

cognizant®

Cognizant AI red teaming Hub: turning AI risk into AI confidence

AI adoption is accelerating at an unprecedented pace. Organizations are deploying AI copilots, customer service agents, underwriting assistants, claims processors, coding assistants and enterprise agents across critical business functions. Many of these systems are already in production.

The question is, have they been tested against real-world attacks?

While organizations are investing heavily in model selection, agent development, integration and deployment, they also need to invest in understanding how those systems behave when they are deliberately challenged. History shows that every enterprise technology becomes a target once it creates value, and AI is no exception.

As AI agents and models gain access to enterprise data, business workflows, applications and decision-making processes, vulnerabilities quickly become business risks. Organizations need to ensure those agents and models can be trusted when exposed to malicious inputs, unexpected scenarios or real-world business pressures.

This is why AI red teaming is rapidly becoming one of the most important disciplines in enterprise AI.



The risks begin after deployment

Organizations often assume that if an AI system performs well during testing, it is ready for production. In reality, production introduces entirely different risks:

> A customer service agent that performs accurately during demonstrations may reveal sensitive information when confronted with carefully crafted prompts.

> A claims processing agent may appear reliable until it encounters instructions designed to manipulate outcomes.

> A banking assistant may generate compliant responses most of the time but make high-risk recommendations under specific conditions.

These failures rarely surface during traditional testing. They often emerge only when AI systems are exposed to the unpredictability of real users and real business environments.

The result can be data leakage, policy violations, hallucinations, unauthorized actions, compliance breaches, financial loss and reputational damage.

While AI is not inherently unsafe, many organizations are deploying AI systems that have never been tested under realistic attack conditions.

What is AI red teaming?

AI red teaming is the practice of deliberately challenging AI models and AI agents to expose vulnerabilities before they become production incidents.

The objective is simple: identify weaknesses before they become business risks.

Red teaming simulates adversarial behavior and evaluates whether an AI system behaves safely, securely and reliably under stress. This includes testing for jailbreaks, prompt injections, system prompt leakage, memory poisoning, hallucinations, policy violations, unauthorized actions, reasoning failures and guardrail bypasses.

Beyond merely identifying weaknesses, the goal is to understand their business impact, strengthen controls and improve trust in AI systems.

Red teaming for real business workflows

Effective AI red teaming evaluates systems within the context of the business processes they support.

Rather than asking: "Can I jailbreak the model?" for example, the more important questions are

01 "Can I convince a claims agent to approve a fraudulent claim?"

02 "Can I manipulate a procurement agent into issuing an unauthorized purchase order?"

03 "Can I bypass controls in a customer service workflow?"

04 "Can I influence a lending agent to make a policy-violating recommendation?"

These are the attacks that create real business risk.

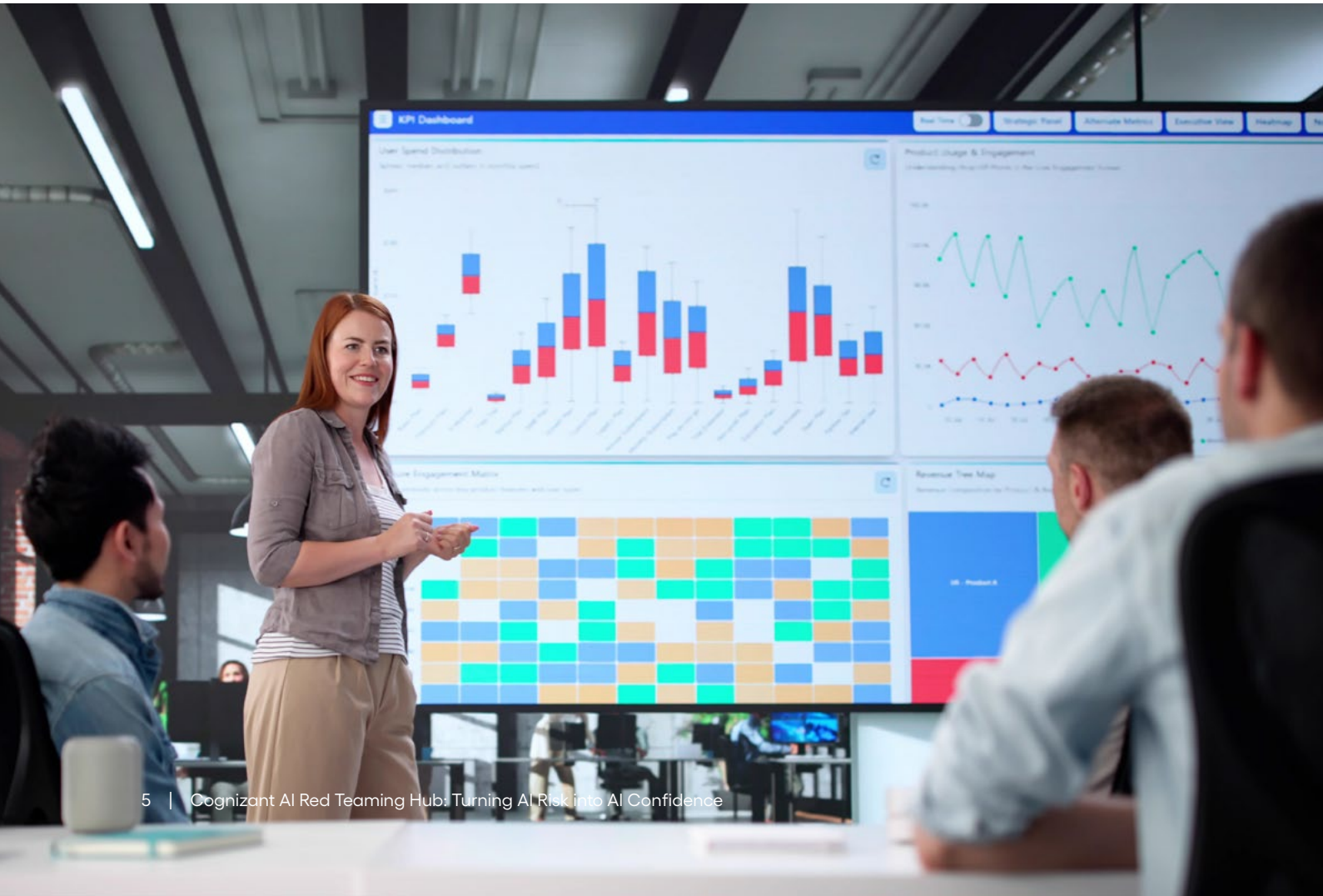
This is where advanced AI red teaming becomes differentiated. Instead of focusing only on technical vulnerabilities, it tests how AI agents behave when interacting with business rules, enterprise policies, approvals, workflows, tools and real-world operational scenarios.

