Manage AI Risks Before They Manage You

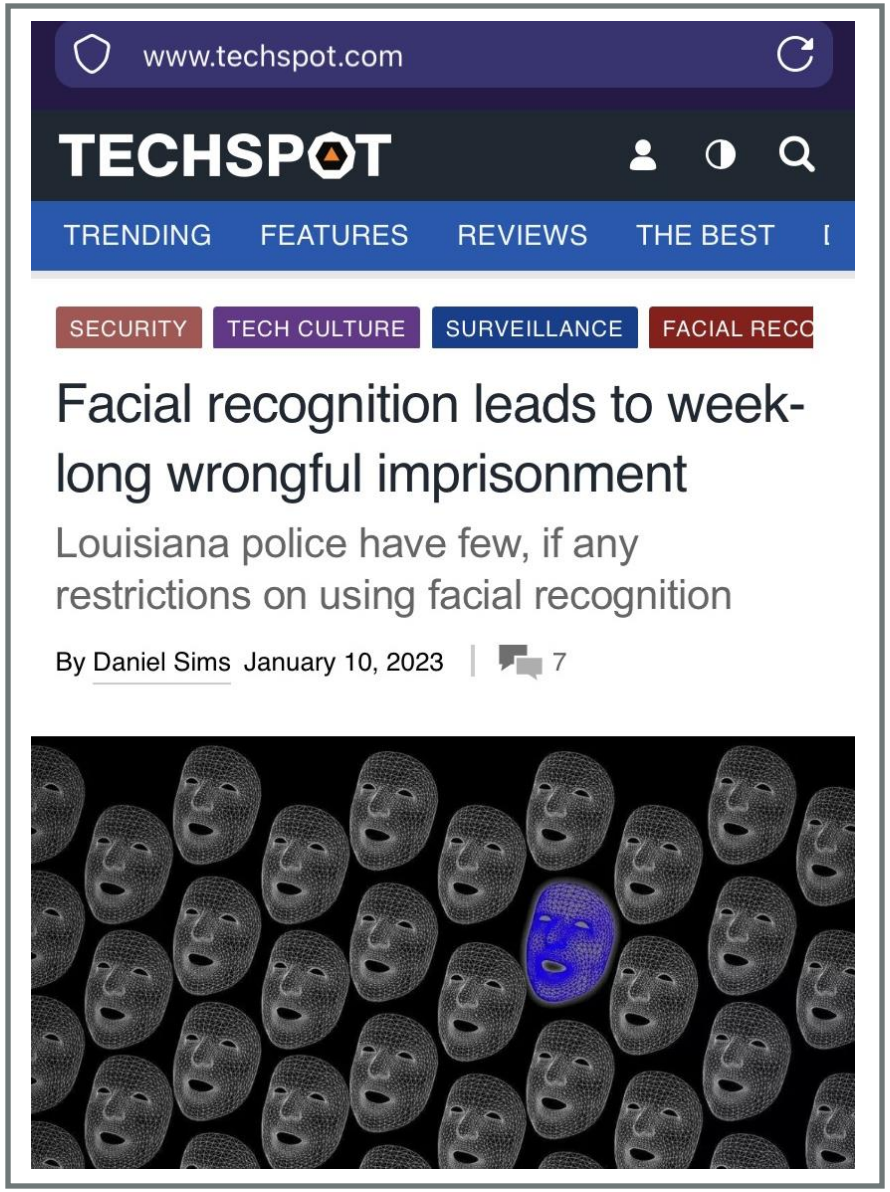
Bern Elliot

Gartner®

AI ‘Misperformance’ Can Threaten Human Life

How About:

- Exam Proctoring
- Uncorroborated Information Use
- Made-Up Data Relied on as Truth
- Failing Self-Driving Cars
- Deepfake-Infused Fraud
- Employment Opportunity
- Imprisonment
- Bidirectional Automation Bias



Source: Facial Recognition Leads to Week-Long Wrongful Imprisonment, TechSpot

Gartner

Key Issues

1

Where, when and how can AI be compromised?

2

What you need to do about new AI risks?

Gartner®

Key Issues

1

Where, when and how can AI be compromised?

2

What you need to do about new AI risks?

