

THESE NOTES WERE MADE BEFORE I HAD ACCESS TO THE SLIDES

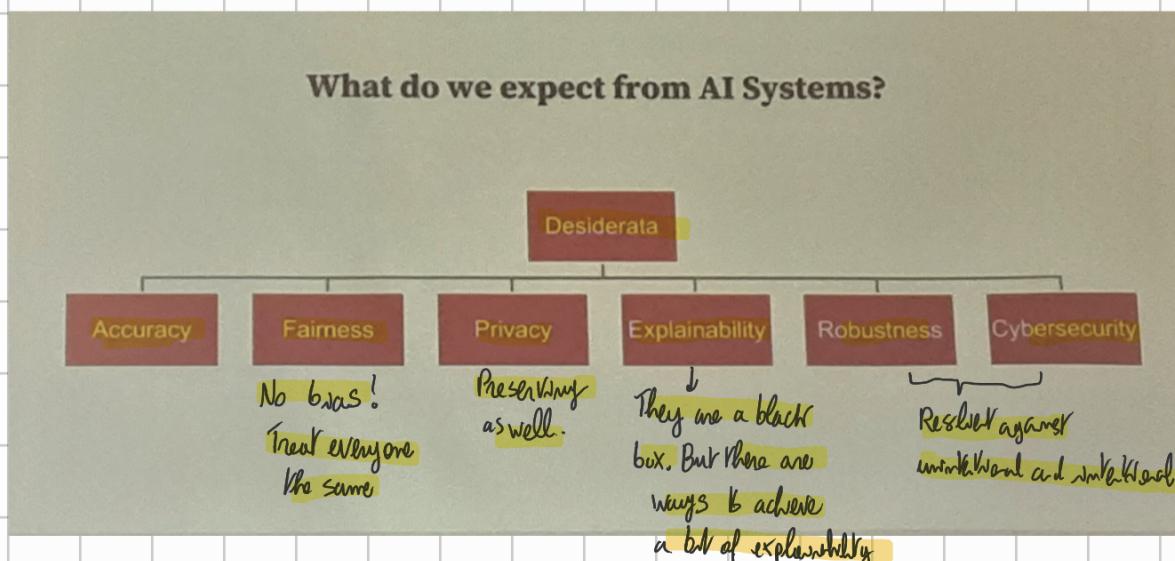
- Pixels of 2 melanoma: the pixels are of the same melanoma, but the AI system sees them as different. 99% benign, 99% malignant. But the pixel on the right is a few pixels different. So that's scary and dangerous.

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 Leyen with plans to improve EU digital relevance.

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

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



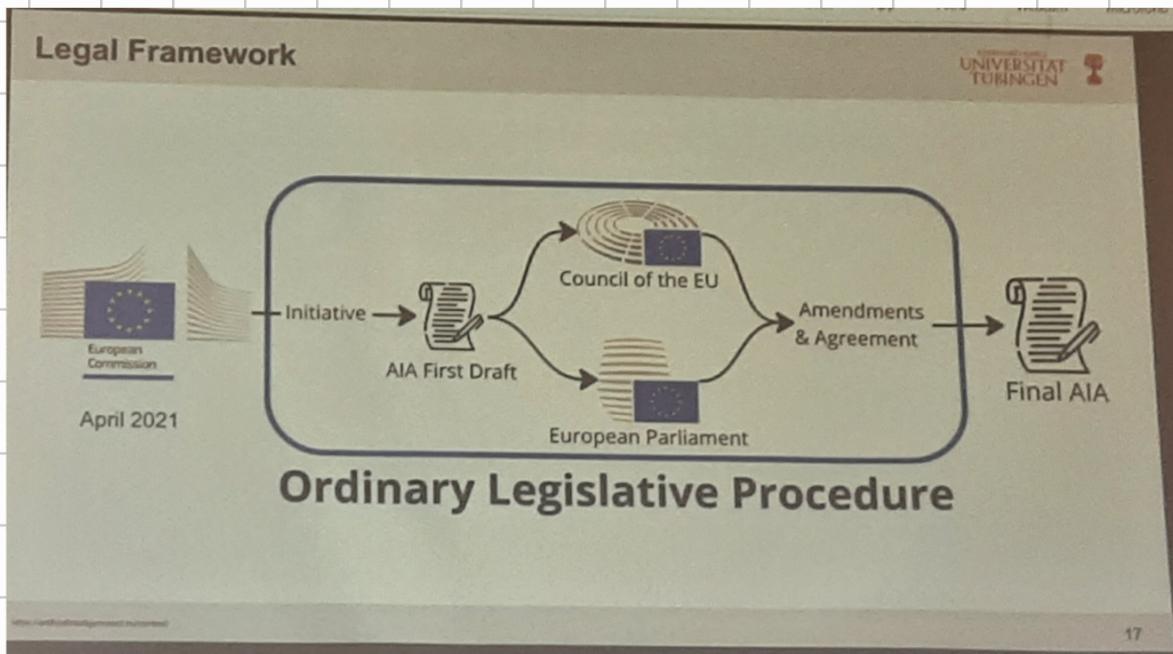
This played a key role in shaping the Act.

