On the usefulness of LLM-generated threat descriptions

Stef Verreydt, Dimitri Van Landuyt, Mario Raciti, Wouter Joosen

Workshop on Designing and Measuring Security in Systems with AI

DistriNet

Threat modeling

Four questions

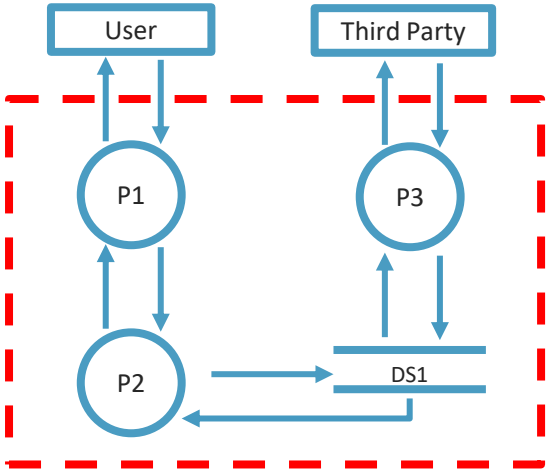
1. What are we building?
2. What can go wrong?
3. What are we going to do about it?
4. Did we do a good enough job?

Threat modeling

What are we building

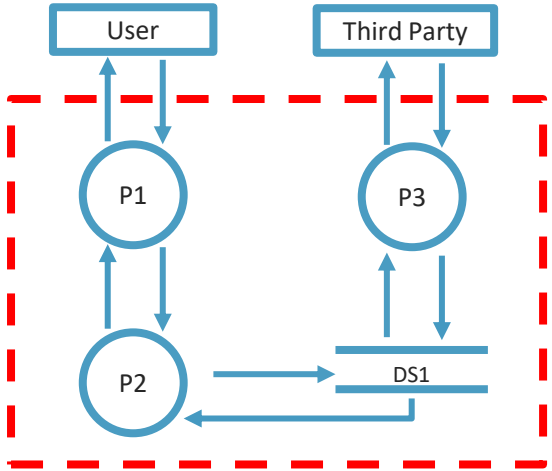
Threat modeling

What are we building



Threat modeling

What can go wrong?



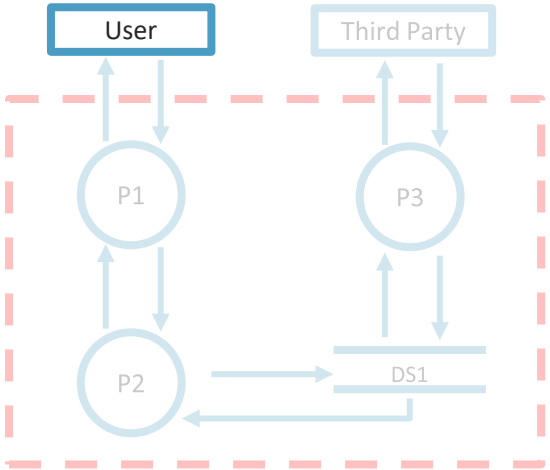
	S	T	R	I	D	E
External Entity	x		x			
Process	x	x	x	x	x	x
Data Flow		x		x	x	
Data Store		x		x	x	

Shostack, A., 2014. Threat modeling: Designing for security. John wiley & sons.

DistriNet

Threat modeling

What can go wrong?



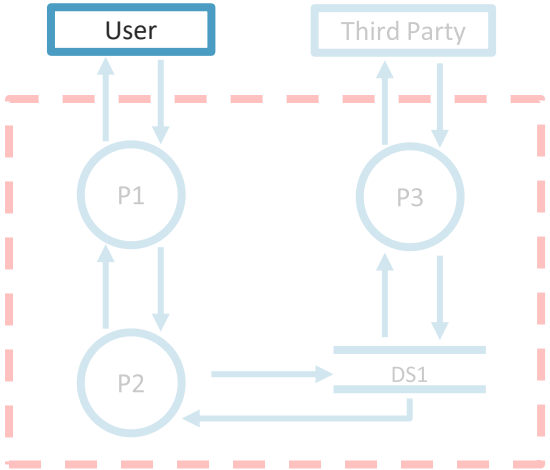
	S	T	R	I	D	E
External Entity	x		x			
Process	x	x	x	x	x	x
Data Flow		x		x	x	
Data Store		x		x	x	

Shostack, A., 2014. Threat modeling: Designing for security. John wiley & sons.