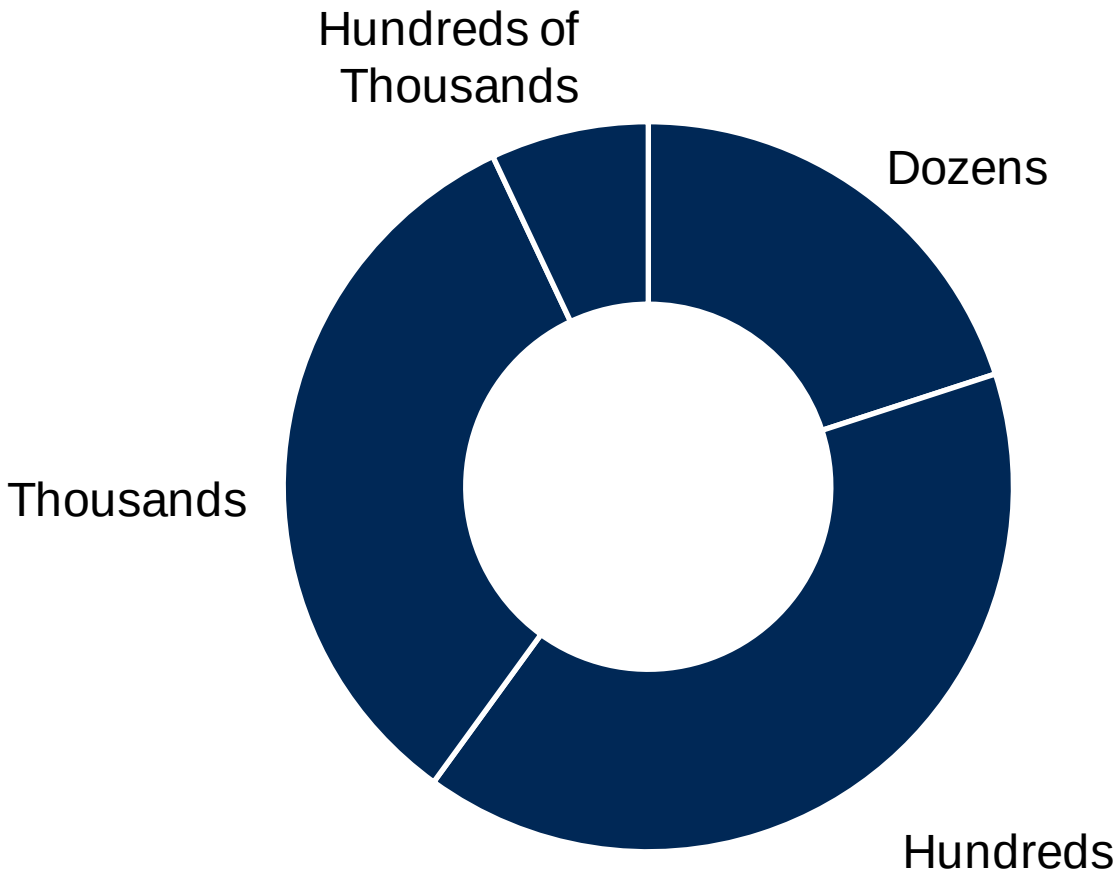
Gartner®

Plenty of Models to Compromise and Attack

Number of AI Models Deployed to Date

73%

of Organizations Have Hundreds or Thousands of Models Deployed



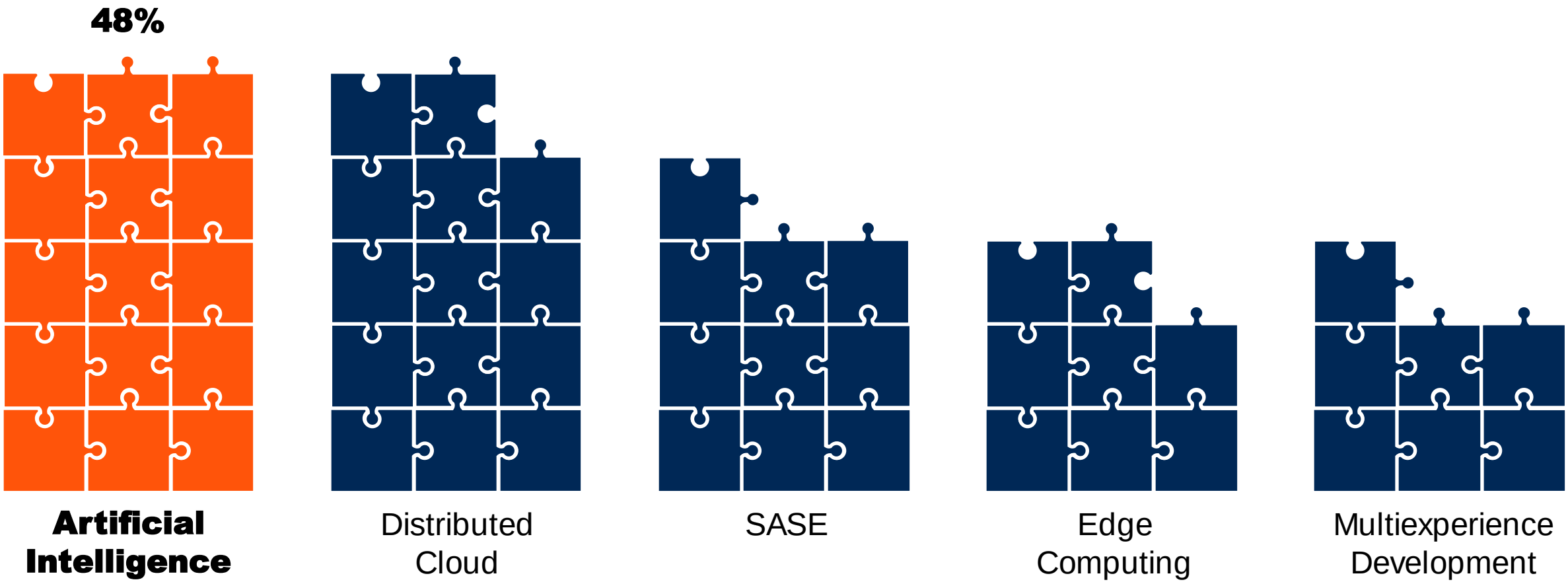
n = 324; Base: Using AI (S08), Excludes Unsure
Q13A: How many AI models has your organization deployed to date?
Source: 2021 Gartner AI in Organizations Survey

Gartner®

AI Solutions Are the Top Emerging Technology …

… To Be Deployed or Close to Deployment Across Enterprises

Emerging Technologies Deployed or Planned to Deploy in Next 12 Months



n = 2,186; CIOs and Technology Executives

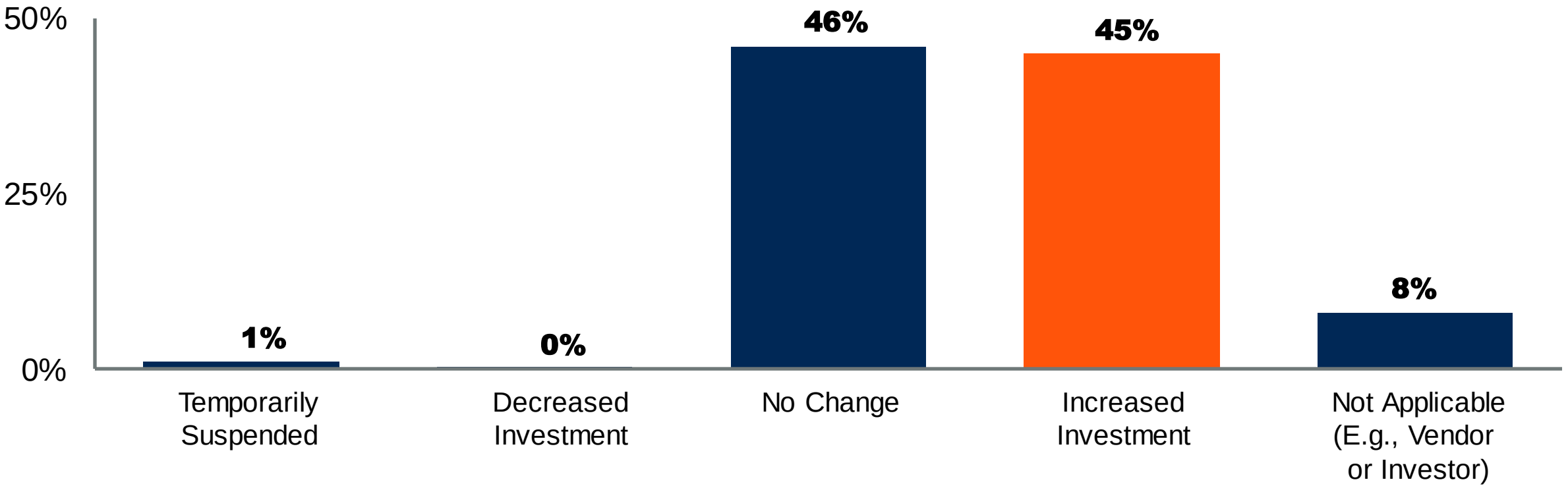
Source: 2023 Gartner CISO and Technology Executive Survey

Gartner®

ChatGPT Spurred a Major Increase in AI Investment

Change in AI Investment Since ChatGPT

Percentage of Respondents



n = 2,554 (30 March and 21 April)

Q: How have your AI investment strategies changed since the recent publicity of ChatGPT?

Source: Beyond the Hype: Enterprise Impact of ChatGPT and Generative AI

Gartner®

Summary: Key Findings From AI Enterprise Survey

Two in five organizations had an AI breach.

One in four were malicious attacks.