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**.

May 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 come 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So the first draft was already outdated!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!

How does the AI Act work?

UNIVERSITÄT
TÜBINGEN

The AI Act in a Nutshell:

- Regulation: Directly applicable in all EU Member States
- Horizontal Regulation: Applies to all sectors
- Vague requirements = New Legislative Approach + Harmonized Standards (DIN, ISO, IEC ...)

22

1. Rules need to be followed directly. 2. Not sector specific! (problem about applicability). 3. Built on the New Legislative Approach. Problem with digital products as 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 hrs) and sector about GP AI Models.



Def: highly criticized...

What is an AI System?

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."

25

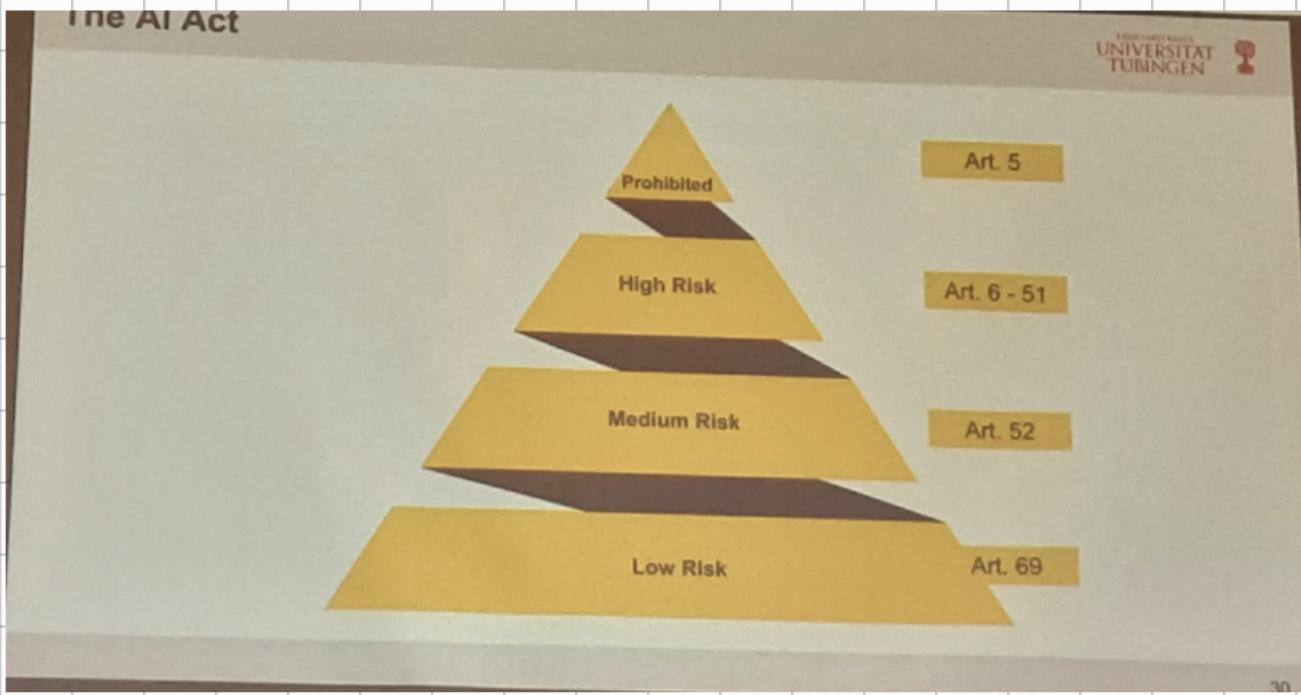
Key word: inference. Knowledge: It's about gaining knowledge without a specifically defined function. They want to exclude current SW systems we are using. What is inference?

What is an AI System?

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."

29



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

Prohibited AI Systems

Art. 5 AI-Act:

- Subliminal, Manipulative, or Deceptive Techniques
- Untargeted Scraping of Facial Images to Create Facial Recognition Databases
- Inferring Emotions in Workplaces and Education
- Biometric Categorisation Systems
- Social Scoring
- „Real-Time“ Remote Biometric Identification (RBI) in Public
- Predictive Policing
- (...)

