Building Trust in AI:

Cognitive Resilience for Modern Companies

DARKSTACK7

Joshua R Nicholson

Managing Partner & CISO at Darkstack7

Joshua Nicholson is a cybersecurity executive with over 20 years of experience leading enterprise security, risk management, and incident response across financial services, consulting, and defense industries. He has held senior leadership roles with Surefire Cyber, DeepSeas, Booz Allen Hamilton, Cofense, Wells Fargo, Ernst & Young, and Hancock & Whitney Bank.

U.S. Marine Corps veteran and holds multiple industry certifications including **CISSP, CISM, SANS-GCIH, CCNP, and GCWN**. He is also the host of the **Cyber Security America** podcast, where he brings global thought leaders together to discuss the evolving intersections of AI, cybersecurity, and resilience.

Artificial Intelligence is no match for natural stupidity.

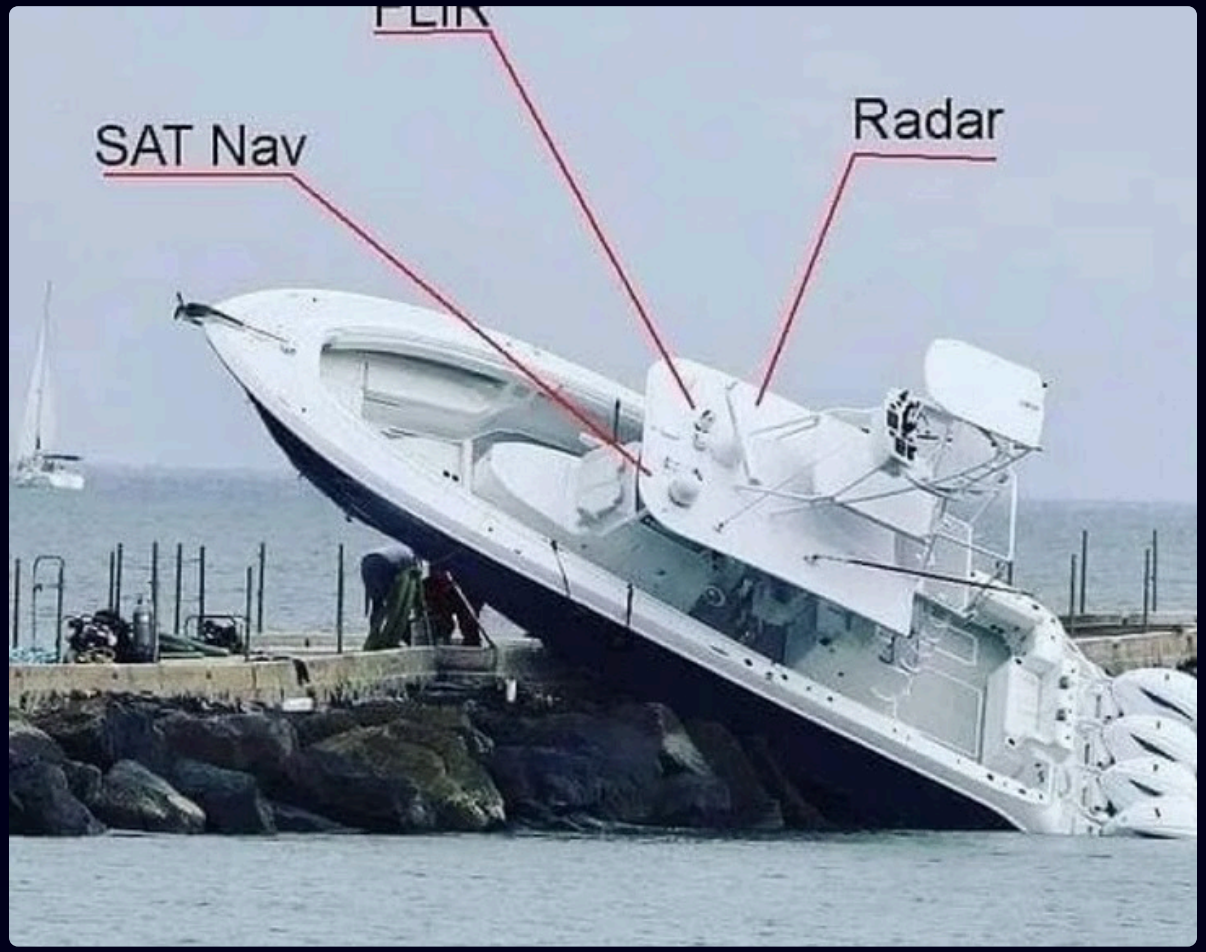


Table of Content

Real-World AI Applications

Cybersecurity Defense

Threat detection, anomaly identification, and automated incident response systems that enhance security posture and reduce response times

Regulatory Compliance

Automated compliance monitoring, policy enforcement, and audit trail generation for regulatory frameworks and industry standards

Operations Optimization

Process automation, predictive maintenance, resource allocation, and workflow optimization to drive operational efficiency

Building the AI-Powered Enterprise

SaaS Foundation

Scalable cloud infrastructure supporting AI workloads and business applications

MDR Integration

Managed detection and response capabilities enhanced by AI-powered threat intelligence

AI Cognition

Intelligent automation that learns, adapts, and optimizes business processes continuously

vCISO Oversight

Virtual security leadership providing strategic guidance and governance for AI initiatives

The convergence of these technologies creates unprecedented opportunities for business acceleration while maintaining enterprise security and compliance standards.

AI Technology Landscape

The AI landscape is rapidly evolving, driving innovation across industries. Understanding its core components is crucial for effective adoption and strategic implementation within enterprises.

