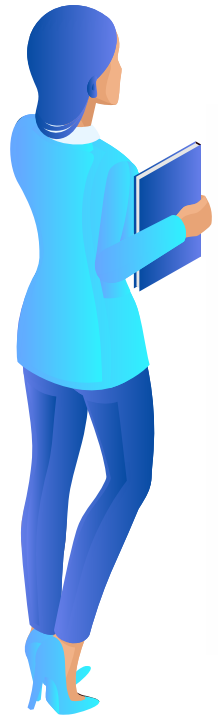
The Future of AI is Hybrid

2025 TECHnalysis Research
Hybrid AI Survey

technalysis
RESEARCH

Survey Purpose and Methodology



Survey Objectives

The survey aimed to assess Hybrid AI adoption, deployment, and strategic direction in U.S. organizations across industries.

Target Respondents

1,026 IT decision-makers from medium and large enterprises across 10 industries participated in the survey.

Data Collection Method

Data was collected via a web-based questionnaire during mid-2025 to gather detailed Hybrid AI insights.

Executive Summary: 3 Key Takeaways

The Great AI Repatriation is Here

The 'cloud-first' era is evolving. Driven by cost and security, a massive wave of organizations is actively moving AI workloads *back* from the public cloud to private infrastructure.

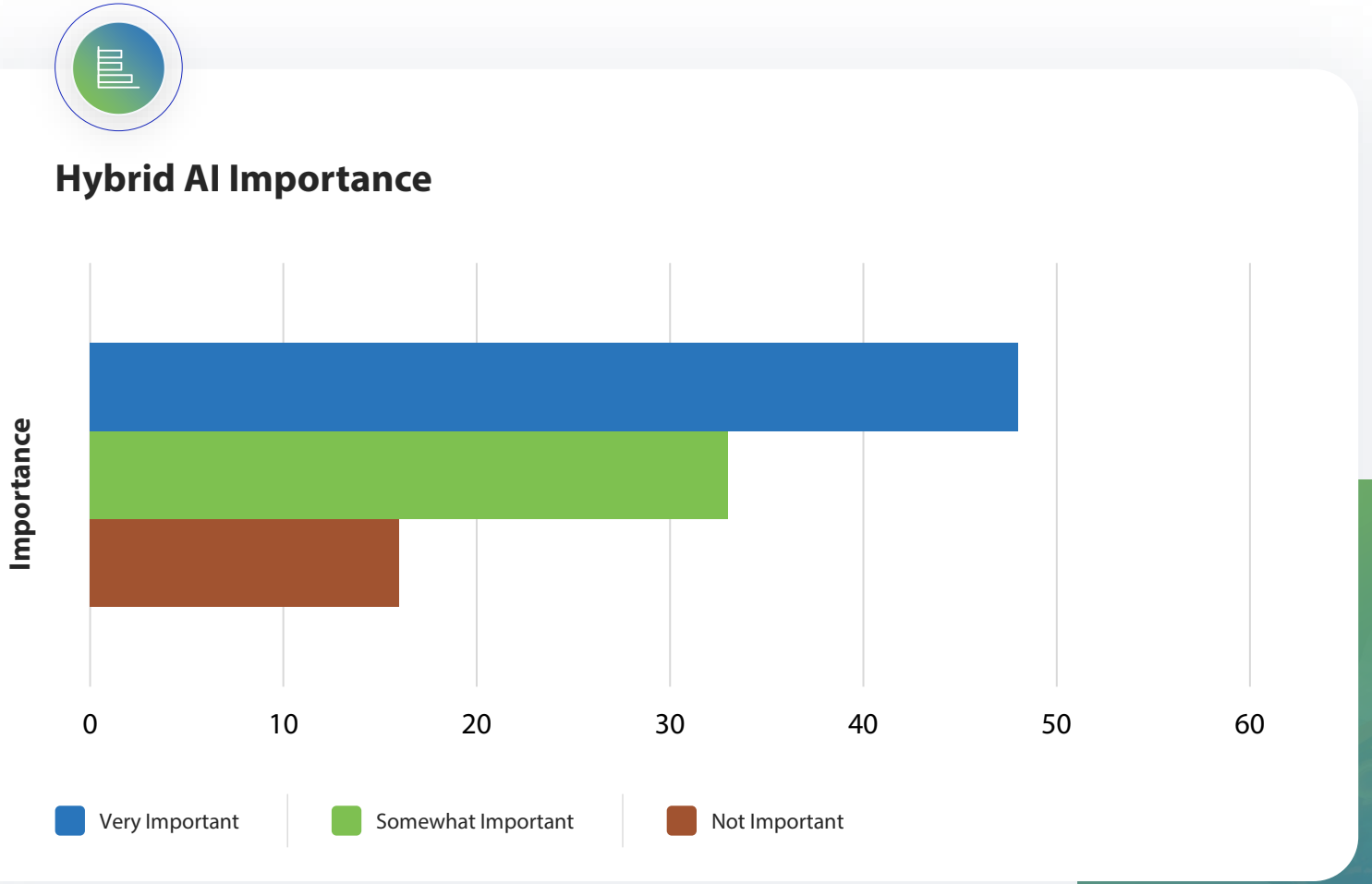
The Future is a 3-Way Hybrid Computing Split

The era of public cloud dominance for AI is ending. Within two years, organizations expect a balanced split of AI workloads between Public Cloud, Private Cloud, and Edge devices.

The AI PC is Making an Impact

While initial adoption has been a bit sluggish, a staggering 85% of decision-makers believe on-device NPUs are important for their AI applications, signaling a massive market pull for edge computing.