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

High-Risk AI Systems

ERICH MÜLLER
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)

Thus is vague, so we provide guidelines to identify them. And this is very important.

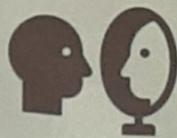
High-Risk AI Systems

ERICH MÜLLER
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. Requirements and then evaluation.

High-Risk AI Systems

FRIEDRICH-ALEXANDER
UNIVERSITÄT
TÜBINGEN



Annex II:

- Machinery
- Toys
- Medical Devices
- Lifts
- Protection Gear
- (...)

Annex III:

- Emotion Recognition
- Critical Infrastructure
- Education and Vocational Training
- Law Enforcement
- Migration, Asylum and Border Control Management
- (...)

AI used in this area
can be considered as
High Risk.

36

What do you really have to do? 7 rules:

Obligations for High-Risk AI Systems

FRIEDRICH-ALEXANDER
UNIVERSITÄT
TÜBINGEN



Art. 9: Risk Management System

Art. 10: Data and Data Governance AI Systems run on data. Is this data fair, representative, etc?

Art. 11: Technical Documentation Every step taken during development needs to be documented

Art. 12: Record-Keeping (Logging) Need to implement a function that logs every single interaction I have with AI system

Art. 13: Transparency and Provision of Information to Deployers

Art. 14: Human Oversight *

Art. 15: Accuracy, Robustness and Cybersecurity

37

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

* Always have a human in the loop!

Medium-Risk AI Systems

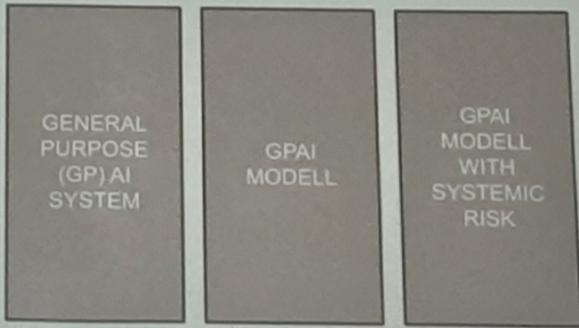
Art. 50 AI-Act:

- NOT High-Risk AI Systems
- Transparency: Watermarking + Inform User

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

38

General-Purpose AI Systems



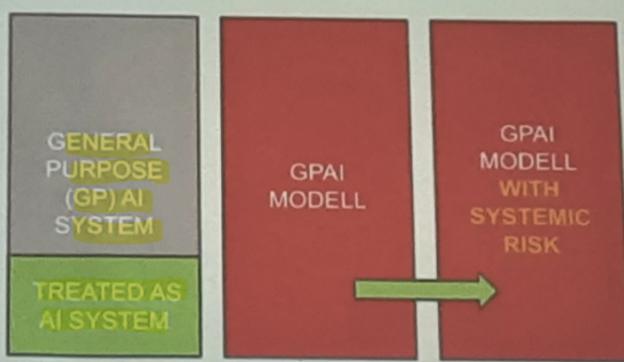
(1)

39

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



40

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

GPAI Models

General-Purpose AI Models:

AI models that display significant generality and is capable of competently performing a wide range of distinct tasks 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)

①
②
③

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

41

General-Purpose AI Models:

AI models that display *significant generality* and is capable of competently performing a wide range of distinct tasks 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)

Obligations:

- Transparency and Provision of Information to Deployers Same for AI Systems AIW
- Technical Documentation
- Copyright
- Contractual Obligations for Downstream Provider ①

42

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 buying.

General-Purpose AI Models with Systemic Risk: ①

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)

- Presumption of „Systemic Risk“ if Model 10^{25} FLOPs

43

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

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)

- Presumption of „Systemic Risk“ if Model 10^{25} FLOPs
- Additional Obligations:
 - Perform Model Evaluation
 - Risk Management System
 - Cybersecurity Tests / Red Teaming

44

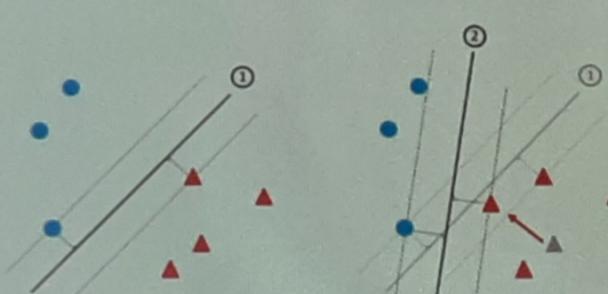
If you make a product on the market it has to comply with the ACT.
Article 6 has an exemption for research.

