THESE NOTES WERE MADE BEFORE I HAD ACCESS TO THE SLIDES

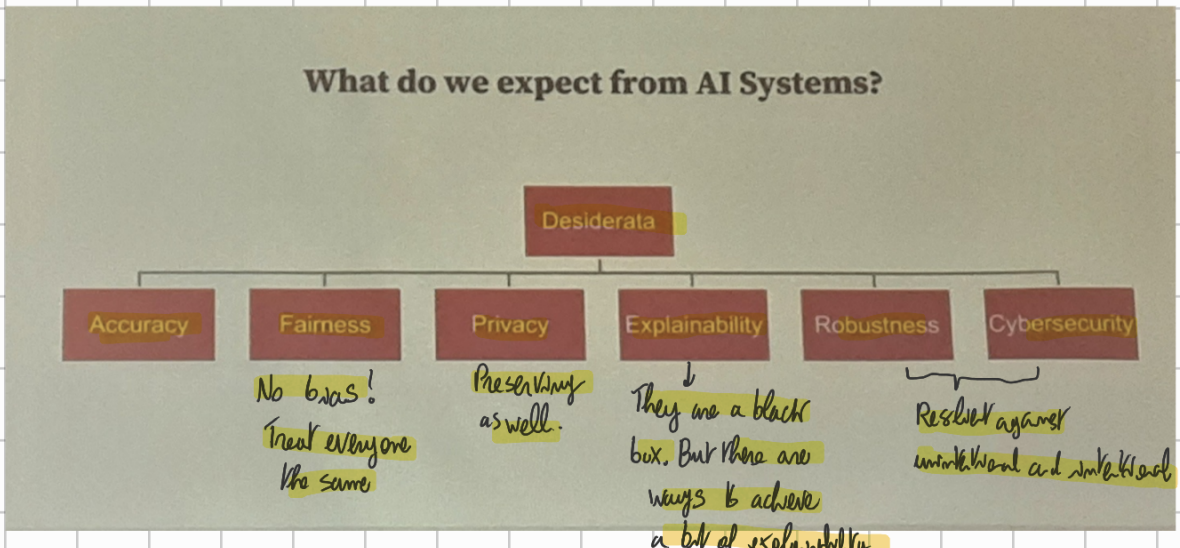
- Pics of 2 melanoma: the pics are of the same melanoma, but the AI system seems them as different. 99% benign, 99% malignant. But the pixel on the right is a few pixel different. So that's scary and dangerous.

1. EU artificial Intelligence Act 2. Robustness and Cybersecurity in the AI Act

Background: in 2019 an idea of digital strategy was set out by Von Der Lyne with plans to improve EU digital relevance.

AI Act: small brick in the wall of EU legislations.

1. What do we expect from AI Systems? They are a complex system.



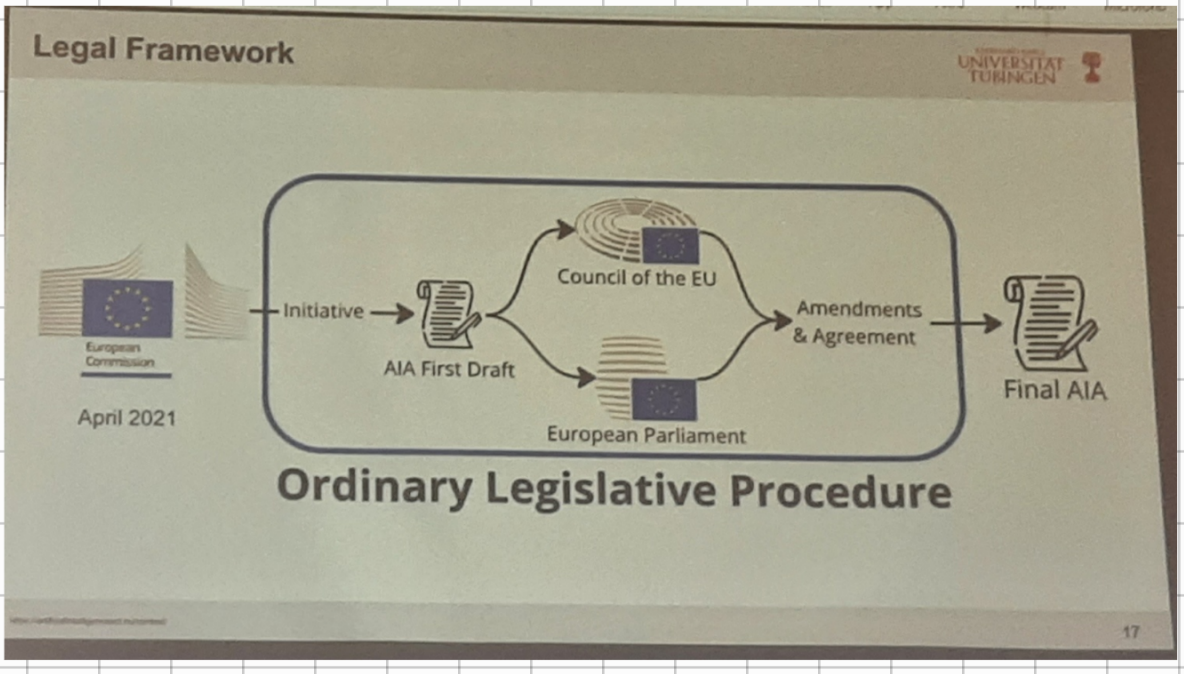
This played a key role in drafting the Act.

desiderata

Article 1(1) AI-Act:

The purpose of this Regulation is to **improve the functioning of the internal market** and promote the uptake of **human-centric and trustworthy artificial intelligence (AI),** while ensuring a high level of protection of **health, safety, fundamental rights** enshrined in the Charter, including democracy, the rule of law and environmental protection, against the harmful effects of AI systems in the Union and supporting **innovation.**

Might seem really broad, and it is actually: it tries to cover a very broad field. So the AI act is trying to achieve this desiderata.



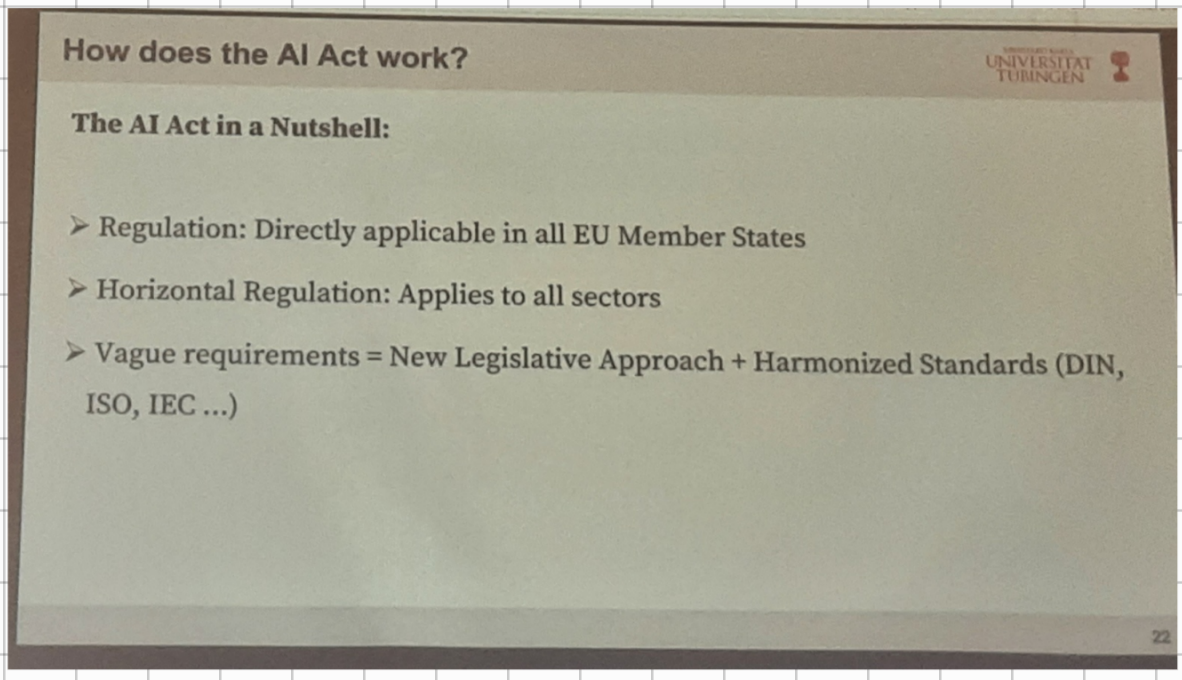
Providers have some months or years to adapt, based on what they need to do. First rules came out in April 2021 and the commission created the first draft sent to ①, ② that started talking and negotiating to achieve an agreement. They came to an agreement in December 2023.