Machine Learning & Deep Learning

Foundational algorithms enabling systems to learn from data and make predictions.

Natural Language Processing (NLP)

Allows AI to understand, interpret, and generate human language for communication and analysis.

Computer Vision

Empowers machines to "see" and interpret visual information, from images to videos.

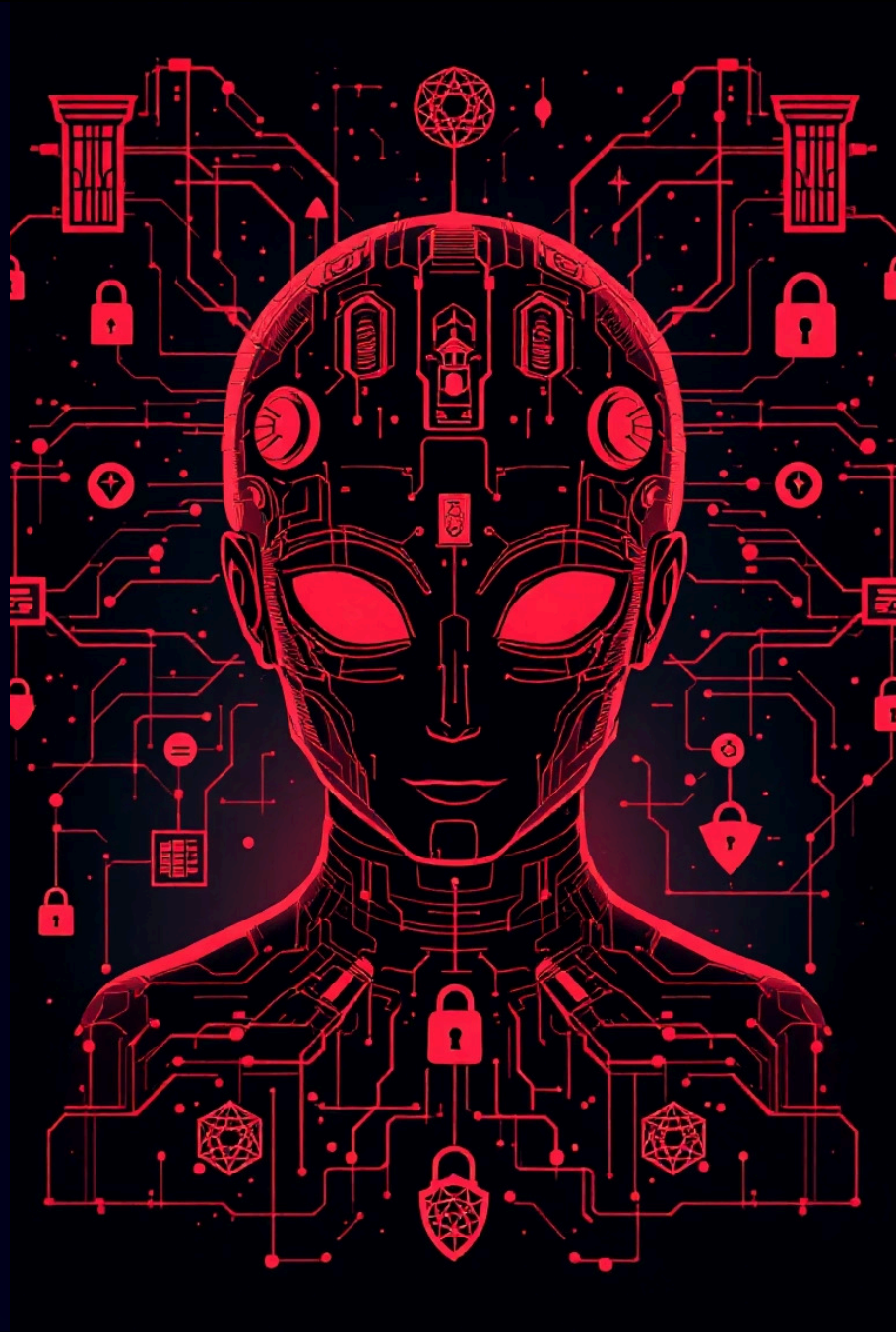
Generative AI

AI models capable of creating new content, from text and images to code and designs.

