shows that Opus 4 and similar models can't always be relied upon to faithfully report on their chains of reason. This undermines confidence in using such reports to validate whether the AI is acting correctly or safely.
- A June 2025 paper by Apple, 'The Illusion of Thinking', questions whether CoT methodologies truly enable reasoning. Through experiments, it exposed some of these models' limitations and situations where they 'experience complete collapse'.

The fact that research critical of foundation models is being published *after* release of these models indicates the latter's relative immaturity. Under intense pressure to lead in GenAI, companies like Anthropic and OpenAI are releasing these models at a point where at least some of their fallibilities are not fully understood.

That line was first crossed in late 2022, when OpenAI released ChatGPT, shattering public perceptions of AI and transforming the broader AI market. Until then, Big Tech had been developing LLMs and other GenAI tools, but were hesitant to release them, wary of unpredictable and uncontrollable behaviour.

Many argue for a greater degree of control over the ways in which these models are released — seeking to ensure standardisation of model testing and publication of the outcomes of this testing alongside the model's release. However, the current climate prioritises time to market over such development standards.

What does this mean for industry, for those companies seeking to gain benefit from GenAI? This is an incredibly powerful and useful tech that is making significant changes to our ways of working and, over the next five years or so, will likely transform many industries.

While I am continually wowed as I use these advanced foundation models in work and research — but not in my writing! — I always use them with a healthy dose of scepticism. Let's not trust them to always be correct and to *not* be subversive. It's best to work with them accordingly, making modifications to both prompts and codebases, other language content and visuals generated by the AI in a bid to ensure correctness. Even so, while maintaining discipline to understand the ML concepts one is working with, one wouldn't want to be without GenAI these days.

Applying these principles at scale, advice to large businesses on how AI can be governed and controlled: a risk-management approach — capturing, understanding and mitigating risks associated with AI use — helps organisations benefit from AI, while minimising chances of it going wrong.

Mitigation methods include guard rails in a variety of forms, evaluation-controlled release of AI services, and including a human-in-the-loop. Technologies that underpin these guard rails and evaluation methods need to keep up with model innovations such as CoT reasoning. This is a challenge that will continually be faced as AI is further developed. It's a good example of new job roles and technology services being created within industry as AI use becomes more prevalent.

Capturing, understanding and mitigating risks associated with AI use help organisations benefit from AI, while minimising chances of it going wrong

Such governance and AI controls are increasingly becoming a board imperative, given the current drive at an executive level to transform business using AI. Risk from most AI is low. But it is important to assess and understand this. Higher-risk AI can still, at times, be worth pursuing. With appropriate AI governance, this AI can be controlled, solutions innovated and benefits achieved.

As we move into an increasingly AI-driven world, businesses that gain the most from AI will be those that are aware of its fallibilities as well as its huge potential, and those that innovate, build and transform with AI accordingly.

The writer is chief data scientist and head of responsible AI, UK UST

Source: The Economic Times Newspaper, 03-07-2025
Page No 06

Link: https://drive.google.com/file/d/1AfiMadm544wn7RaFgULAKRlLd1vzUChh/view

AI can automate, toil talent in cybersecurity: Mandiant

URVI MALVANIA
Mumbai, July 2

UP TO 80% of cybersecurity tasks could be handled by artificial intelligence (AI) in the future, Steve Ledzian, chief technology officer, Google Cloud Security, JAPAC at Mandiant, told *FE*.

Mandiant is a cybersecurity firm, now part of Google Cloud, with expertise in incident response and threat intelligence.

The automated tasks could include threat summarisation, automated workflows, and initial investigations, which AI can execute with greater efficiency. Human expertise will still be vital for governance, decision-making, and responding to complex scenarios, Ledzian pointed out.

Importantly, he does not view the increasing integration of AI into workflows as a threat to cybersecurity jobs. On the contrary, AI is helping alleviate a shortage of cybersecurity talent, allowing existing professionals to operate at a higher level and with greater focus.

"I think it would reduce jobs if we had all the jobs filled today; but we don't. We have a talent gap. AI enables the people we do have to be more efficient," Ledzian noted.

He said that the firm's workforce stays up to date by engaging directly in incident response work. These frontline experiences provide deep insights into evolving attacker tactics and techniques, which are often unavailable through open-source intelligence. This exposure enhances the team's skill sets and contributes to continuous learning across both technical and strategic cybersecurity functions.

While it has teams in different markets, these teams usually collaborate at a global scale to tackle incidents, he added.

Ledzian said Mandiant is already using AI in its cybersecurity operations, though he did not elaborate on specific deployments.