[INSERT SLIDE WITH CHATGPT, META AI]

Commission created first draft in 2021, so they had something different than general purpose models like chatGPT. They had in mind medical devices, self driving cars etc.

So the first draft was already outdated! The draft needed to be changed accordingly.
AI act enforced in June 2024.

NOTE: this is the first AI regulation created in the world!



1. Rules need to be followed directly. 2. Not sector specific! Horizontal about applicability. 3. Built on the New Legislative Approach. Problem with digital products is that we can't keep up. So we start with vague requirements that will be specified by harmonized standards.

- What is regulated?

AI SYSTEMS (first draft was all about this) and section about GP AI Models.
↓
Def: highly criticised...

What is an AI System?

UNIVERSITÄT TÜBINGEN

Article 3(1) AI-Act:

"An AI system is a machine-based system designed to operate with varying levels of autonomy and that may exhibit adaptiveness after deployment and that, for explicit or implicit objectives, infers, from the input it receives, how to generate outputs such as predictions, content, recommendations, or decisions that can influence physical or virtual environments."

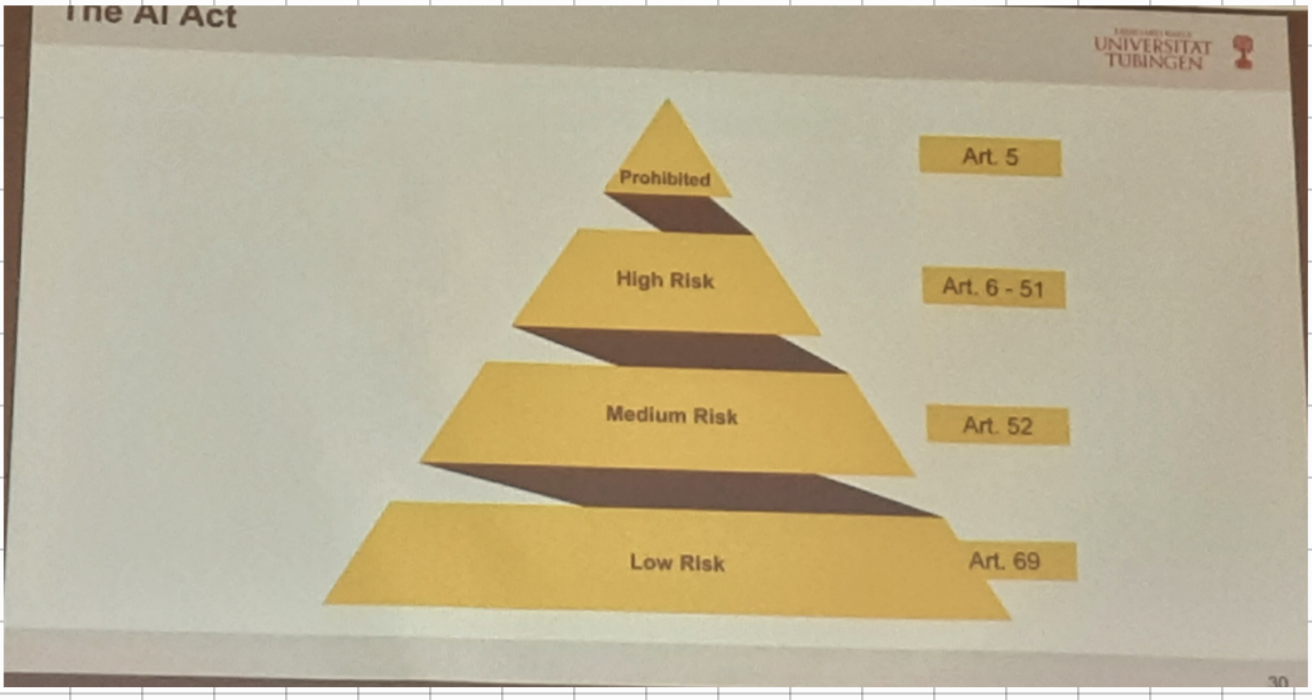
Key word: inferring knowledge: it's about gaining knowledge without a specifically defined function. They wanted to exclude current SW systems we are using. What is inference?

What is an AI System?

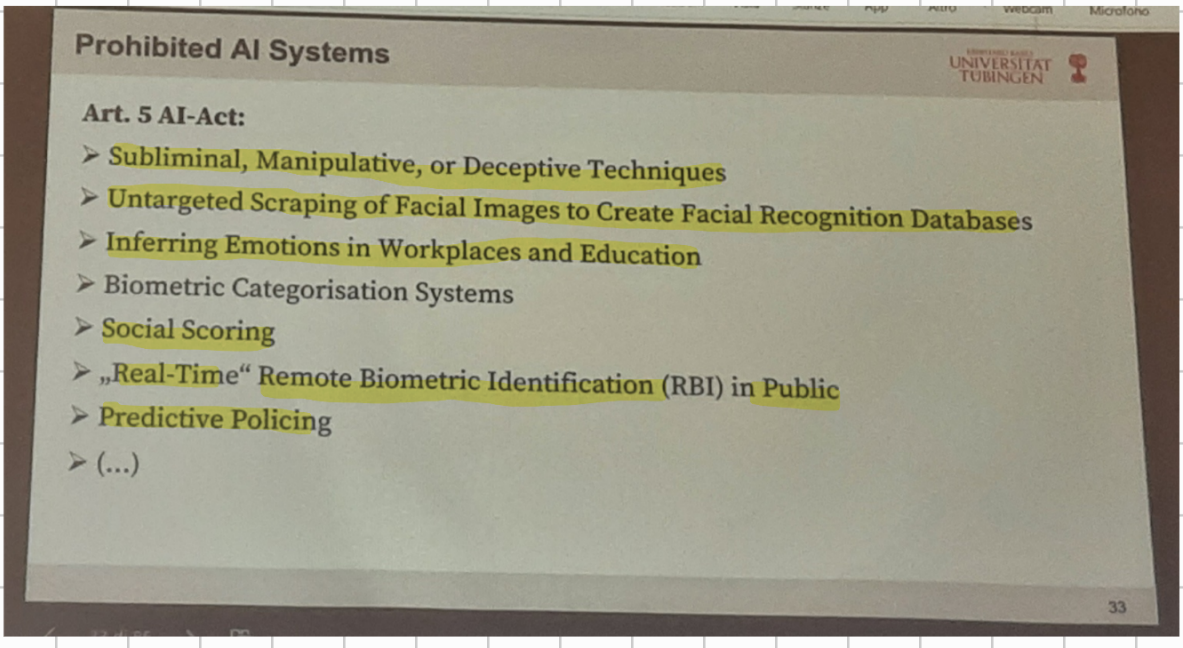
UNIVERSITÄT TÜBINGEN

Recital (12) explains 3(1) AI-Act:

"The techniques that enable inference while building an AI system include **machine learning approaches** that learn from data how to achieve certain objectives, and **logic- and knowledge-based approaches** that infer from encoded knowledge or symbolic representation of the task to be solved."



The AI systems are divided into these 4 categories. For low risks you don't need to fullfil any legal req. The AI act is actually dealing with HR systems. And keep in mind that this is general in the sector.



We do not want systems that are defined as deceptive (like AI algorithms for FB, IG...) exc.

High-Risk AI Systems

UNIVERSITÄT TÜBINGEN

Art. 6 AI-Act:

They are systems that:

High-Risk AI Systems pose risks to people's health, safety or fundamental rights

- List of High-Risk AI Systems in Annex II and III
- List can be modified by EU Commission (Art. 7 AI Act) (2)

This is vague, so we provide guidelines to identify them. And (2), this is very important.