AI Risk Management

DARKSTACK7

`

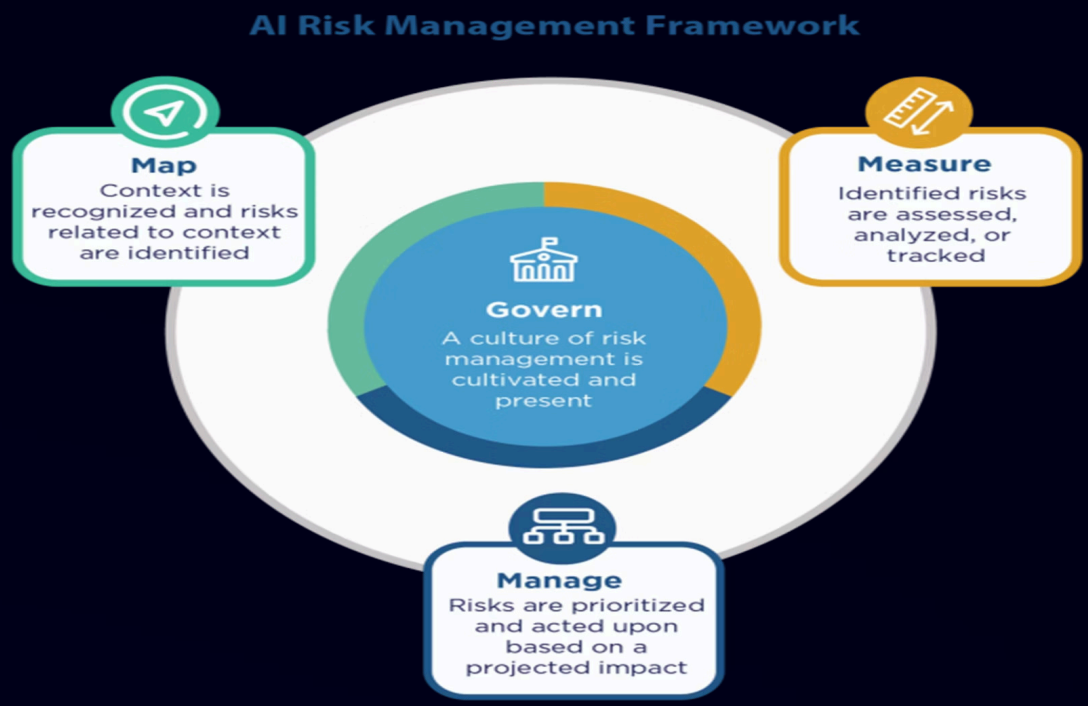


Governance Frameworks Drive Success

The NIST AI Risk Management Framework provides structured approach to identifying, assessing, and mitigating AI risks across the enterprise lifecycle.

Duty of Care Risk Analysis (DoCRA) standard establishes legal and ethical responsibilities for AI system deployment and ongoing management.

- Risk identification and assessment protocols
- Continuous monitoring and validation requirements
- Stakeholder accountability frameworks



NIST AI Risk Management

1

Map function establishes the context to frame risks related to an AI system. The AI lifecycle consists of many interdependent activities involving a diverse set of actors (See In practice, AI actors in charge of one part of the process often do not have full visibility or control over other parts

2

Govern is a cross-cutting function that is infused throughout AI risk management and enables the other functions of the process. Aspects of **govern**, especially those related to compliance or evaluation, should be integrated into each of the other functions.

3

Measure function employs quantitative, qualitative, or mixed-method tools, techniques, and methodologies to analyze, assess, benchmark, and monitor AI risk and related impacts.

It uses knowledge relevant to AI risks identified in the **map** function and informs the **manage** function.

4

Manage function entails allocating risk resources to mapped and measured risks on a regular basis and as defined by the **govern** function. Risk treatment comprises plans to respond to, recover from, and communicate about incidents or events.

What is Duty of Care Risk Analysis

Implementing AI responsibly requires adherence to core principles that balance innovation with human impact and risk mitigation.

Prioritizing Impact

Paying as much attention to those you can harm as to your own risk, ensuring that societal and individual well-being are paramount in AI development and deployment.

Achieving Acceptable Risk

Targeting an acceptable risk threshold where no party would need repair, meaning systems are designed to prevent harm rather than just react to it.

Proportionate Safeguards

Using safeguards that are no more burdensome than the risks they reduce, fostering efficient and effective security measures without impeding progress.

Deep Dive into DoCRA Principles

Understanding the core components and calculations of Duty of Care Risk Analysis is essential for responsible AI governance.

Calculating AI Risk

In DoCRA, risk is defined as the potential harm to **all parties**. This is quantified by evaluating the cumulative **Impact** (on mission, objectives, and obligations) multiplied by the **Likelihood** of that impact occurring.

`Risk = Impact (Mission + Objectives + Obligations) x Likelihood`

Understanding Impact Components

Impact considers potential adverse effects on an organization's mission, its strategic objectives, and its regulatory or ethical obligations. This holistic view ensures all facets of potential harm are considered, moving beyond just financial loss.

Proportionality of Safeguards

A fundamental principle of DoCRA is that the burden of implementing a mitigating safeguard must be less than the risk it reduces. This ensures cost-effective and justifiable security measures without over-engineering.

`Burden of Safeguard < Reduced Risk`

AI Fundamentals for Executives

DARKSTACK7

`

Common Prompting Pitfalls

Ambiguous Instructions

Vague or unclear directives lead to unpredictable outputs and inconsistent results across different AI models

Information Overload

Excessive context or too many simultaneous requests can overwhelm AI systems and degrade performance quality

Missing Context

Insufficient background information prevents AI from understanding nuanced requirements and business constraints

The 5-Step Prompt Engineering Framework

01

Task Definition

Clearly articulate the specific objective, desired outcome, and success criteria for the AI system

02

Context Setting

Provide relevant background information, constraints, and environmental parameters that guide execution

03

Reference Integration

Include authoritative sources, examples, and standards that inform decision-making and output quality

