Deepfakes – příležitost nebo hrozba?

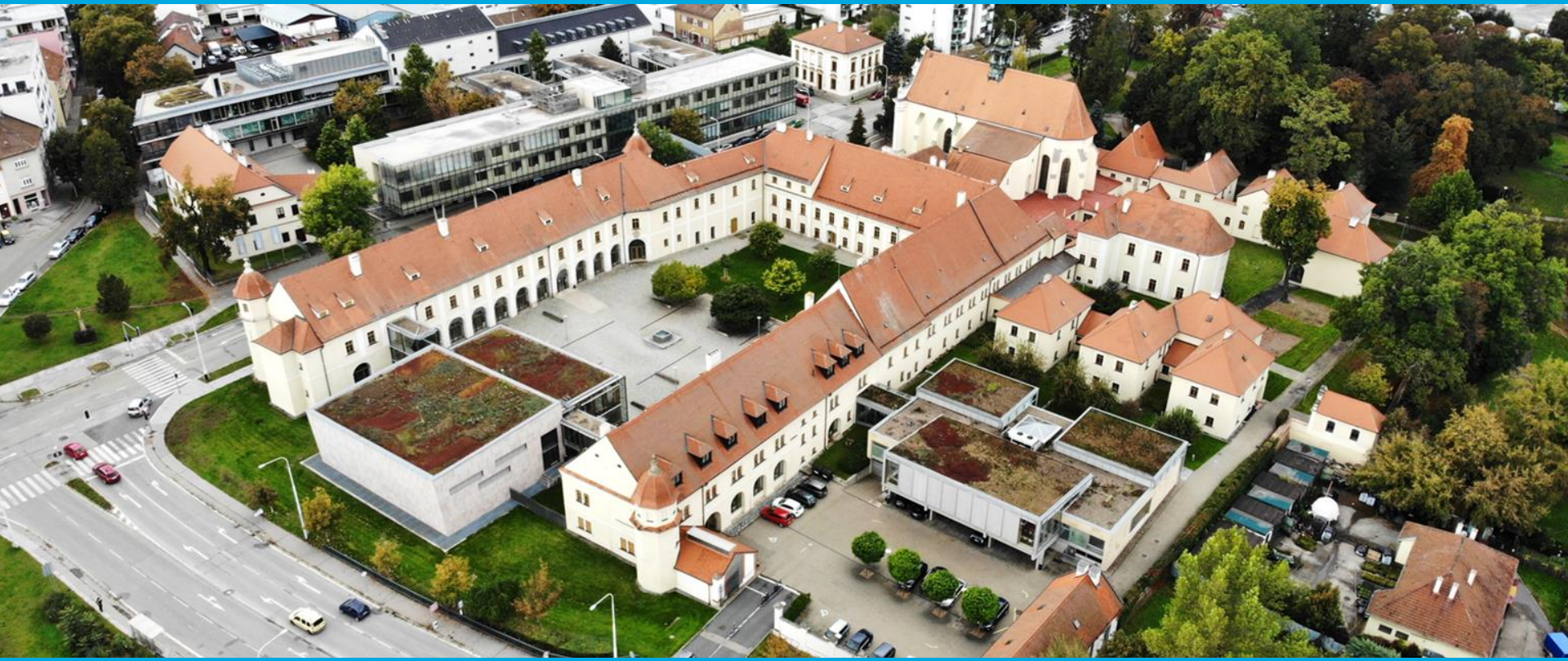
Kamil Malinka a Anton Firc

malinka@fit.vut.cz, ifirc@fit.vut.cz

Security@FIT

BRNO UNIVERSITY OF TECHNOLOGY
FACULTY OF INFORMATION TECHNOLOGY

www.fit.vutbr.cz

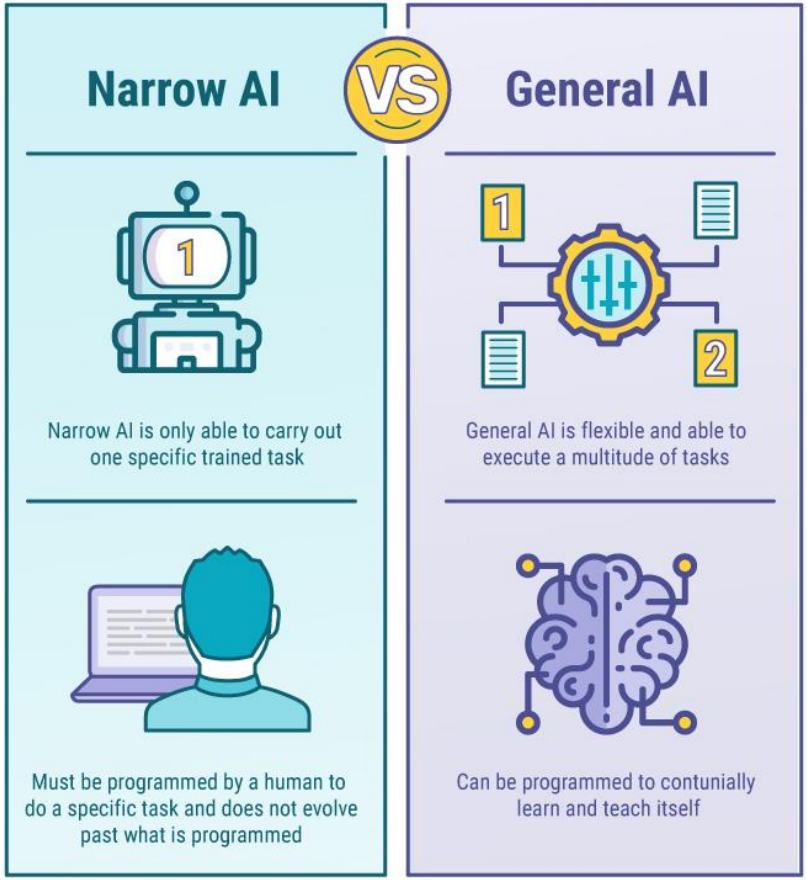


Vzdělávání - Výzkum - Společnost