Hybrid AI is a Business Imperative



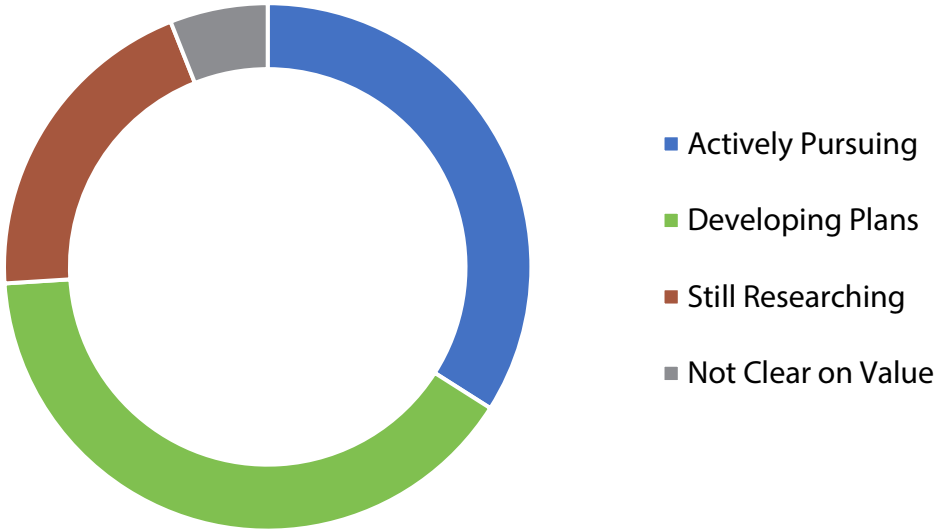
"Currently, most AI workloads run on the public cloud, but we are expanding into private clouds and edge computing to enhance privacy, reduce latency, and support real-time applications."

Survey Respondent

Over 80% of all decision-makers believe a Hybrid AI architecture is important for their organization

Organizations are Moving Beyond Research

Current Interest/Activity in Hybrid AI



Nearly 75% of organizations are either actively pursuing or developing plans for Hybrid AI. The time for research is over; the time for planning and deployment is now.

technalysis RESEARCH

The Top 3 Drivers for Hybrid AI: Cost, Security, Privacy

Cost Reductions

The number one driver. Running all AI workloads, especially GenAI, in the public cloud is proving to be expensive.

Data Security & Compliance

Organizations need to maintain control over their data, especially in regulated industries.

Enhanced Data Privacy

Keeping sensitive data on-premise or on-device is a non-negotiable requirement for many new AI applications.

The Top 3 Challenges for Hybrid AI: Data, People, Cost

Data Preparation & Integration

The classic 'garbage in, garbage out' problem, now scaled up. Getting data ready for AI is the top bottleneck.

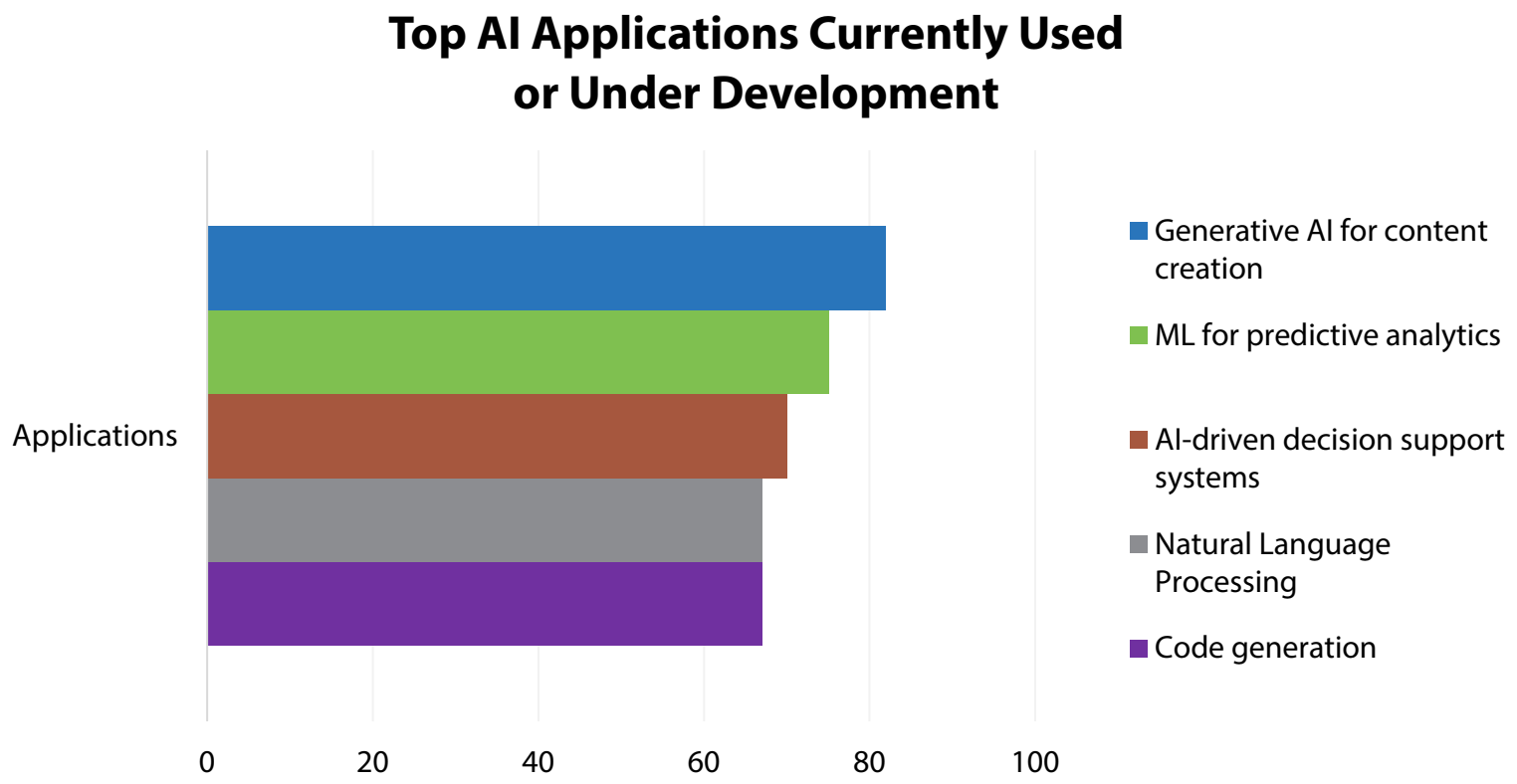
Lack of Skilled Personnel

The talent gap is real. Finding people who can build, manage, and integrate these complex systems is a major hurdle.

High Costs of Infrastructure

While 'Cost Reduction' is a driver, the upfront capital expense for on-prem AI servers (e.g., GPUs) is a significant barrier.