04

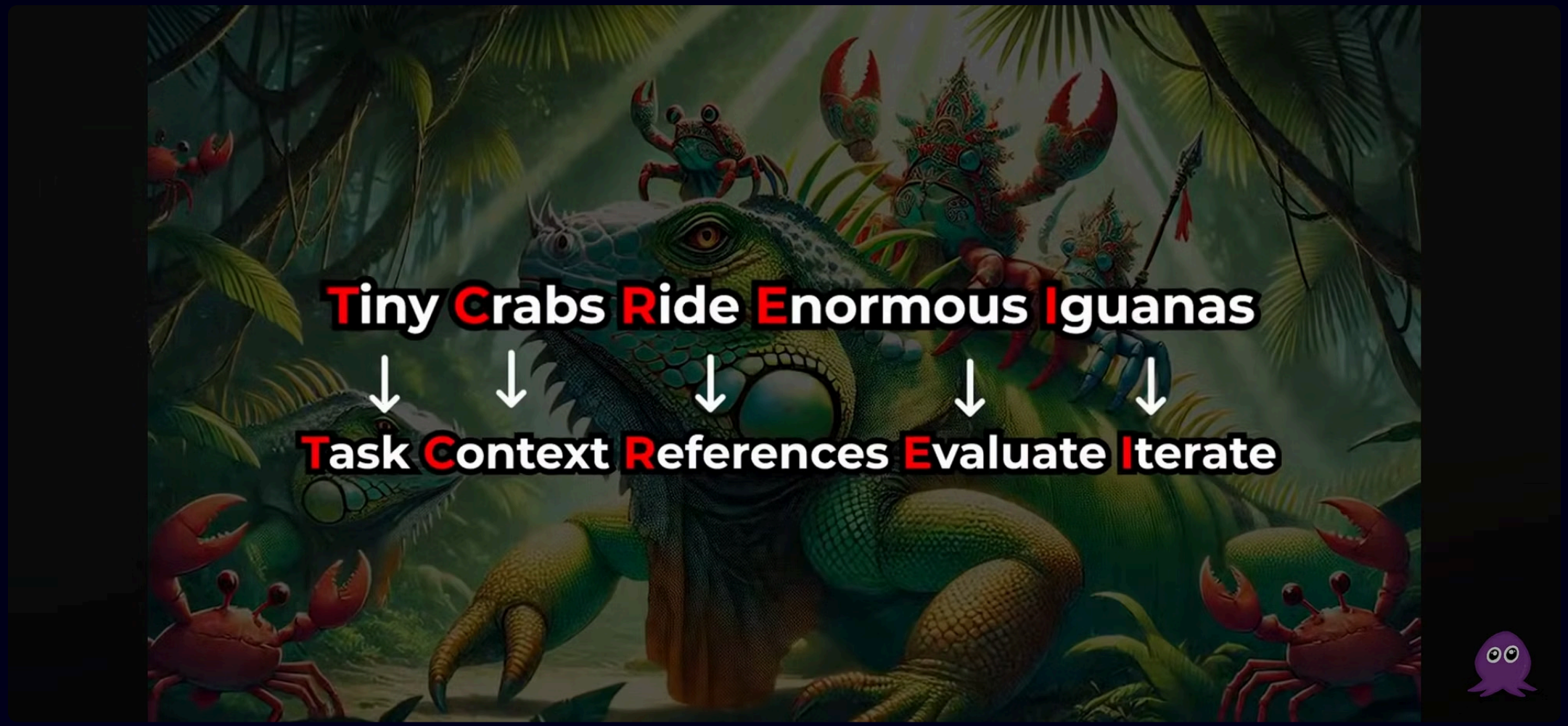
Evaluation Criteria

Establish measurable benchmarks and validation methods to assess performance and accuracy

05

Iterative Refinement

Continuously improve prompts based on results, feedback, and changing business requirements


Tiny Crabs Ride Enormous Iguanas
Task Context References Evaluate Iterate

Advanced Prompting Techniques

Prompt Chaining

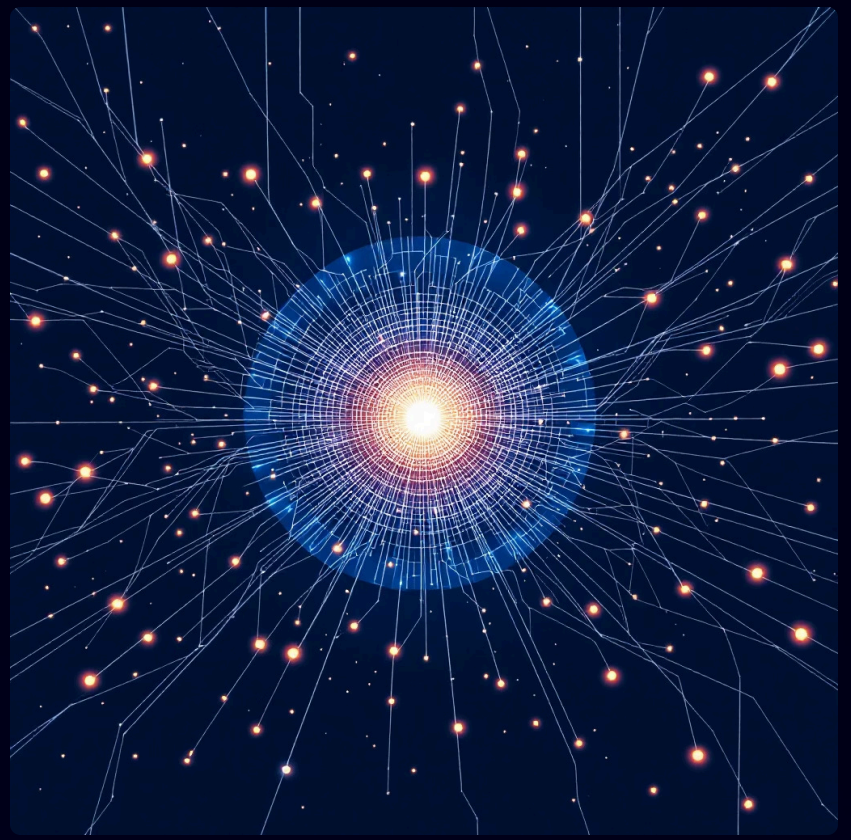
Sequential prompts where each output becomes input for the next step

Chain-of-Thought

Step-by-step reasoning that shows the AI's decision-making process

Tree-of-Thought

Multiple reasoning paths explored simultaneously for complex problem-solving



Agent frameworks enable sophisticated workflows that combine multiple AI capabilities for enterprise-grade solutions.