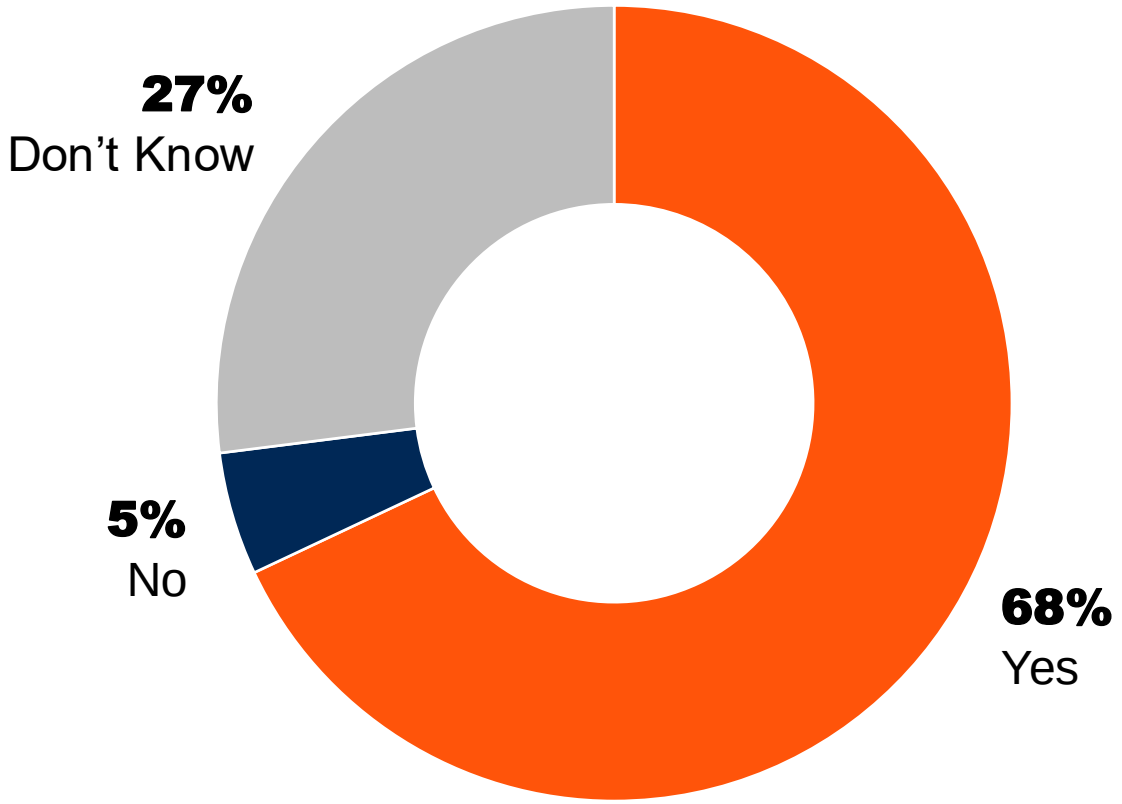
Two in three organizations have a task force on AI risks.

Active controls for AI risk, privacy and security **achieve better AI project results.**

Gartner®

Most Say — Benefits Outweigh the Risks

Percentage of Respondents Who Agree



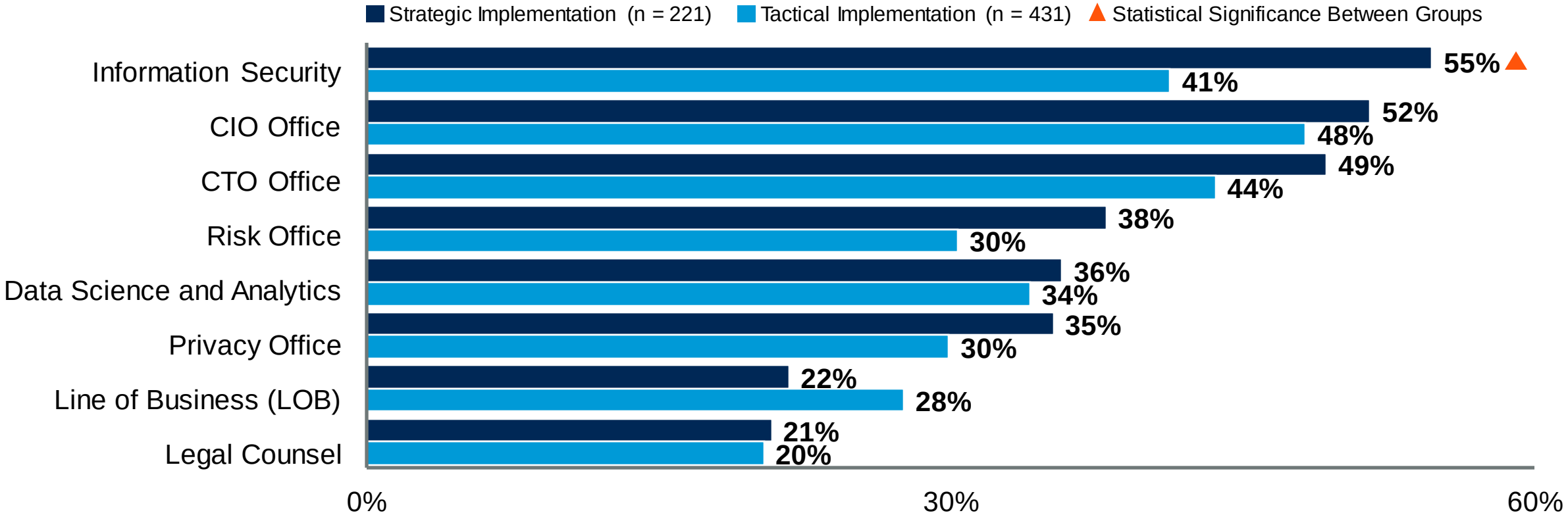
n = 1,079
Q: Do you believe the benefits of generative AI outweigh the risks?
Source: Beyond the Hype: Enterprise Impact of ChatGPT and Generative AI Webinar Polls, 21 April 2023

Gartner®

Organizations Deploying AI Have Prioritized Information Security to Ensure AI Privacy, Security and/or Risk

Responsibility for Implementing Programs to Ensure AI Privacy, Security and/or Risk

Multiple Responses Allowed



Base: Have or Plan to Have a Task Force (Q21), Excludes Unsure

Q22A: Which parts of your organization are part or will be part of the task force responsible for implementing programs that ensure AI privacy, security and/or risk?

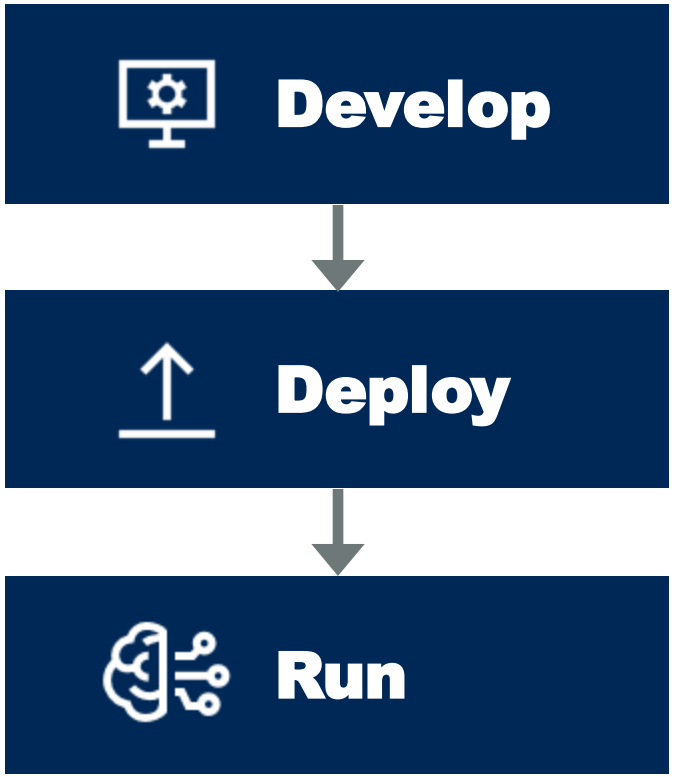
Q07: How widespread is or will be the use of AI in your organization?

