CHALLENGES AND STRATEGIES IN DETECTION OF AI-GENERATED CONTENT

Martin Steinebach

CONTENT

Image

- Neural Edition
- Text-to-Image

Video

- Deepfakes
- Text-to-Video

Audio

- Text-to-Speech
- Voice Cloning

Text

- LLMS

Image-to-Video

Lipsync

Overdub

SECURITY CHALLENGES

Image

- Desinformation, Cyber Mobbing, Insurance Fraud

Video

- Desinformation, Identity Theft

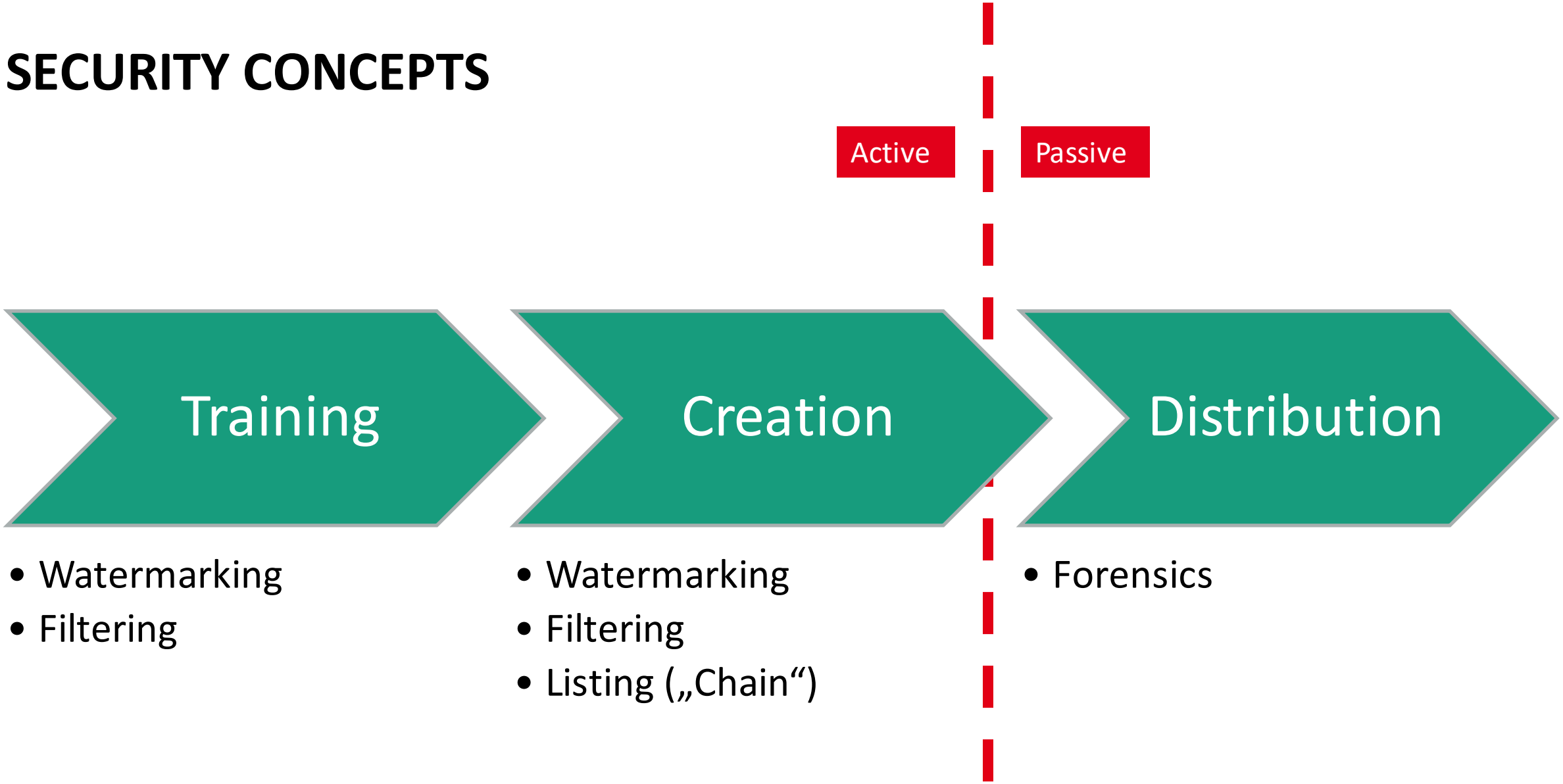
Audio

- Social Engineering, Fraud

Text

- Desinformation, Phishing, Fraud

