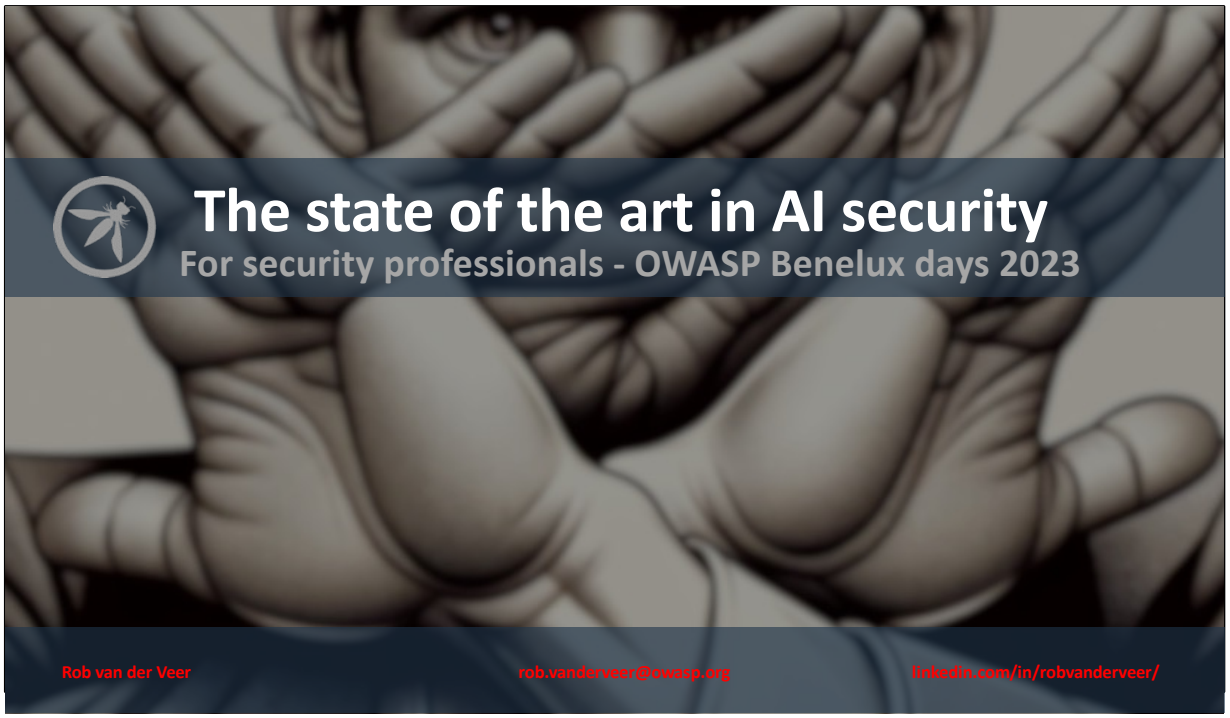



Let's talk about AI.
Like if we are not talking about AI enough, right? Even my mother is talking about AI.
So let's focus this talk on AI for YOU. For YOU as a security professional. What does it mean? What is the latest greatest?
Let's talk about AI for security professionals.
By raise of hands, how many of you are in an organization that is creating systems using AI models?
(one third raise their hand)
Next year this is going to be 75% or maybe a 100%. That's the time we are living in.
So…


31 years ago. In that time you really shouldn't mention the term Artificial Intelligence to clients: it wouldn't sell. Nobody understood what I was doing.
Just to illustrate to you: in that time: connecting computers was not common. We walked around with floppy disks in the office. There were no self driving cars, face recognizing phones, chatgpts. AI wasn't at all pervasive like it is today. There were just a few models, for example in marketing, selecting the best group of consumers to send a letter to,,,,, remember letters?

And over the years AI got better and better, because of Moore's law, because the democratizing of tools with open source, and because of innovations -
most notably the inception of transformer networks, marked by the paper 'Attention is all you need' in 2017. Transformers are neural network at the heart of generative AI….

Now, 31 years later, everybody is using AI. It has become much more accessible and much more capable in all those years. It used to be better than people in a few things, but now it can even write better than us. With AI we are more effective and efficient humans, and its effect is quite profound.

To illustrate this: have you seen the latest product launch of OpenAI? It featured a video with a woman who used ChatGPT to find the right words to express her love to her father.  Have you seen this?
I think it's a good illustration of the profound effect on us as humans.

The talk of the town is the question what this all means for our future, and the future of the world…


Are we safe?
Back in the day I was also building AI models for the police. To predict criminal careers, and to optimize police work. Back then it really helped the police. Today, this is actually ethically frowned upon, and it will even become illegal with the upcoming EU AI Act. Predictive policing vendors will be out of business. Well, at least in Europe. So first technology arrives, and then we as humanity learn how it can be harmful, and how we should protect ourselves against it.
Let's go through the key things to protect.

Privacy.
Back in the 90s when we built our police models we collected data from various police systems into a datawarehouse, combining all types of data: weather data, sensus data. People loved it. Today, this would have been illegal. By now, we have more rules for privacy, regarding what you can use data for. We've become more responsible.

Purpose –See the Clearview case: a system that used people's photos on the internet to identify indiviuals, which was forbidden because these photes were collected for a different purpose.

Transparancy – See the Dutch Syri case, where the court ruled this fraud-predicting system was illegal because the makers did not want to disclose what data the model used.
Equality – See the Amazon job application case, where an AI model was not looking at gender directly, but picked up on resume words that male applicants use more than female applicants, and copied the bias that was present in the example set of hired employees.
Accuracy & Safety: self-driving cars? Are robots accurate enough to carry weapons?
Security: how are we going to protect these models and the underlying data, in order to protect ourselves?

This is a lot. **So where do you come in here as a security professional?** All these sings concern you as human, right? But to what extent is this your professional concern? Is all of this your responsibility? Do you have to master everything.? Where should you focus on in your work?

> I would like to help you with that. With this talk I want to make your life easier as a security professional.
> There are two buckets of AI things for you. One is: Mastery, the other is Awareness. Awareness means you need to know that is there, stay alert about it, and maybe warn the people who call the shots: for example the executive with the new riskful business idea, or the over enthusiastic data scientist. Maybe the ignorant product owner. Try to help, but make sure you at least take care of your Mastery bucket.

I've marked privacy NOT with a full circle. It is half your direct responsibility: where it concerns the security of personal data: protecting it, minimizing it, anonymizing it. The other half is about the human rights mentioned. If you want to know more, see the OWASP AI security and privacy guide at owaspai.org.
Security is also marked with half a circle. This may surprise you. But part of AI security is not your task. It IS a concern. But there is a part of AI security for data scientists. I'll show you later.