Common Logical Fallacies in Cyber + AI Contexts

AI systems often produce confident but false outputs that can mislead investigators or decision-makers. Understanding logical fallacies helps detect flawed reasoning in both AI and human reports.

Ad Hominem

Attacking the messenger instead of the message

Example: Dismissing a security alert because it came from a junior analyst.

Strawman

Misrepresenting a position to make it easier to attack

Example: "AI is just perfect, we don't need human oversight."

False Cause (Post Hoc)

Assuming correlation equals causation

Example: Assuming that new software update for a breach just because it happened the same week.

Appeal to Authority

Accepting claims without question because they come from an authority figure

Example: Trusting an AI output without verifying the data source.

Slippery Slope

Arguing one small step will inevitably lead to extreme consequences

Example: "If we use AI for threat detection, it will eventually replace human investigators entirely."

Confirmation Bias

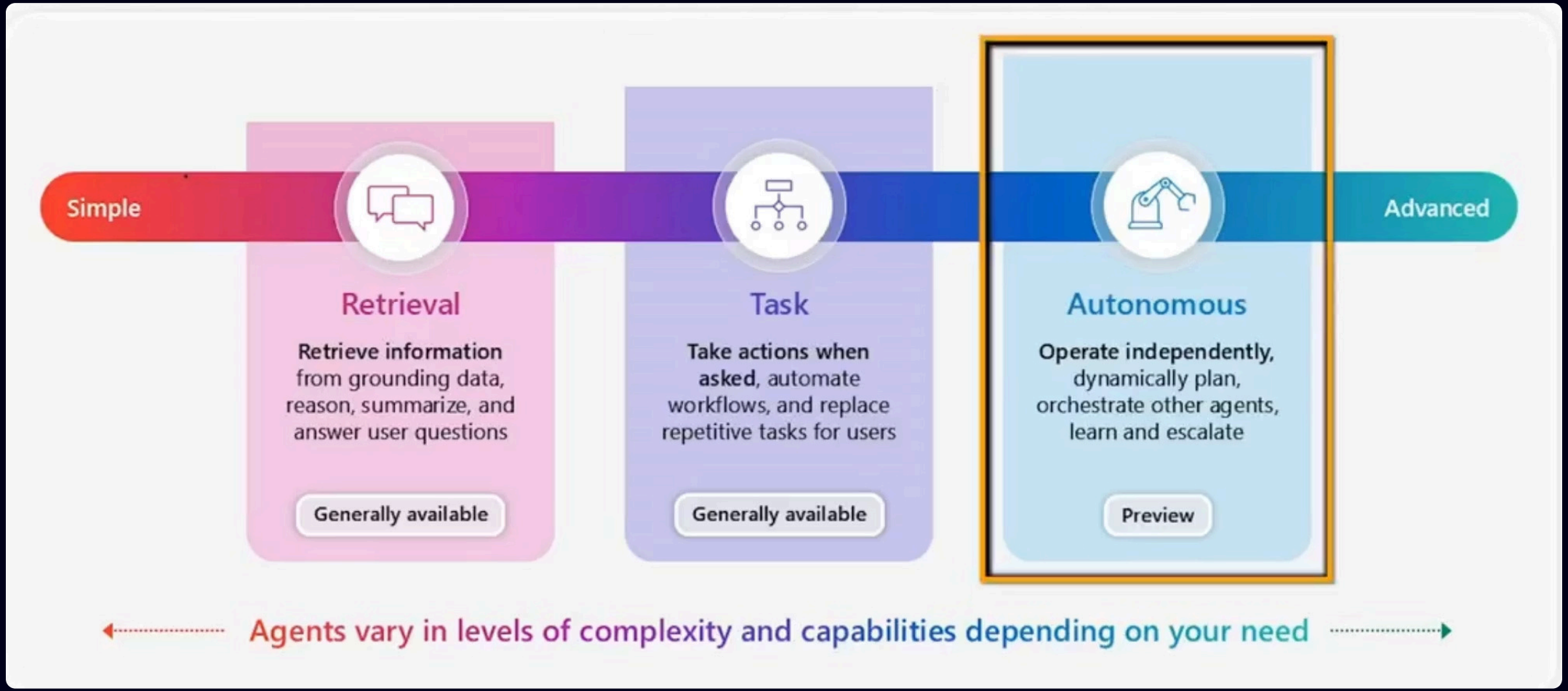
Only accepting evidence that supports what you already believe

Example: Focusing only on indicators of ransomware when it might actually be insider fraud.