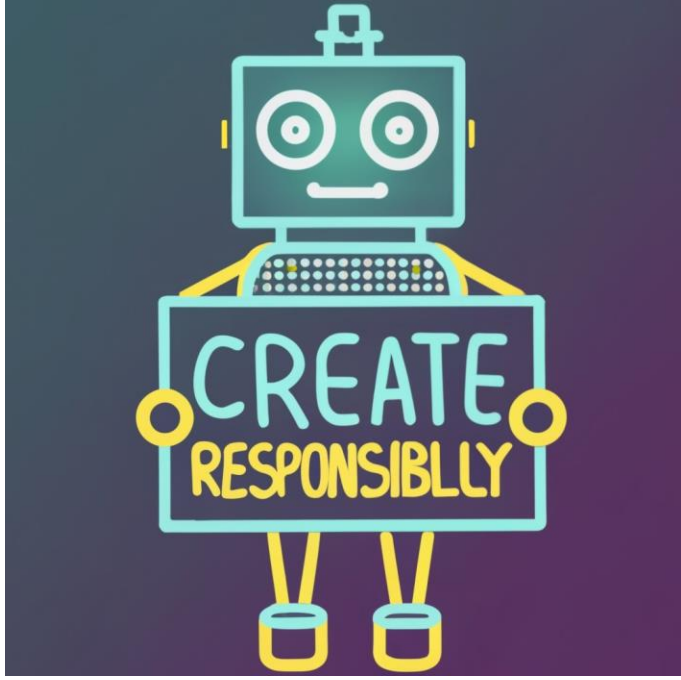
SECURITY CONCEPTS



SECURITY CONCEPTS – TRAINING

- Watermarking
  - Imperceptible watermarks, no „logo“
  - “If we only provide watermarked training data, the system will only create watermarked content” (no)
- Filtering
  - Remove problematic training data by filtering
    - Persons of public interest, NSFW, violence,…
    - Reliability is limited
  - Remove copyrighted material
  - Transfer Training easily „retrains“ this