Source: 2021 Gartner AI in Organizations Survey

Gartner®

Compromises Span All Stages of AI Operations

AI Life Cycle



Compromises & Attacks

- Data poisoning or compromises (any stage).
- Model outcome manipulation or deterioration at runtime.
- Model or data misuse, compromise or theft.

AI TRiSM is a team sport; involves privacy, security, dev, AI & more.

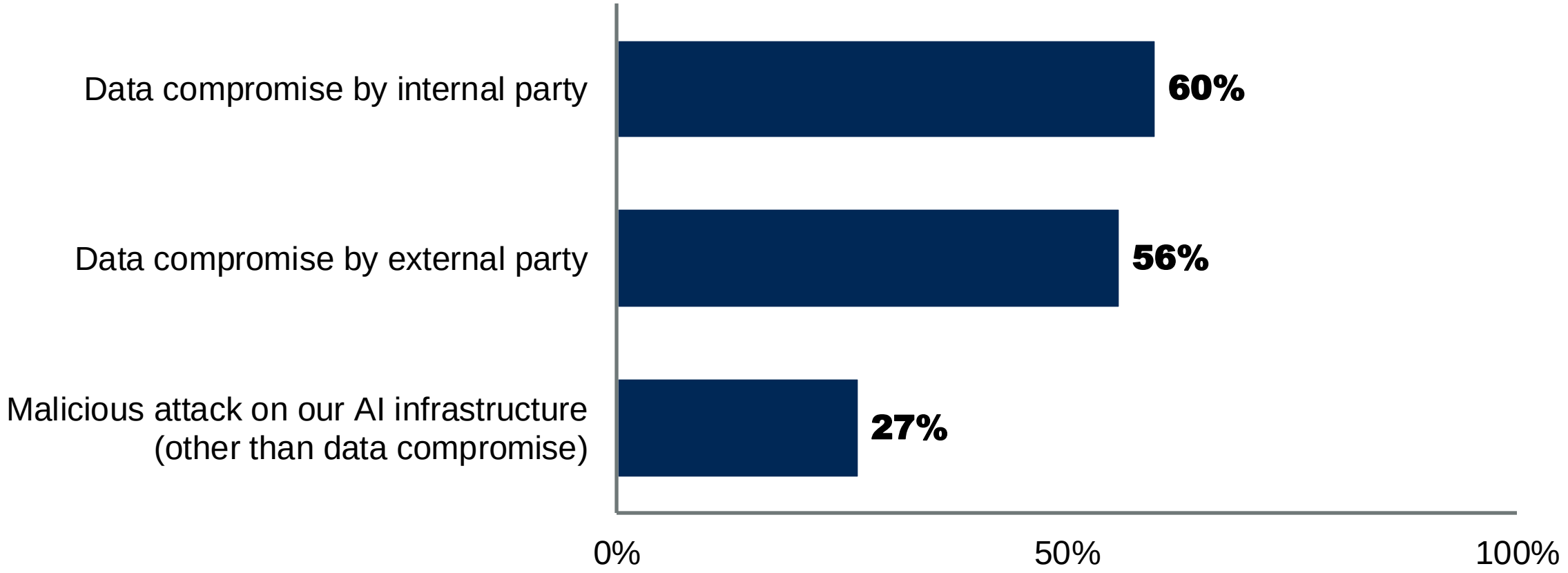
AI TRiSM = AI Trust, Risk and Security Management

Gartner

Many AI Breaches Are Caused by Insiders

Actual Types of Breaches

Multiple Responses Allowed



n = 131; Base: Team of AI Privacy Breach or Security

Q26: What types of AI privacy breaches and/or security incidents were those?

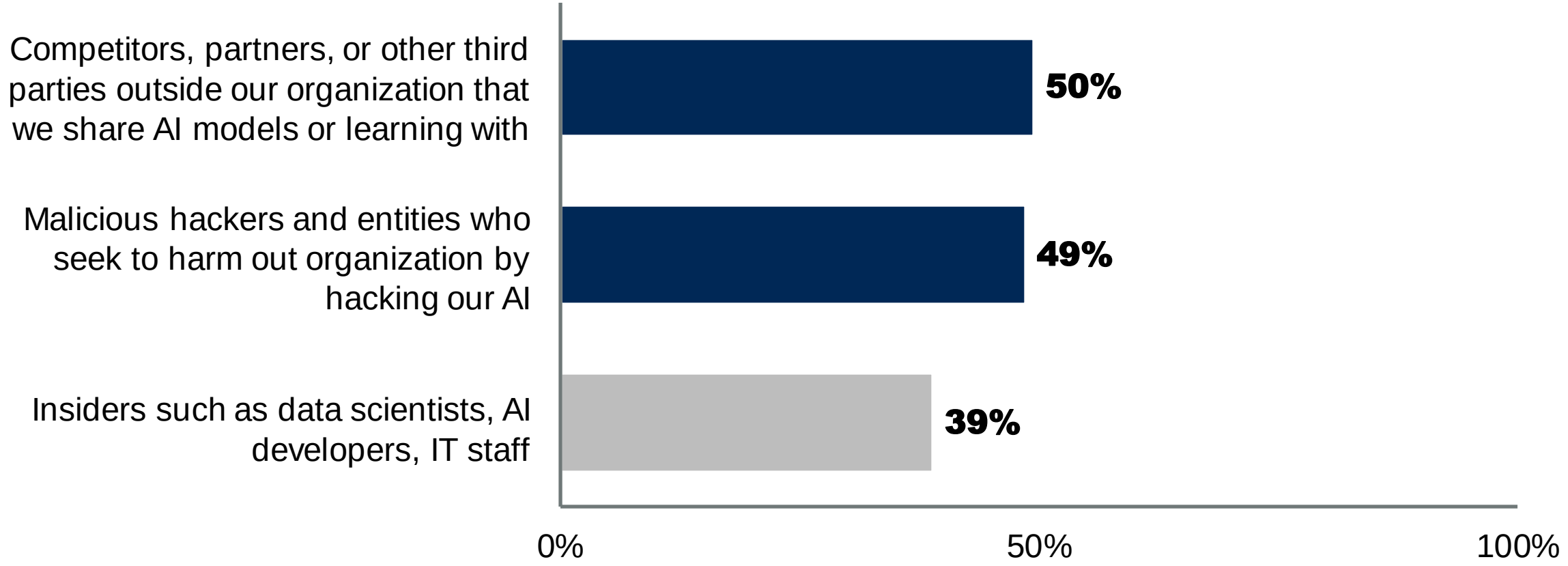
Source: 2021 Gartner AI in Organizations Survey

