OWASP GLOBAL AppSec

VIENNA '26 JUN 25-26

25 years of open source security

AI and the Threat Modeling Manifesto: Conflicts, Failure Modes, and Better Patterns

BY: VIKRAMADITYA NARAYAN

Quick Poll

- Raise your hand if you've tried AI to create a threat model
- How many of you thought that the threat model was great?

Structure of This Talk

01 Background

02 Threat Modeling: Humans Vs AI

03 AI Failure Modes

04 Anti-Patterns of Applying AI in Threat Modeling

05 Emerging Best Practices

06 Extend the Manifesto for AI?

Part 1: Background

About

- Creator of OWASP Precogly - an open-source alternative to commercial threat modeling tools
- Chapter Lead for Threat Modeling Connect Bangalore
- Focused conversations with 100+ AppSec leads about threat modeling challenges
- Formerly:
  - Contributor to a machine learning library
  - Built 50+ RAG based chatbots for NYC surgeons

LinkedIn

The Threat Modeling Manifesto (TMM)

A set of values and principles for effective threat modeling.

Threat Modeling Manifesto Values*

1. A culture of finding and fixing design issues over checkbox compliance.
2. People and collaboration over processes, methodologies, and tools.
3. A journey of understanding over a security or privacy snapshot.
4. Doing threat modeling over talking about it.
5. Continuous refinement over a single delivery.

"while there is value in the items on the right, we value the items on the left more."

* Source:https://www.threatmodelingmanifesto.org

Patterns and Anti-Patterns

Varied Viewpoints

Assemble a diverse team with appropriate subject matter experts and cross-functional collaboration.

Hero Threat Modeler

Threat modeling does not depend on one's innate ability or unique mindset; everyone can and should do it.

What I Learned from 100+AppSec Conversations

Priorities:

1. Derive simplified abstractions of reality
2. Surface surprising threats and toxic combinations
3. Add controls before things go into production
4. Keep threat models updated as systems evolve

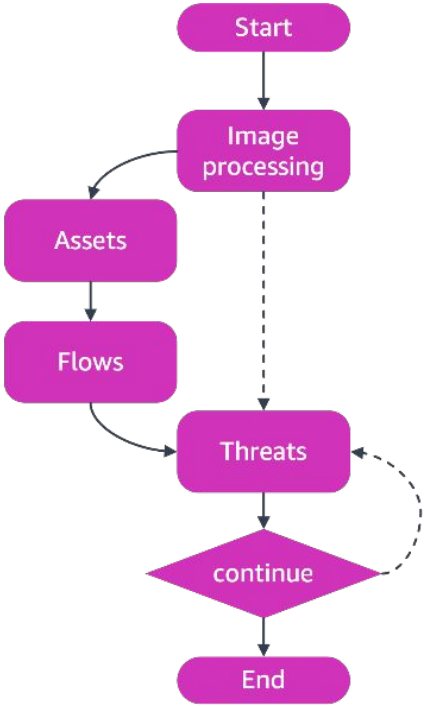
But it's getting hard:

It takes 1 to 8 days to produce a threat model; But systems are beginning to go from idea to production in weeks.

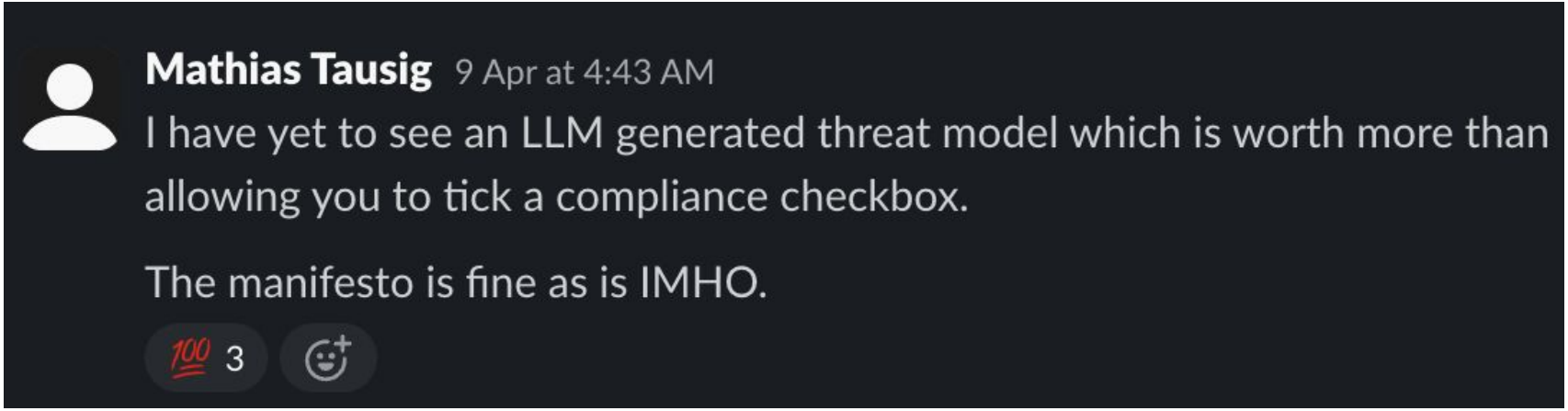
Threat Modeling with AI is a No Brainer

discovery. The co-pilot has driven 20% efficiency in our threat modeling process enabling faster models of new systems and broader scale. In addition the AITMC has uncovered an average of nine additional novel threats per model created, creating safer and more resilient

Threat Designer workflow



"Has AI quietly broken the Threat Modeling Manifesto?"