In order to protect human righs, we already have privacy regulations. AI regulation and standardization are very similar, in the sense that they are meant to guide us in protecting the world against AI. It creates clarity, allowing people to think better before they start, and making it easier to go to court and make fair and consistent decisions whether something is allowed or not. Such rules are important for AI.

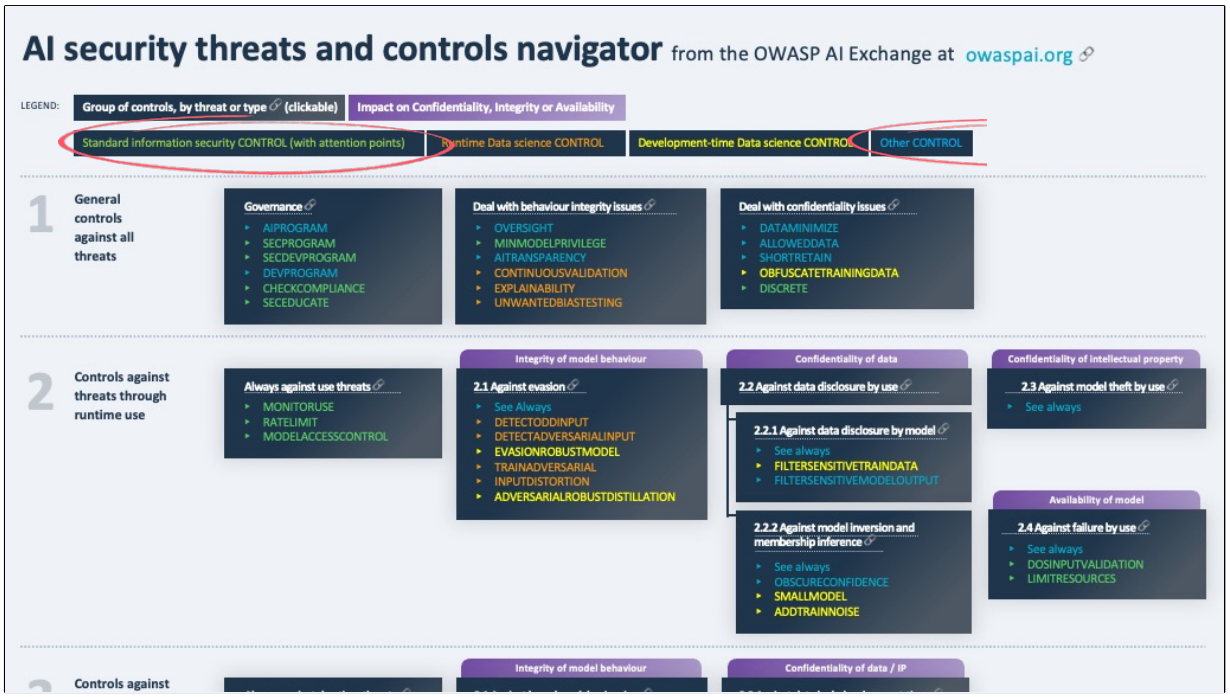

So lets look at what developments there are with rules on AI.
First of all there's the **ISO 5338**. The global standard on AI lifecycle. It sets the standard on how to engineer AI systems.
I was the lead author and worked with a team on it for the last 3 years. Dealing with about 600 comments from all over the world.
In November 2023 all the ISO countries of the world approved of the final version and it was released in December.
I am super happy and proud of that. We'll have a brief look later on.

Then, there's the upcoming **European AI Act**. People expect it to be just as trendsetting as the European privacy regulation has been. The GDPR.
The AI act identifies application areas that are not allowed, and situations in which an AI system is high risk. And for those high risk situations it sets several requirements, like for example conformance to AI security standards.
AI security standards that are not there yet.
So I am part of **CEN/CENELEC Working group 5** and THAT is our assignment: make sure that there are standards to help build and verify secure AI.
For that I'm also in the working group for **ISO 27090** on AI security and **27091** on AI privacy. These standards are just starting and are expected to be ready 2025+.

And there are many organizations working on AI security frameworks, like MITRE, NIST, ENISA, ETSI, Microsoft, IEEE. But also OWASP with the LLM top 10 and the ML top 10.
Plus I started an initiative a year ago, called the OWASP AI security and privacy guide. I presented on it in Dublin at the Global appsec conference.

There's a lot, and at the same time this is not an easy topic. It's multidisciplinary and there are just a few experts in the world on the combination of things. Every framework is incomplete, sometimes flawed and they're not aligning with eachother. They use different terms and sometimes different paradigms.
So I made a decision: I wanted to create a platform to exchange knowledge and for any expert in the world to contribute. I created the **OWASP AI Exchange**. It's an open source document that collects AI security threats and controls. Let's look at it.