SECURITY CONCEPTS – CREATION

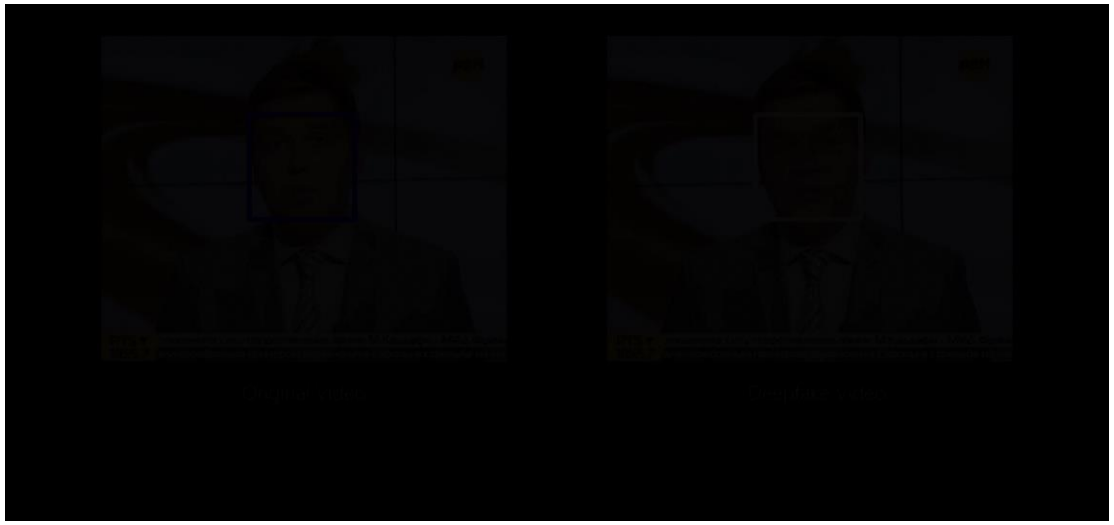


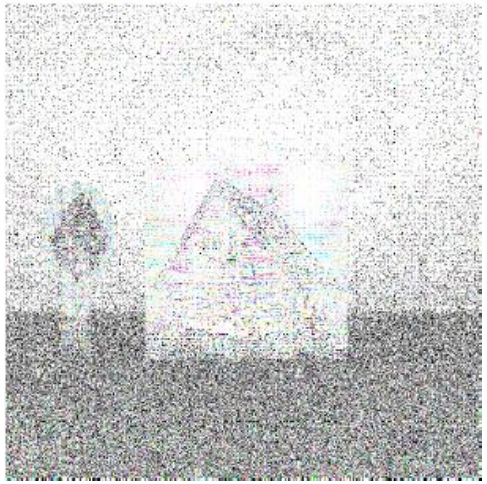
- Watermarking
  - Embedding of watermark during or after content creation
    - Visible watermarks
    - Imperceptible watermarks
- Filtering
  - Detection of problematic content after creation
- Listing
  - Created content is added to publicly available list
    - Hash (robust, fuzzy or cryptographic)
- **General problem: Open-Source implementations allow to remove this**

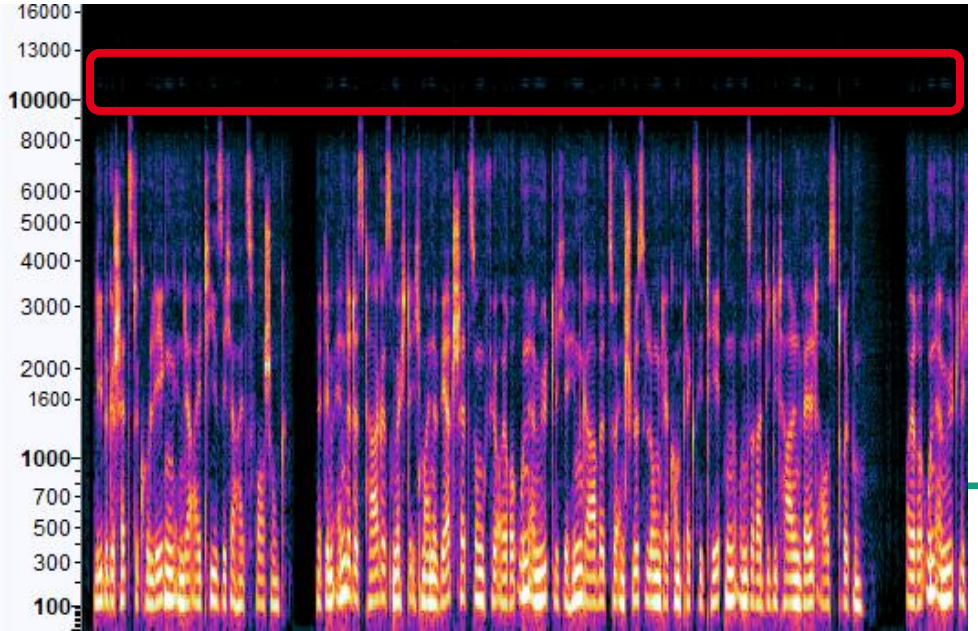
SECURITY CONCEPTS – DISTRIBUTION

- Forensics
  - Detection of creation traces
    - Upscaling patterns
    - High-frequency audio patterns
  - Detection of typical styles
    - NLP plain vanilla ChatGPT detection
  - Detection of creation errors
    - Skin color mismatch (deepfakes)
    - Sharpness mismatch (text-to-image)
  - Challenges:
    - Error rates (false alarms)
    - Counter forensics

SECURITY CONCEPTS – FORENSICS (EXAMPLES)







SECURITY CONCEPTS – INVERSION

- If we cannot reliably identify synthetic content, maybe we can identify real content?
- Two approaches
  - Trusted list: Signature of trusted creator plus searchable abstraction (~hash)
    - Speeches
    - News content
  - Signed content
    - “Trustworthy camera"
- Who decides about trusted creators?
- How to deal with user generated content?

CHALLENGES

Vast amount of unstructured training data

Open Source Implementations

Rapid Development

Mix of real and artifical data

STRATEGIES

Classification and filtering of training data

Watermarking infrastructures

Creation of lookup services for content

Research in forensic tools

THANK YOU!

Contact

- Prof. Dr. Martin Steinebach
- Head of Media Security and IT Forensics
- Tel. +49 6151 869-349
- martin.steinebach@sit.fraunhofer.de

https://www.sit.fraunhofer.de

Fraunhofer SIT