Part 2: Humans Vs AI - Live Threat Modeling Experiment

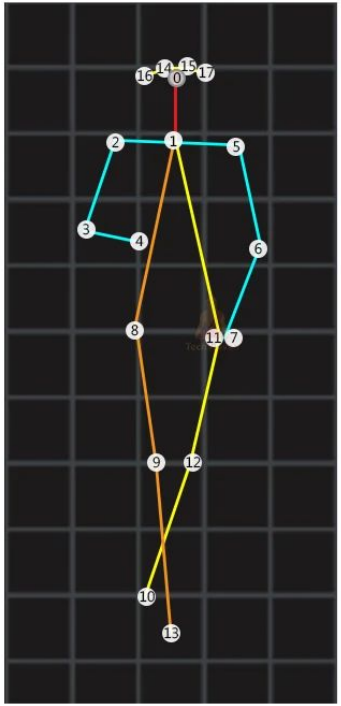
WiFi Sense for Hotels

WiFi Sense uses existing WiFi signals (IEEE 802.11 bf standard) in your hotel to:

1. Track room occupancy without door sensor
2. Track mini-bar and bathroom usage patterns
3. Provide additional services based on pose estimation

Benefits:

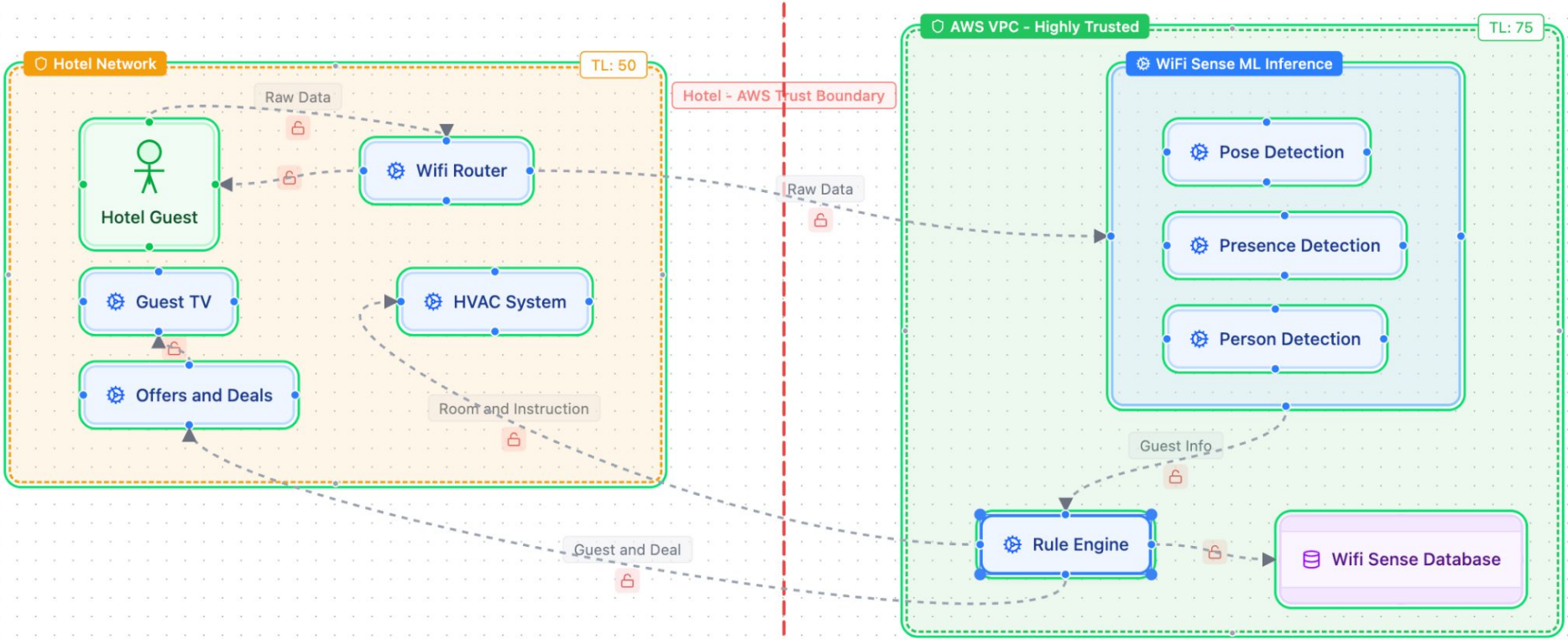
1. 15 to 30% savings on heating and AC
2. Increase in revenues of 10%



