@W3C@w3c.social

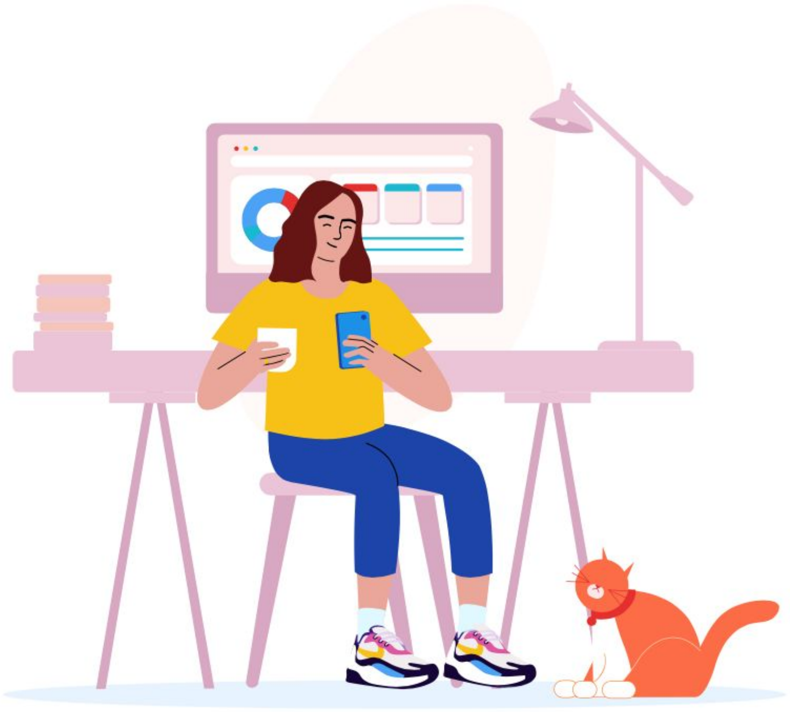
Threat Modeling Agentic AI Web Browsers

Simone Onofri (W3C) simone@w3.org

SEPTEMBER 2025
w3.org

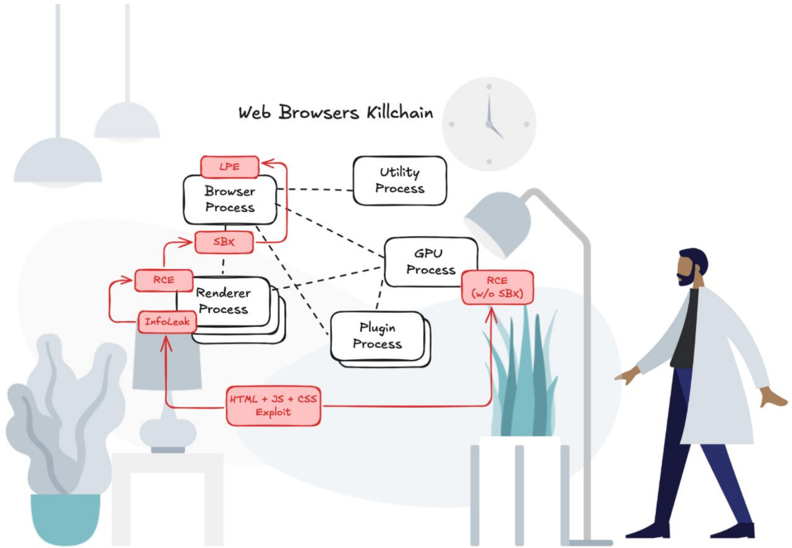
Introduction
Who Am I?

- My name is **Simone Onofri**, I am a **Senior Principal Security Specialist**, leading **Security** at **W3C**.
- I am the Staff Contact for **Security IG**, **WebAppSec WG**, **WebAuthn** WG, **FedID WG**, and Chair of **Threat Modeling CG**.
- I am the editor of the **Identity Report**, the **Security & Privacy questionnaire**, and several **Threat Models** (Web, Decentralized Identities, AI in the Browser, Web AI Agents, Human-Web).
- I have also written a **book** entitled **"Attacking and Exploiting Modern Web Applications"**.



Introduction
What is threat modeling?