Gartner

But Organizations Are More Worried About Outsiders

Perceptions of Breaches: Most Worried About Outsiders

Multiple Responses Allowed



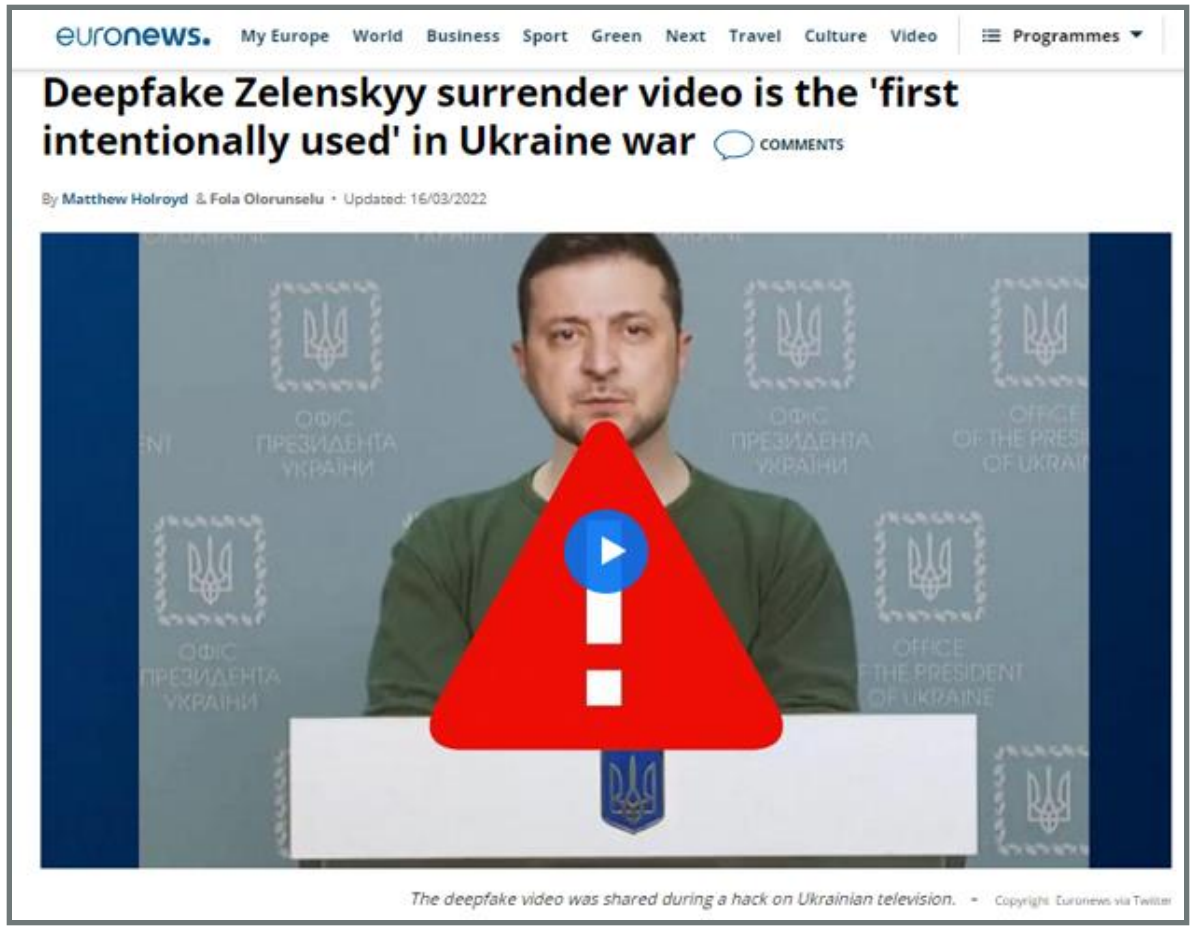
n = 218; Base: AI Privacy, Security and Risk Management (Q18)

Q20: Which parties is your organization most worried about when it comes to AI privacy, security and/or risk?

Source: 2021 Gratner AI in Organizations Survey

Gartner®

Generative AI (GenAI) Attacks Leverage Social Engineering



March 2019: First fraud using a fake AI-generated CEO voice to embezzle €220,000 ($243,000).

Now takes three seconds of voice to create a voice fake.

Source: Fraudsters Used AI to Mimic CEO's Voice in Unusual Cybercrime Case, The Wall Street Journal

Gartner

AI Trust Risk & Security Management

By 2026, global regulations curtailing unsafe GenAI use will not exist despite significant financial losses and incidents that cause physical harm to humans.

Gartner®

In 2024, 15% of successful account takeover attacks will use deepfakes to socially engineer users to turn over sensitive data or move money into criminal accounts.

Source: Gartner

Gartner®

Malicious Inputs to AI Models; Digital and Physical

Video

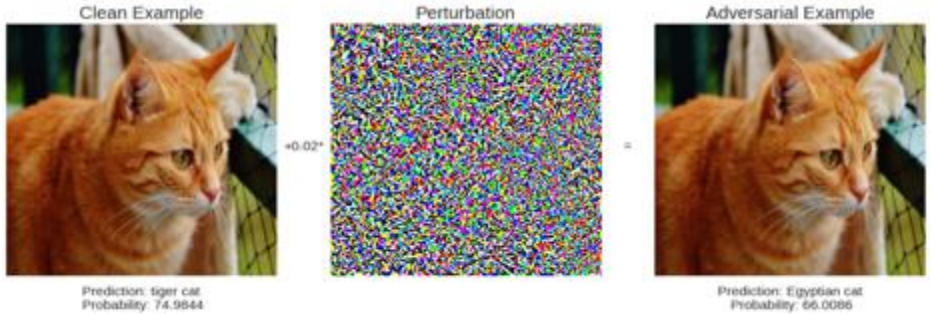
Text

Image

Voice

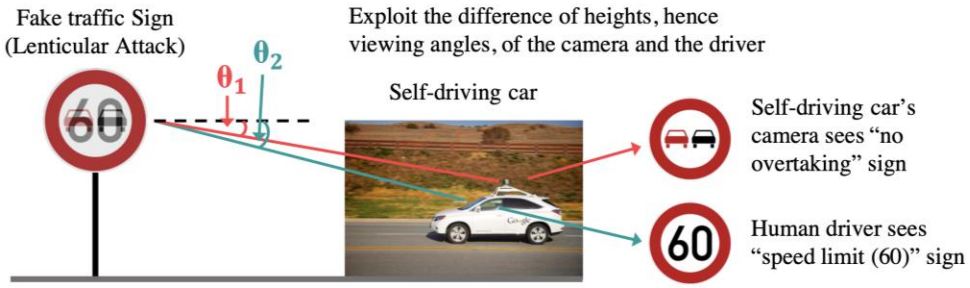
Digital Media Types

Manipulated Digital Images



Criminal adds perturbations to digital image to fool AI model

Manipulated Physical Signs



Sticker changes sign from "speed limit" to "no overtaking"; fools self-driving AI model