2 ROBUSTNESS AND CYBER SECURITY IN THE AI ACT

There are 4 types of attacks on AI systems:

1. Poisoning Attacks

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

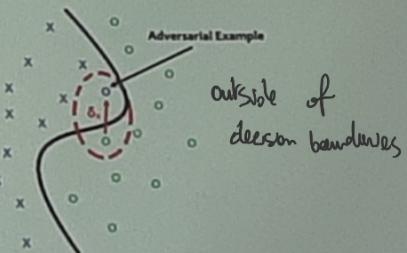
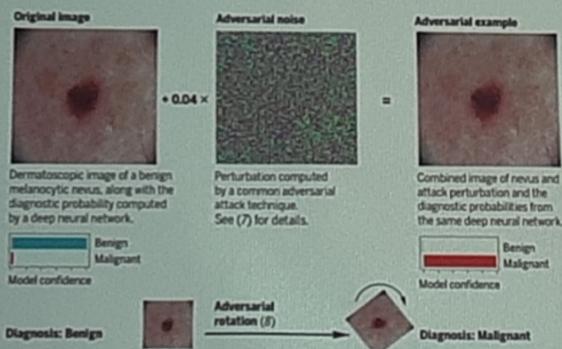


50

2. Adversarial Attacks



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



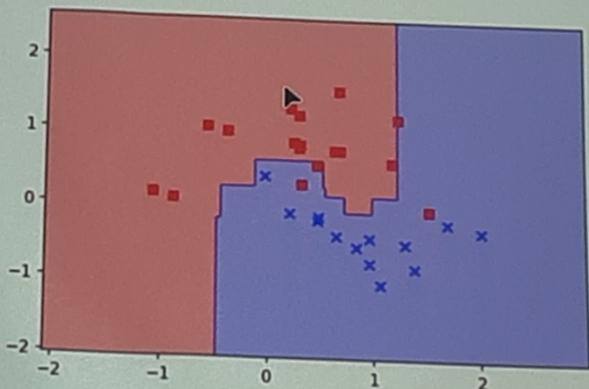
Finlayson et al., Science 2019, 1287
Marchaisse/Goldreich; ACM Computing Surveys 2023, 1 (7).

51

3. Model Stealing



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



<https://stackoverflow.com/questions/50301423/finding-data-points-close-to-the-decision-boundary-of-a-classifier>

52

4. Privacy Theft



- ↳ Possibly to extract data; ex. 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 my data



The New York Times

OpenAI

1. How does EU deals with AI?

Legal Challenges

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.

2. (...)

3. Accuracy

4. Robustness

5. Cybersecurity

56

Legal Challenges

Article 15 AI Act: Robustness

High-risk AI systems shall be as resilient as possible regarding errors, faults or inconsistencies that may occur within the system or the environment in which the system operates, in particular due to their interaction with natural persons or other systems. Technical and organisational measures shall be taken in this regard. (...)

High-risk AI systems that continue to learn after being placed on the market or put into service shall be developed in such a way as to eliminate or reduce as far as possible the risk of possibly biased outputs influencing input for future operations (feedback loops) (...).

↳ Some AI systems learn as they are on the market

57

Legal Challenges

Article 15 AI Act: Cybersecurity 3rd party attacks vs correctness (Robustness)

High-risk AI systems shall be resilient against attempts by unauthorised third parties to alter their use, outputs or performance by exploiting system vulnerabilities. The technical solutions aiming to ensure the cybersecurity of high-risk AI systems shall be appropriate to the relevant circumstances and the risks.

The technical solutions to address AI specific vulnerabilities shall include, where appropriate, measures to prevent, detect, respond to, resolve and control for attacks trying to manipulate the training data set (data poisoning), or pre-trained components used in training (model poisoning), inputs designed to cause the AI model to make a mistake (adversarial examples or model evasion), confidentiality attacks or model flaws.

58

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.