What's Driving the Need? GenAI



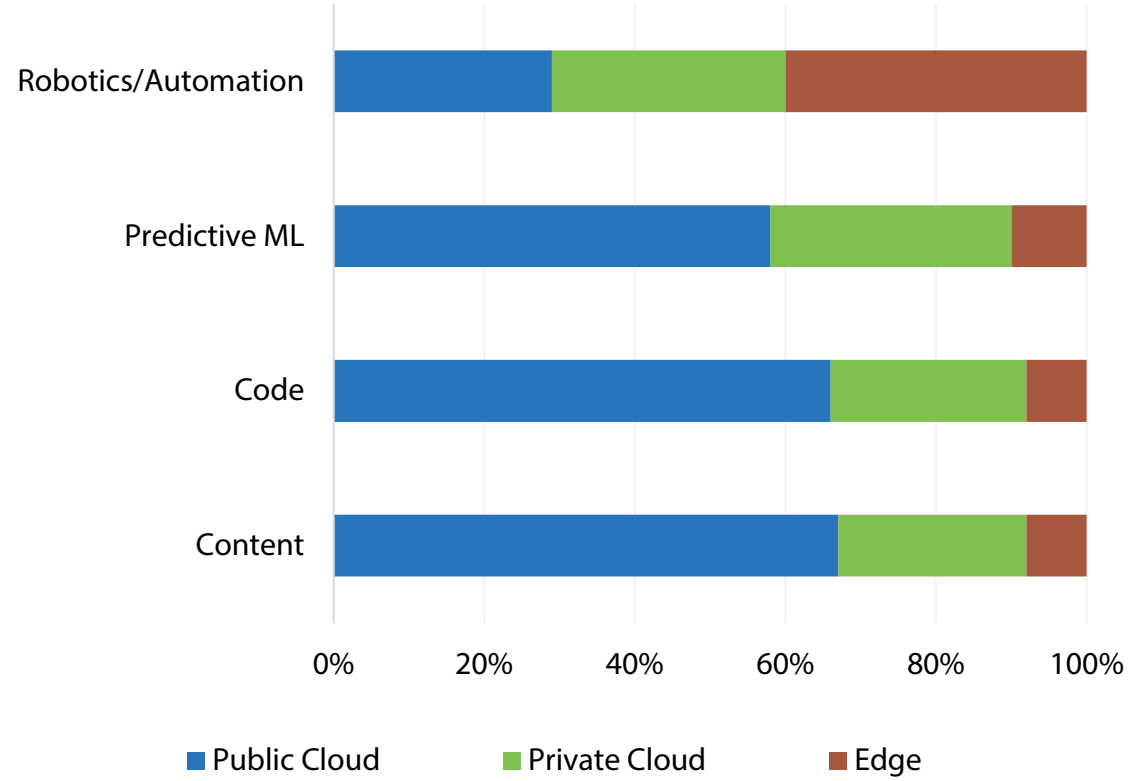
Generative AI for content creation is the #1 application being used or developed, followed closely by traditional ML-powered data analytics.

More organizations are developing AI decision support and chatbot-style NLP systems, but Code generation is the third highest application already in use.

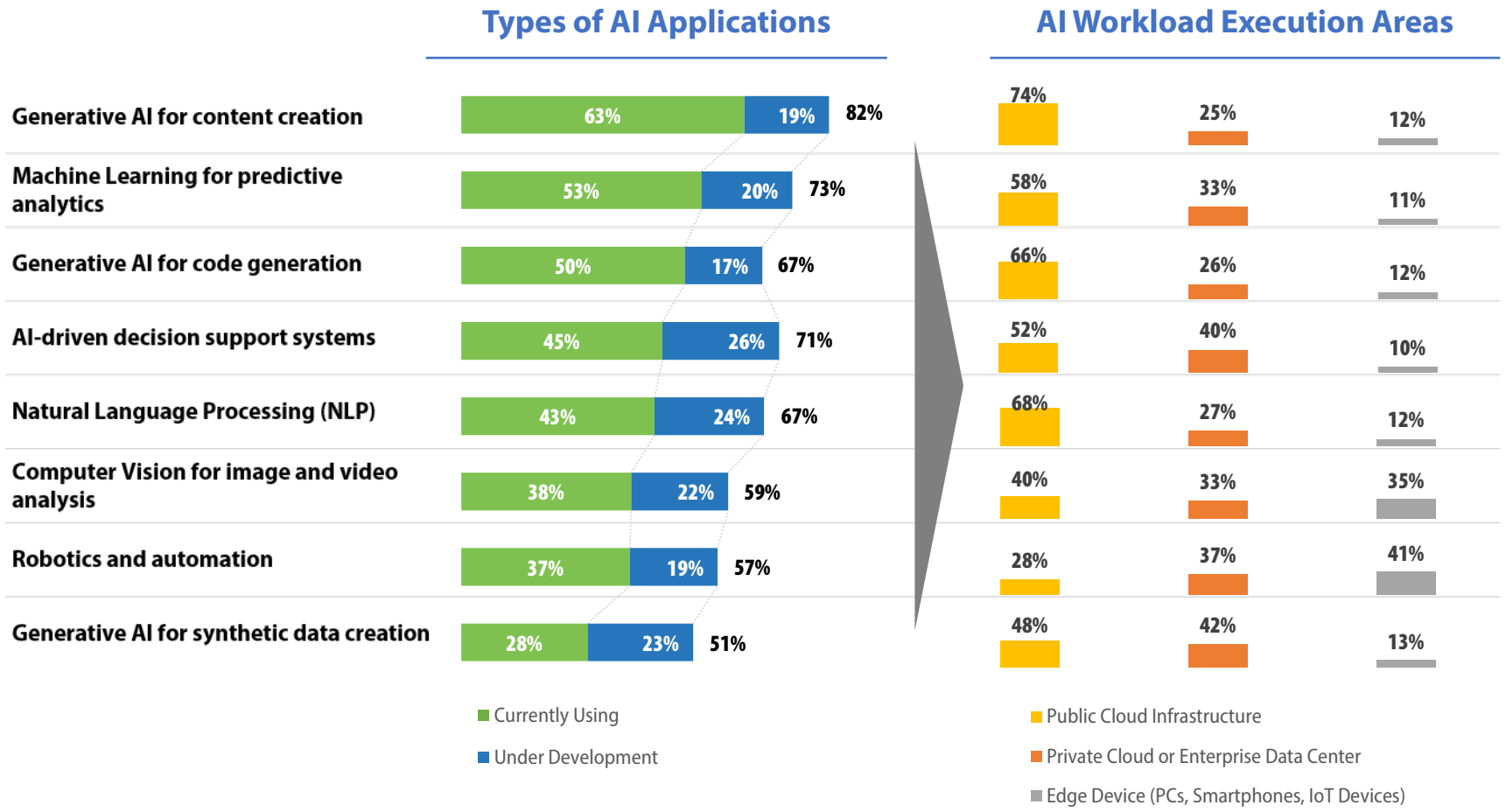
Where AI Workloads Run Today

For most popular GenAI and ML apps, the Public Cloud is the dominant venue today. The major exception is Robotics & Automation, which already leans heavily on the Edge