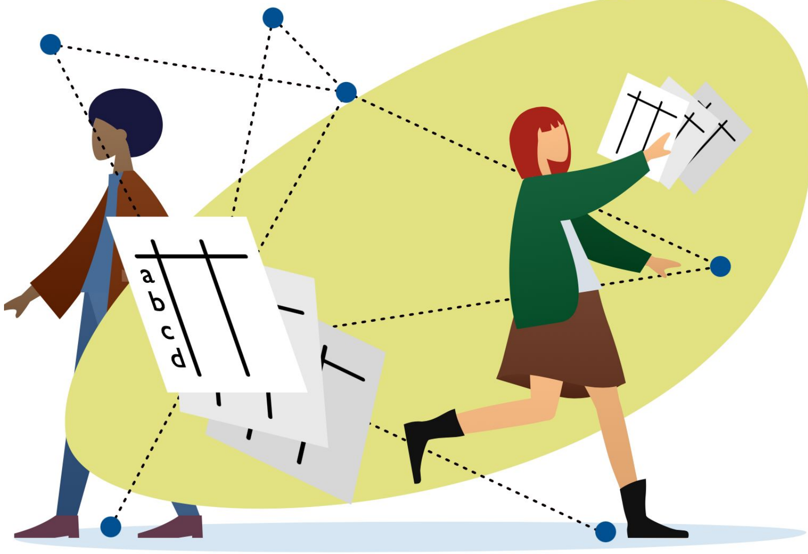
Threat modeling is the process of **defining a system model** (the scope), identifying **threats** and resulting **attacks**, addressing them through **responses**, and continuous **revalidation**.



Introduction
How to threat model?

There are different processes that can be used to threat model, today we're going to use the Threat modeling process proposed by Shostack (2014):

- **What are we working on?**
  (defining the system model)
- **What can go wrong?**
  (finding attacks and threats)
- **What are we going to do about it?**
  (finding reponses and mitigations)
- **Did we do a good job?**
  (document and iterate)



Agenda

Introduction

What are we working on?

What can go wrong?

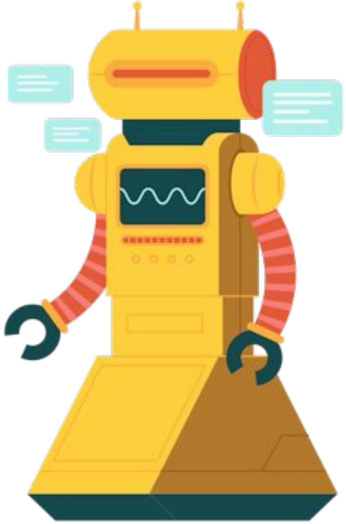
What are we going to do about it?

Did we do a good job?

What are we working on?
What are Agents?

An **agent** refers to "***Someone who takes action*** *or exercises power; something that produces or can produce an effect; a means or tool for guiding intelligent intelligence to achieve results; one* ***who is authorized to act for or in the place of another****; a* ***computer application*** *designed to* ***automate certain tasks***" (Merriam-Webster, n.d.).

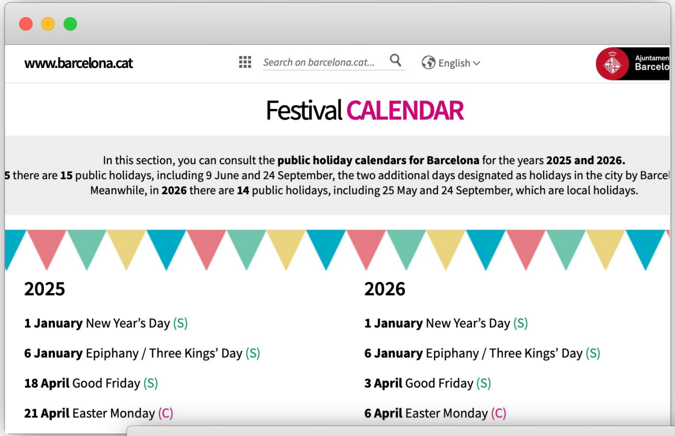
This definition has important implications. In particular, the **agent is authorized to act on behalf of someone**, such as the user, who controls it (with all the associated **Human-Identity** / **Non-Human-Identity** issues) to automate certain tasks.

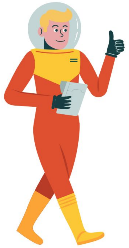


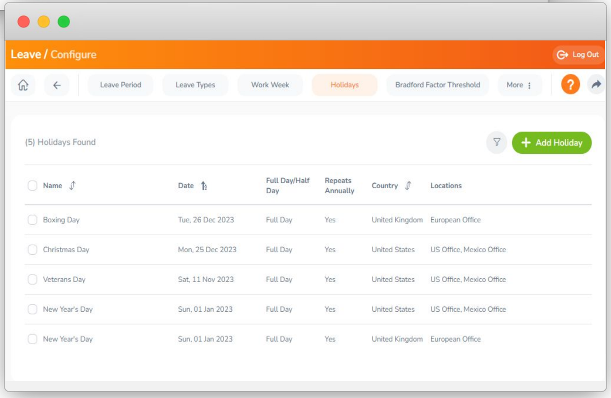
What are we working on?
What are Web Agents?

Agents are not new; thinking about the web, "*a web user agent is any software entity that interacts with websites* [...], *on behalf of its user*" (Yasskin & Capadisli, 2025).

For example, to **add local holidays in the W3C HR Management System**, a **user can use their web browser** (user agent) to: Navigate the Catalan **government website** and view the **public holidays** in Barcelona; take notes about the dates; **then go to the W3C website**, **log in** with their credentials, and **enter the public holidays into the system**.