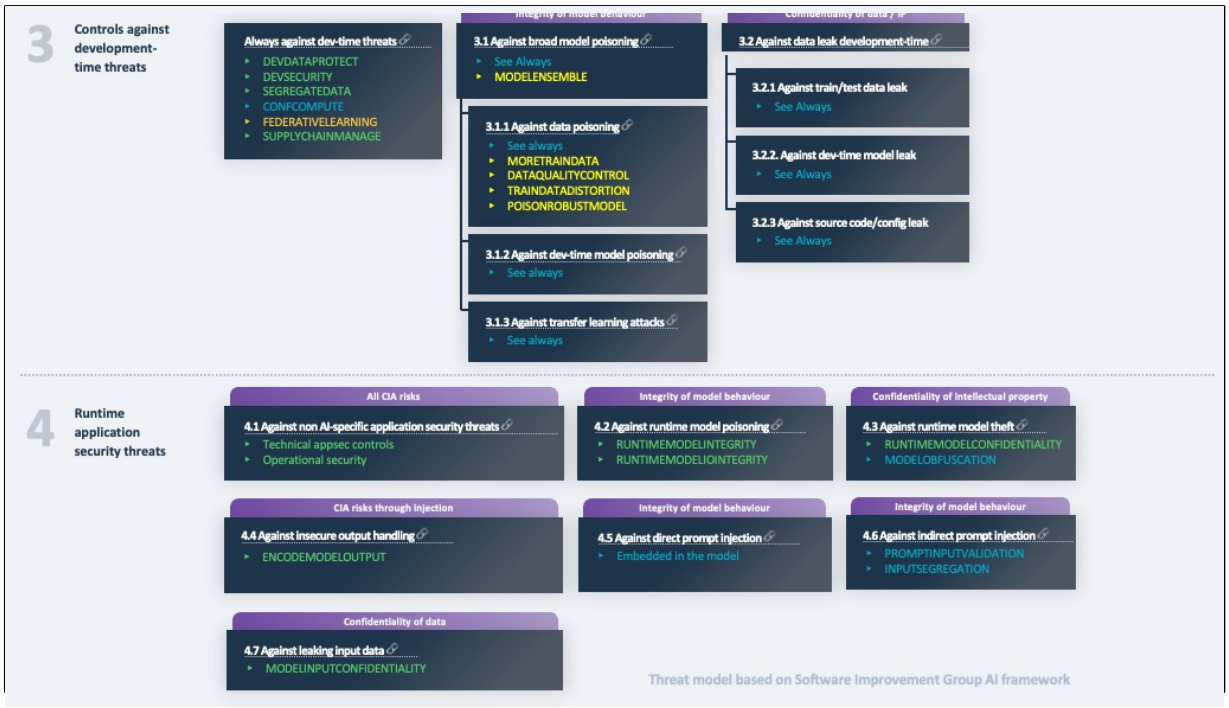

Here it is. The OWASP AI Exchange. You can find it at owaspai,org. It's a markdown document in an OWASP repository at Github.
You can see the document zoomed out on the right hand side. It's 38 pages, and counting.
On the left you see what we call the Navigator: it is one overview of all the threats, risks, controls, and who should take care of them.
And when you click on a section it takes you to the corresponding part of the AI Exchange document.
An increasing number of experts is contributing to it. A professor from Spain, a machine learning researcher from the US, a woman from Iran, now working for Accenture in Germany. The list goes on and on.
We've assigned it a CC zero license, which means that anybody can use its content without attribution.
I believe that only by being truly open and altruistic, we can achieve the knowledge sharing, the consistency, and the tempo that is so critical right now.
And the AI exchange is actually being used as input to the standardization efforts that I mentioned.
I love it when a plan comes together.



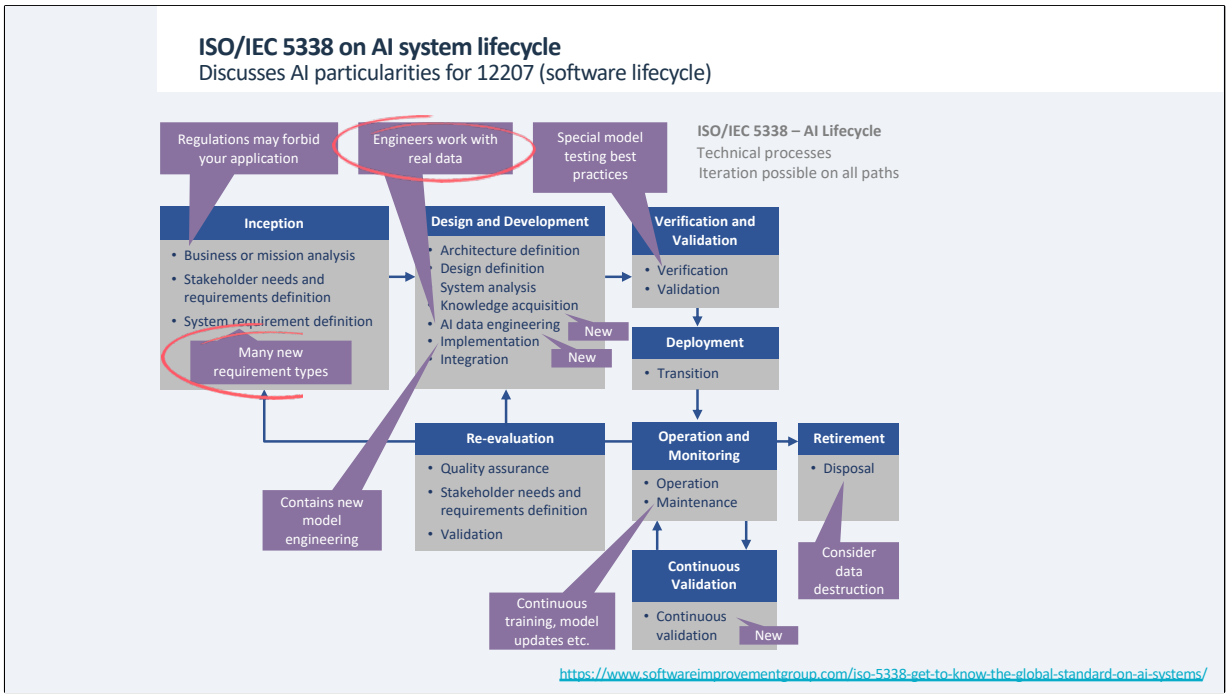

So Please have a look at the Exchange at your own leisure. We don't have the time right now to go into it deeply,
But let me highlight two things for you.

First of all: there are a couple of things that you always need to take care of for AI security. One of them is proper governance.
You need to have an AI program that keeps an inventory of initiatives, using for example ISO/IEC 42001.
You need to involve the work of AI engineers into your security program, and your secure software development program.
And educate these people in information security. I get to that later

Second: the controls are mentioned here in capitals. As you can see they have different colors.
The orange and the yellow ones are for data scientists. They are not your concern. I'll show you later why not.
Your concern are the green controls: they are information security controls. And to a large part the blue ones, the OTHER controls. They are also important for you.