WiFi Sense for Hotels - System Description

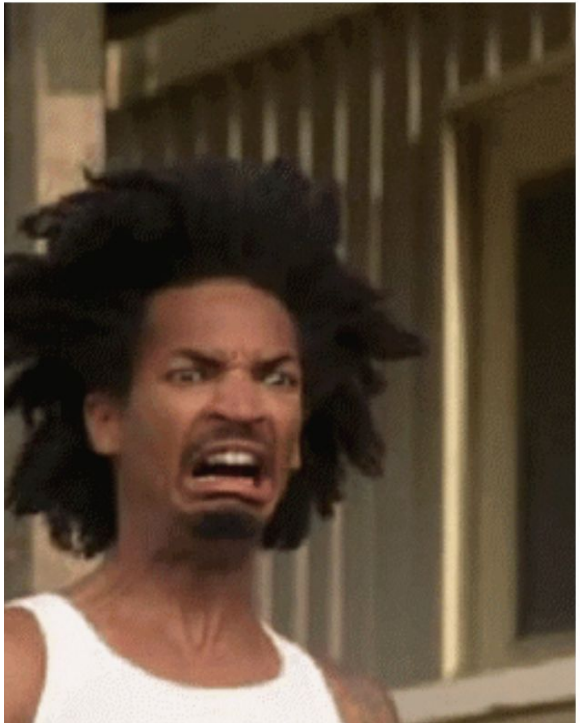
- WiFi system sends tracking information in raw data format (channel state information or CSI) to the machine learning inference system on AWS
- ML Inference System returns pose, person and presence.
- Rule engine stores this info in the DB
- Rule engine sends info back to hotel network where the HVAC system controls the heating, ventilation and AC
- The offers and deals system then provides offers to guests via the guest TV

DFD of the System - What Can Go Wrong?



AI Vs Human

Rank	Threat
1	Guest Occupancy & Behavioral Data Exposure
2	Rule Engine Compromise
3	Spoofed or Manipulated CSI Data
4	Unauthorized HVAC Control
5	ML Model Poisoning / Inference Manipulation



Humans live outside the context window

Part 3: (Silent) Failure Modes

Silent Failure Mode #1: Hallucination

Confident but incorrect threats

Invented mappings

False sense of rigor

Dangerous *because* it's indistinguishable from the real thing

Silent Failure Mode #2: Automation Bias

The tendency for people to overtrust automated systems and accept their outputs without sufficient critical thinking.

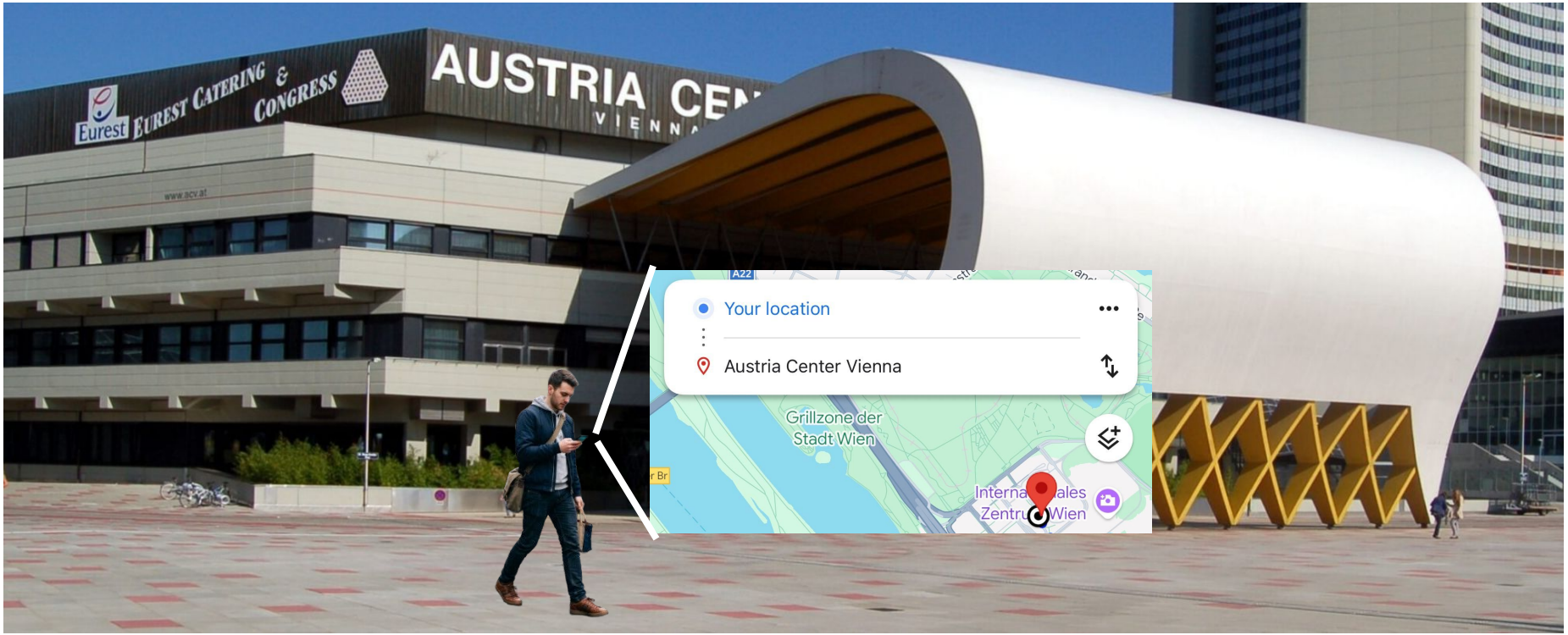
Tesla driver killed while using autopilot was watching Harry Potter, witness says

A "crash" in AppSec is a breach that happens 6 months later because the reviewer accepted the AI threat model without sufficient critical thinking.

Silent Failure Mode #3: The Illusion of Completeness

Threat Model Looks Complete	Reality
✅ All boxes filled	❌ Assumptions unexamined
✅ All threats listed	❌ Nobody understands
✅ All mitigations mapped	❌ Nobody cares

Silent Failure Mode #4: Threat Modeling Skill Atrophy



The Psychology Behind Threat Modeling Skill Atrophy

Desirable Difficulty

Learning requires struggle

Cognitive effort aids our journey of *understanding* the threat model.