Current Execution Location by Application Type



GenAI Workloads by Location



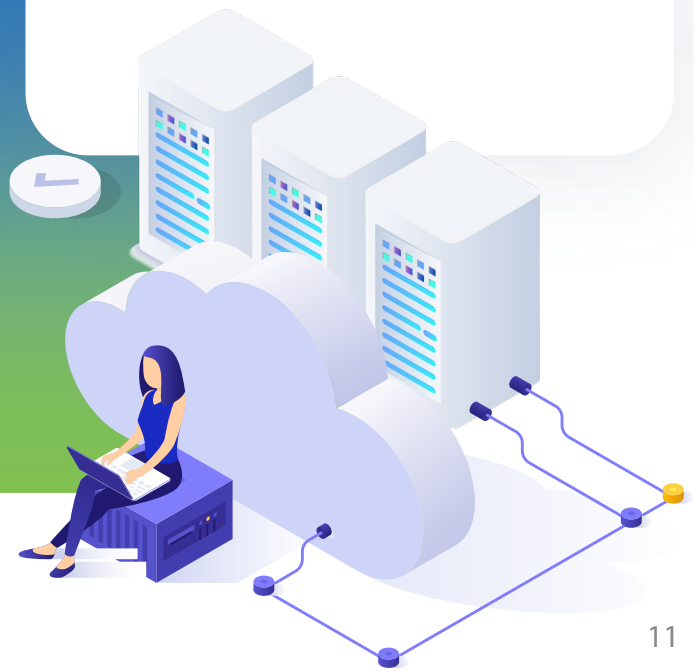
Talking Point: The 'Cloud-First' Era For AI is Over

The narrative of the last decade was 'move everything to the cloud.' The data shows this trend is now reversing when it comes to AI workloads.

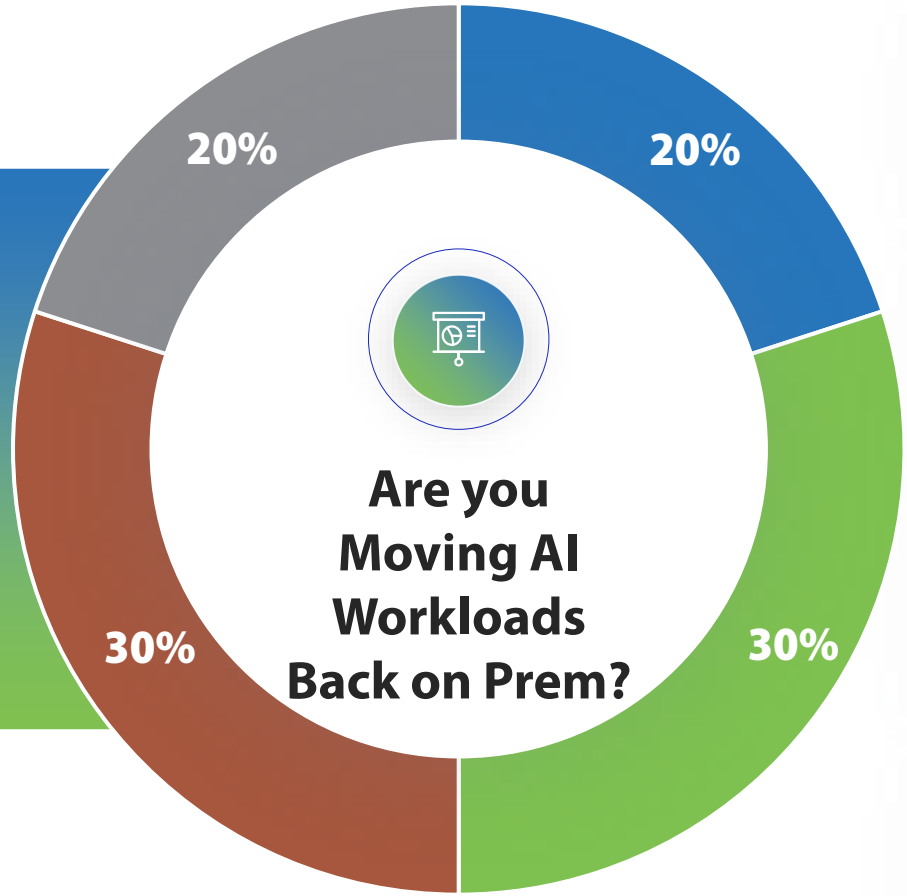
Organizations have hit a tipping point where the drivers for leaving the public cloud—cost, security, and privacy—are beginning to outweigh the benefits of staying—again, in the case of AI applications.

This isn't a failure of cloud; it's a maturation of the market. The new strategy is 'cloud-smart', not 'cloud-first'.