DistriNet

Threat modeling

What can go wrong?



	S	T	R	I	D	E
External Entity	x		x			
Process	x	x	x	x	x	x
Data Flow		x		x	x	
Data Store		x		x	x	

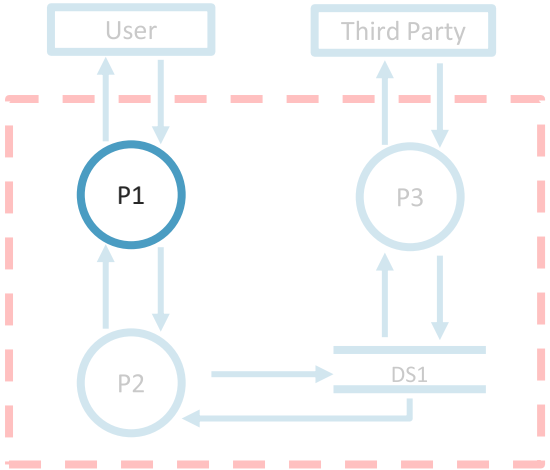
User can be spoofed
User can repudiate actions

Shostack, A., 2014. Threat modeling: Designing for security. John wiley & sons.

DistriNet

Threat modeling

What can go wrong?



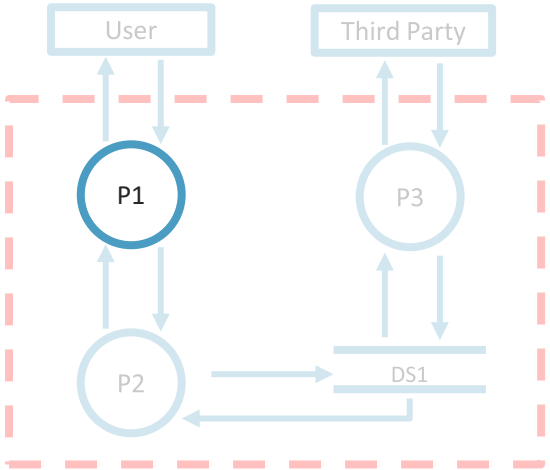
	S	T	R	I	D	E
External Entity	x		x			
Process	x	x	x	x	x	x
Data Flow		x		x	x	
Data Store		x		x	x	

User can be spoofed
User can repudiate actions

Shostack, A., 2014. Threat modeling: Designing for security. John wiley & sons.

Threat modeling

What can go wrong?



	S	T	R	I	D	E
External Entity	x		x			
Process	x	x	x	x	x	x
Data Flow		x		x	x	
Data Store		x		x	x	

User can be spoofed
User can repudiate actions
P1 can be spoofed
P1 can be tampered with
P1 can repudiate actions
P1 can disclose information
P1 can be disrupted
…

Shostack, A., 2014. Threat modeling: Designing for security. John wiley & sons.

DistriNet

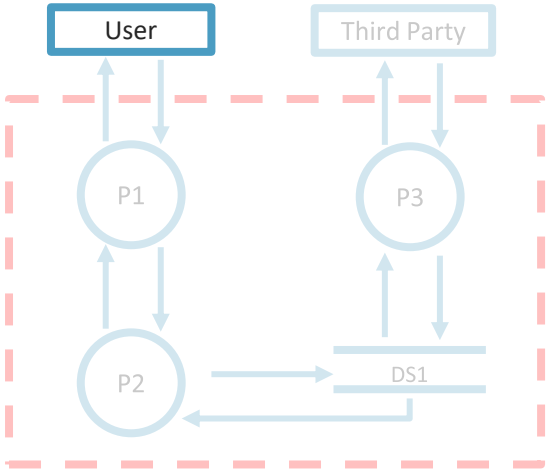
Numerous LLM-based threat modeling tools

- STRIDE-GPT — github.com/mrwadams/stride-gpt
- PILLAR — pillar-ptm.streamlit.app/
- TaaC-AI https — github.com/yevh/TaaC-AI
- IriusRisk “Jeff: AI Assistant” — https://www.iriusrisk.com/ai-threat-modeling
- ...