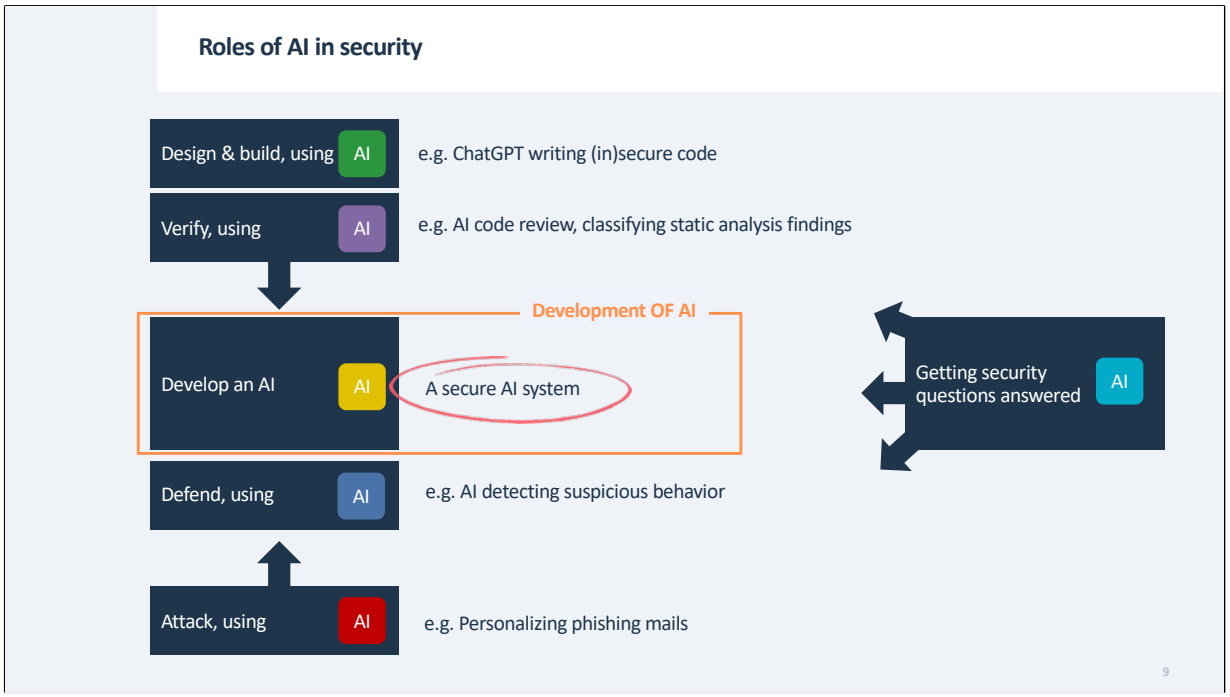



And here's the bottom part of the navigator.
Before we go furher into this. Let's look at the 5338.




ISO/IEC 5338 is built on the AI framework that we developed at Software Improvement Group.
The great thing about it is that it doesn't introduce a completely new lifecycle.
It builds on the existing software lifecycle.
Organizations already have an understanding of how to create software. And processses for it in place.
And AI systems are software systems.
There are just a few processes that are new. And AI systems have a number of particularities. Those we discuss in 5338.

Talking about lifecycle, let's look at the role that AI can play in it.

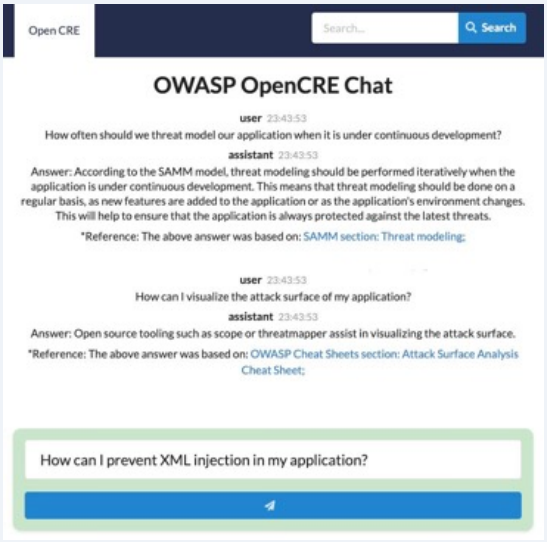



Sidestep on generating code: It hopefully speaks for itself, But please be careful when writing code using LLM such as ChatGPT or Copilot. You write a prompt, but it is not the prompt that you will be maintaining. That will be the large blob of code that you generated. You need to check it because even if it does exactly what you want it to, it may need to be improved first to make it more maintainable, more scalable or more secure.  You can't just turn your head. You OWN that code now. And here's the thing: it requires more skill to review unfamiliar code and find all the flaws then to write the code yourself. It will be quicker to let the AI write it, but if you don't have the review skills, you are being very productive at creating technical debt, which is the same as creating programmer jobs. Handle with care.

Let's look at 'Getting security questions answered with AI' by diving into OpenCRE-Chat