*"Hybrid AI has enabled us to balance performance, security, and flexibility by using public cloud for training, private cloud for sensitive data, and edge for real-time inference."—**Survey Respondent***



The Great AI Repatriation is Happening Now



- Yes, we already have
- Yes, we are planning to
- Not yet, but thinking about it
- No, and no plans

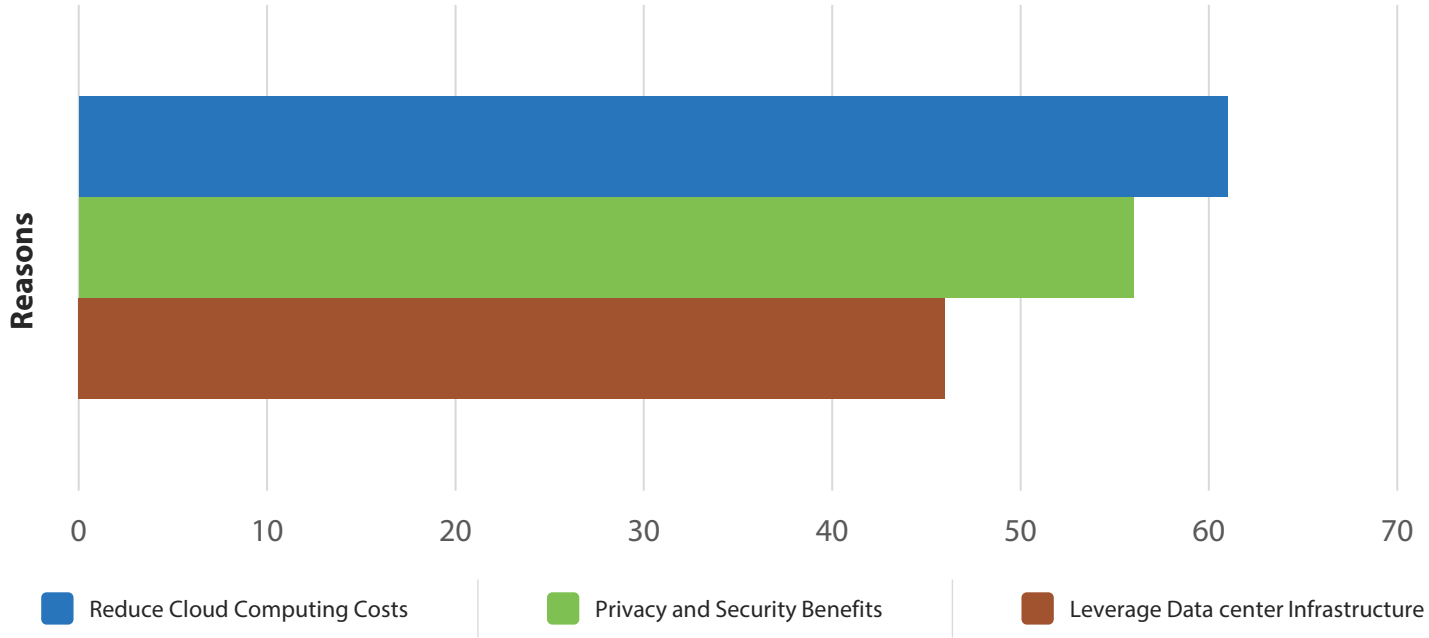
80% of all organizations have either already repatriated AI workloads, are actively planning to, or are considering it.

Only **20%** have no plans to do so.

technalysis RESEARCH

Why Are They Leaving? (It's Not Complicated)

Top 3 Reasons for Repatriation



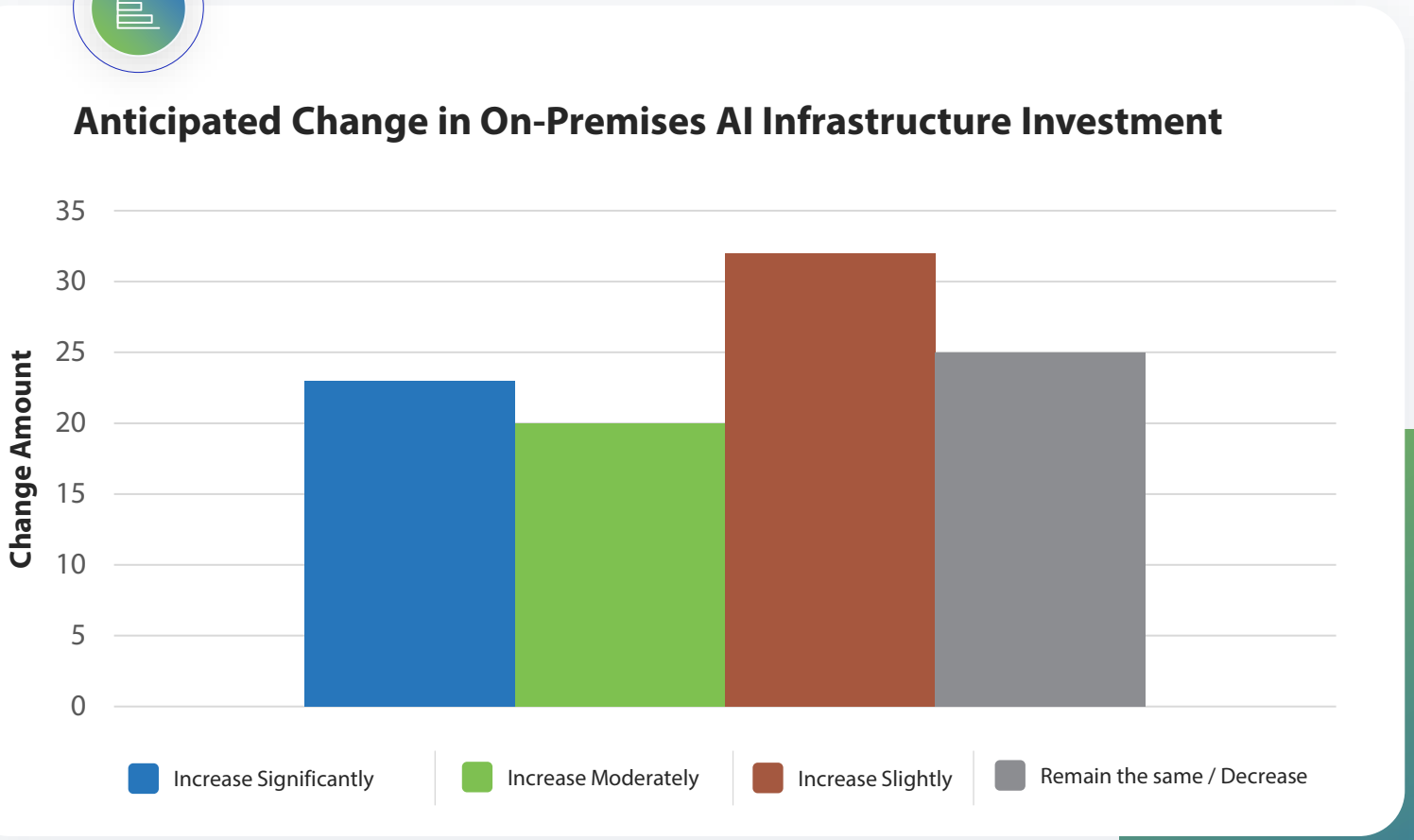
"Initially we aimed to cut cloud bills, and now our entire roadmap has changed."