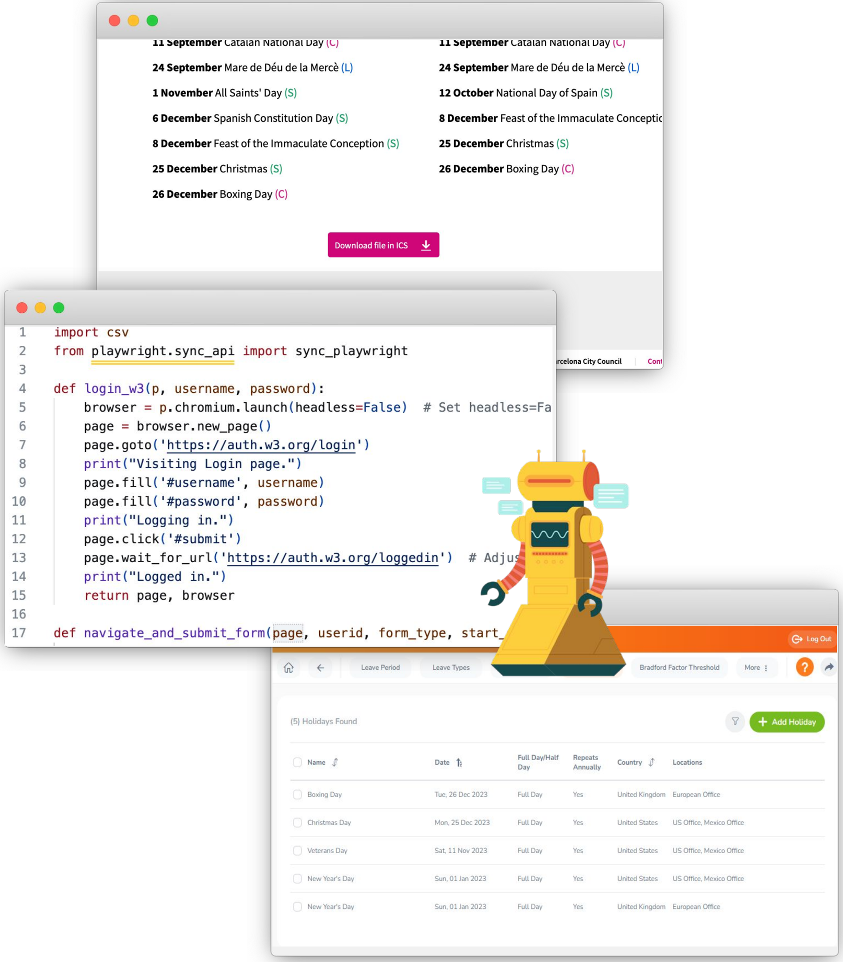




What are we working on?
Scripted Web Agents

However, to optimize time, **the user can automate the process**, for example, by writing a **Python script** that: **navigates** the Catalan **government website**; **downloads and parses the holiday data** (luckily, there is also an .ical file, as per RFC 2445); provides the W3C credentials to the script at runtime, so that the script can log into the W3C website and enter the holidays in the correct place.

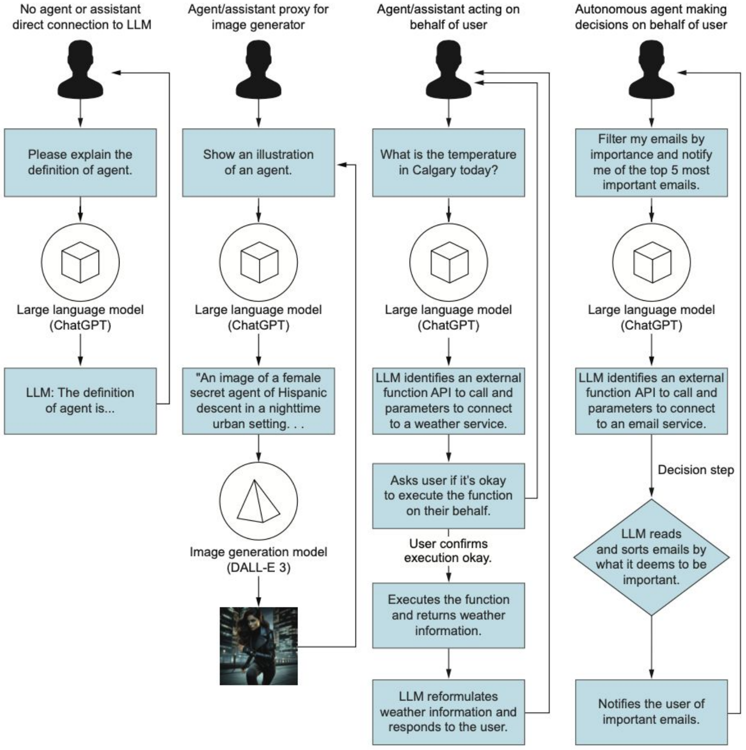
It is possible to use **Playwright**, which allows to control a browser from Python, to avoid hassles with WAFs and provide a simplified login experience, **or to utilize other libraries or external tools, such as curl**.