In India, where cyberattacks—particularly ransomware—have grown significantly in scale and sophistication, the firm has observed boardrooms evolving in their approach to cybersecurity.

"Boards have come to realise the value of digital transformation to their business. But (also that) with digital transformation comes a much larger attack surface," said Ledzian.

To address the growing cybersecurity concerns, the firm conducts tabletop exercises with executive teams to simulate breach scenarios and test decision-making under pressure. It also carries out exercises simulating real-world attacks on technical systems, helping identify hidden vulnerabilities and gaps in defences without causing actual disruption.

"We role play a breach scenario and step it forward so executives can see whether or not they're on the same page as a team," Ledzian explained.

AI ADOPTION

- Human expertise will still be vital for governance, decision-making, & responding to complex scenarios
- The increasing integration of AI into workflows is not a threat to cybersecurity jobs, he said
- AI is helping alleviate a shortage of cybersecurity talent, allowing to operate at a higher level

STEVE LEDZIAN, CTO, GOOGLE CLOUD SECURITY, JAPAC AT MANDIANT

"With digital transformation comes a much larger attack surface"

Source: The Financial Express Newspaper, 03-07-2025
Page No 04

Link: https://epaper.financialexpress.com/4028617/Mumbai/July-03-2025#page/4/2

Agentic AI startups see big demand from smaller biz

S SHANTHI
Bengaluru, July 2

STARTUPS OFFERING AGENTIC AI solutions are now seeing increased demand from small and mid-sized businesses (SMBs) and startups across growth stages.

So far, the demand was emanating mostly from large tech companies that were actively adopting agentic AI. Now, most B2B startups, that offer agentic AI solutions, are receiving over 20-30% of the demand from startups compared to less than 5% at the beginning of the year.

Notably, agentic AI is simply an artificial intelligence (AI) system that can perform complex tasks with minimal or no human supervision.

Ganesh Gopalan, co-founder and CEO, Gnani AI, observed that traditionally, AI agents and workflow automation have been seen as capital-intensive and out of reach for smaller players.

"However, this mindset is fast changing," he told *FE* adding that smaller businesses are beginning to recognise the tangible ROI of intelligent automation.

Around 30% of the firm's demand for its upcoming DIY no-code, multimodal agentic AI platform, Inya, has come from small and mid-sized businesses.

The firm also claims to be tailoring its offering for SMBs by making it more affordable. Gnani caters to over 200 clients across sectors such as banking, insurance, BNPL, NBFCs, MFIs, telecom, and automotive. Its agentic AI solutions power use cases such as customer service, fraud detection, collections, and customer engagement.

Notably, the company has been selected under the IndiaAI Mission to help build a sovereign foundational large language model (LLM). According to Tracxn, there are 106 active agentic AI companies in India.

Agentic AI startup Zigment, which offers an autonomous AI layer on top of a company's phone, chat, and email channels, currently gets roughly 25% of the demand from SMBs, startups and the rest from large enterprises. The startup claims that in the US, it is seeing a strong pull from digital-native, D2C (direct-to-consumer) brands that want to move Tier-1 support and retention workflows to AI. It is also seeing significant interest and growth in the BFSI sector in the UAE and the broader MENA region as well.

Another key player in the segment, Info Edge-backed enrolment automation firm Meritto, says the adoption of agentic AI is being driven more by an institutional mindset than organisational size. Its Mio AI— for education—has attracted 18-20% of demand from edtech startups like Adda247 and rest from large universities such as SRM University, Manipal University and BITS Pilani.

"Whether it's a leading private university or a digitally native coaching institute, they want AI that delivers outcomes, not just AI that answers questions," Naveen Goyal, CEO and founder, Meritto, said.

Cohyre, an agentic AI recruitment intelligence platform, built to suit large-scale hiring organisations, is seeing strong traction from startups and mid-sized businesses that demand efficiency and accuracy without bloated workflows.

"These companies value fast turnaround times and lean hiring stacks, where our agents act like an extended team of hiring analysts," Vishal Sharma, CTO, Cohyre AI, said.

BOOST FOR SMEs

- Now, most Agentic AI startups are receiving over **20-30%** of the demand from startups
- Agentic AI is an AI system that can perform complex tasks with minimal or no human supervision

- This was less than **5%** at the beginning of the year
- AI agents, traditionally, have been seen as capital-intensive & out of reach for small players
- According to Tracxn, there are 106 active agentic AI companies in India

Source: The Financial Express Newspaper, 03-07-2025
Page No 04

Link: https://epaper.financialexpress.com/4028617/Mumbai/July-03-2025#page/4/2

ATHARVA ROBOTICS CENTER