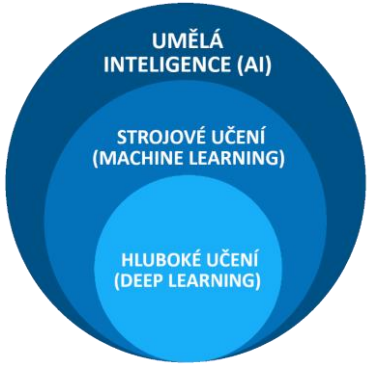
- **Výzkumné směry**
  - Bezpečnostní dopady AI
  - Blockchainové technologie
  - Analýza malware
  - IoT technologie
  - Použitelná bezpečnost
  - Finanční protokoly

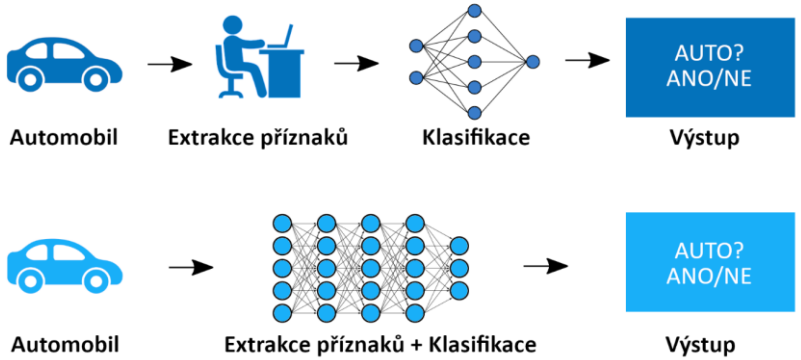
Umělá inteligence: co to vlastně je?