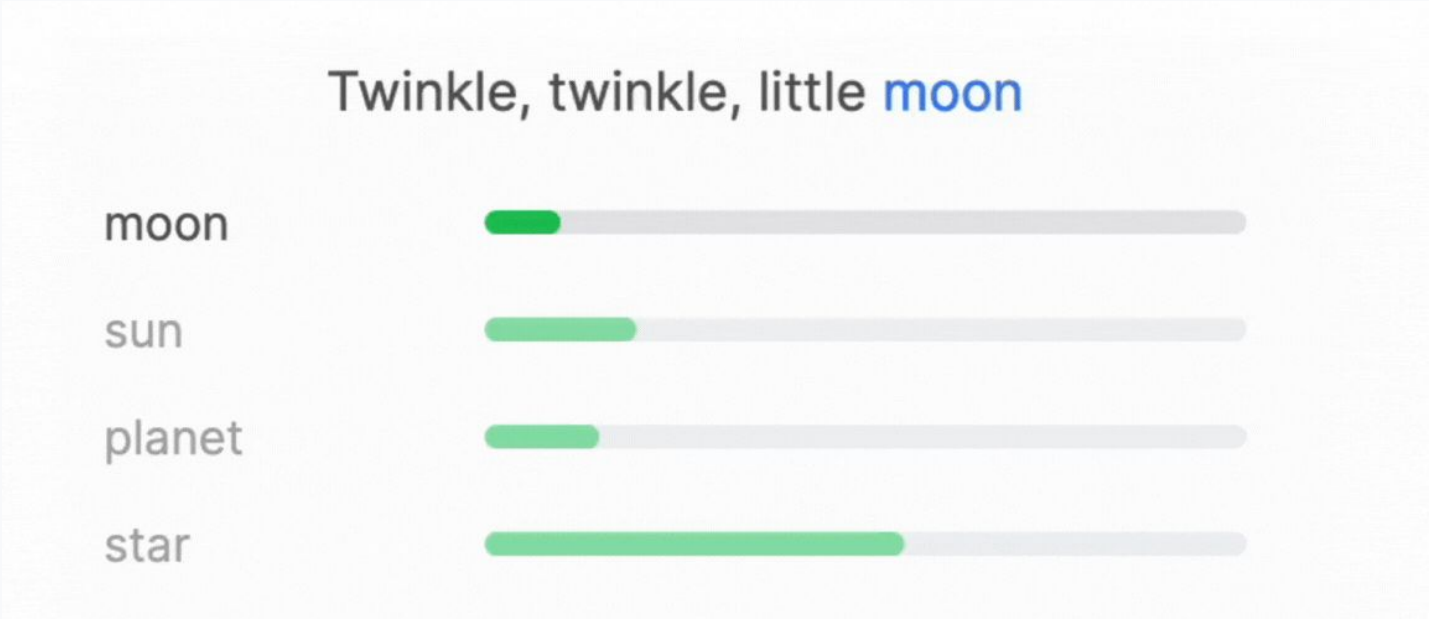
The Dunning–Kruger Effect

It ain't that hard!

When AI does the threat modeling we believe that it's "not that difficult"

Part 4: Anti-Patterns of AI Threat Modeling

LLMs 101 or How Most "AI" Today Actually Works



Anti-Pattern #1: The Hero Threat Modeler

These anti-patterns inhibit threat modeling:

Hero Threat Modeler

Threat modeling does not depend on one's innate ability or unique mindset; everyone can and should do it.