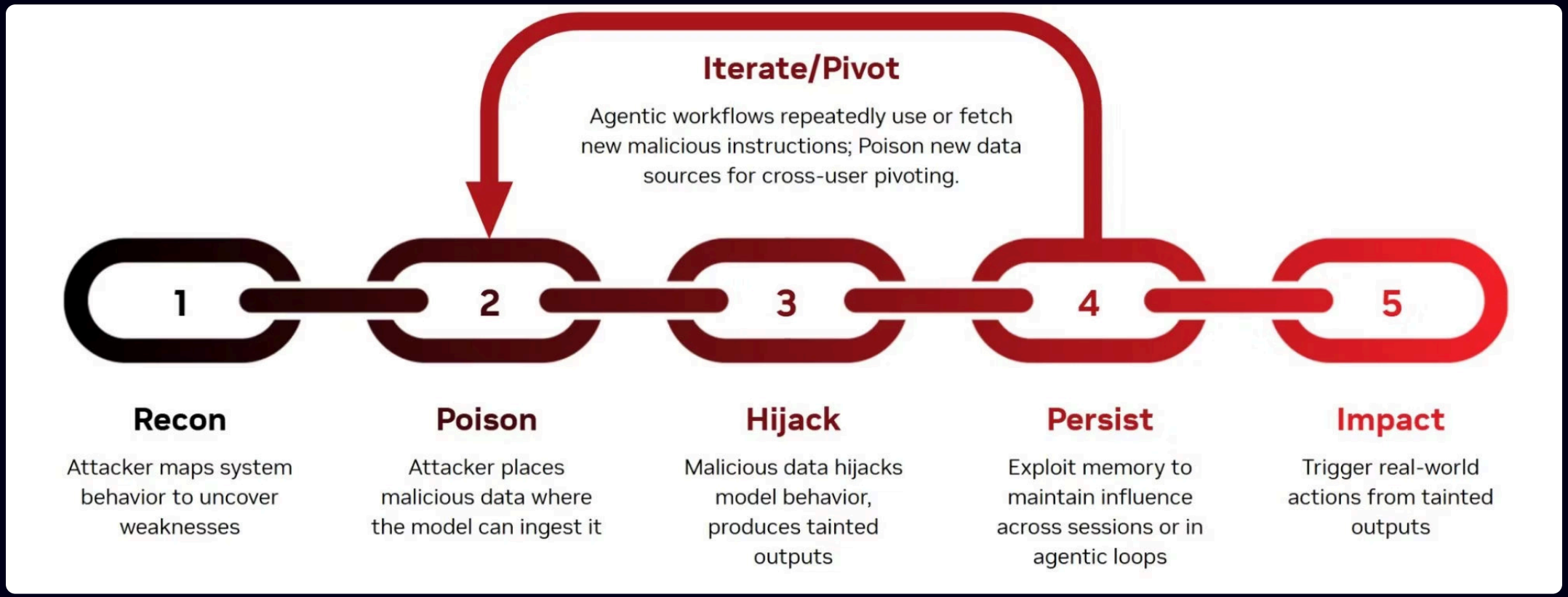
Attacking / Hardening

Secure Agentic Deployment

Agent Complexity & Capabilities



AI Attack Kill-chain



Anthropic

Hardening Guidelines for Secure Agentic Deployment

Access Control

- Use least privilege principles. Agents should only have permissions needed for their tasks.
- Employ role-based access control (RBAC) and separate environments for testing.

Data Security

- Enable data loss prevention (DLP) policies, encryption at rest and in transit.
- Avoid exposing sensitive documents to AI agents unnecessarily.

Audit & Logging

- Maintain comprehensive logs for user approvals and changes made.
- Use SIEM integration to monitor anomalies.

Human-in-the-Loop

- Implement approvals for high-risk actions.
- Avoid fully autonomous execution on sensitive workflows without oversight.

Input/Output Validation

- Sanitize inputs to search and validate outputs before execution.
- Maintain documentation to satisfy regulatory requirements.

Network & Endpoint Security

- Restrict agent communication to approved APIs/services.
- Use VPNs, firewalls, and network segmentation when feasible.

Patch & Update Management

- Keep AI platforms and dependencies updated.
- Apply vendor-provided security patches promptly.

Model Security

- Protect models against model inversion attacks.
- Avoid exposing local secrets or API keys in prompts.

Prompt Security

- Prevent malicious prompt injection attacks.

Operational Policies & Compliance

- Define acceptable use policies, retention policies for AI-generated data, and incident response procedures involving AI systems.
- Mask or anonymize sensitive data when possible.