- **AI** (umělá inteligence): Stroj se schopností vyhodnocovat data a činit rozhodnutí stejně dobře nebo lépe než člověk. Rozumí datům a výstupům
- **Narrow AI** (úzce zaměřená AI): AI zaměřená na jeden konkrétní úkon/téma
- **General AI** (všeobecná umělá inteligence): Umělá inteligence schopná jako člověk přizpůsobit své fungování na jakékoliv téma/úkoly



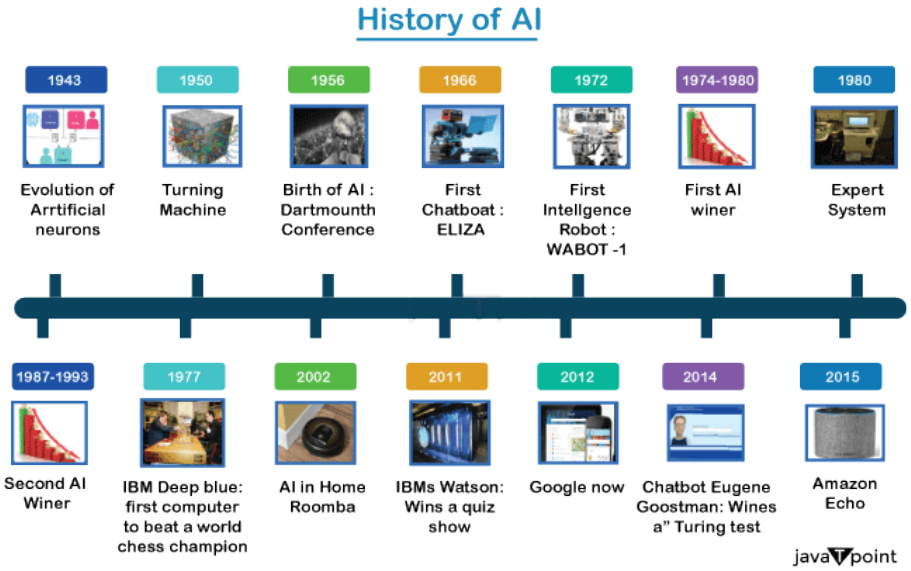
- **Machine Learning** se zabývá tvorbou a studiem statistických algoritmů schopných se učit z dat a aplikovat naučené na nová data, a tím vykonávat úkoly bez nutnosti konkrétních pokynů.
- **Neural network** (neuronová síť) je matematický model pro zpracování dat inspirovaný funkcí lidského mozku.
- **Deep learning** (hluboké učení) je varianta strojového učení využívající neurální sítě s mnoha vrstvami k spracování složitých dat a k učení se složitým vzorcům bez nutnosti ručního nastavování





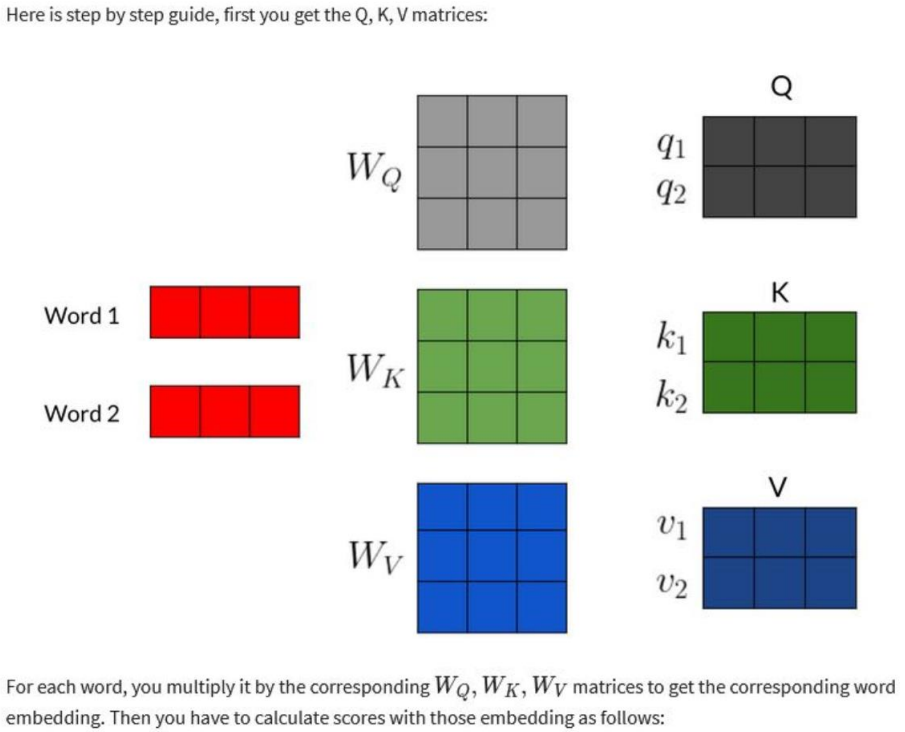
Výmysl současné doby?