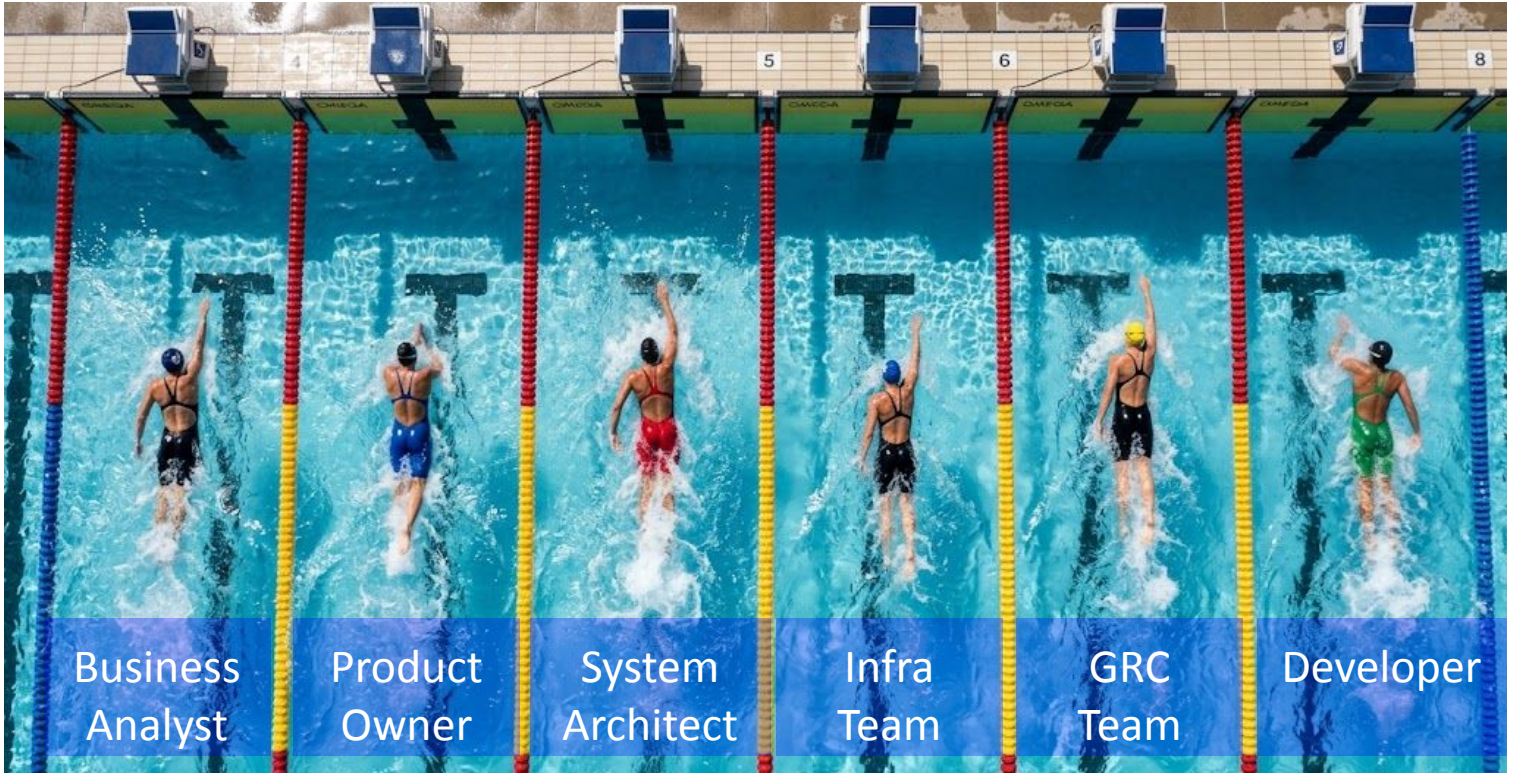
The Claim

"We built an AI threat modeling tool which automatically completed 150+ threat models in 2 weeks."