Security Hardening Guidelines

Access Control

Role-based permissions and zero-trust architecture for AI system access

Data Protection

Encryption, anonymization, and secure data handling throughout AI pipelines

Audit & Logging

Comprehensive monitoring and traceability of all AI system interactions

Human-in-the-Loop

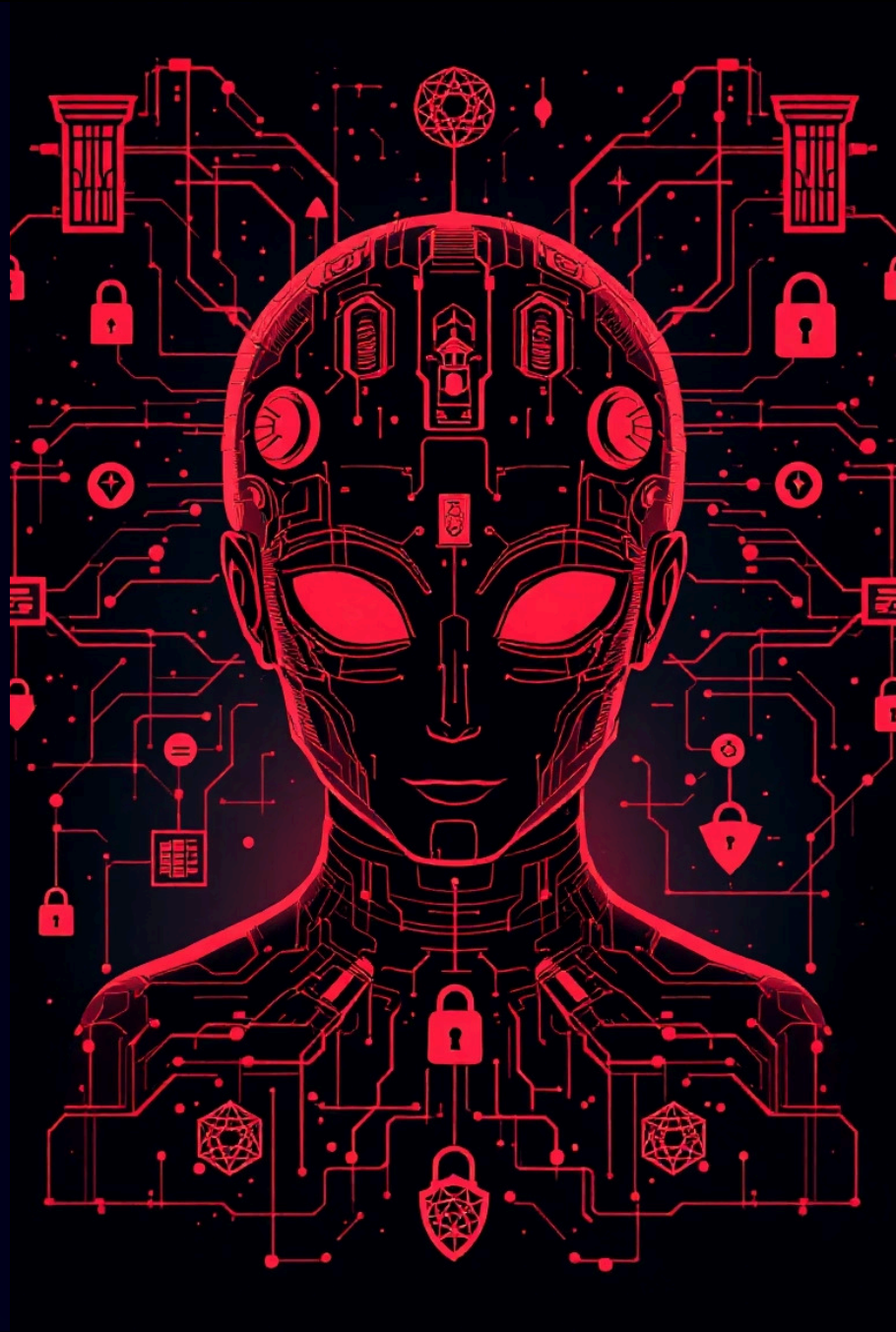
Critical decision checkpoints requiring human validation and approval

⚠ Remember: Prompt injection attacks can compromise AI systems. Implement input validation and output filtering.

Cutting Edge Tools

DARKSTACK7

`



Leading AI Platforms & Tools

Enterprise AI Platforms

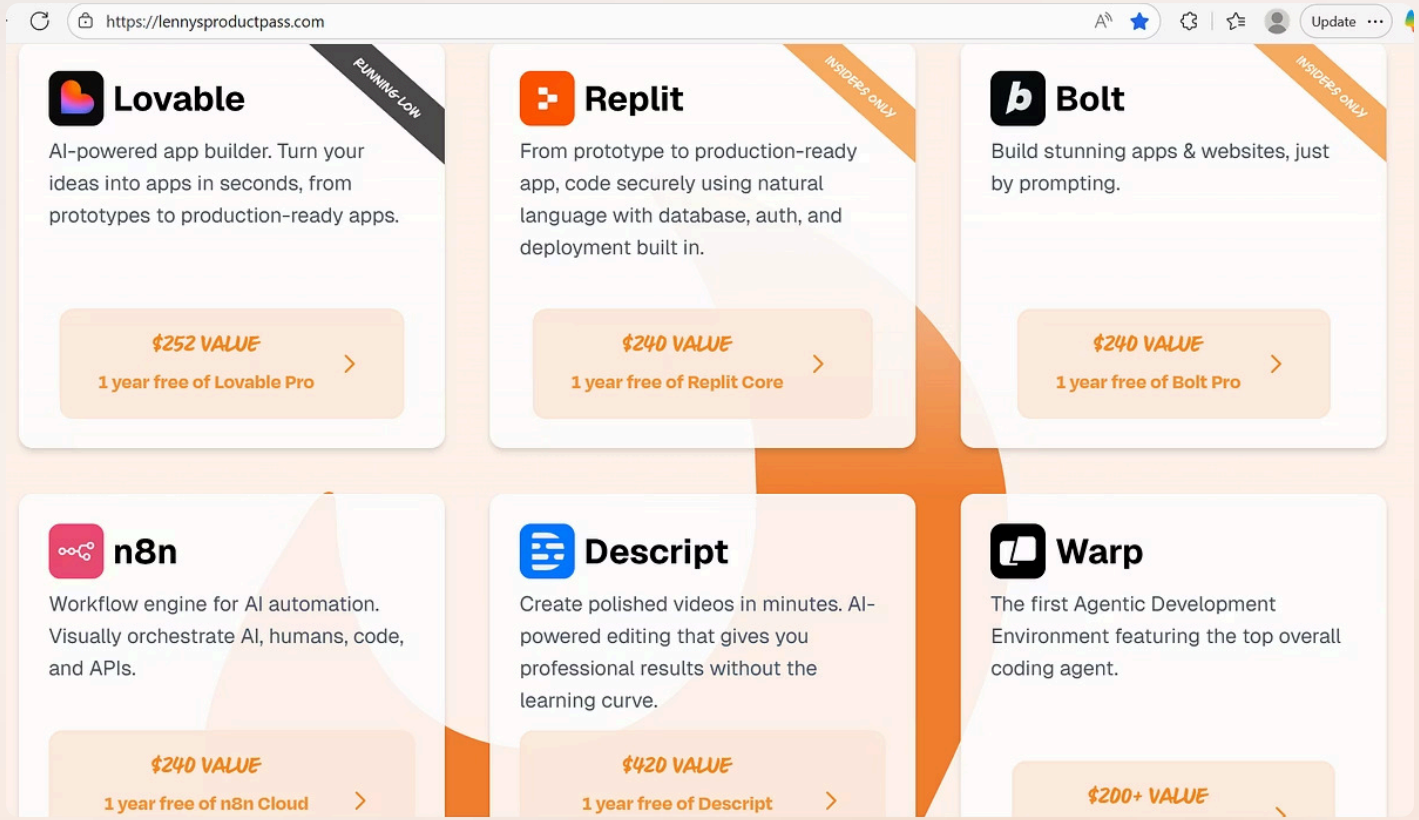
- Microsoft Copilot for business automation
- OpenAI GPT for content and code generation
- Anthropic Claude for analytical reasoning
- Google Gemini for multimodal processing