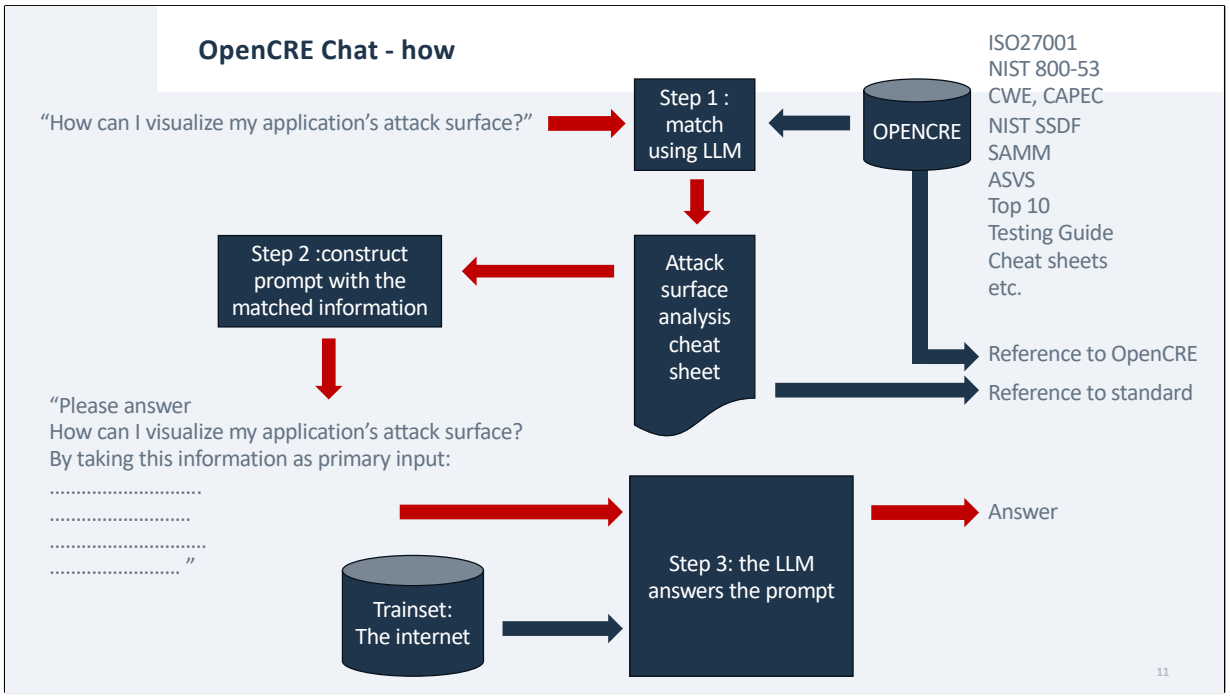
## OpenCRE.org Chat

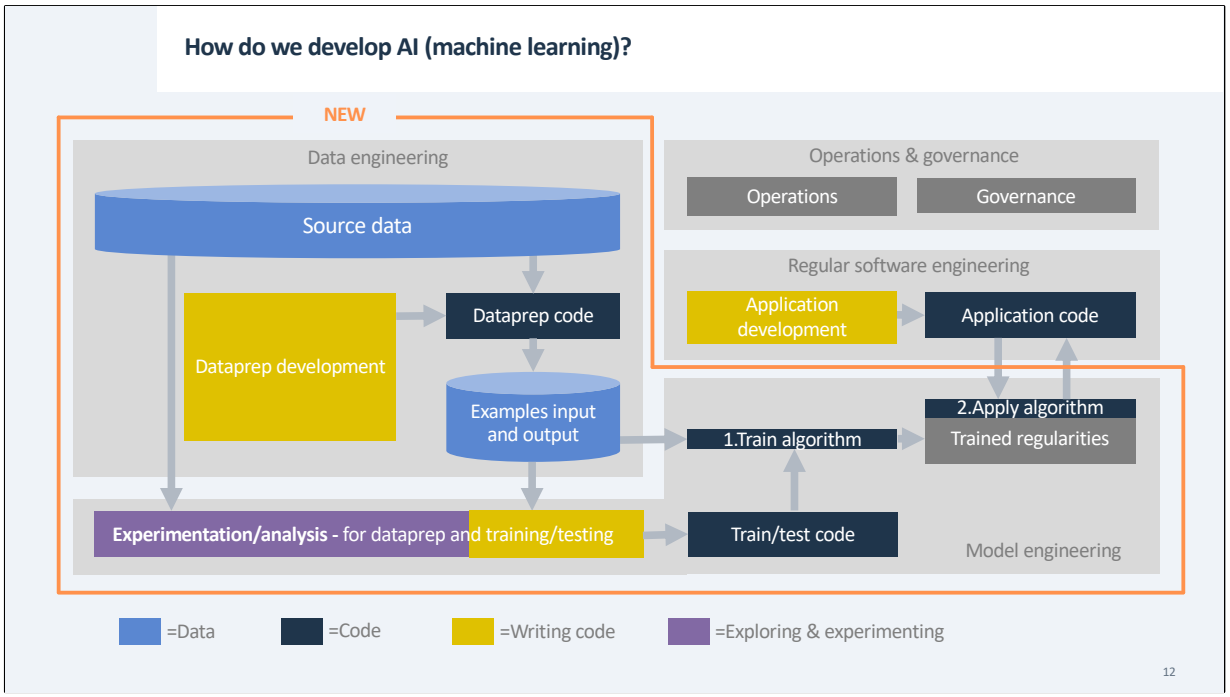
https://opencre.org/chatbot




10

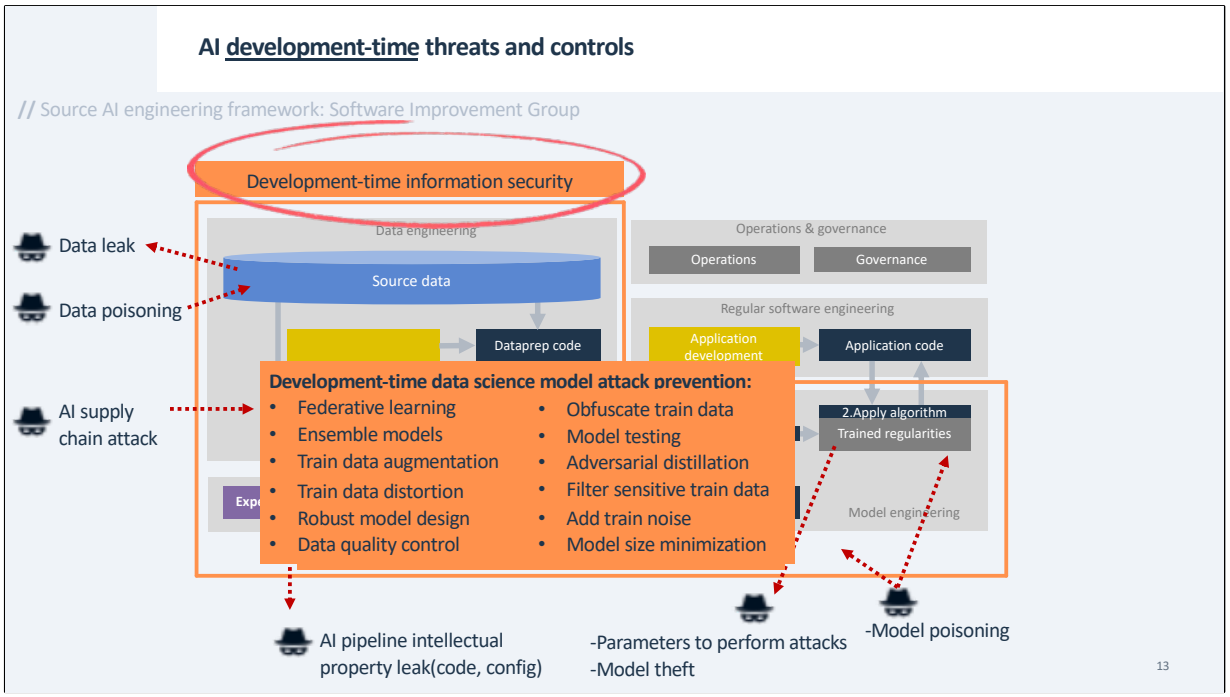
I started OpenCRE about 4 years ago with Spyros Gasteratos, and it's a platform to unlock all security standards, and integrate them into one source. Just this week we launched a new feature called Map analysis to automatically map two standards to eachother. Check it out.
To learn more, please have a look at opencre.org.
For now I want to focus on its Chat feature.
This was an idea from Sherif Mansour. Because we collect so many security standards, we actually have a knowledge base that we can put a large language model on it, from Google, to answer security questions. And that's what we built. The world's first and only security chatbot.
It's pretty cool, you ask it a security question and it answers you while primarily using the security standards, which is the most reliable information available. And it also provides you the link to the right section in the right standard.
The first response we got was from Japan. People were really happy that they could ask questions and get answers in their own language. So they get Japanese answers, with sample code that also has Japanes in the comments. How unbelievably cool is that?