High-Risk AI Systems

UNIVERSITÄT TÜBINGEN

Art. 6 AI-Act:

High-Risk AI Systems pose risks to people's health, safety or fundamental rights

- List of High-Risk AI Systems in Annex II and III
- List can be modified by EU Commission (Art. 7 AI Act)

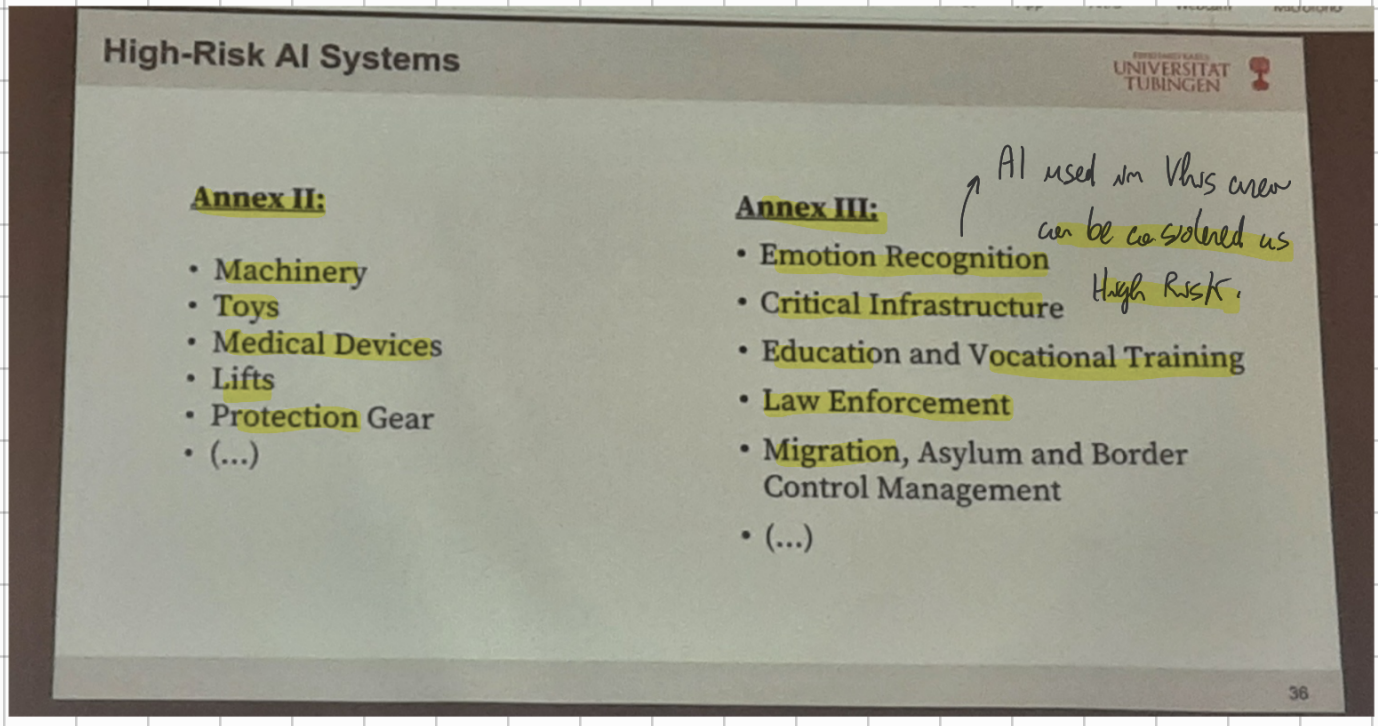
Self Conformity Assessment

Third Party Conformity Assessment

35

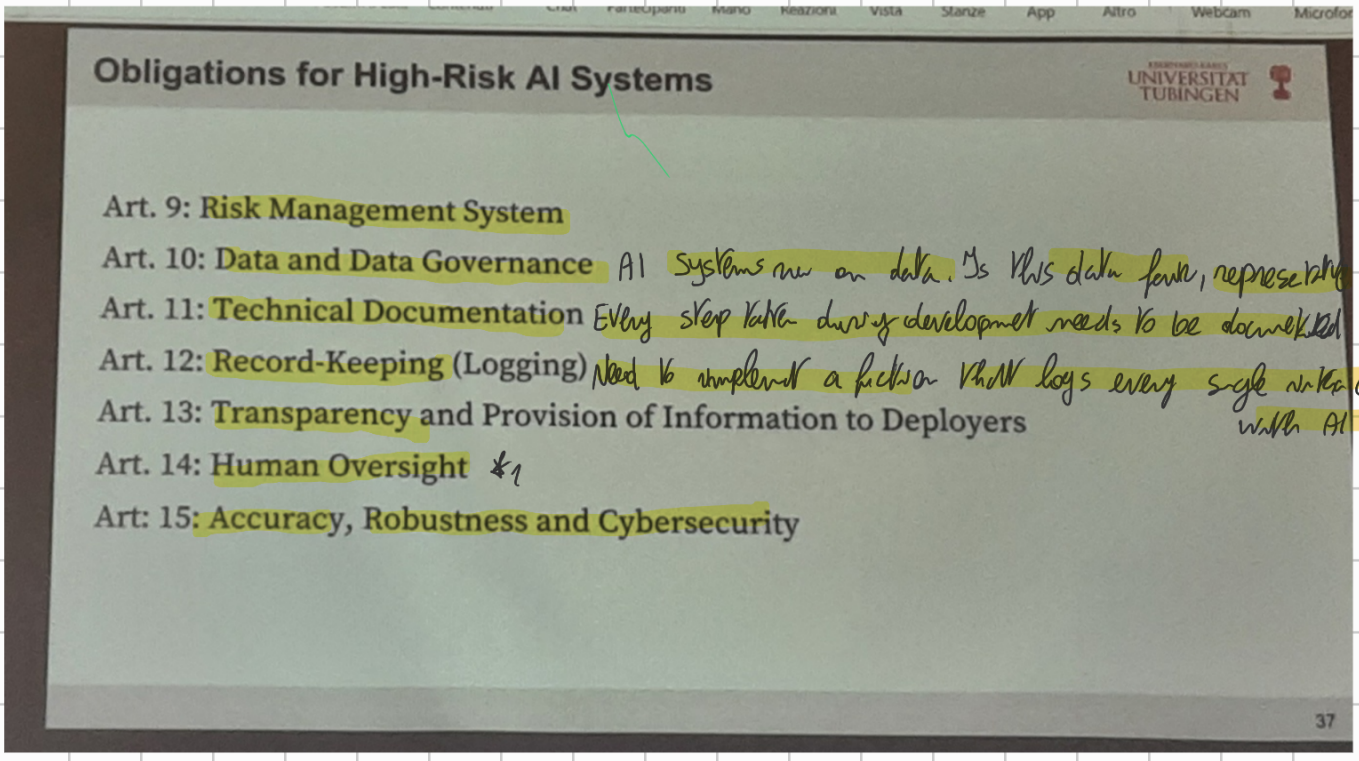
You can either carry out a self conformity assessment for the AI act requirements, or (not always mandatory) you can get the job done by a third Party Conformity Assessment.

NOTE: same structure seen in the Cyber Resilience act and in the Cybersecurity Act! These types of interventions follow the same pattern. Requirement and then evaluation.



AI used in VRUs crew can be considered as High Risk.