Physical Object

Sensors

Camera

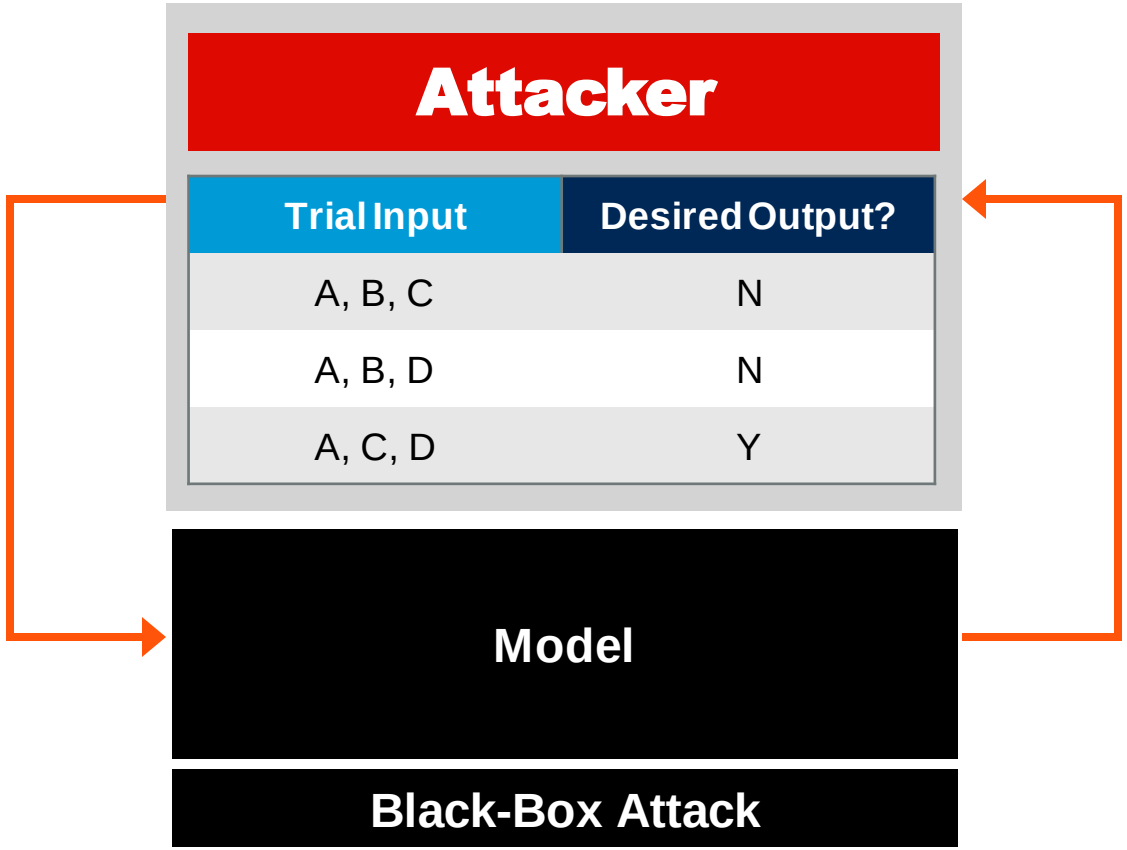
Data Feed

Physical Vector

Source: iProov
Source: DARTS: Deceiving Autonomous Cars With Toxic Signs, Cornell University (arXiv)