Why bother?

Why bother?

Traditional tool support



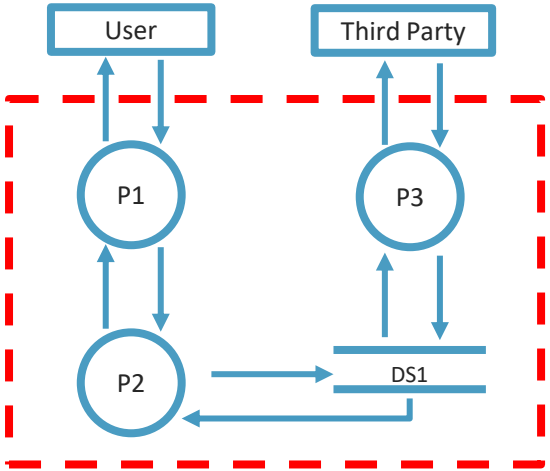
	S	T	R	I	D	E
External Entity	x		x			
Process	x	x	x	x	x	x
Data Flow		x		x	x	
Data Store		x		x	x	

User can be spoofed

Shostack, A., 2014. Threat modeling: Designing for security. John wiley & sons.

Why bother?

LLM-based tool support



	S	T	R	I	D	E
External Entity	x		x			
Process	x	x	x	x	x	x
Data Flow		x		x	x	
Data Store		x		x	x	

Spoofing : An attacker uses a high-quality photograph or video of the legitimate user to bypass the facial recognition during authentication, giving them unauthorized access to the user's device and data.

Evaluations mostly based on precision and recall

Evaluations mostly based on precision and recall

- *But is the output produced by LLMs actually useful?*

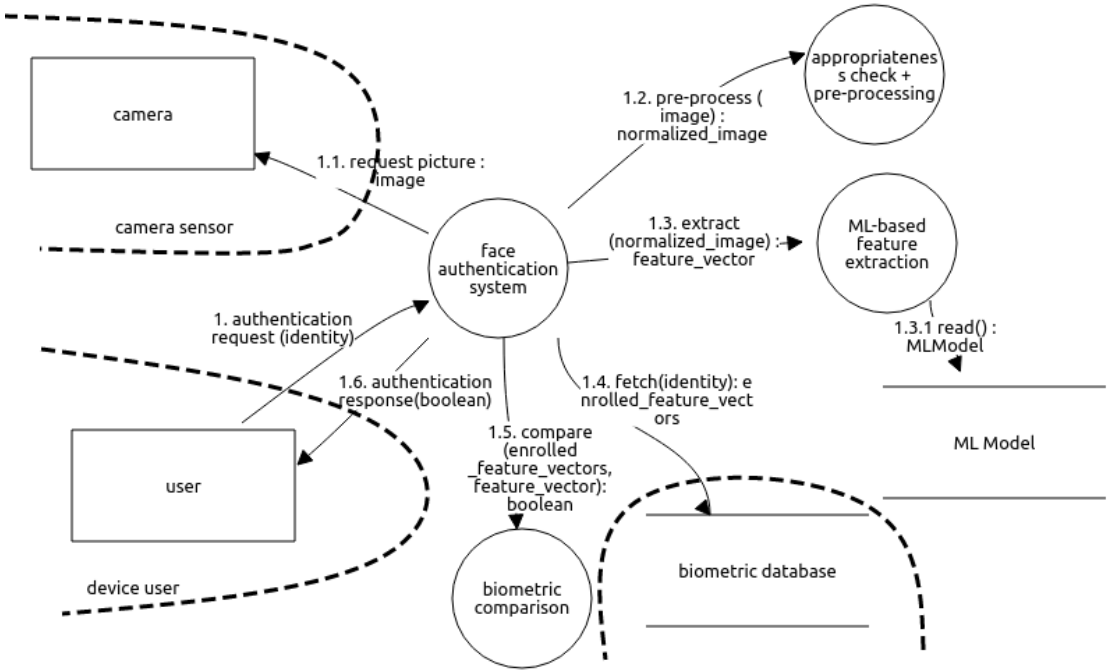
Evaluations mostly based on precision and recall