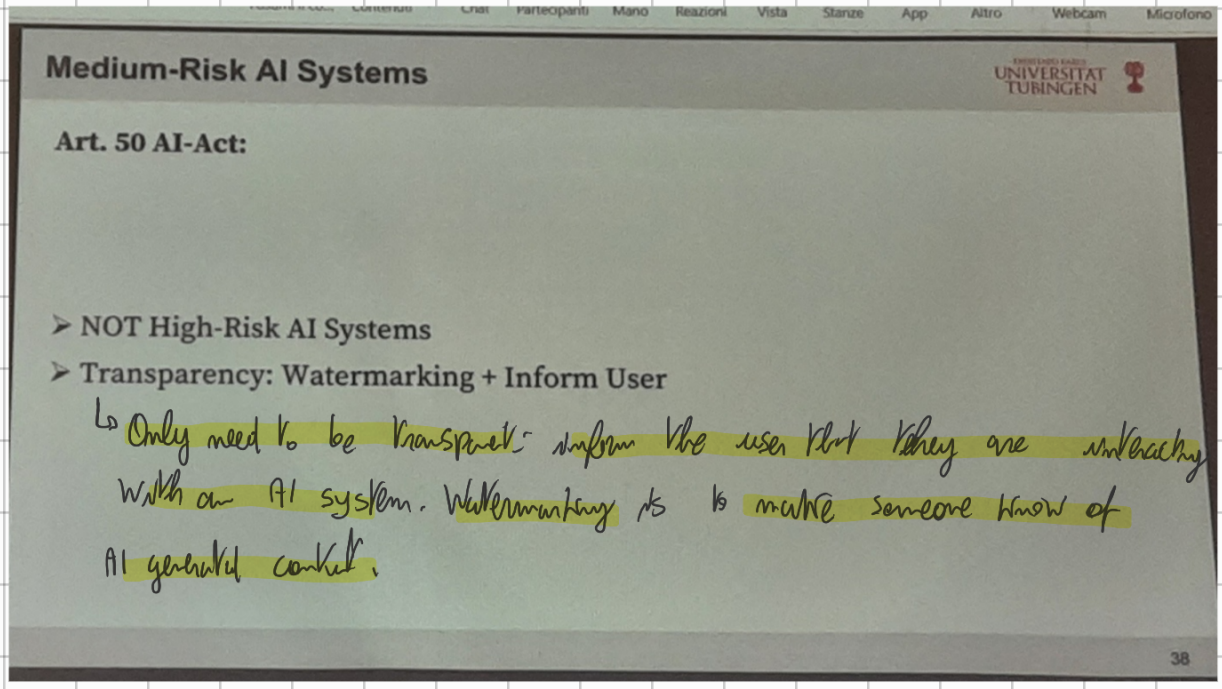
What do you actually have to do? 7 rules:



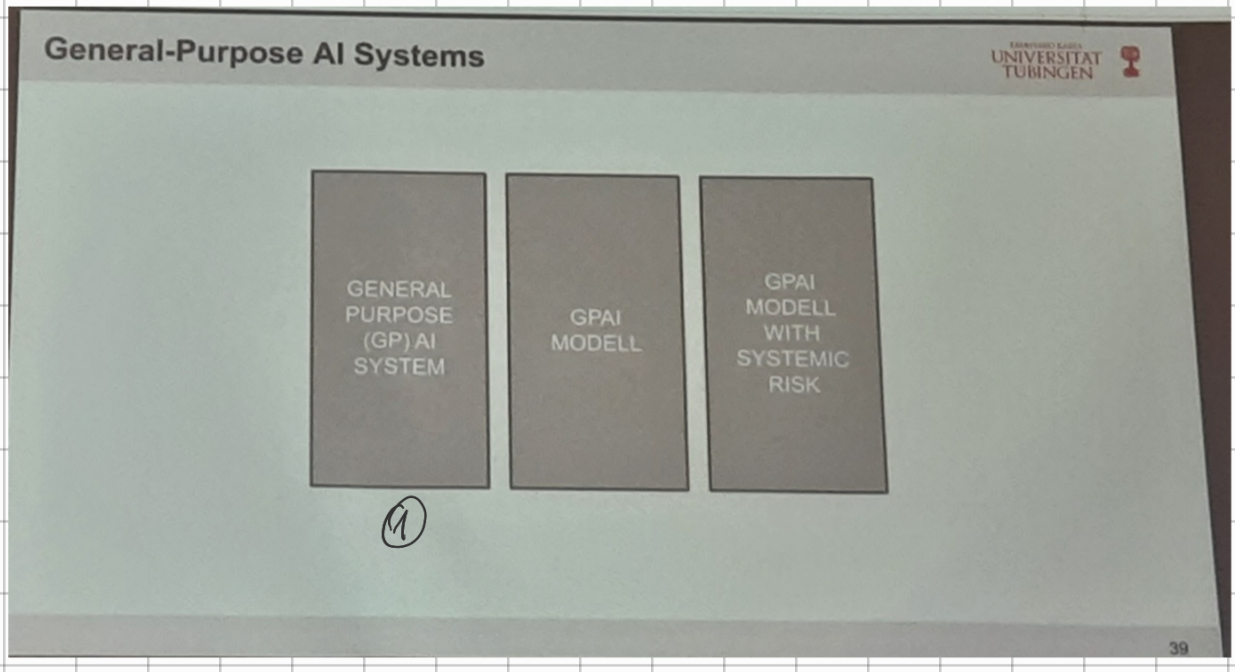
AI systems run on data. Is VRUs data fair, representative etc?
Every step taken during development needs to be documented.
Need to implement a function which logs every single interaction I have with AI system

13. Technically not easy to achieve. Also about providing info to someone that deploys the system + how everything works.

*1 Always have a human in the loop!



↳ Only need to be transparent: inform the user that they are interacting with an AI system. Watermarking is to make someone know of AI generated content.

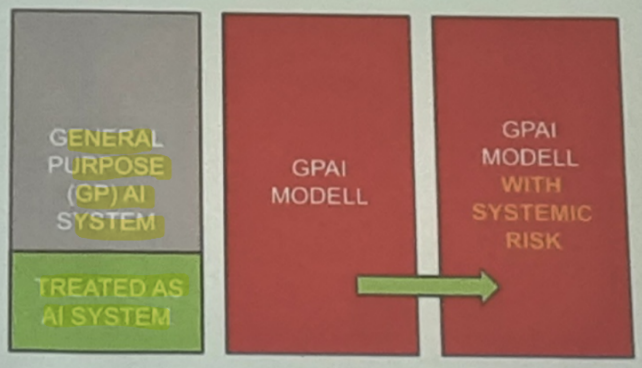


They also apply risk based approach. They applied the same risk approach

① System that can perform a big variety of tasks. System: product built around a model. A model itself is not a system.

General-Purpose AI Systems

UNIVERSITÄT TÜBINGEN



40

We have two kinds of Models: GPAI models (regular) and GPAI with systematic risk.

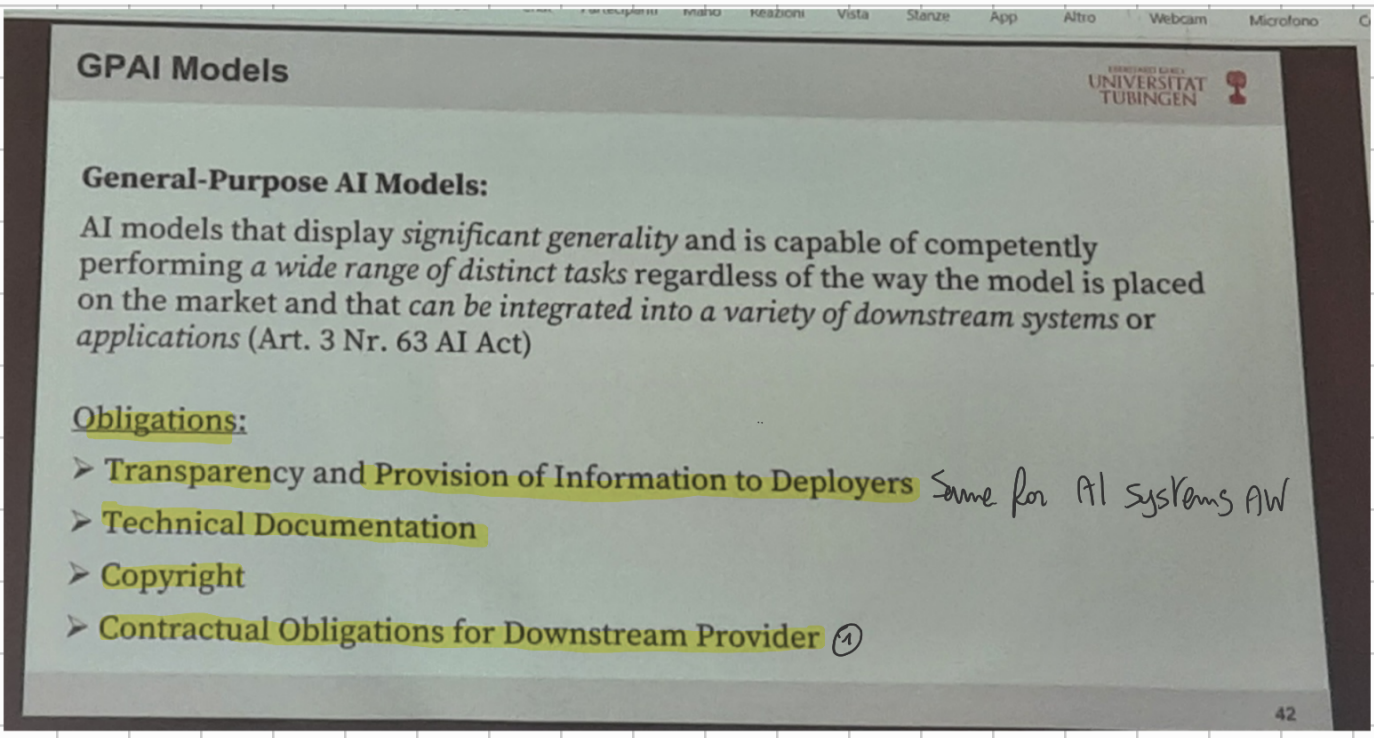
GPAI Models

UNIVERSITÄT TÜBINGEN

General-Purpose AI Models:

AI models that display *significant generality* ① and is capable of competently performing *a wide range of distinct tasks* ② regardless of the way the model is placed on the market and that *can be integrated into a variety of downstream systems* or *applications* (Art. 3 Nr. 63 AI Act) ③

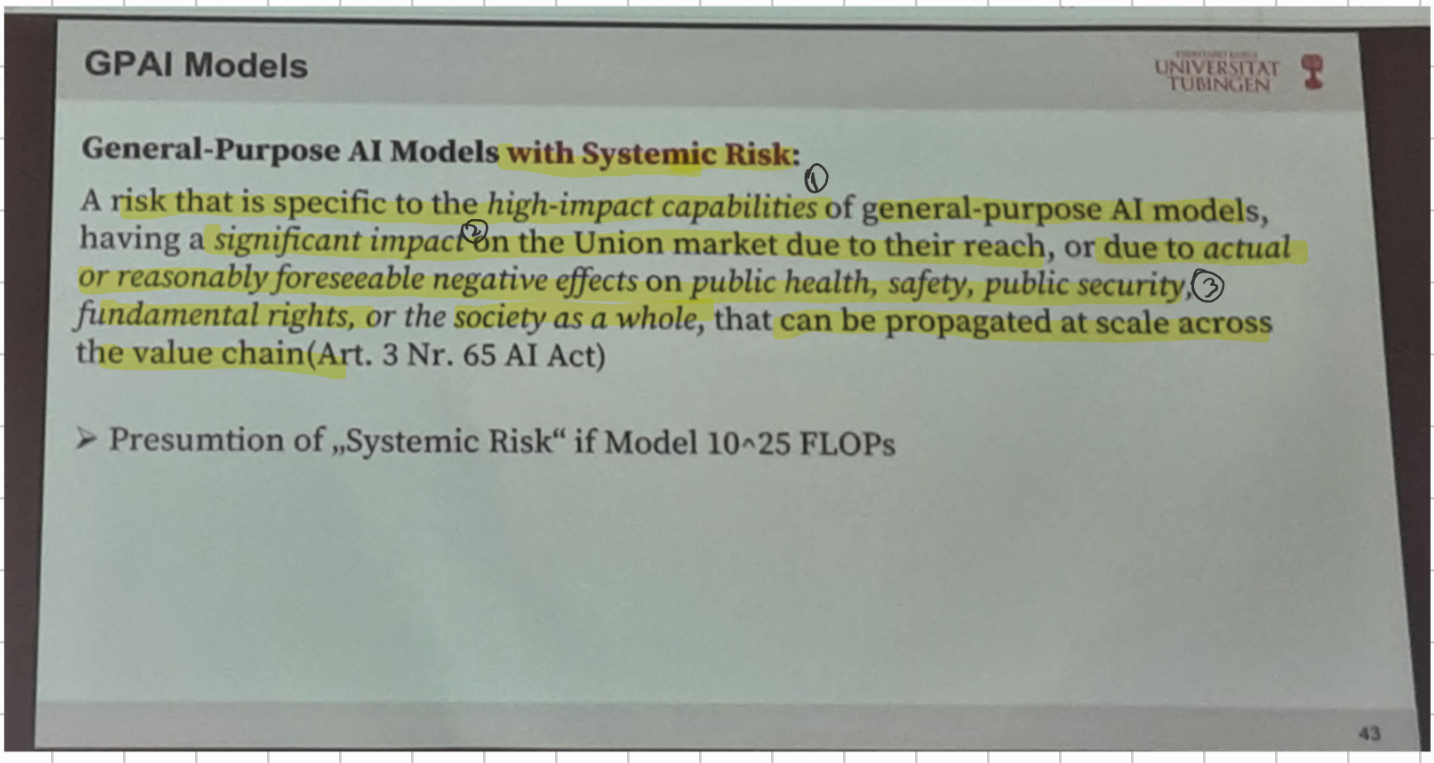
- GPAI System: ChatGPT
- GPAI Model: GPT-3, GPT-4o1, DALL-E, Google Bard, LlaMA, PaLM



Same for AI systems AW

Copyright issues: the models are being trained on the whole internet. Complex problem.

① If you are training a model and use it privately, there are obligations for the one selling model and the one using.



Huge debate as to whether ChatGPT helps you build biological weapons easily. This is considered high risk. It poses a systemic risk.

EU came up with a presumption: as soon as your model has 10^{25} floating point op/s,

it is considered complex enough to be dangerous.

GPAI Models

UNIVERSITÄT TÜBINGEN

General-Purpose AI Models with Systemic Risk:

A risk that is specific to the *high-impact capabilities* of general-purpose AI models, having a *significant impact* on the Union market due to their reach, or due to *actual or reasonably foreseeable negative effects* on *public health, safety, public security, fundamental rights, or the society as a whole*, that can be propagated at scale across the value chain(Art. 3 Nr. 65 AI Act)

- Presumtion of „Systemic Risk" if Model 10^25 FLOPs
- Additional Obligations:
  - Perform Model Evaluation
  - Risk Management System
  - Cybersecurity Tests / Red Teaming

44

If you make a product on the market it has to comply with the ACT. Article 6 has an exemption for research.

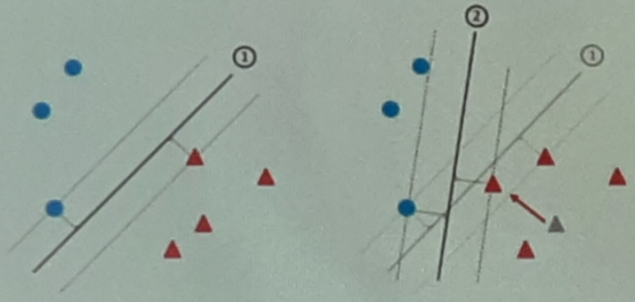
2. ROBUSTNESS AND CYBER SECURITY IN THE AI ACT

- There are 4 types of attacks on AI systems:

1. Poisoning Attacks

UNIVERSITÄT TÜBINGEN

- Data is added or modified during Training
- *Consequence*: Shift of the AI model's decision boundary

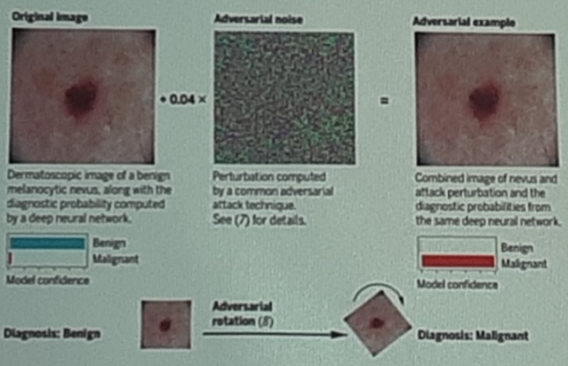


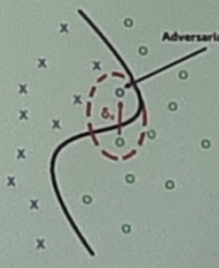
50

2. Adversarial Attacks

UNIVERSITÄT TÜBINGEN

- Inputs are manipulated and intentionally directed to the 'wrong' side of the decision boundary





outside of decision boundaries

Finlayson u. a., Science 2019, 1287
Machado/Silva/Goldschmidt, ACM Computing Surveys 2023, 1 (7).