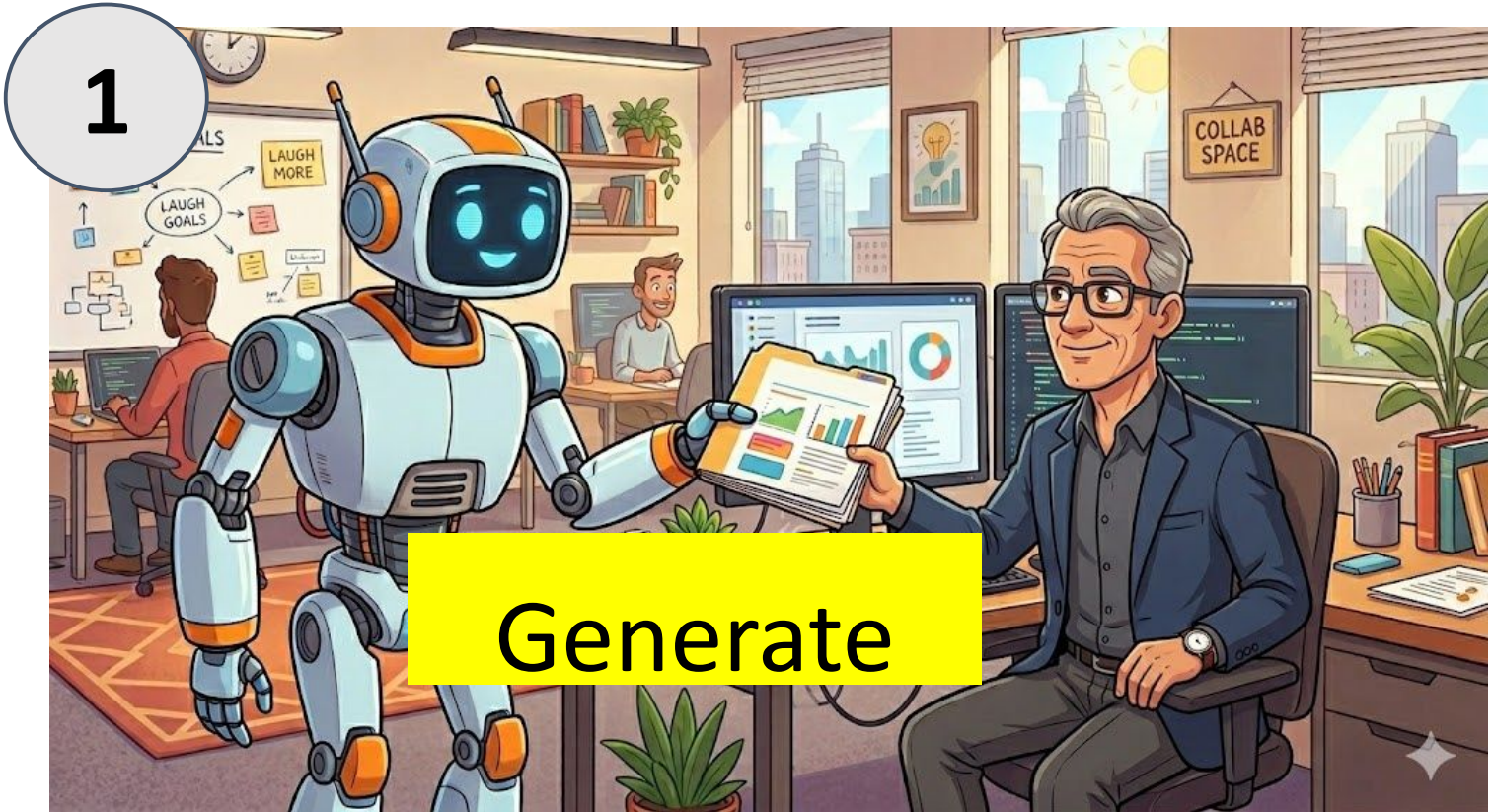
What This Results In

- Centralized expertise
- Low shared ownership
- Checkbox exercise

Anti-Pattern #2: De-emphasizing Collaboration and People



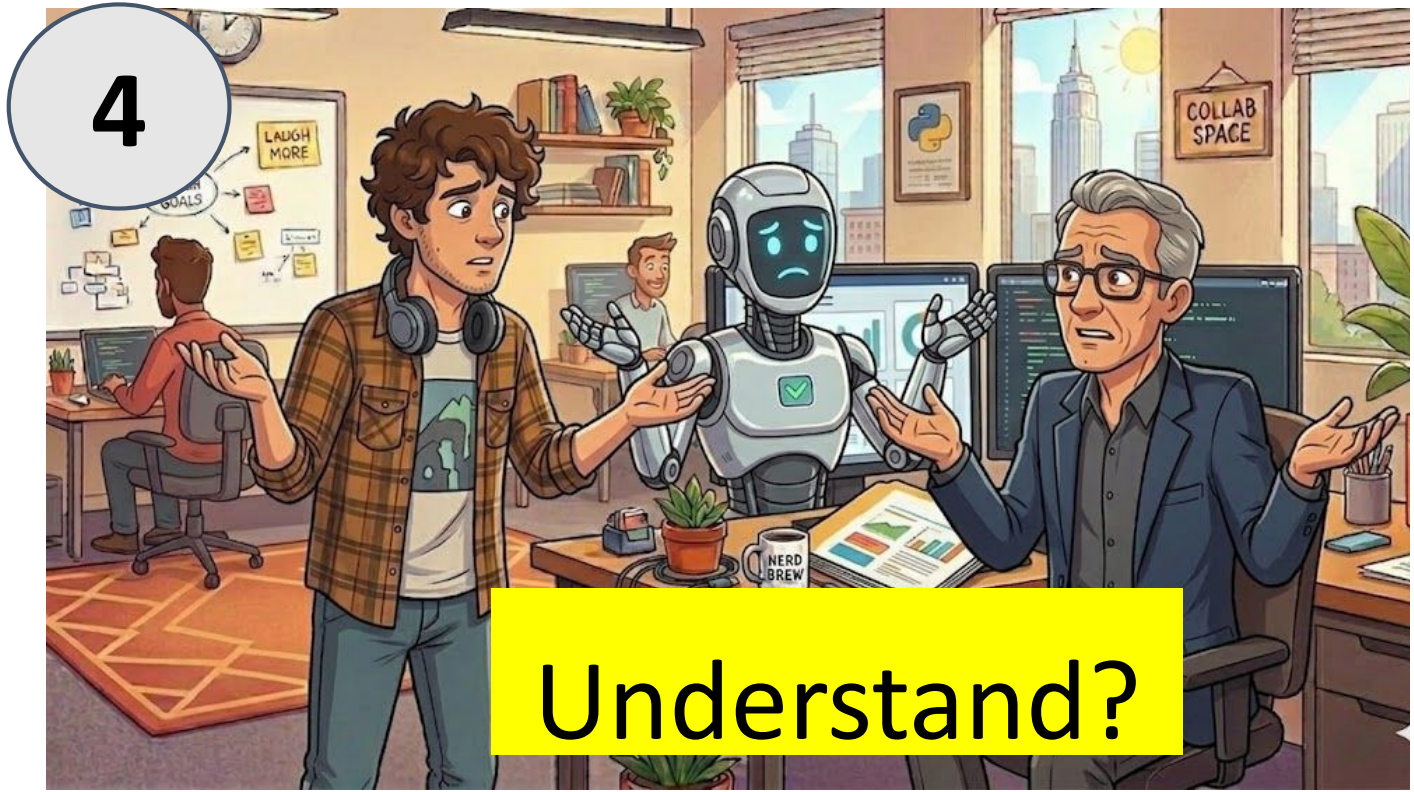
Anti-Pattern #3: Snapshots Over Journey of Understanding



