TLP:GREEN
Contains AI

Trust, not just Defenses

Rethinking Data Security as a Catalyst for AI Adoption

Candid Wüest

June 2025

Would you trust any AI Agent?

Book me a holiday with my credit card

Do you truly TRUST AI to help you?

AI is not a Future Concept - it's happening Now

Why is AI adoption stalling?

- ❖ Excitement is high
- ❖ Technology is ready
- ❖ There are use-cases
- ❖ But trust is missing

Market Reality

- ❖ Lack of visibility / poor data quality
- ❖ Fragmented / disconnected data
- ❖ Fear and uncertainty / regulations
- ❖ Unprotected data pipelines

Common Pain Points

- ❖ Frustrating user experience
- ❖ Shadow AI usage
- ❖ Privacy risks / Data leaks
- ❖ Risk of hallucinations
- ❖ Lack of transparency

Security ≠ Trust ≠ Compliance

Clarity

Classify

Control

Govern

Transparency

User-centric

Do you know what's happening with your data?

OpenAI slams court order to save all ChatGPT logs, including deleted chats

OpenAI defends privacy of hundreds of millions of ChatGPT users.

ASHLEY BELANGER – JUN 4, 2025 9:56 PM | 174

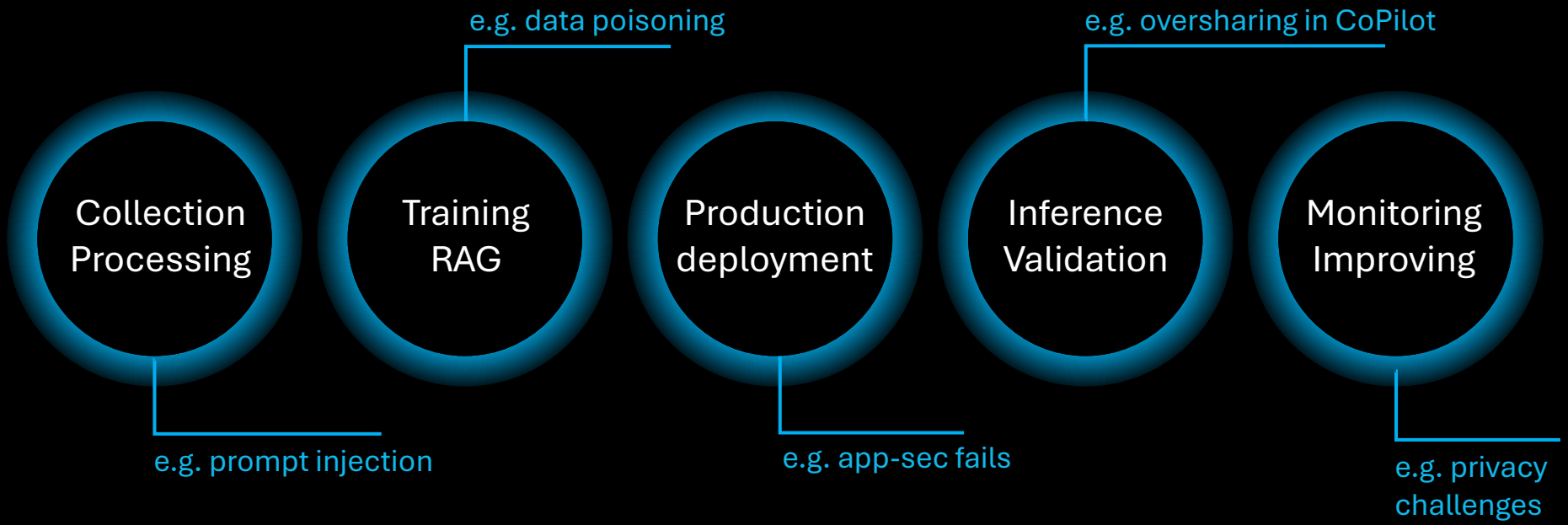
ars TECHNICA

Meta to use Facebook and Instagram content for AI training: How to opt-out

WORLD / 23 APRIL 2025

NEWS FROM FINLAND HELSINKI TIMES

Securing Data Throughout the AI Lifecycle



Holistic Data Governance Strategy

Well-governed, trusted data enables:

- ❖ Stakeholder confidence and trust in AI
- ❖ Faster AI development
- ❖ Safer & more reliable AI integration

Apply zero-trust principles to AI pipelines:

- ❖ Continuous authenticated data access
- ❖ Monitor model inputs/outputs
- ❖ Isolate training environments

Cross the bridge to growth

Data security as an enabling force for AI

Thank you
for your attention!

Get the slides

Candid Wüest

My LinkedIn

Candid Wüest

- 25+ years in cyber security
- Currently Security Advocate for xorlab AG (email security)
- Ex-Vice President of Global Research @ Acronis
- 17 years @ Symantec's Global Security Response Team
- ETH Zürich, many useless certificates, 15+ patents, author and advisor
- Organizer of AREA41 conference, BSidesZH, Defcon Switzerland,...
- Likes "breaking" things ;-)

MITRE ATLAS™

Adversarial Threat Landscape for Artificial-Intelligence Systems

Reconnaissance[&]	Resource Development[&]	Initial Access[&]	ML Model Access	Execution[&]	Persistence[&]	Privilege Escalation[&]	Defense Evasion[&]	Credential Access[&]	Discovery[&]	Collection[&]	ML Attack Staging	Exfiltration[&]	Impact[&]
5 techniques	7 techniques	6 techniques	4 techniques	3 techniques	3 techniques	3 techniques	3 techniques	1 technique	4 techniques	3 techniques	4 techniques	4 techniques	6 techniques
Search for Victim's Publicly Available Research Materials	Acquire Public ML Artifacts	ML Supply Chain Compromise	ML Model Inference API Access	User Execution [&]	Poison Training Data	LLM Prompt Injection	Evade ML Model	Unsecured Credentials [&]	Discover ML Model Ontology	ML Artifact Collection	Create Proxy ML Model	Exfiltration via ML Inference API	Evade ML Model
Search for Publicly Available Adversarial Vulnerability Analysis	Obtain Capabilities [&]	Valid Accounts [&]	ML-Enabled Product or Service	Command and Scripting Interpreter [&]	Backdoor ML Model	LLM Plugin Compromise	LLM Prompt Injection		Discover ML Model Family	Data from Information Repositories [&]	Backdoor ML Model	Exfiltration via Cyber Means	Denial of ML Service
Search Victim-Owned Websites	Develop Capabilities [&]	Evade ML Model	Physical Environment Access	LLM Plugin Compromise	LLM Prompt Injection	LLM Jailbreak	LLM Jailbreak		Discover ML Artifacts	Data from Local System [&]	Verify Attack	LLM Meta Prompt	Spamming ML System with Chaff Data
Search Appli Repositories		Exploit											
Active Scanning [&]													

AI TRiSM

Google's Secure AI Framework (SAIF)

OWASP Top 10 for LLM Applications