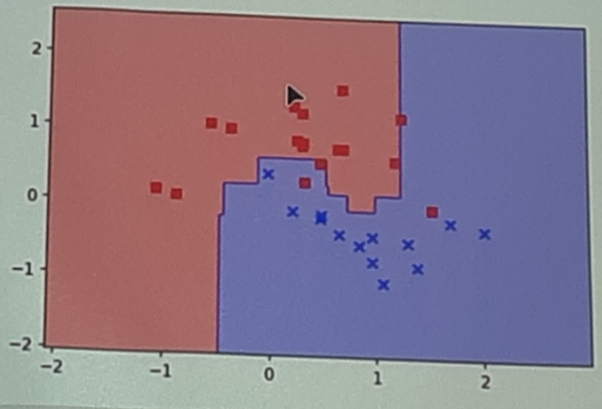
51

Chat Partecipanti Mano Reazioni Vista Stanze App Altro Webcam Microfono

3. Model Stealing

UNIVERSITÄT TÜBINGEN

- Replication of the AI model and its decision boundary by observing the correlation between input and output



Understand how to invert AI and replicate a model just by observing how AI behaves.

https://stackoverflow.com/questions/60891423/finding-data-points-close-to-the-decision-boundary-of-a-classifier

52

52 di 85

Chat Partecipanti Mano Reazioni Vista Stanze App Altro Webcam Microfono

4. Privacy Theft

↳ Possibility to extract data i.e. my face is used to train a Parkinson model.

- Replication and extraction of training data and parameters

↳ The new york times regarding getting information about training data

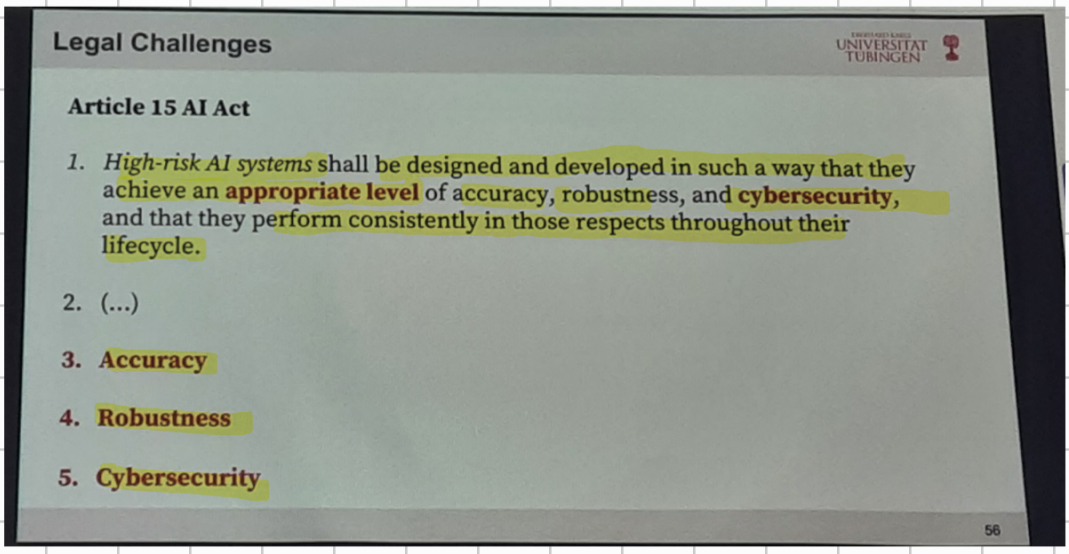
The New York Times

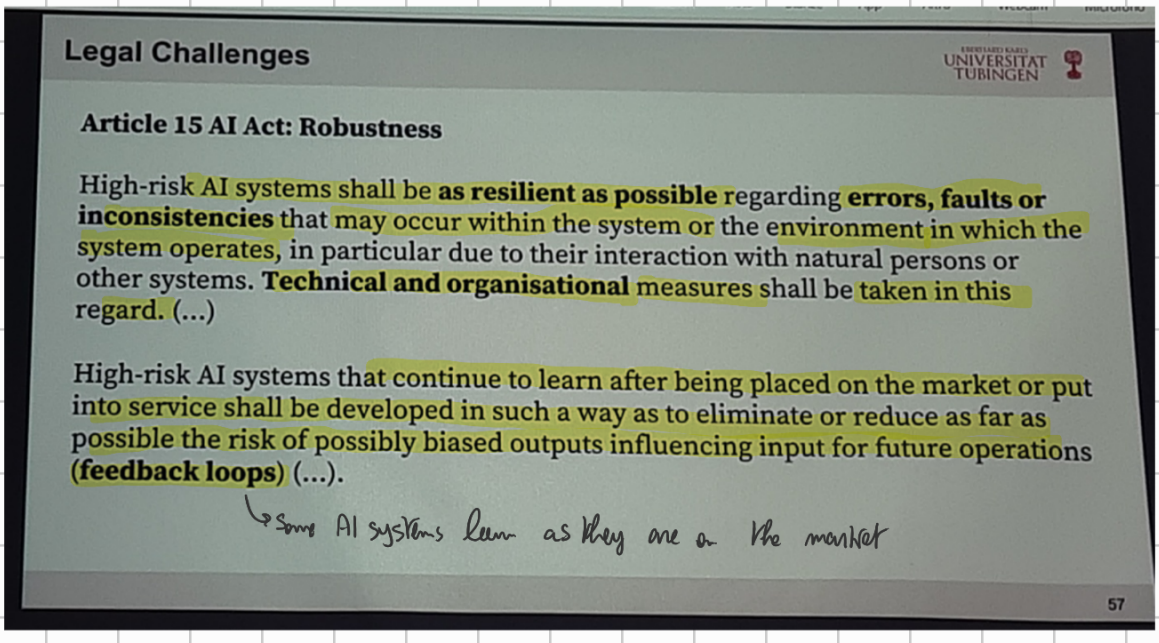
OpenAI

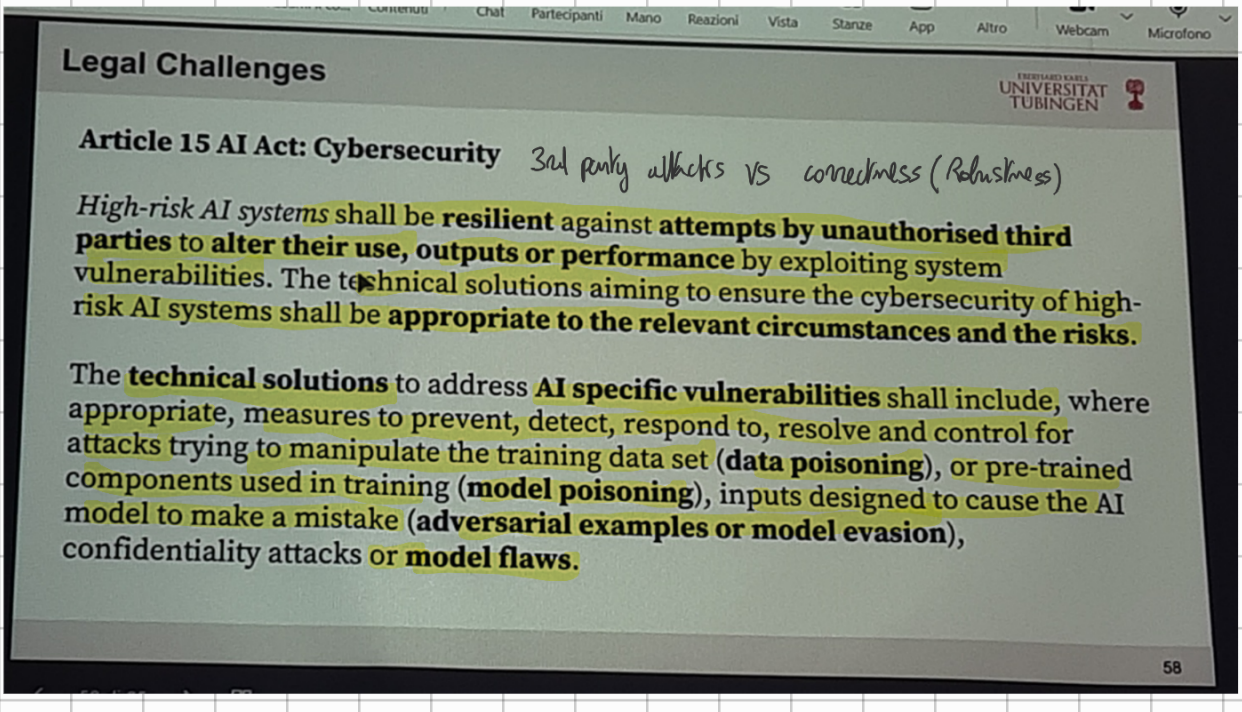
Fredrikson M, Jha S, Ristenpart T (2015), Proceedings of the ACM SIGSAC conference on computer and communications security

53

1. How does EU deals with AI?







Legal Challenges

UNIVERSITÄT TÜBINGEN

Article 15 AI Act

1. High-risk AI systems shall be designed and developed in such a way that they achieve an **appropriate level** of accuracy, robustness, and **cybersecurity**, and that they perform consistently in those respects throughout their lifecycle.