OpenCRE Chat - how
"How can I visualize my application's attack surface?"
Step 1 : match using LLM
OPENCRE
ISO27001
NIST 800-53
CWE, CAPEC
NIST SSDF
SAMM
ASVS
Top 10
Testing Guide
Cheat sheets
etc.
Step 2 :construct prompt with the matched information
Attack surface analysis cheat sheet
Reference to OpenCRE
Reference to standard
"Please answer
How can I visualize my application's attack surface?
By taking this information as primary input:
.....
.....
.....
....."
Trainset: The internet
Step 3: the LLM answers the prompt
Answer
11




This whole development-time part with its orange boundary is complete new attack surface. Data engineering and model engineering.
By the way, in the recent Biden executive order to help assure AI trustworthiness, this entire attack surface is a blind spot.

Let's see what security threats there are to AI development-time




Training data can leak, or it can be poisoned to change the model behaviour.
AI systems have a complex supply chain, with new frameworks and sometimes with models and train data that are obtained from elsewhere.
Is your AI model valuable intellectual property? Better protect your data, code, configuration and model parameters.
These model parameters also deserve protection because if an attacker changes them, the model behaves differently.

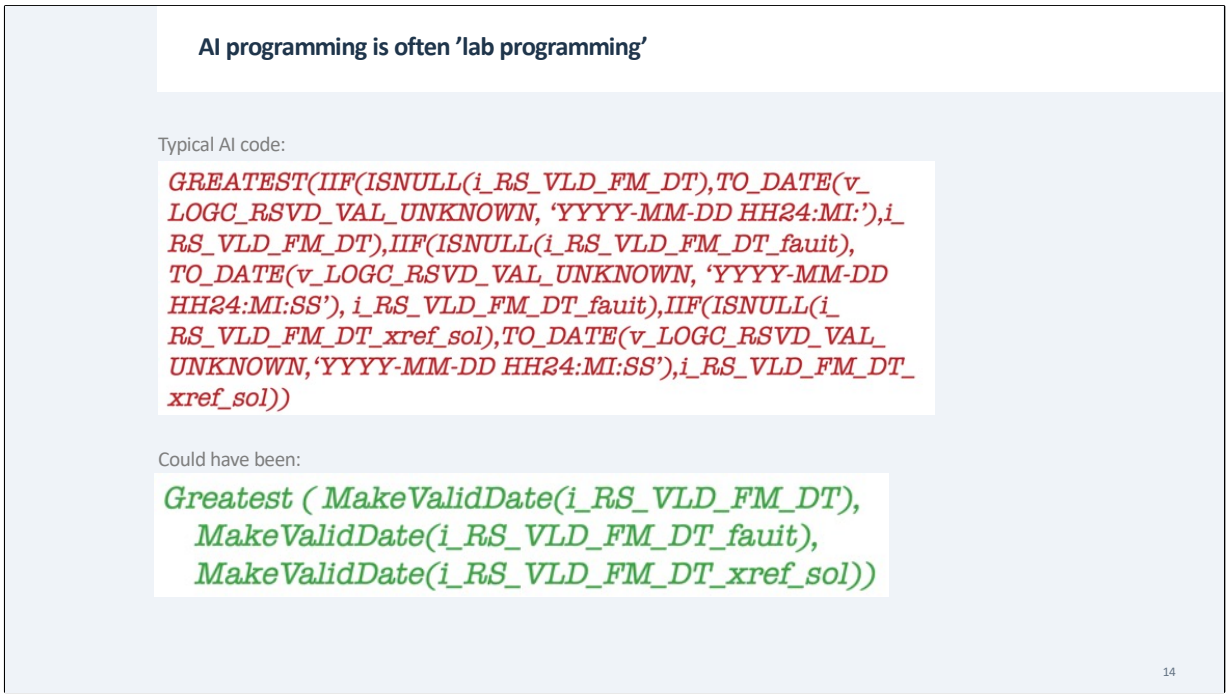
So what are the main controls against these threats (the orange blocks)? There are two categories:

First: Development-time information security of this new attack surface: protecting that data, those model parameters, code, have access control, and supply chain management
Second: Development-time data science model attack prevention. For example adding noise to training data. This is mostly very advanced AI terrain.

And what is your main concern: The information security. Why not the data science stuff? I mean, look at it.

If you’re a security professional and you master all that datascience stuff. Congratulations, you belong to the 50 people in the world that can do both security and datascience in depth. For normal people: let’s stick to what we know and have affinity for.

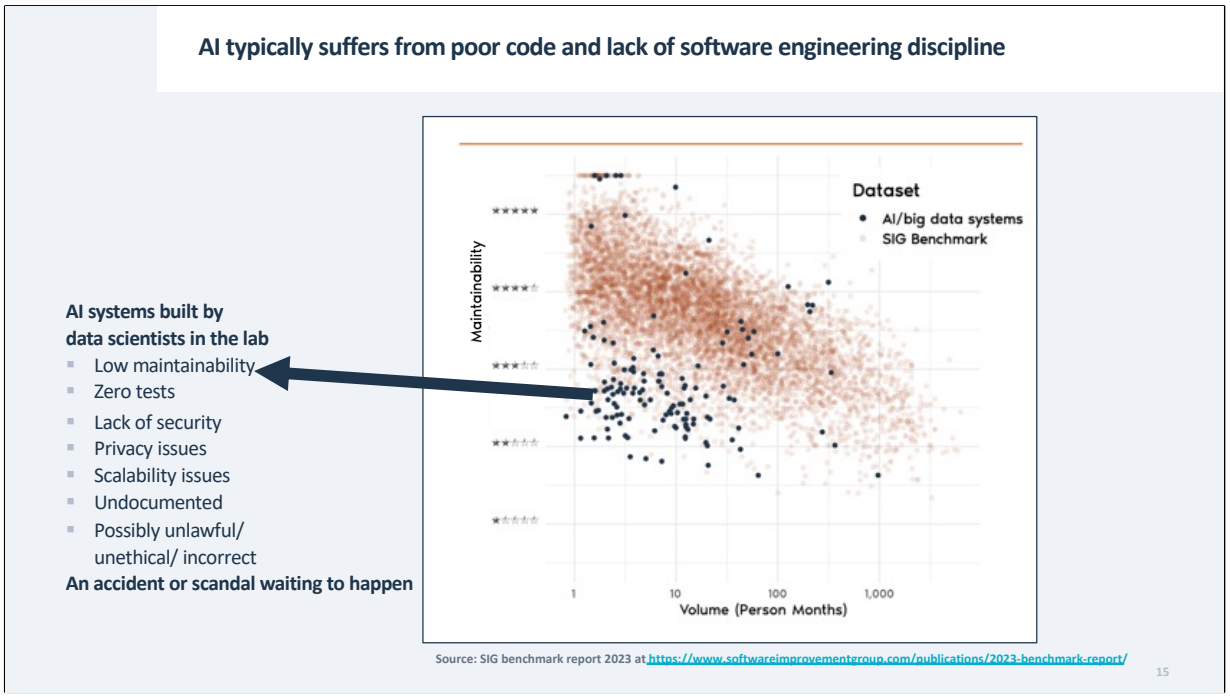
So the data science world is a bit of a different world. It is like they are not bothered too much by security. And it’s also like they’re not bothered too much by software engineering best practices.
Because, this is what we see a lot:




This is typical datascience code that we encounter at clients.