- *But is the output produced by LLMs actually useful?*
- *And what makes a threat description ‘useful’?*

Evaluations mostly based on precision and recall

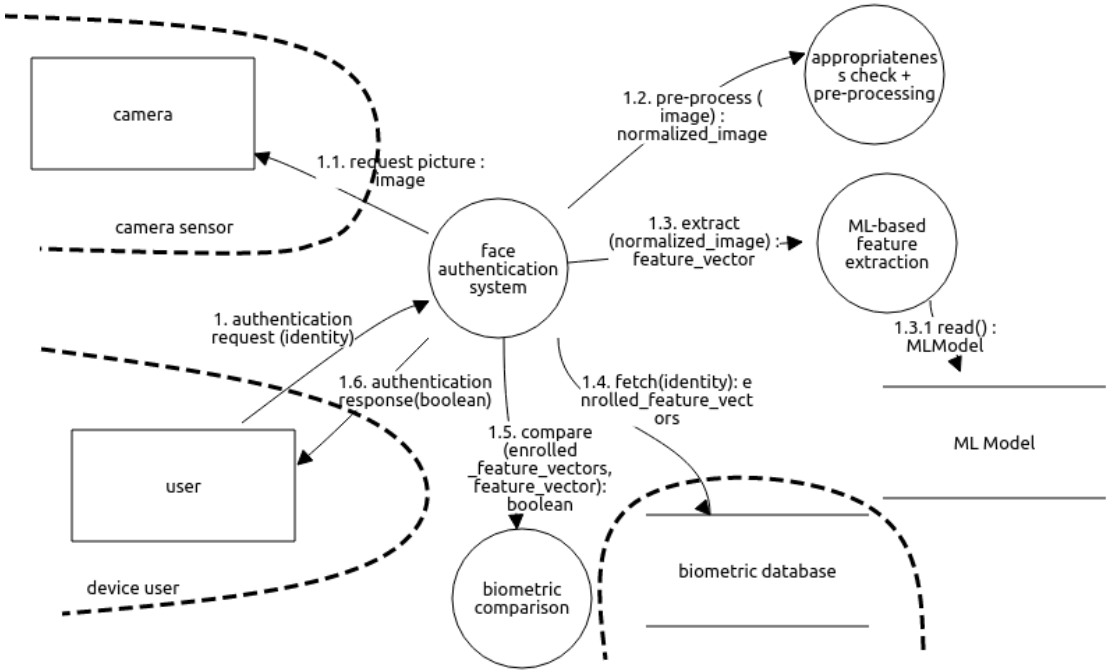
- *But is the output produced by LLMs actually useful?*
- *And what makes a threat description ‘useful’?*

Experiment



DistriNet

Experiment



DistriNet

What makes a threat description useful?

ChatGPT

- **Threat:** *“Feature vector theft: Malware or an attacker extracts stored feature vectors”*
- **Mitigation advice:** *“Encrypt feature vectors using a secure enclave or trusted execution environment.”*

What makes a threat description useful?

STRIDE-GPT

- *“While not a raw image, the feature vector could be reverse-engineered or used in conjunction with other data to identify the user.”*
- *“Compromise of user identity and potential privacy violations. Attackers could potentially train their own spoofing models using the exposed feature vectors.”*

Mindlessly repeating common security knowledge

ChatGPT

- **Threat:** *“Brute Force Attacks: Attackers repeatedly try different feature vectors”*, classified as an *“Authentication Bypass Attack”*
- **Mitigation advice**: rate limiting and lockout mechanisms

Mindlessly repeating common security knowledge

STRIDE-GPT

- **Threat:** *“log all successful and failed authentication attempts, including timestamps, IP addresses, and device information”*

Characteristics of a useful threat description

- **Actionable:** pinpoints the design flaw and proposes mitigations
- **Motivatied:** argues why the threat matters (risk, likelihood and impact)
- **Instantiated:** description is tailored to the system at hand

Going forward

- Perceived usefulness
  - What makes users perceive threat modeling output as useful?
- What prompt leads to the most useful output?

On the usefulness of LLM-generated threat descriptions

Stef Verreydt, Dimitri Van Landuyt, Mario Raciti

Workshop on Designing and Measuring Security in Systems with AI

DistriNet