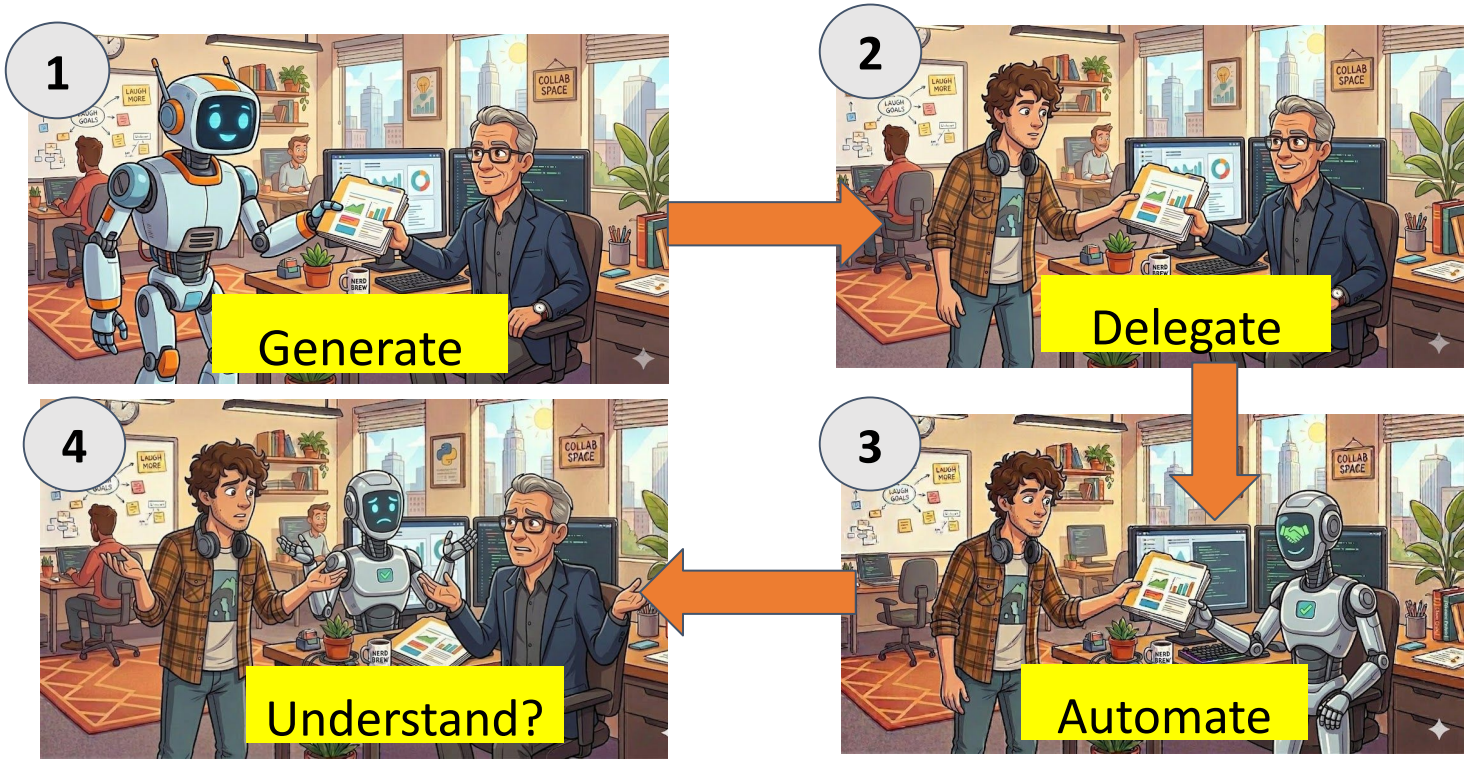
Anti-Pattern #3: Snapshots Over Journey of Understanding

Anti-Pattern #3: Snapshots Over Journey of Understanding

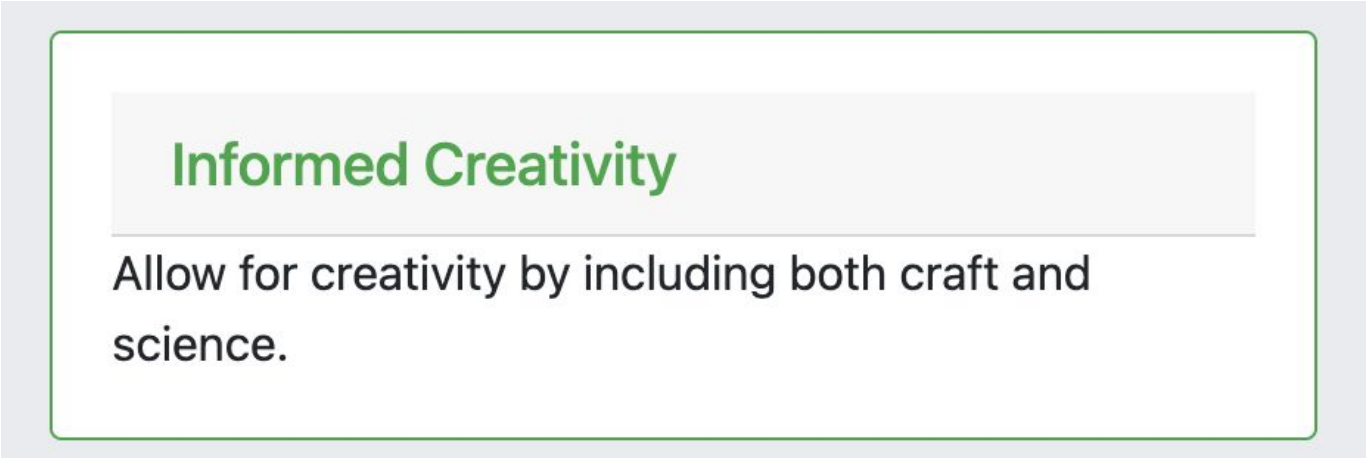
Anti-Pattern #3: Snapshots Over Journey of Understanding



Anti-Pattern #3: Snapshots Over Journey of Understanding



Anti-Pattern #4 - Delegating Creativity to AI



- Creativity is an emergent property of human-interaction

Anti-Pattern #5 - Choosing Exhaustive Enumeration Over Deliberation

- Enumerated threats ≠ secure design

Mathias Tausig 5:43 PM

Slightly more detailed opinion:

What you are getting from threat modeling automation tools (LLM base or not) are very generic security statements like "Mitm on unencrypted network connection". Those findings are not worthless, especially not if you are in an organization with a very low maturity level. But the countermeasures can all be summarized by "Implement ASVS", there a no product specific design flaws in there.