What is Cybersecurity?

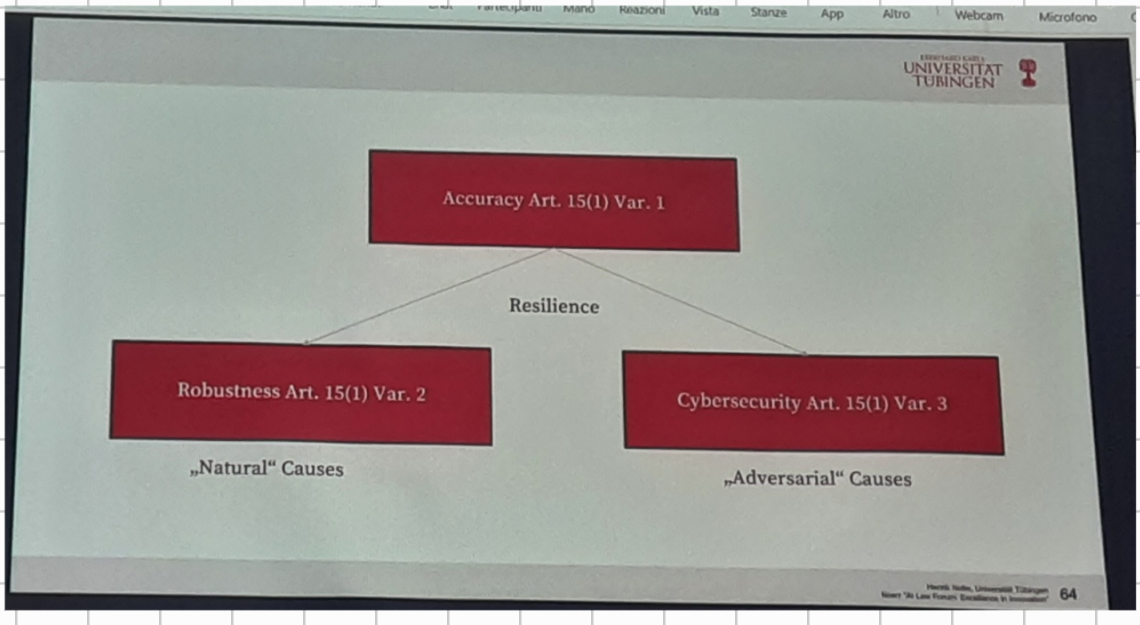
„Activities necessary to protect network and information systems, the users of such systems, and other persons affected by cyber threats" (Art. 2 Cybersecurity Act)

↓

CIA basically! And protection of cyberspace at whole

62

Cybersecurity and robustness are not defined in the act!

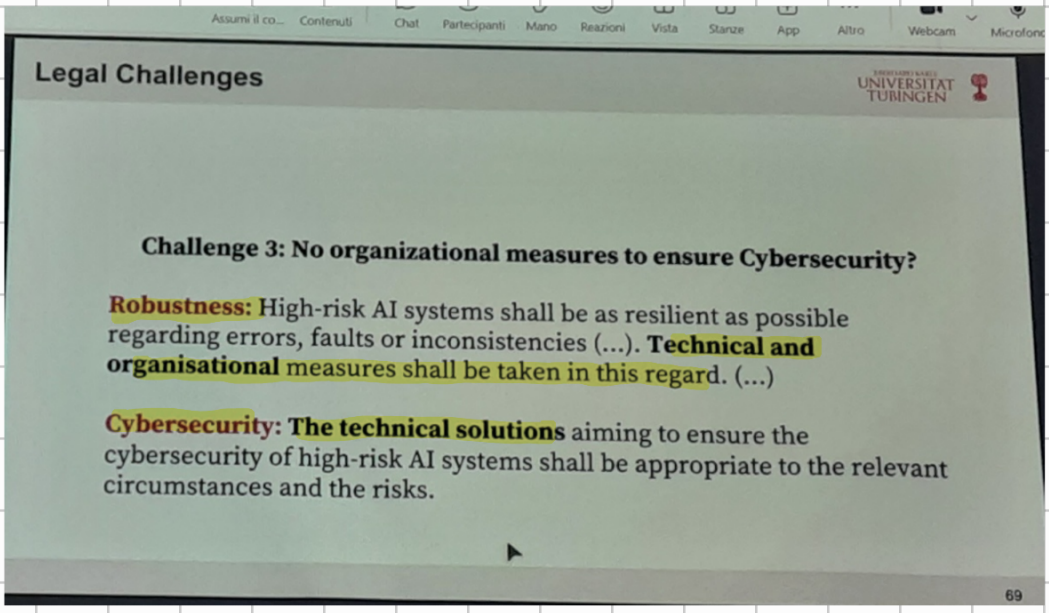


Cybersecurity and Robustness to some extent they mean the same. Those two desiderata should work towards Accuracy. Not being influence by Natural and Adversarial causes.

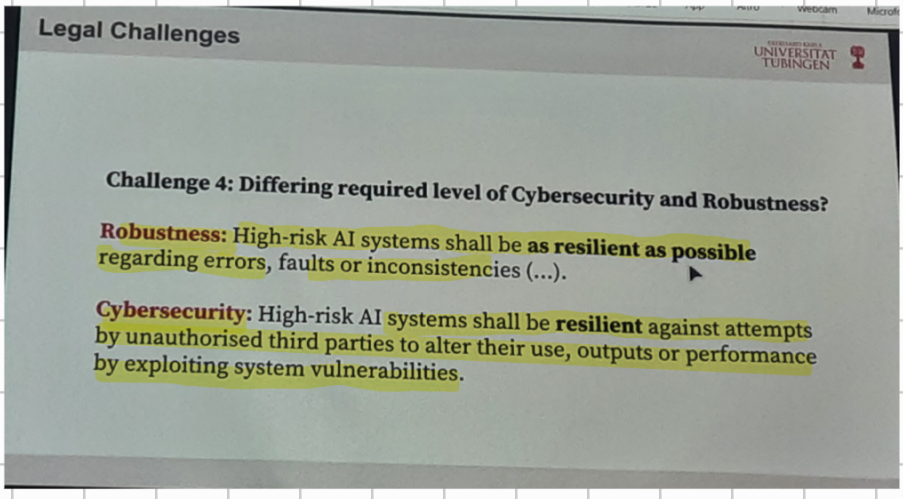
If you see in the article, you have technical and org. measures mentioned in the robustness part. While we only talk about technical measures for the cybersec.

Challenge 2: AI System vs. ML-Model? These terms are used to describe a model but not about a system. Robot with sensors: how do we make systems compliant with model requirements? It's hard!