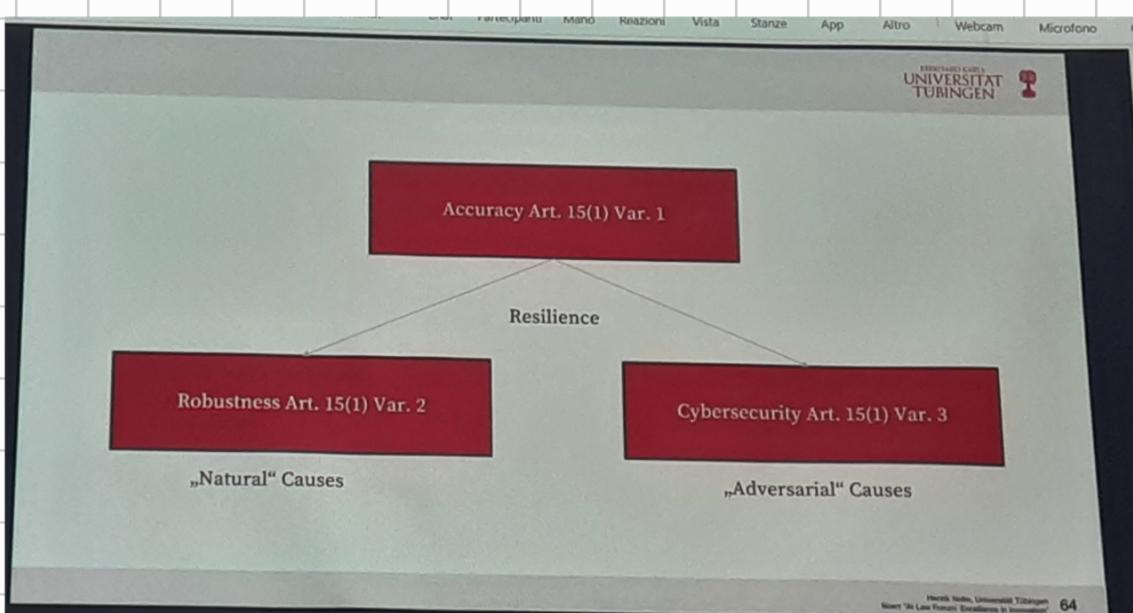
What is Cybersecurity?

„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 national level

62

Cybersecurity and robustness are not defined in the act!



Henrik Nettekoven, Universität Tübingen
Newer 'AI Law Forum: Excellence in Innovation' 64

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 cybersecurity.

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 than ML parts?

Assumi il co... Contenuti Chat Partecipanti Mano Reazioni Vista Stanze App Altro Webcam Microfono

Legal Challenges

FRIEDRICH-ALEXANDER UNIVERSITÄT TÜBINGEN

Challenge 3: No organizational measures to ensure Cybersecurity?

Robustness: High-risk AI systems shall be as resilient as possible regarding errors, faults or inconsistencies (...). **Technical and organisational** measures shall be taken in this regard. (...)

Cybersecurity: The technical solutions aiming to ensure the cybersecurity of high-risk AI systems shall be appropriate to the relevant circumstances and the risks.

69

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

Assumi il co... Contenuti Chat Partecipanti Mano Reazioni Vista Stanze App Altro Webcam Microfono

Legal Challenges

FRIEDRICH-ALEXANDER UNIVERSITÄT TÜBINGEN

Challenge 4: Differing required level of Cybersecurity and Robustness?

Robustness: High-risk AI systems shall be as resilient as possible regarding errors, faults or inconsistencies (...).

Cybersecurity: High-risk AI systems shall be **resilient** against attempts by unauthorised third parties to alter their use, outputs or performance by exploiting system vulnerabilities.

Why is it different?

Assumere il controllo Contenuti Chat Partecipanti Mano Reazioni Vista Stanze App Altro Webcam Microfono

Legal Challenges

EICHENHORN KARLS UNIVERSITÄT TÜBINGEN

↗ In different languages?

Challenge 5: Inconsistent Terminology for AI Systems?

Lifecycle vs. Lifetime
Technical Robustness vs. Robustness
AI-Specific Vulnerability?

↙ What makes a Vulnerability AI-specific?

71

Assumere il controllo Contenuti Chat Partecipanti Mano Reazioni Vista Stanze App Altro Webcam Microfono

Legal Challenges

EICHENHORN KARLS UNIVERSITÄT TÜBINGEN

Challenge 6: Cybersecurity only for GPAI Models with Systemic Risk?

Why cybersecurity ^{not} also for small models?

Assumere il controllo Chat Partecipanti Mano Reazioni Vista Stanze App Altro Webcam Microfono

Legal Challenges

EICHENHORN KARLS UNIVERSITÄT TÜBINGEN

Challenge 7: Different level of cybersecurity for GPAI Models and AI Systems?

Cybersecurity AI System: 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.

↗ finding the right balance

Cybersecurity GPAI Model with Systemic Risk: ensure an adequate level of cybersecurity protection for the general-purpose AI model with systemic risk and the physical infrastructure of the model.

↗ finding the minimum requirement.

73 di 85

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

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