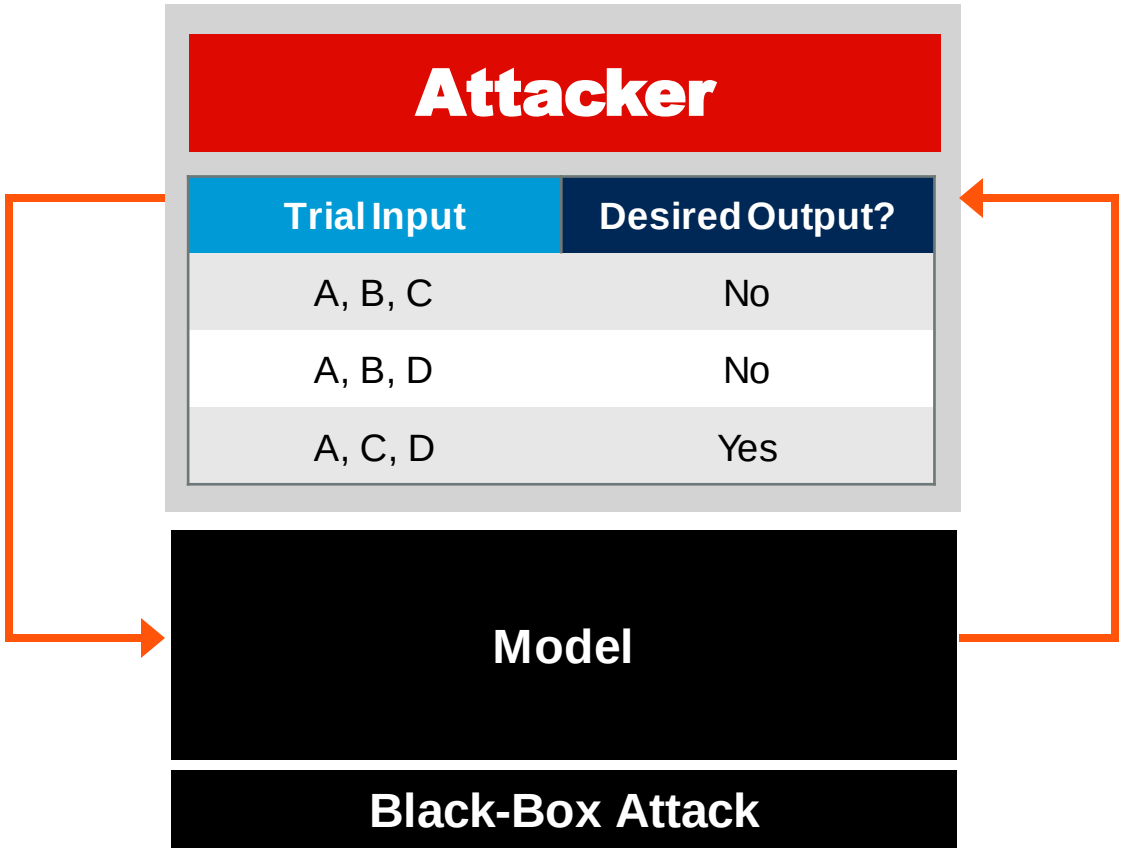
Gartner

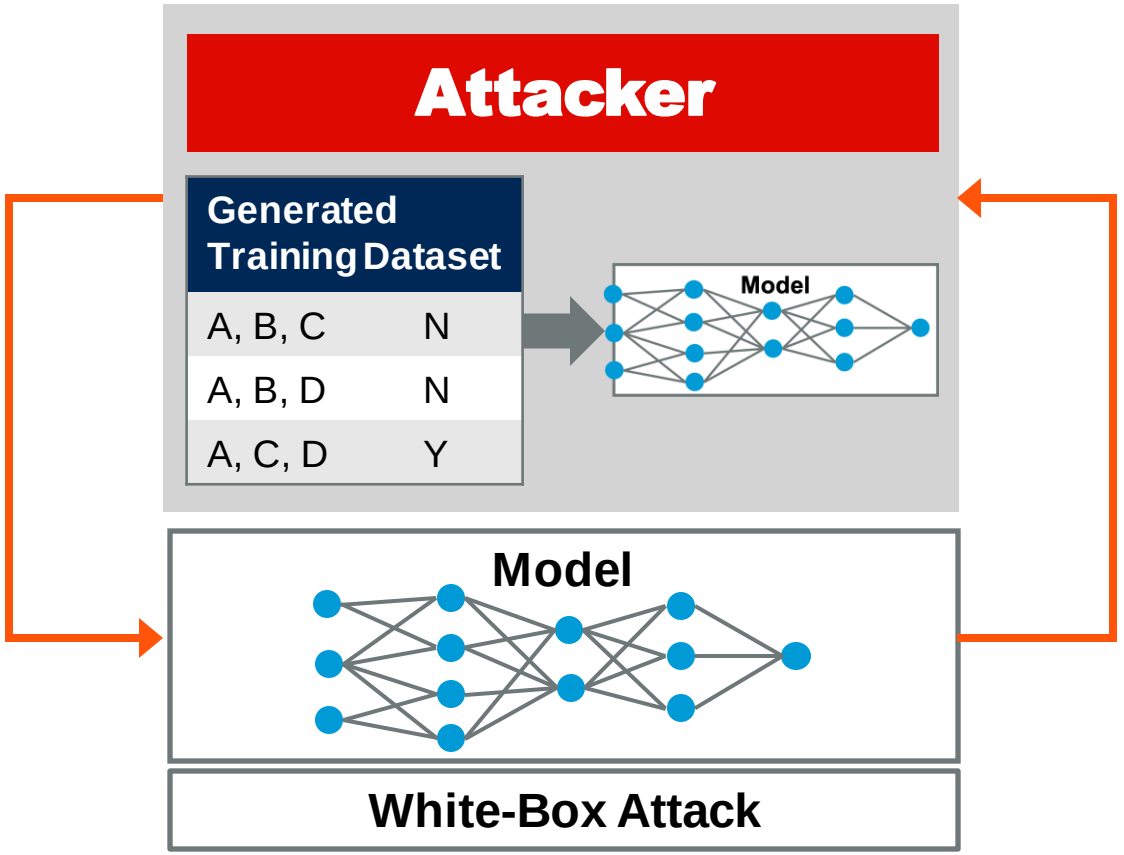
Query Attack Against AI Model: Black Box



Gartner®

Query Attack Against AI Model: White Box





Gartner®

Key Issues

1

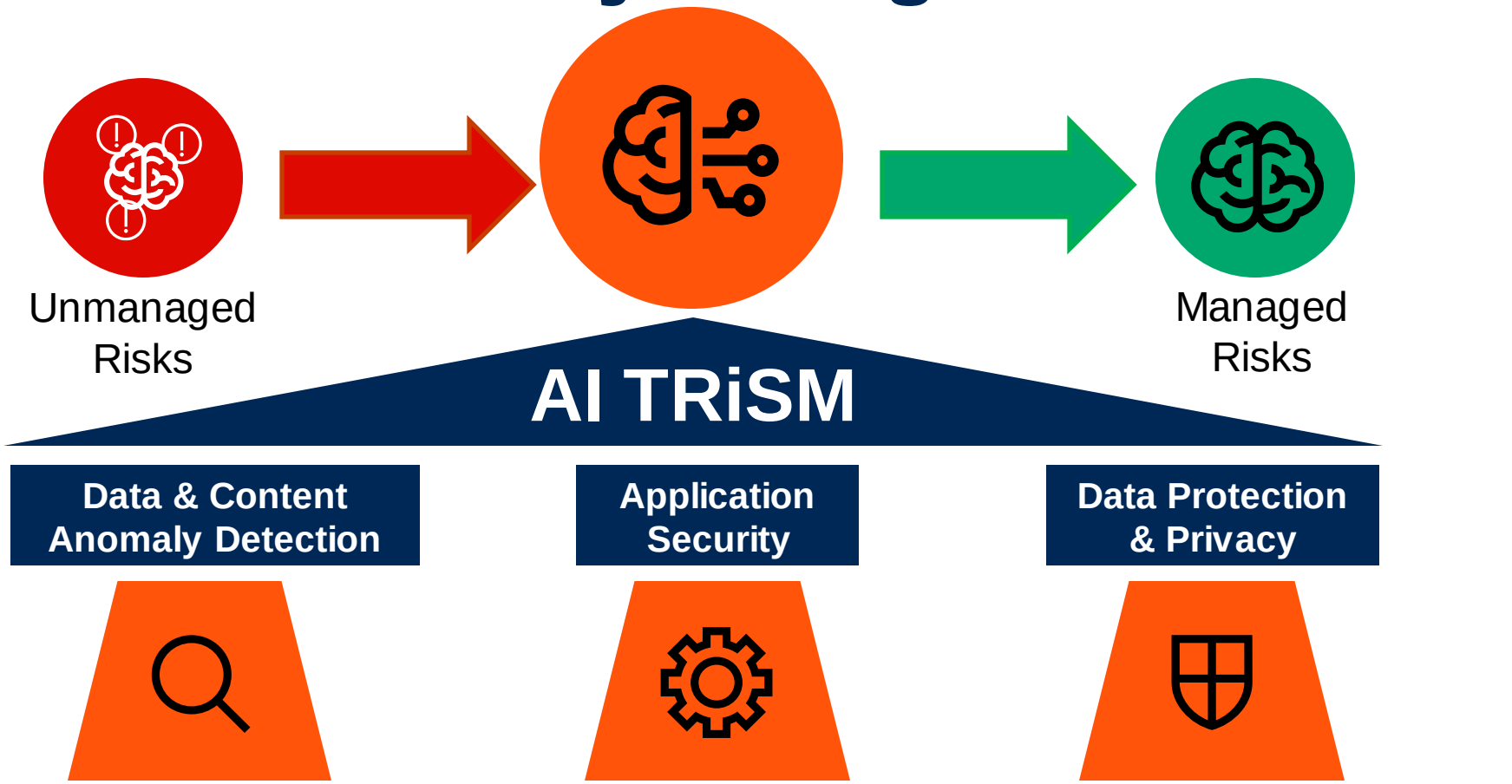
Where, when and how can AI be compromised?

2

What you need to do about new AI risks?