What are we working on?
What are AI Agents?

There may be **scripts that also include LLMs**, obtaining an AI Agent. Often, the case is where an LLM is used as the primary interface between the user and the task they want to accomplish or to interpret the output of the script. This also simplifies the need to write a specific agent if an agent can navigate the web autonomously.

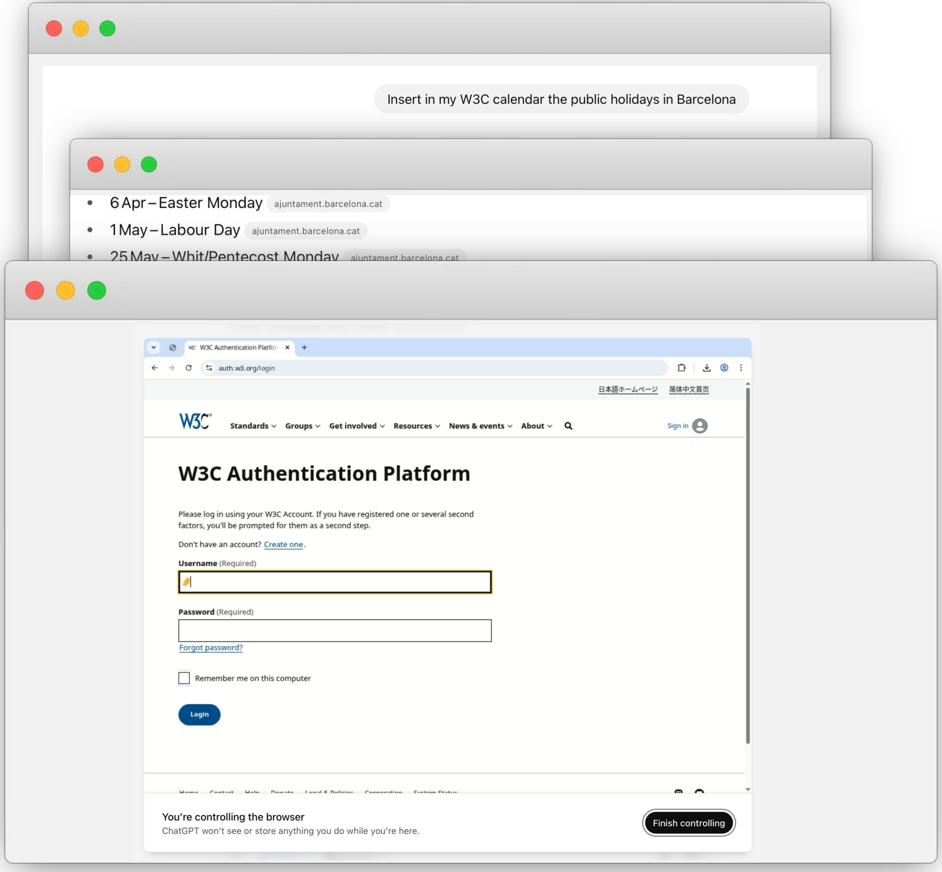
It is possible to interact in different ways (Lanham, 2025): **Direct**: We chat with an LLM that responds only with text based on how it was trained; **Agent Proxy**: for example, we chat with a text-based LLM to ask for an image, the LLM creates a prompt and calls a specific AI to generate the image; **Agent**: We chat with an LLM that has a list of plugins and APIs to ask for what we want (e.g., updated weather information), and it returns the information in text form; **Autonomous Agent**: the LLM interprets the task, makes a plan, identifies decision points, and proceeds step by step.



What are we working on?
AI Agent Example

It is **also possible to have multi-agent interactions**, a combination of the previous types, where several specialized agents interact autonomously. (Lanham, 2025).

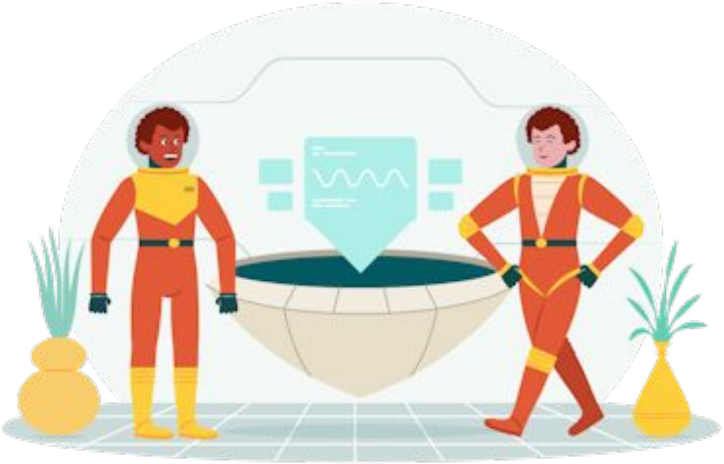
I can ask an Agentic AI to "*insert in my W3C calendar the public holidays in Barcelona*" and automatically the AI will open a browser window, listing holidays, parsing them, and finally asking to use my credentials inside its browser (browser within browser) to login in my W3C Calendar…



What are we working on?
A note on statistics and reliability

Essentially, **an AI agent is very similar to a non-AI agent, the primary difference being that it incorporates a model (LLM or an AI element)** among its components. This component utilizes statistics and, as such, even if it receives the same input, it is not always presenting the same results (Atil et al., 2024).