Survey Respondent

The top two reasons are identical to the top drivers for Hybrid AI in general: Cost Control and Privacy/Security. Many also want to leverage on-prem investments they've already made.

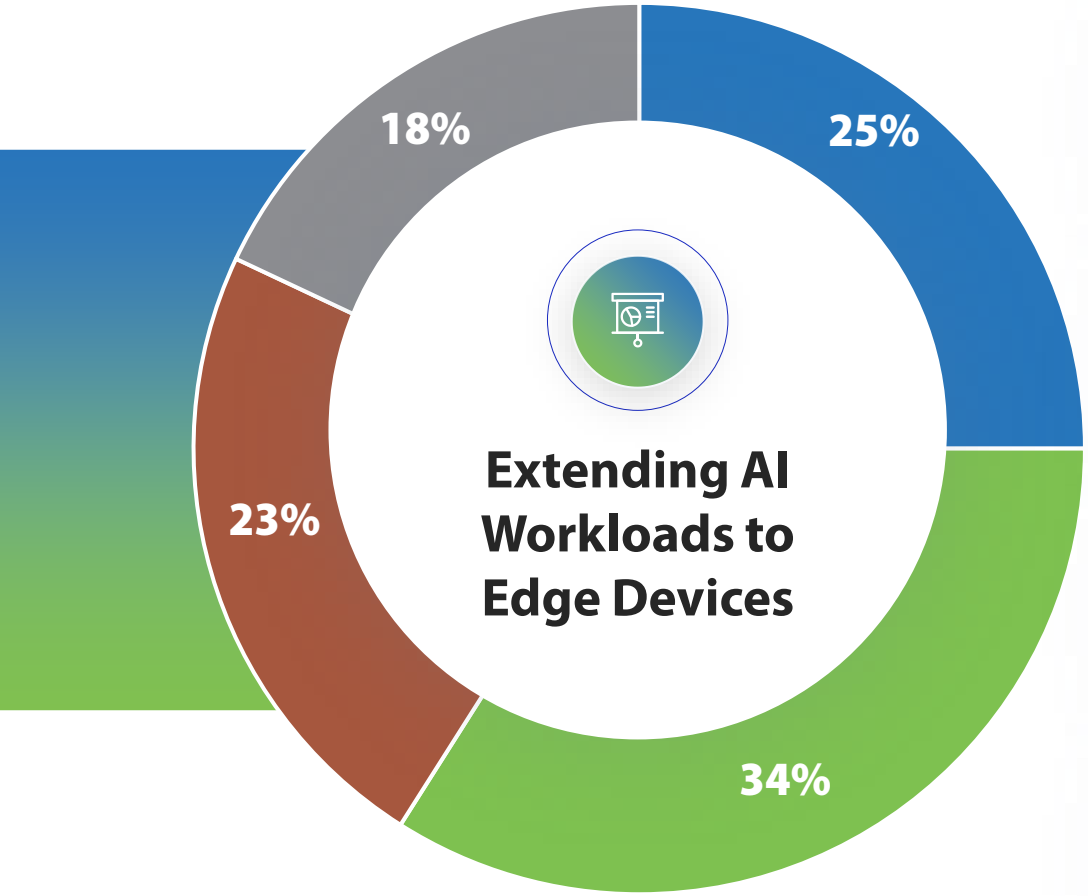
technalysis RESEARCH

Investment is Shifting to On-Prem



Over 77% of organizations expect their investment in on-premises AI infrastructure to increase over the next 1-3 years. This is where the budget is flowing.

The Next Frontier: AI to the Edge



- Already have them in place
- Yes, in the planning stages
- Under consideration
- No

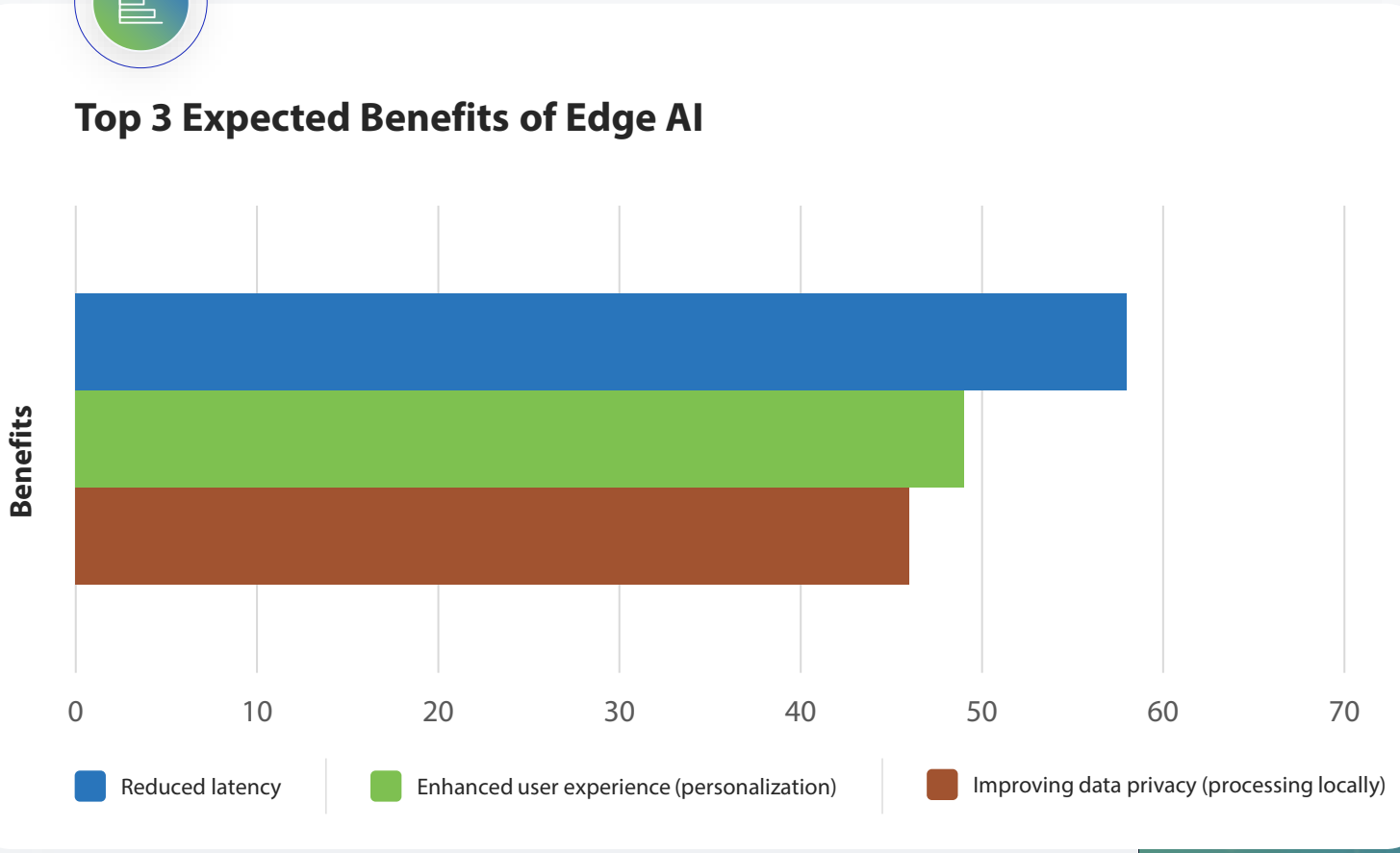
"Edge delivers instant action; cloud learning keeps our models sharp over time."