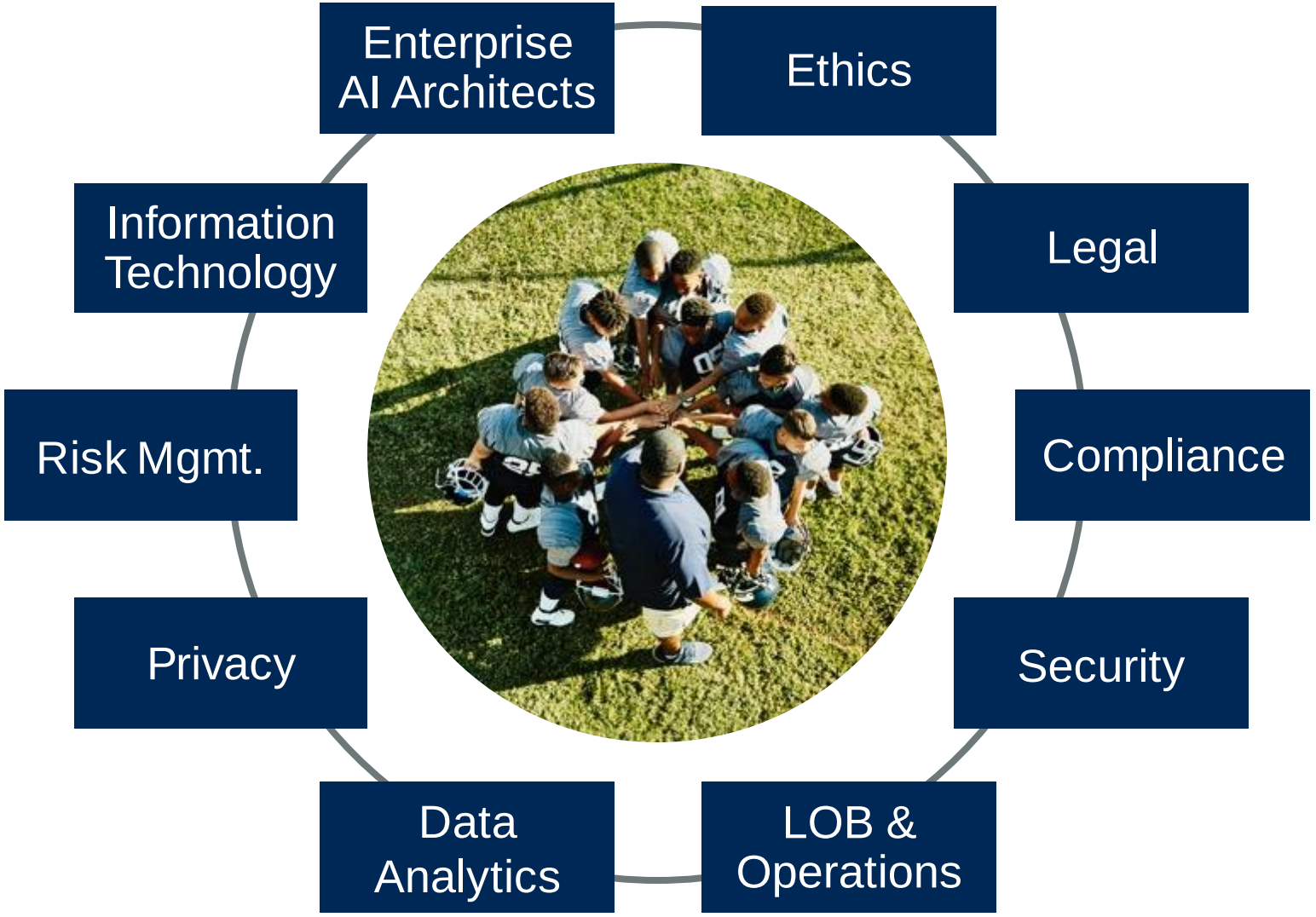
Gartner®

AI Trust Risk & Security Management



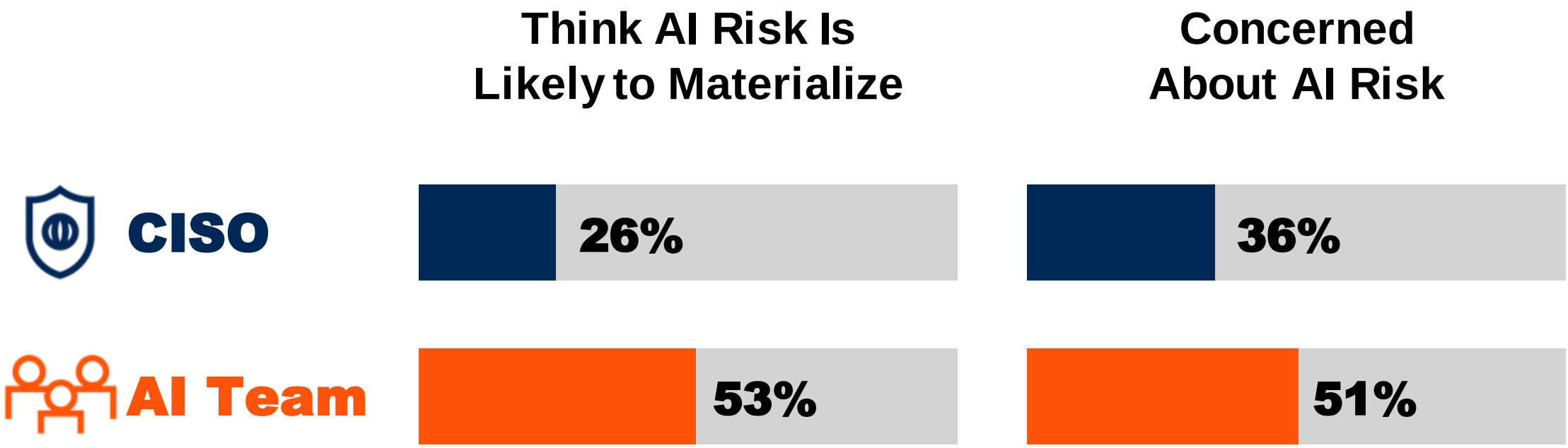
Gartner

First Get Organized: AI TRiSM Is a Team Sport



Gartner®

CISOs Need to Speak With Their AI Teams



n = 64 Base: CISOs in Enterprises at Least Piloting AI Solutions
n = 49 Base: AI Workers
Q: What is the likelihood that information risk stemming from AI solutions will materialize in the next 12-18 months in your enterprise?
Q: How concerned are you about information risk in your enterprise associated with AI solutions?
Source: 2021 Gartner State of AI Cyber Risk Management Study

Once Organized: Set Your Priorities

1

AI Inventory: Explainability & Interpretability

2

AI Risk Awareness

3

Privacy and Data Protection

4

Robust ModelOps

5

AI Security and Resilience

Gartner®

Predicts: AI TRiSM Improves AI Results

By 2026,
organizations that operationalize AI transparency, trust and security will see 150% improved adoption, business goals and user acceptance compared to 2022 results.

Source: Gartner

Gartner®

Example Case Study: The Danish Cancer Society

- Mathematically Explainable Models
- Identify Gene Combinations
- Develop Better Drugs

Explainability

Gartner

Recommendations

- Get on the AI teams — ensure multidisciplinary input.
- Start with purpose — most flows from that starting point.
- Continuously monitor across all stages.
- Embed controls by default, not “bolt on.”

Recommended Gartner Research

- **What Executives Need to Do to Support the Responsible Use of AI**
  Bart Willemsen and Jim Hare
- **Quick Answer: How Can Executive Leaders Manage AI Trust, Risk and Security?**
  Avivah Litan and Bart Willemsen
- **Quick Answer: Privacy Basics for a Digital Twin of a Customer**
  Bart Willemsen
- **AI in Organizations Survey Results: Managing AI Risk Leads to Positive Business Outcomes**
  Avivah Litan, Farhan Choudhary, Jeremy D'Hoinne and Bart Willemsen
- **Market Guide for AI Trust, Risk and Security Management**
  Avivah Litan, Jeremy D'Hoinne, Bart Willemsen and Sumit Agarwal

Gartner®