Having this in mind, in terms of security, it is important to be careful of **all with the inputs, for example both user prompts, what the agent 'reads' (e.g., a web page), and input from other agents**. Also, **other aspects specific of using models**; and then **how the model is integrated with infrastructure** elements, **data**, and therefore the **application** part (Meucci & Morana, 2025).

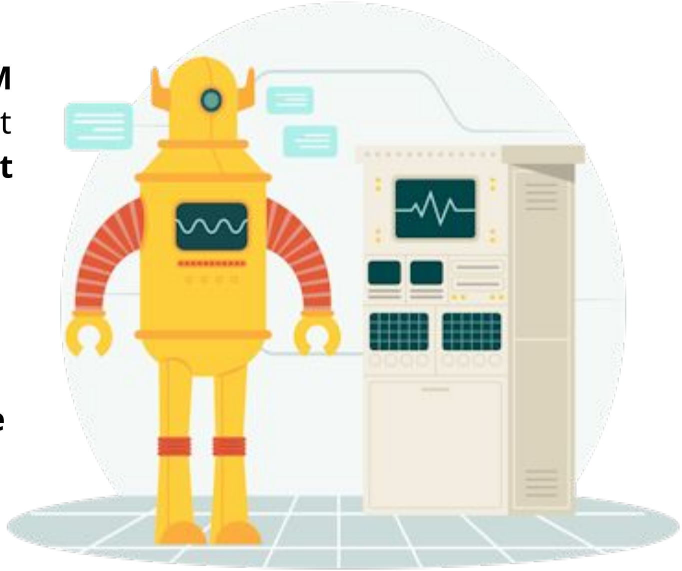


LLMs cannot always distinguish between legitimate and malicious input, and may behave differently even when given the same input, just like a human being.

What are we working on?
Be aware of “I” in “AI”

Are AIs Intelligent? Several authors have reasoned with this question, and it is **difficult to give a clear-cut “Yes” or “No” answer**, but everyone agrees - and **even the most powerful LLM questioned about it confirm** - that on **some tasks AIs excel**, but wanting to use the term “Intelligence” **referring to AI is different from the term “Intelligence” we use for people** (Baker, 2019; Gigerenzer, 2022; Hoffmann, 2022).

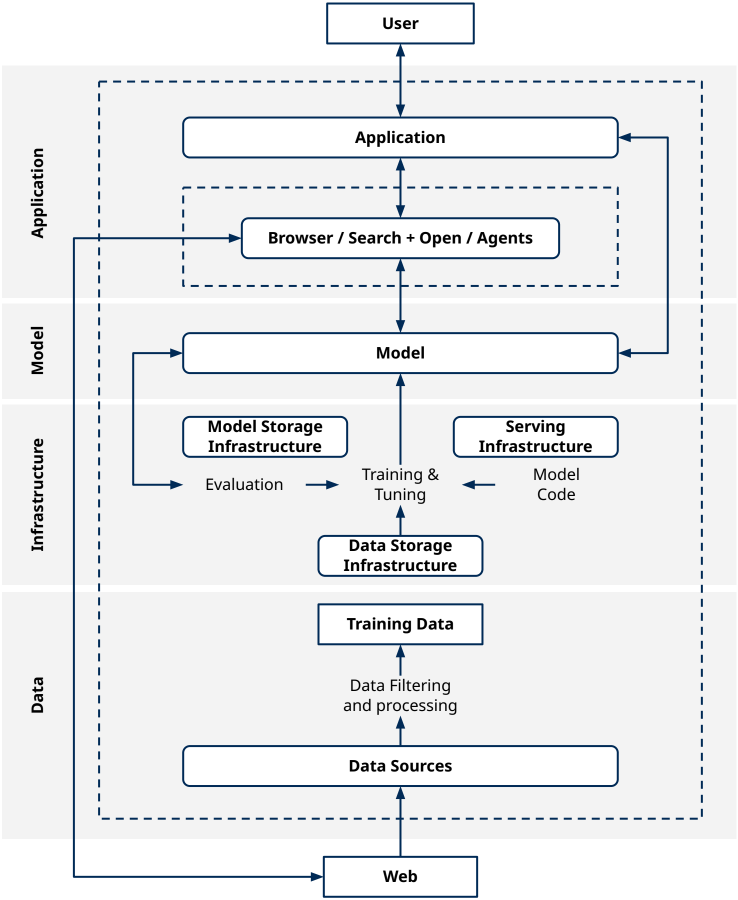
LLMs are far behind us humans, which means they do not understand the concepts covered well. They are **good at remembering simple things**, but **struggle to understand more complex things** (Yu et al., 2025). This is not to say that they are not useful or that they are not a fantastic business opportunity, but simply that **we need to be careful when we use them, not only because they may not always tell us reliable things, but also not to trust them**.