Survey Respondent

Nearly **60%** of organizations have already extended AI to the edge or are in the planning stages.

technalysis RESEARCH

Why the Edge? Latency, UX, and Privacy



"Hybrid AI Enables us to balance public cloud scalability with private infrastructure for sensitive data while edge deployment supports real time use. In [the] future, we anticipate deeper integration of hybrid workflows-leveraging public cloud for training [and] on-premises for low-latency inference to optimize costs, ensure compliance and unify governance across environments by 2026."

Survey Respondent

The benefits of Edge AI are distinct from cloud. It's about real-time performance (latency), personalization, and keeping user data on the device (privacy).

The AI PC Enigma

"We currently use hybrid AI for cost efficient, compliant workloads across cloud and edge, and plan to expand with AI agents and NPU-powered devices to optimize performance and scalability."

Survey Respondent

85% Today

An overwhelming 85% of decision-makers believe it is *currently* important (Essential, Very, or Somewhat) for PCs to have NPUs.

90%+ in 2 Years

That number jumps to over 90% when asked about the next 2 years. This is a powerful, fast-moving consensus.

This is one of the most surprising findings. Despite a slow start, the awareness of the importance of AI-accelerated client hardware already exists. The Education sector (96%) sees the highest current importance.

Top Challenges for Edge AI

Device Resource Constraints

- Power, memory, etc.

Managing & Updating Models

- How to update thousands of distributed devices

Developing Edge-Optimized AI Models

- Models must be smaller and more efficient

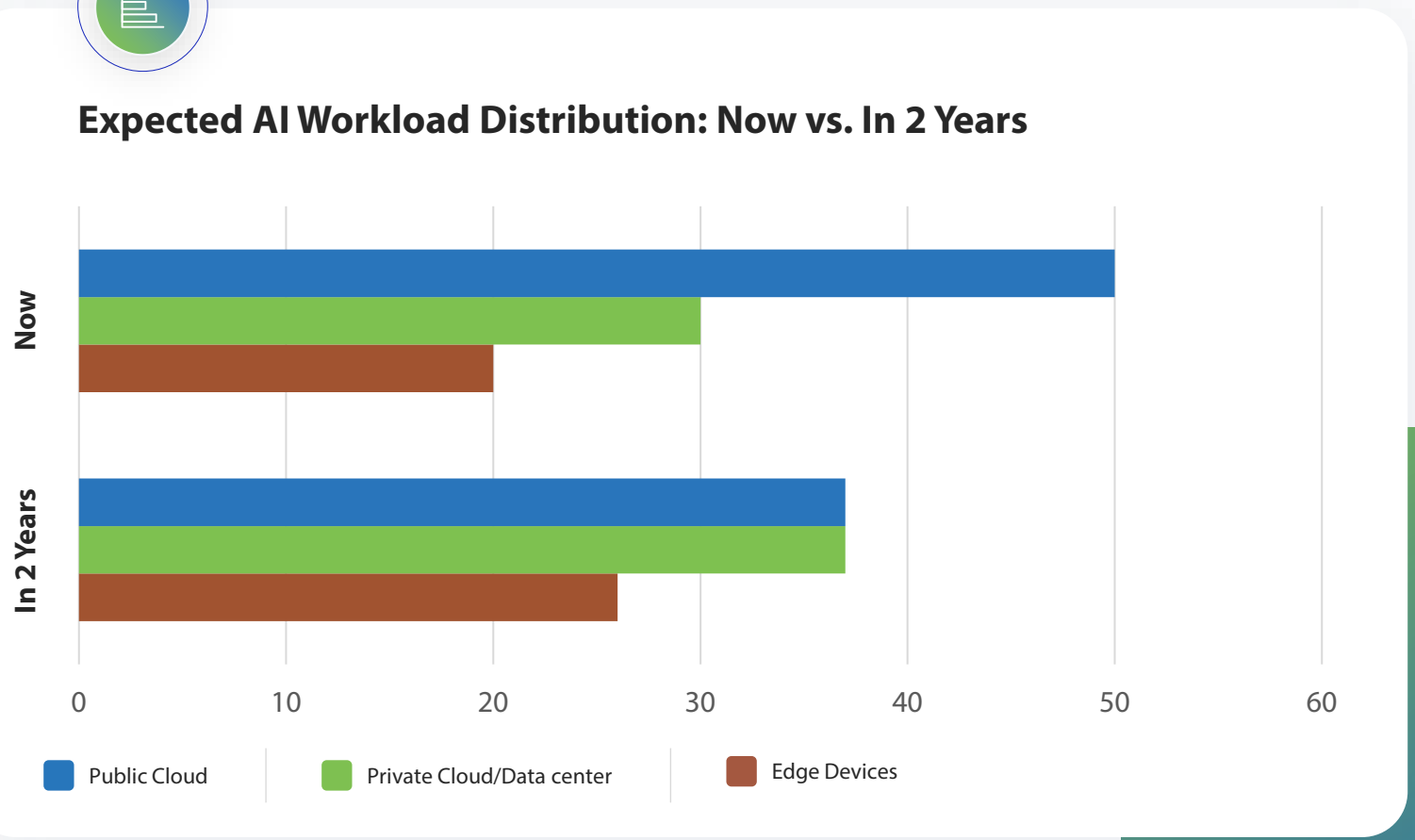
Data Security & Privacy at the Edge

- Protecting data on vulnerable client devices

The challenges of the edge are very different from the cloud. It's less about data prep and more about managing a large, distributed, and resource-constrained fleet of devices.