It's like the normal rules of software engineer don't apply to data scientists
It lacks abstraction, readability, testing, and reuse

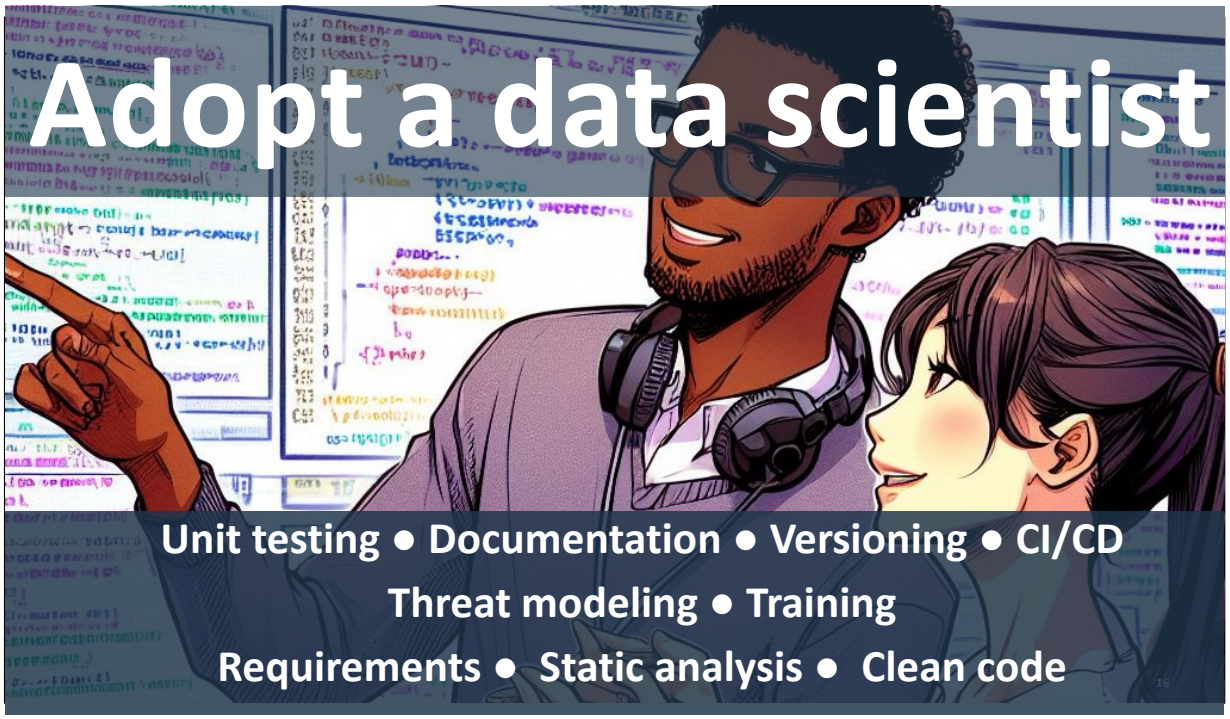
We did some research into this phenomenen and this is the result:




Now, why is this?
Data scientist are very focused on creating working models, not on creating future-proof software.
These systems are developed in the lab by data scientist and then when the model finally works, it needs to go live.
And then you run into trouble.
Because then it has to conform to the rules of the real world: it's going to be attacked, it will have to scale, it will have to be maintained, and changed.
But it will have no tests, people will not understand it or be able to make a change without breaking it.

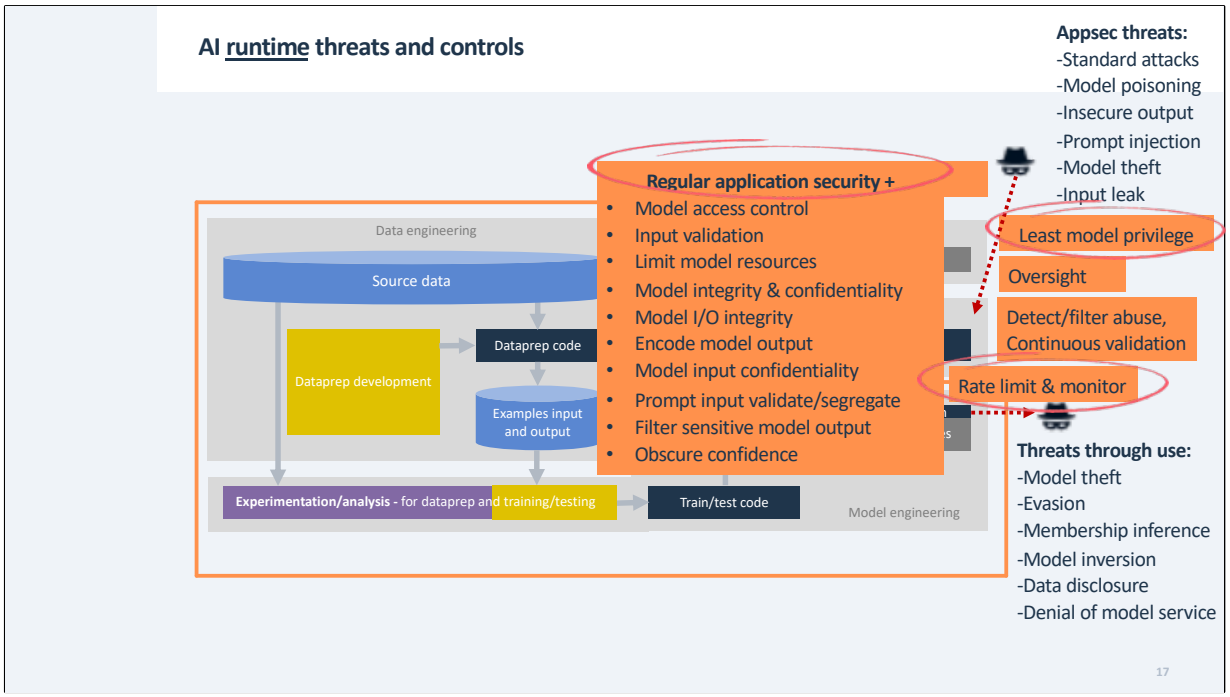
It's better to get AI engineering first time right.