- ✅ Input validation
- ✅ Output encoding
- ✅ TLS 1.3 everywhere
- ✅ AES-256 at rest
- ✅ MFA enabled
- ✅ RBAC implemented
- ✅ WAF deployed
- ✅ Rate limiting
- ✅ Audit logging
- ✅ Dependency scanning
- ✅ Secrets in vault

Part 5: Towards Manifesto Aligned AI-Patterns

"Has AI quietly broken the Threat Modeling Manifesto?"

The Manifesto tried to solve a previously human focused effort. Could an AI focused effort still implement the principles of the Manifesto? Maybe, if Agents are instructed to do so, but would it give the same meaning…

- Matthew J. Coles on OWASP Slack

Would An AI Threat Model Carry the Same Meaning?

"Threat modeling's value is not just in the list of threats it produces. It is in what happens to the people in the room while they produce it. The arguments about trust boundaries. The moment someone from the backend team realizes the payments team has been assuming something that isn't true …

These are not inefficiencies to be optimized away. They are the point. The journey of understanding is the threat model, not a side quest of it."

- Izar Tarandach

With AI Are We Just Shifting Things Right?

Before AI

"we need security expertise to do threat modeling"

With AI

"we need security expertise to validate the threat modeling our AI did."

https://threatmodeling.dev/team-werewolves-wins/

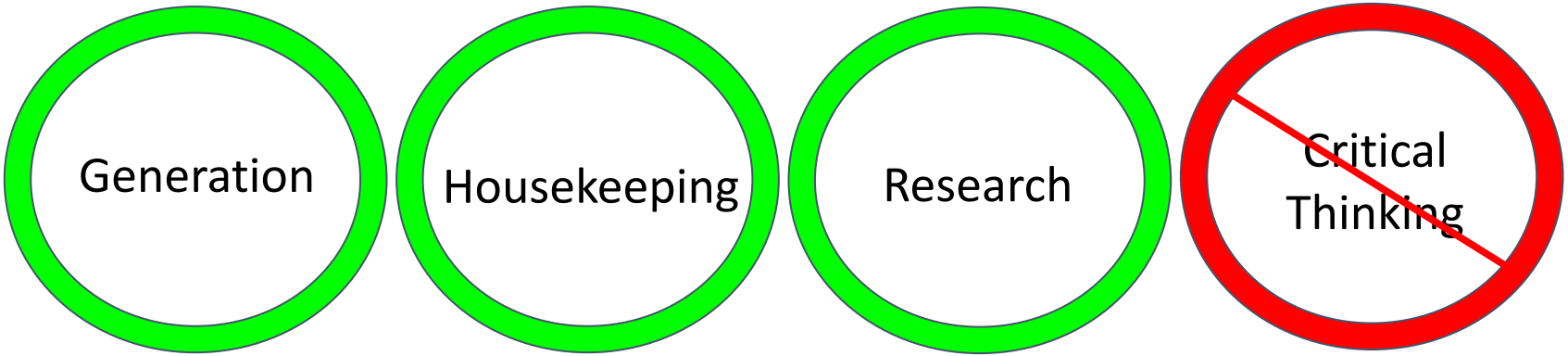
Threat Modeling is a Soft Skill

The hardest part of **scaling a threat modeling program** isn't the technical analysis—**it's securing buy-in**. Threat modeling is fundamentally a **soft skill** rooted in human interaction and shared vision.
- Sebastien Deleersnyder

<u>But what if AI does all the work in the SDLC? Do we still need soft skills?</u>

- Can the AI allocate a budget to mitigate a risk?
- Can the AI pay the fine if there is a breach?

Pattern #1 - AI as Accelerant



Key Mental Model: AI accelerates the process. Humans own judgment.

Pattern #2 - Design for Disagreement and Multiple Viewpoints

1. Keep the visual aids - the DFD becomes more important with AI
2. Make humans explicitly accept or reject AI output.

Don't Just Hit Enter

```
● Update(.github/workflows/release-please.yml)
  └ Added 1 line, removed 1 line
     12    release-please:
     13      runs-on: ubuntu-latest
     14      steps:
     15 -      - uses: googleapis/release-p
        -lease-action@v4
     15 +      - uses: googleapis/release-p
        +lease-action@v5
     16          with:
     17            release-type: simple

● Done. The change is just @v4 → @v5. Want me to
  commit and push this?

❯ yes, commit and push it
```

This creates automation bias

This keeps the human engaged

```
❯ Actually, I'd like to do the commit myself.
  Just guide me through it.

● Sure. Run these commands:

  git add .github/workflows/release-please.yml
  git commit -m "chore: bump release-please-action
  from v4 to v5"
  git push

  Note the chore: prefix - this is a good habit now
   that you're using Release Please. This commit
  won't trigger a version bump since chore is not a
   fix or feat, but it'll be properly parsed.
```

Pattern #3 - Keep the Human in the Driver Seat

- The human remains the driver and lead author (sign-off as @HumanName + AI not as AI + @HumanName)*

  *Credit to Hendrik Ewerlin for first proposing this idea

- But why are we anthropomorphizing the AI?
  - We don't say "This report was produced by HumanName + Microsoft Word"
  - AI is just another tool regardless of what the AI marketing teams say

Part 6: Extending the Manifesto for the Age of AI

A Proposed Starting Point for a Sixth Value

AI-augmented human reasoning over fully autonomous generation.

Danke schön!

LinkedIn