- Teorie neuronových sítí vznikla již v roce 1943
- První praktické pokusy v šedesátých letech a pokroky v osmdesátkách
- Zlom v 21. století
  - vazba na pokroky v hardwaru
- Kdy bude další průlom?



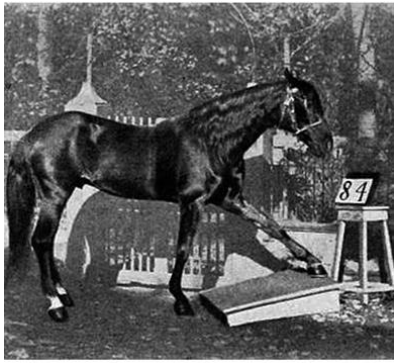
Je AI opravdu AI?

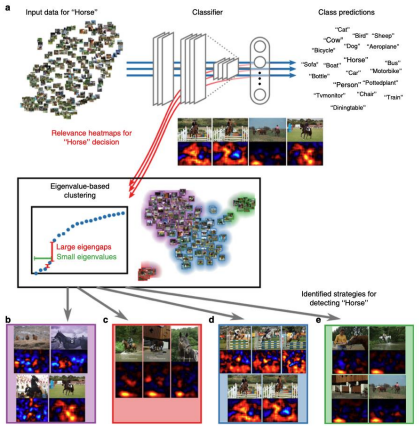
- Pokročilý ML
  - Jak moc ChatGPT reálně rozumí tomu co píše, a kolik jsou jen neskutečná kvanta naučených vzorců?
  - “jednoduché násobení matic“
- To co máme dnes je Narrow AI
- I v případě že máme „jen“ ML – má velký disruptivní potenciál



Důležitost dat

- Většina současné AI technologie využívá ML a primárně DL
- Dataset = data ze kterých se AI učí
- Datasety jsou stejně, ne-li více důležité než AI algoritmy samotné
- “Clever Hans“ efekt
- Obrovské množství možných biasů, chyb, atd.
- Klíčové know-how

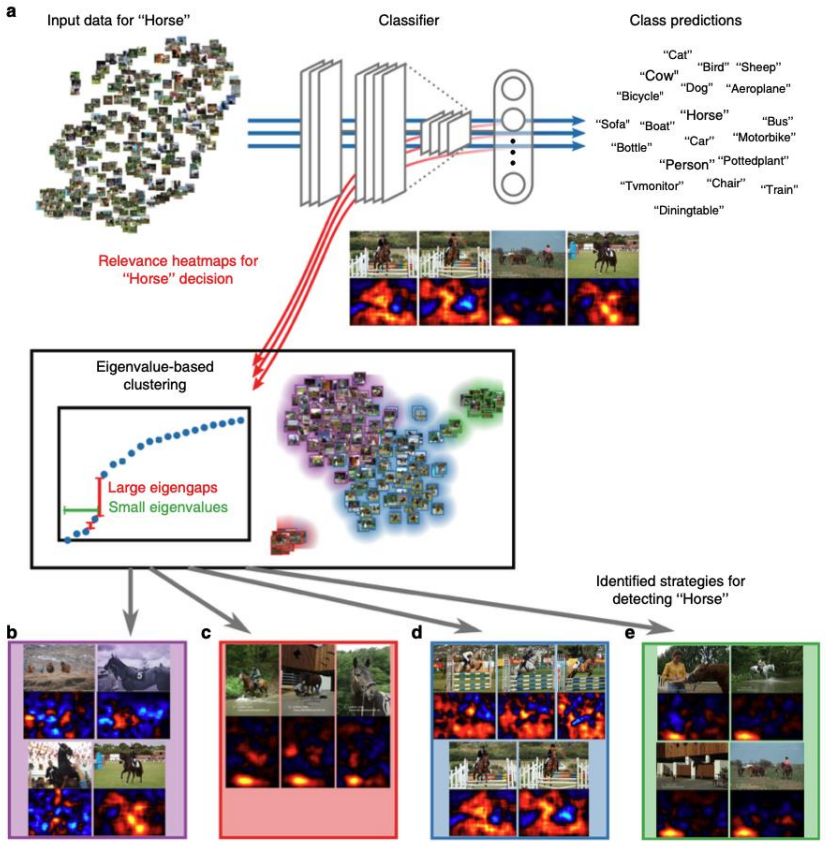




[1] Lapuschkin, S., Wäldchen, S., Binder, A. et al. Unmasking Clever Hans predictors and assessing what machines really learn. Nat Commun 10, 1096 (2019). https://doi.org/10.1038/s41467-019-08987-4.