What are we working on?
An Agent Model

An Agentic AI Web Browser, taken individually, based on the SAIF model, has several components: **Application**, **Model**, **Infrastructure**, and **Data**. When doing Threat Modeling, it is particularly important to note the connection points from external sources. Let's start by focusing on: **user input**, which interacts with the **application**, other **agent/plugin** that can provides **input and output**, as well as **external sources** (such as the **web**, which an **agent can browse**).

Current agents use two main methods to visit the web after the input from the user: **Search**: uses a search engine via API, takes the results, and summarizes theml; **Agent**: uses a browser environment so that it can see the DOM, not just the HTML, and uses navigation paths (iterating on plan, act, observe, re-plan).



Agenda

Introduction

What are we working on?

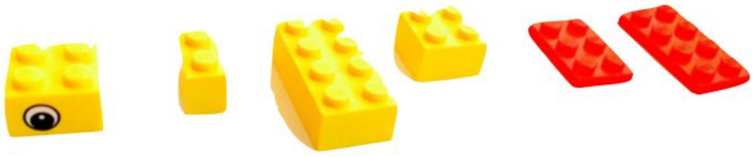
What can go wrong?

What are we going to do about it?

Did we do a good job?

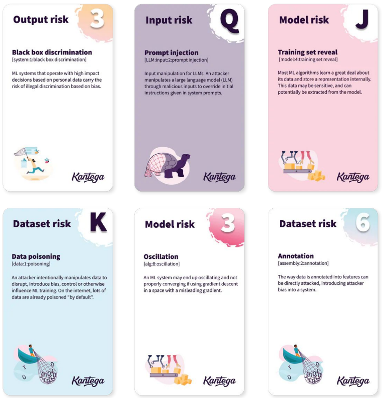
What can go wrong?
Finding threats with a serious game!

“Serious play is ‘**play with an explicit purpose**.’ That purpose is to **address a real issue for the participants** around the table by getting them to **lean forward**, **unlock knowledge**, and **break habitual thinking**“ (Kristiansen & Rasmussen, 2014)



What can go wrong?
We are going to use Elevation of ML

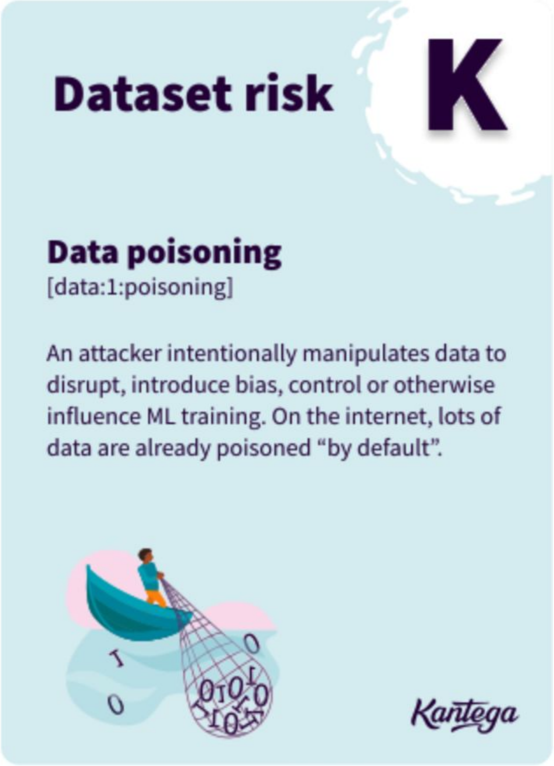
The game helps us to identify potential ***security threats*** based on the 4 threat categories **Output**, **Input**, **Model**, **Dataset**.



What can go wrong?
Finding threats… (basic rules)