SaaS Orchestration

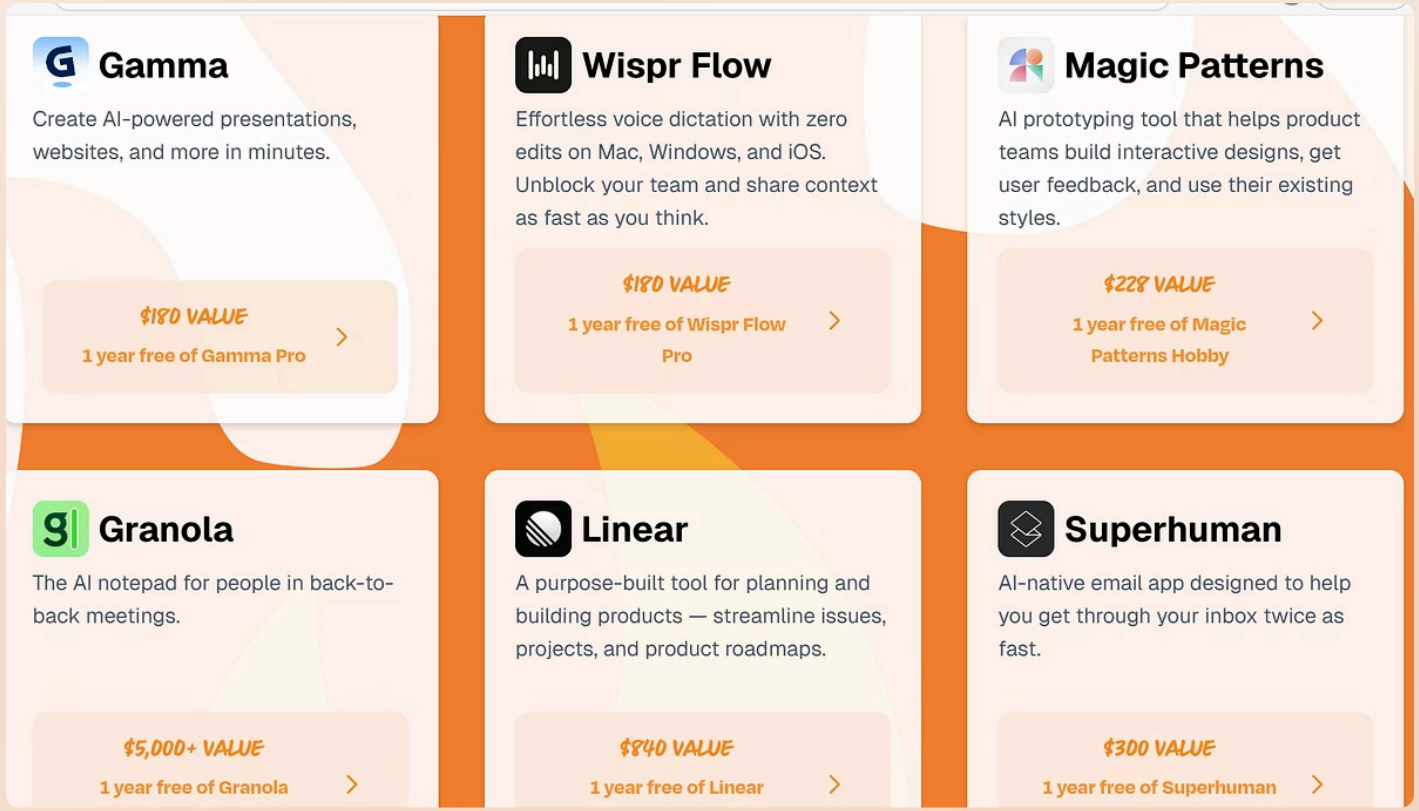
- Zapier and Make for workflow automation
- LangChain for AI application development
- LlamaIndex for knowledge management
- Custom APIs for enterprise integration

Cutting Edge AI Tools

lovable



Cutting Edge AI Tools



Elevate Your Cybersecurity with Darkstack7

Ready to strengthen your defenses and protect your digital assets? Darkstack7 offers comprehensive cybersecurity solutions tailored to your needs. Discover how our expert services can secure your future.

Our Expertise Includes:

- Cybersecurity Consulting
- Virtual CISO Services
- Incident Response
- Tabletop Exercises (TTX)
- Insider Threat Investigations
- Penetration Testing
- Managed Detection and Response (MDR)
- AI Security Assessments
- Compliance Consulting
- Security Training

Visit our website to learn more about our services and schedule a consultation today.

Visit Darkstack7.com

Connect with Us:

Follow Joshua R Nicholson on LinkedIn for ongoing cybersecurity insights and updates. Scan the QR code below:

Cyber Security America Podcast:

Subscribe to the Cyber Security America podcast on YouTube for ongoing cybersecurity insights and expert discussions.

Subscribe on YouTube