Which is why I recommend:




Adopt a data scientist.
Make sure that data scientists are going to apply modern software engineering and security practices.
They will adopt you back, so you can learn about the awesome data science world.
In general we recommend mixed teams, so that these roles can learn from each other.

So we talked about development-time issues with software engineering and security.
How about runtime?




There are two types of run-time threats:

1. Application security threats, with the standard security attacks that harm the confidentiality and integrity of data, and may also be aimed at manipulating model behaviour.
2. Threats through normal use of the model: providing inputs to mislead the model, reconstruct training data, copy the model, or bring it to its kees.

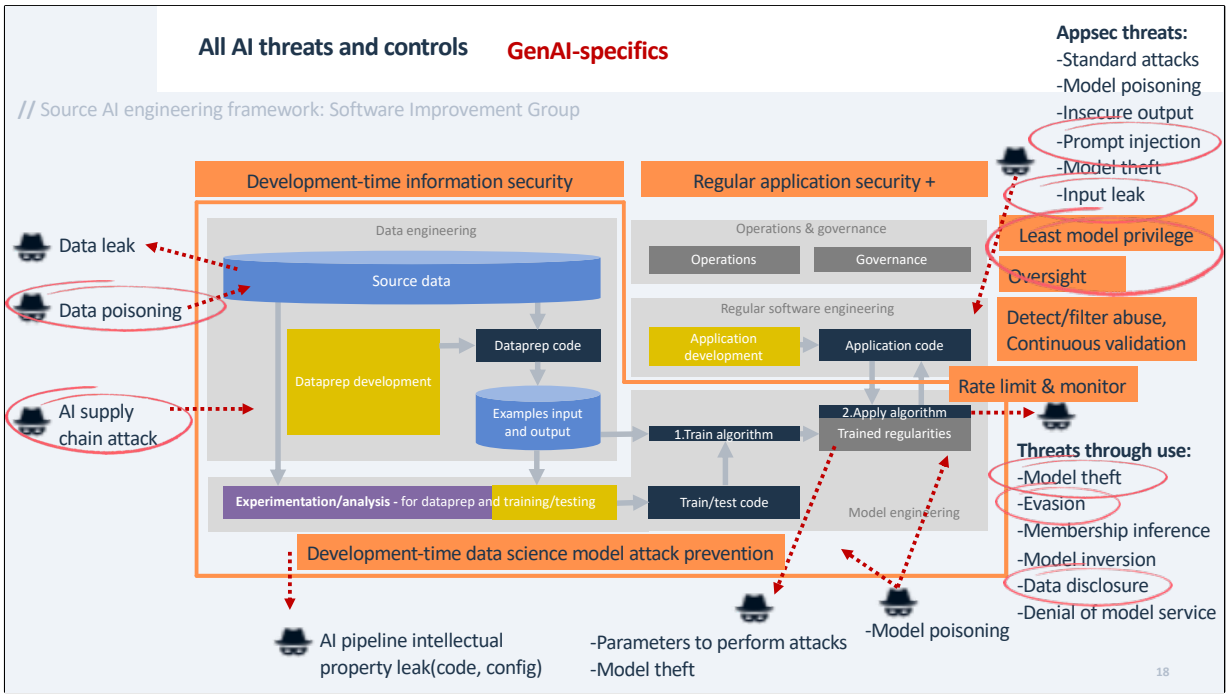
Regarding the controls against these threats (orange blocks):

- Least model privilege/ oversight/: It's important to see the model as a potential attacker. AI is unpredictable by itself AND it can be hacked. These are two good reasons to restrict AI and watch over it. AI models are meant to approximate an answer. Let's take for example OpenCRE Chat. We asked it to explain what Cross site request frobbing is. Which is an attack that doesn't exist at all, and it came up with a complete made up story including example code.
- Detect/filter abuse and continuous validation: these are statistical countermeasures during runtime, and the territory of datascientists
- Rate limiting & monitoring are important to limit and detect threats through use
- Regular application security +: runtime application security of an AI system is just like normal application security, with a few attention points, such as prompt

injection for Large Language Models.

See owaspai.org for more details

Now let’s put both development-time and runtime in one picture to see the complete overview of threats and controls:




And I'm going to use it to go through what is specific about Generative AI security.
It's important to realize that GenAI is AI and your dealing with the same basic issues.
AI in general and GenAI are not wildly different from a security perspecitve. Really
But some risks are lower, some are higher, and there are a couple of new things.

- Evasion attacks for GenAI include specifically evasion of policies that intend to censor (e.g. violent) output
- Unwanted output of sensitive training data is an AI-broad issue, but especially a high risk with systems that output rich content such as GenAI
- Training data poisoning is an AI-broad problem, and with GenAI the risk is generally higher since training data can be supplied from different sources that may be challenging to control, such as the internet
- Overreliance is an AI-broad issue, and in addition Large Language Models can make matters worse by coming across very confident and knowledgeable

- GenAI models mostly live in the cloud - often managed by an external party, which increases the risk of leaking training data and leaking prompts. This issue is not limited to GenAI. Additional risks that are typical for GenAI are: 1) model use involves user interaction through prompts, adding user data and corresponding privacy issues, and 2) GenAI model input (prompts) can contain rich context information with sensitive data (e.g. company secrets).
- Pre-trained models are applied also outside of GenAI, but the approach is quite common in GenAI, which increases the risk of transfer learning attacks
- The typical application of plug-ins in Large Language Models creates specific risks regarding the protection and privileges of these plugins - as they allow large language model to act outside of their normal conversation with the user
- Prompt injection is a GenAI specific threat, listed under Application security threats


And there you have it. The latest greatest in AI security.
I know this was a lot, but that's what you get with the state of the art.
To make it simple: your part as a security professional, your mastery bucket is in the green and blue text at the AI Exchange.
See https://github.com/OWASP/www-project-ai-security-and-privacy-guide/raw/main/assets/images/owaspaioverviewpdfv3.pdf
Have a look. Spread the word on the Exchange, good luck and Thank you for your time.