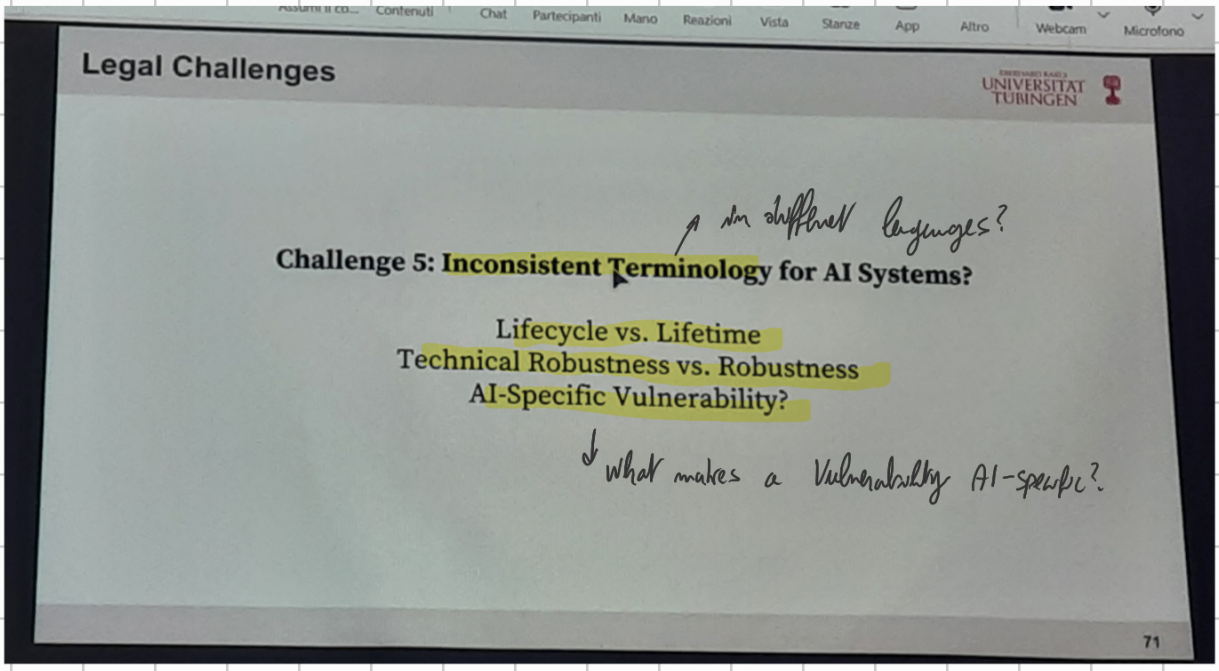
How do we measure accuracy or robustness for other non ML parts?



Possible approach: org. measures are included in cybersecurity? The term is not defined so we think it might include it.

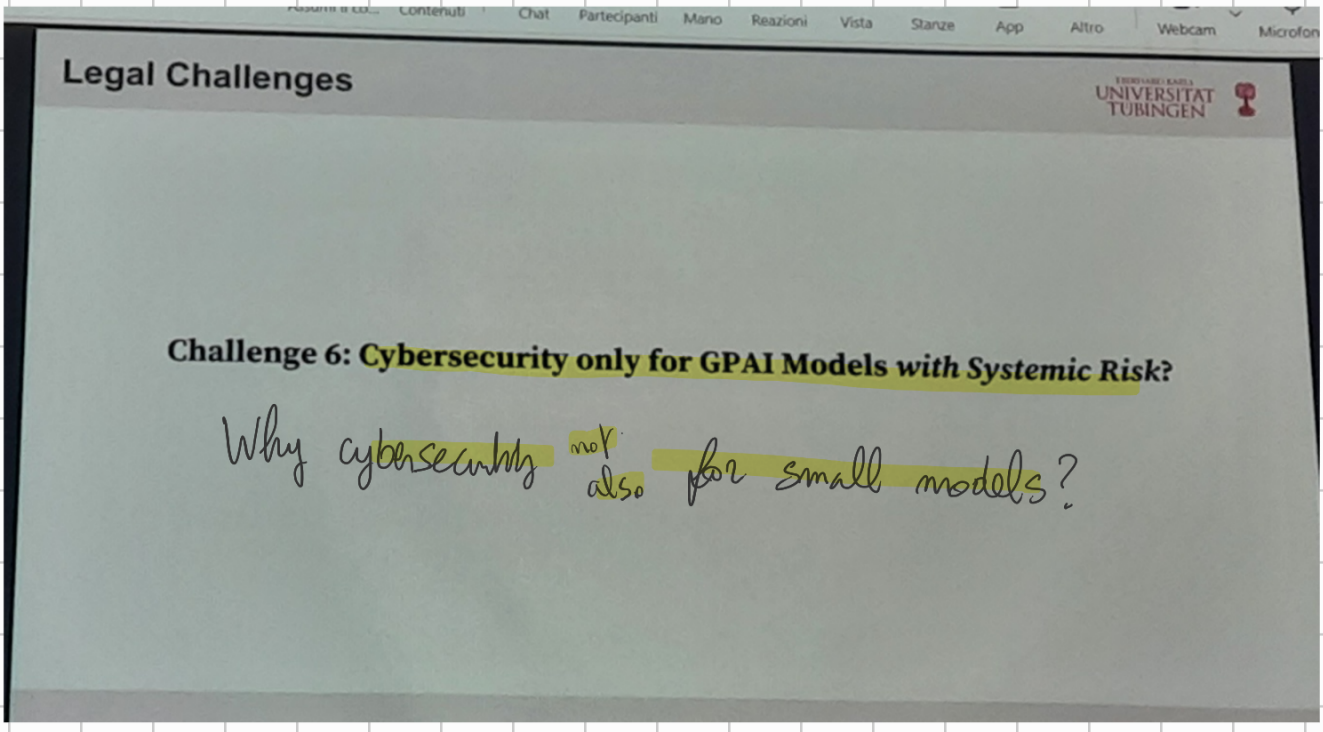


Why is it different?

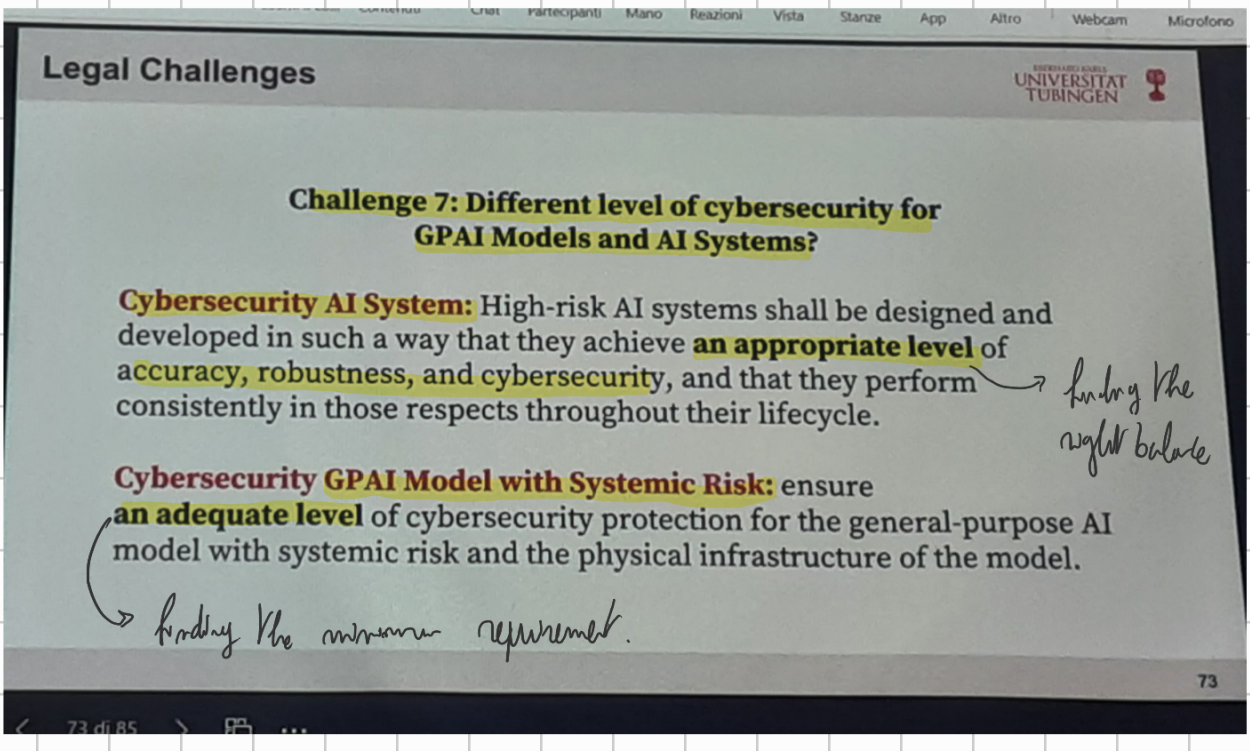


in different languages?

what makes a Vulnerability AI-specific?



Why cybersecurity not also for small models?



finding the right balance

finding the minimum requirement.

How do we balance all the Desiderata? You can't have an accurate and cybersecure data. Who knows!

We will see how the problems get handled in the standards that will be defined.