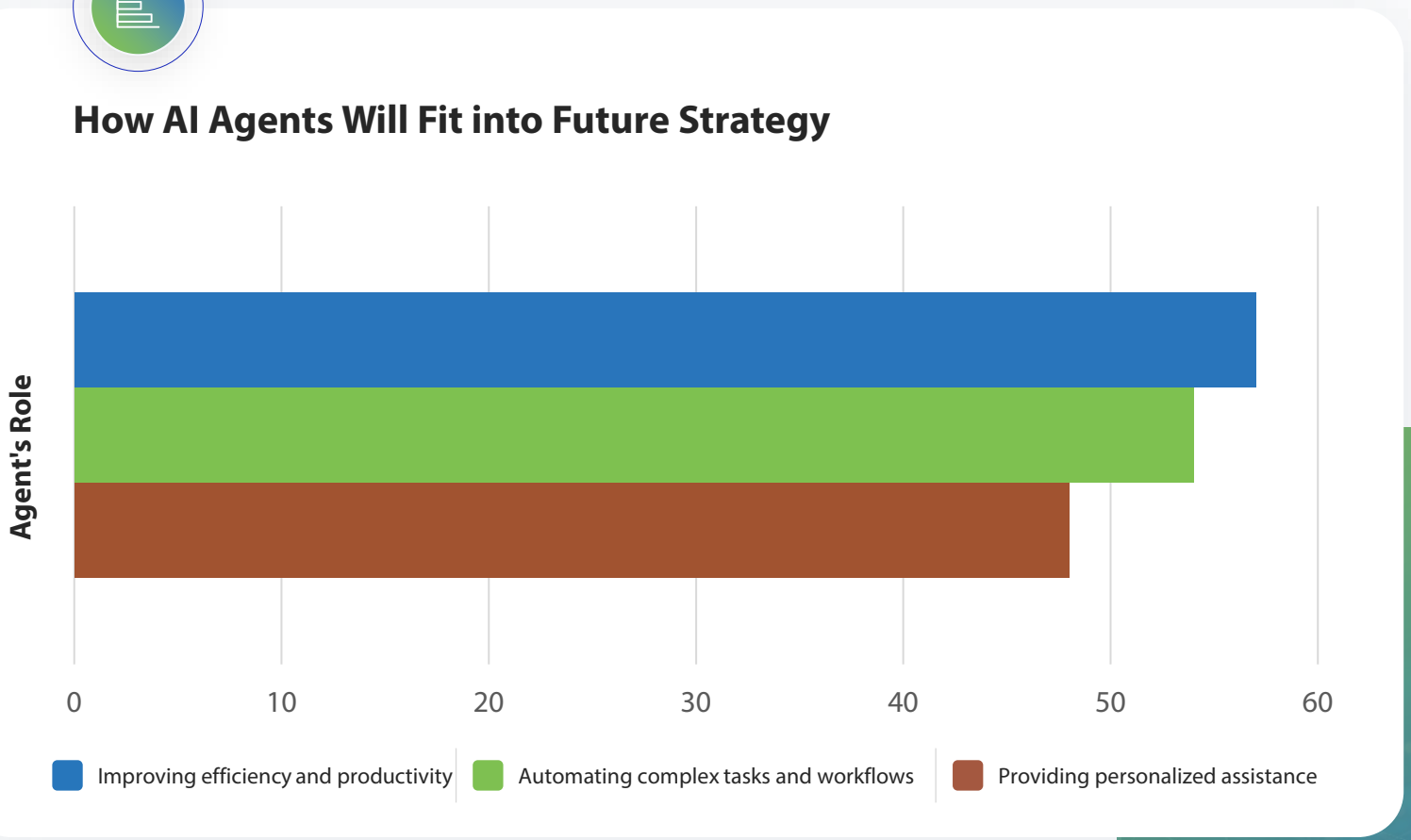
technalysis RESEARCH

Most Important Takeaway: The 3-Way Split



Public Cloud's share is expected to drop by 10-15 percentage points, while Private Cloud and Edge increase to create a balanced, three-pillar architecture.

The Final Piece: The Rise of AI Agents



The vast majority (94%+) of organizations see a significant role for AI agents. The top uses are improving productivity and automating complex tasks.

Top Concerns for AI Agents: Control & Security

While excitement for agents is high, the primary concerns are fundamental: security, privacy, and control. This is especially true in a hybrid model where an agent may span all three workload locations.

Assuring the security and privacy of data accessed by agents

- What can they see?

Maintaining control and oversight of agent actions

- What can they do?

Managing the complexity of distributed deployments

- How do we manage them everywhere?

Ensuring reliability and robustness

- Will they work consistently across cloud, private, and edge?

Conclusions

Hybrid AI is the new default

- The market has moved past 'cloud-first' to 'cloud-smart,' balancing workloads based on specific needs.

The key drivers are universal

- Cost, Security, and Privacy. This trend is a direct response to the high cost and data risks of large-scale GenAI.

The future is a balanced 3-way split

- Workloads will be optimized across Public Cloud, Private Data Centers, and Edge Devices.

The AI PC is the new battleground

- The demand for on-device processing (NPUs) is an immediate and surprisingly strong trend, fueled by the need for low-latency, private, and personalized AI experiences.

Contact

Bob O'Donnell

President and Chief Analyst

TECHnalysis Research, LLC

1136 Halsey Blvd.

Foster City, CA 94404

bob@technalysisresearch.com

(650) 224-2355

@bobodtech

www.technalysisresearch.com