Models are important, but agents are critical

Enterprise risk resides both within the AI model and within the AI agent, itself.

Agents access enterprise data, invoke tools, execute actions, coordinate workflows and make recommendations that influence business decisions. Even when the underlying model is secure, the agent itself may introduce risk through permissions, integrations, workflow design or business logic.

This is why red teaming must extend beyond model testing. Organizations need visibility into how the entire AI system behaves—not just the model, not just the prompt, but the complete operational workflow.



Moving from AI deployment to AI trust: Cognizant AI Red Teaming Hub

The next phase of enterprise AI will not be defined by how many agents organizations deploy. It will be defined by how much those agents can be trusted.

To help organizations build that trust, Cognizant has developed AI Red Teaming Hub, a purpose-built AI assurance Platform as part of its AI Model Training & Data Services portfolio. The platform combines Cognizant's proprietary business-context intelligence, agentic AI assurance capabilities and evaluation frameworks with leading AI security and red-teaming technologies from strategic ecosystem partners. Together, these capabilities enable organizations to assess both AI models and AI agents against a broad spectrum of vulnerabilities, including jailbreaks, prompt injection attacks, system prompt leakage, memory poisoning, hallucinations, reasoning failures, data leakage, policy bypasses, unauthorized actions, and business-process risks.

What differentiates the platform is that it goes beyond generic AI attacks. In addition to testing the core model and agent behavior, it evaluates how agents interact with enterprise tools, data, approvals, workflows and business processes. More importantly, it includes industry-specific attack libraries for banking, insurance, healthcare, telecom, retail and other sectors, enabling organizations to test real-world business risks rather than just technical vulnerabilities.

Instead of simply asking whether a model can be jailbreaked, the platform evaluates whether a claims agent can be manipulated into approving a fraudulent claim, a procurement agent can be tricked into creating an unauthorized purchase order, or a lending agent can be influenced into making a policy-violating recommendation.

The outcome is a comprehensive assessment of AI risk across model behavior, agent reasoning, tool orchestration, workflow execution, governance controls and business outcomes. This enables organizations to strengthen guardrails, validate controls, improve governance and deploy AI with significantly greater confidence.

As AI agents become more autonomous and deeply embedded into enterprise operations, red teaming is rapidly becoming a foundational discipline for responsible AI adoption.

Do you have AI agents in production? Have they been tested against both generic AI attacks and domain-specific business attack scenarios?

If not, now is the time. Reach out to the Cognizant AI Model Training & Data Services team to learn how our AI Red Teaming Platform can help identify vulnerabilities, strengthen trust and reduce AI risk before issues reach production.

Authors



Dr. Venkatesh Upadrista

Global Head of AI Offerings - IOA

Drvenkatesh.upadrista@cognizant.com

Follow

Dr Venkatesh Upadrista leads AI Offerings for Cognizant's Intuitive Operations and Automation (IOA) service line, driving end-to-end growth—from sales and solutioning to delivery and client engagement for AI-led transformations.

With over two decades of leadership experience, he is an active AI researcher and innovator, having authored multiple books and published several research papers in AI and blockchain. He also leads advanced research in Artificial General Intelligence at a leading academic institution in the UK, focusing on next-generation areas such as world models, meta-learning and causal reasoning—advancing the development of human-like intelligence systems for real-world enterprise impact.

cognizant®

Cognizant (Nasdaq: CTSH) is an AI Builder and technology services provider, bridging the gap between AI investment and enterprise value by building full-stack AI solutions for our clients. Our deep industry, process and engineering expertise enables us to build an organization's unique context into technology systems that amplify human potential, drive tangible outcomes and keep global enterprises ahead in a fast-changing world. See how at cognizant.ai or @cognizant.

World Headquarters

300 Frank W. Burr Blvd.
Suite 36, 6th Floor
Teaneck, NJ 07666 USA
Phone: +1 201 801 0233

European Headquarters

280 Bishopsgate
London
EC2M 4AG
England
Tel: +44 (01) 020 7297 7600

India Corporate Office

Siruseri-Software Technology Park
of India (STPI)
SDB Block—Ground Floor North Wing
Plot No H4, SIPCOT IT Park
Chengalpattu District
Chennai 603103, Tamil Nadu
Tel: 1800 208 6999

APAC Headquarters

1 Fusionopolis Link, Level 5
NEXUS@One-North, North Tower
Singapore 138542 Phone: +65
6812 4000

© Copyright 2025–2027, Cognizant. All rights reserved. No part of this document may be reproduced, stored in a retrieval system, transmitted in any form or by any means, electronic,mechanical, photocopying, recording or otherwise, without the express written permission of Cognizant. The information contained herein is subject to change without notice. All other trademarks mentioned herein are the property of their respective owners.

WF 5357200