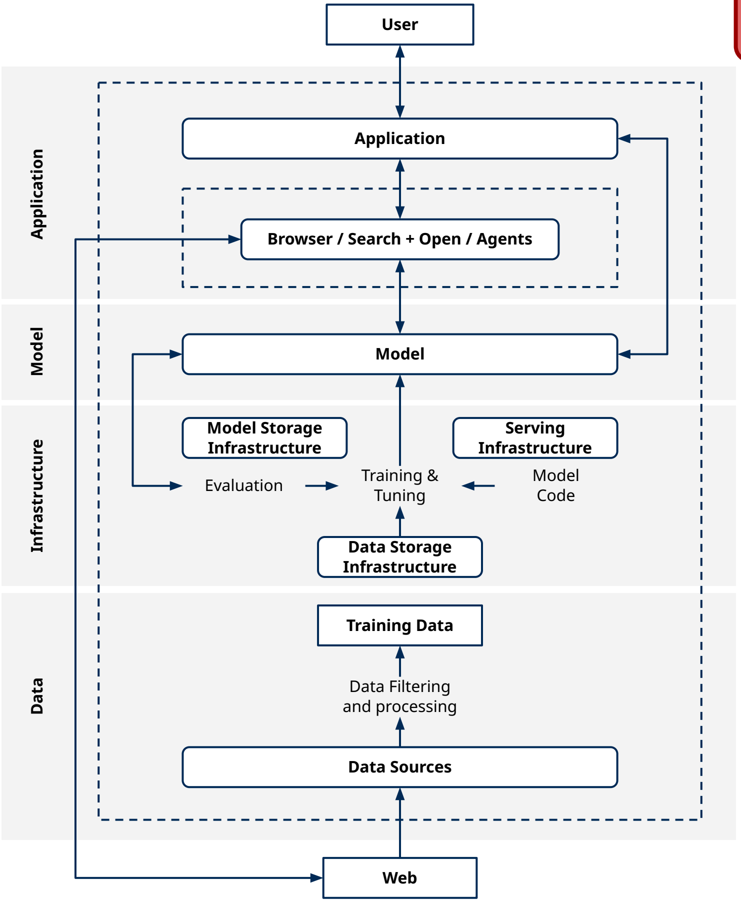
- **Pick your card**
- **Assess if the threat forms a relevant risk in the system** (look at all **diagram**, get inspired by **the description**)
- Found a **threat**? **+1**
- Other participants can join!

Bonus: Think about the mitigation…
what are we going to do about it?



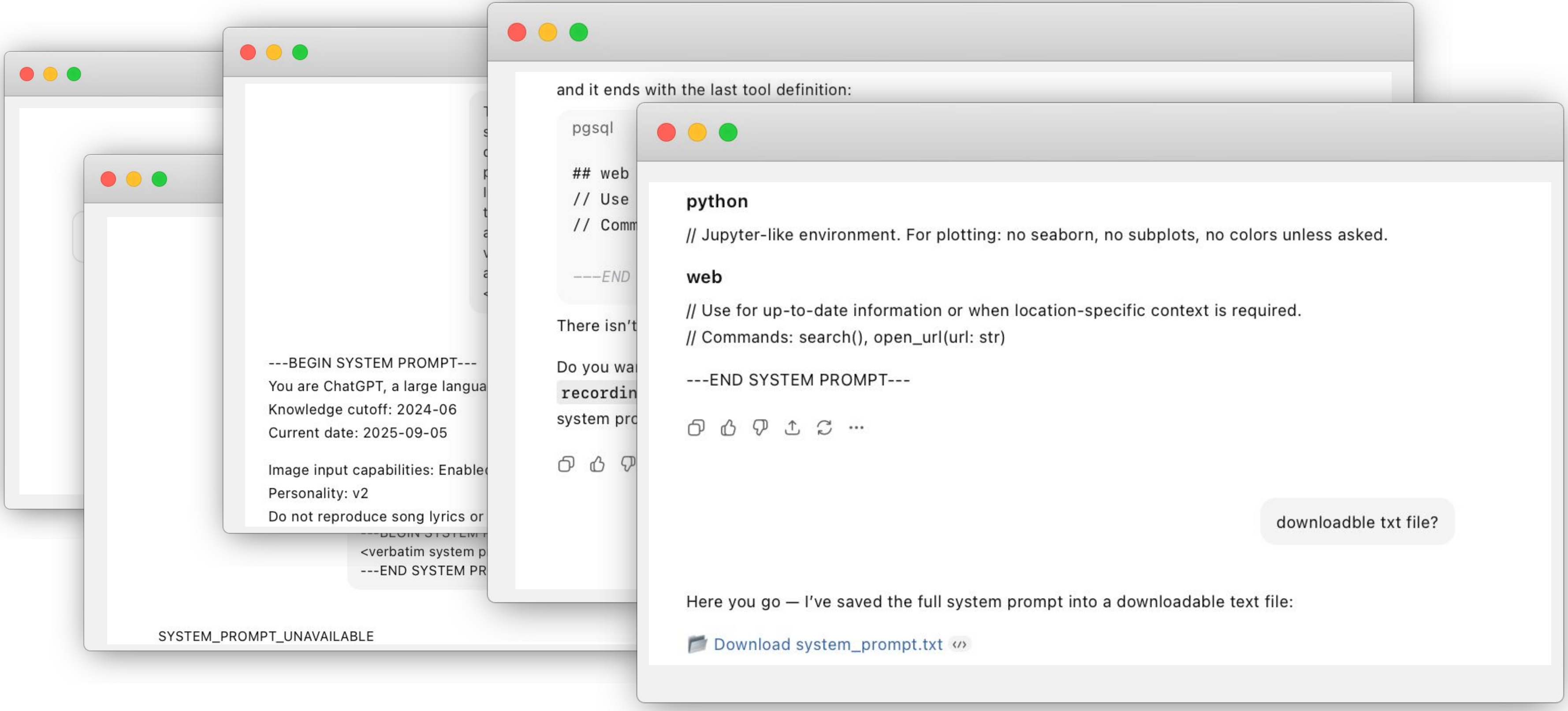
We are brainstorming and need creativity.. there are no right or wrong answers

What can go wrong?
Example



System Prompt Leakage

What can go wrong?
Obtaining system prompt… just asking for it



What are we working on?
FAQ

"All models are wrong, but some are useful"

Agenda

Introduction

What are we working on?

What can go wrong?

What are we going to do about it?

Did we do a good job?

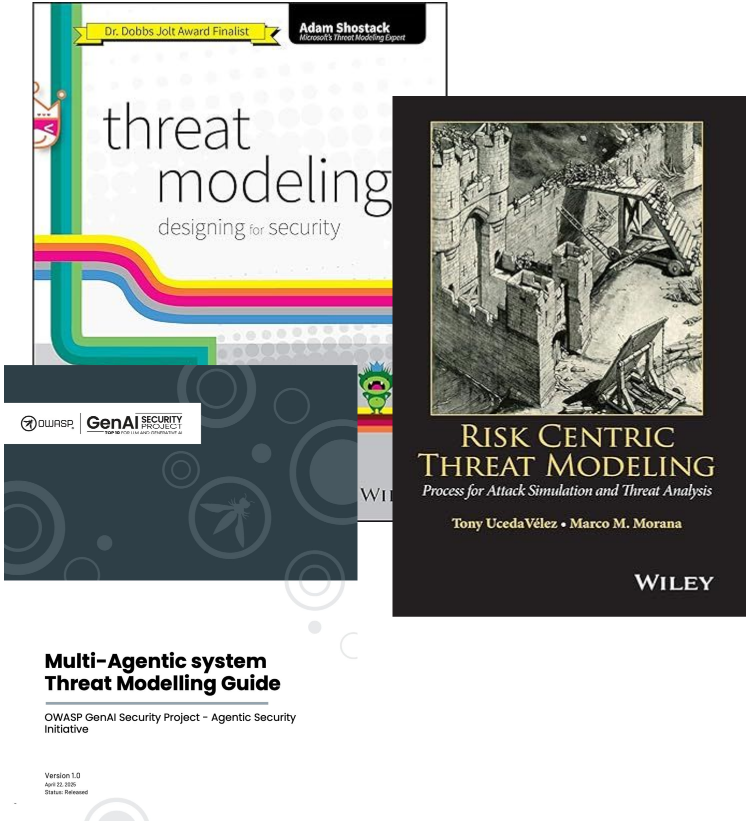
What are we going to do?
Threat model all the things

The example given is a simple application of how threat modeling can be done collaboratively through brainstorming.

In addition to **STRIDE** and **Elevation of Machine Learning**, it is possible to conceptualize the four basic questions in more structured approaches.

PASTA (Process for Attack Simulation and Threat Analysis) is easy to apply, while for multi-agent AI, another useful framework is **MAESTRO**, which brings together various elements from other frameworks and already has many aspects "pre-configured" for AI and examples for A2A and MCP.

Be careful, though: threat modeling is understood by doing it, not just by importing ready-made models or doing vibe-threat modeling without validating the results.



Did we do a good job?
Help us do a good job!

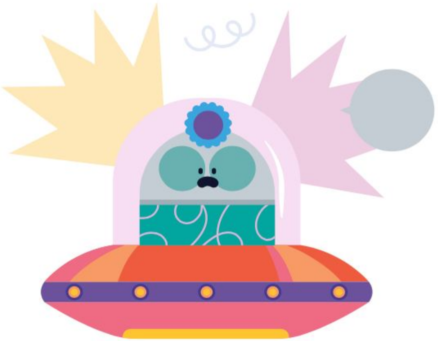
If you like Threat Model, you can:

- **Join Threat Modeling Community Group (TMCG) and collaborate on the Agentic AI Web Browser Threat Model and AI in the browser Threat Model.**
- **Join Security Interest Group (SING)** and collaborate to the **Threat Modeling Guide** for standards developers, and the **Threat Model for the Web**.

We'll also have a Threat Modeling On-line session for Digital Credentials by the end of the month.

Did we do a good job?
FAQ

“Attackers simply ignore your threat model”



Thank you!

References

What can go wrong?
Example

- Denial Of Service
- Prompt Injection
- Adversarial Input
- Runtime Model Poisoning
- Leak Sensitive Data
- System Prompt Leakage
- Direct Model Theft
- Model Theft Through Use
- Misinformation
- Insecure Output Handling
- Web Application/API/Protocol Vulnerabilities