Důležitost dat

- Většina současné AI technologie využívá ML a primárně DL
- Dataset = data ze kterých se AI učí
- Datasety jsou stejně, ne-li více důležité než AI algoritmy samotné
- “Clever Hans“ efekt
- Obrovské množství možných biasů, chyb, atd.
- Klíčové know-how

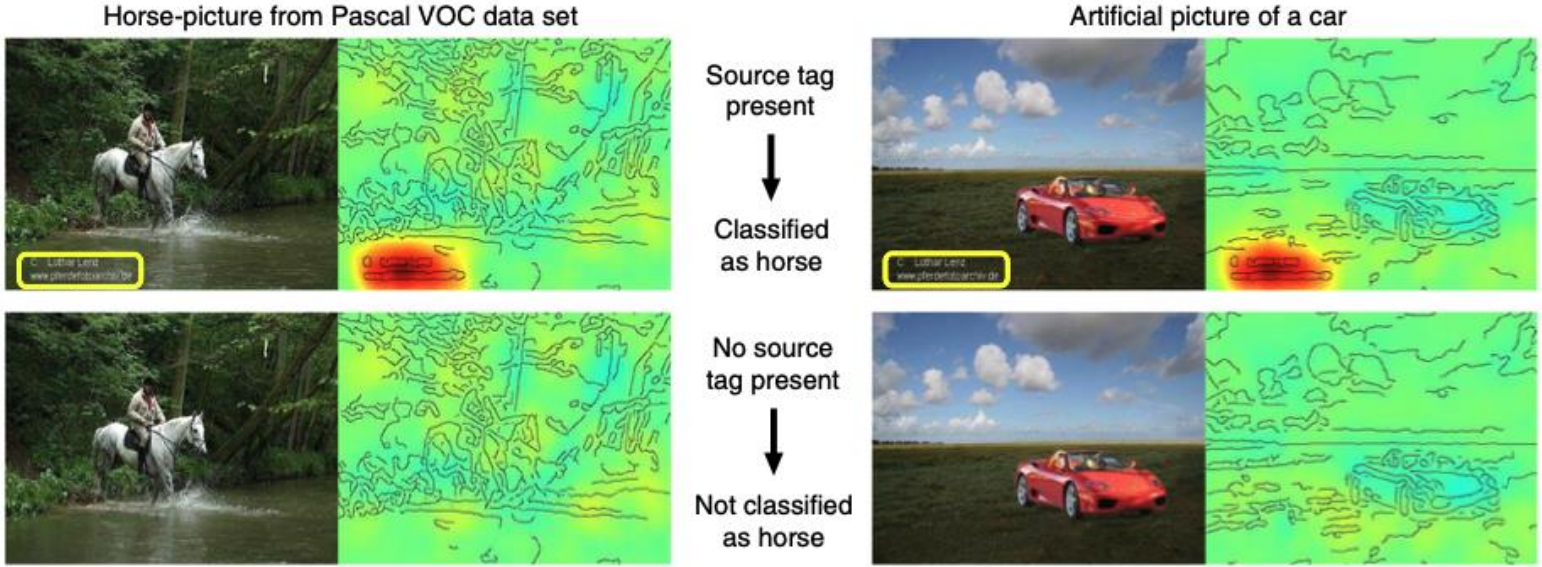


[1] Lapuschkin, S., Wäldchen, S., Binder, A. et al. Unmasking Clever Hans predictors and assessing what machines really learn. Nat Commun 10, 1096 (2019). https://doi.org/10.1038/s41467-019-08987-4.

Důležitost dat – k čemu to vedlo?

- Většina současné AI technologie využívá ML a

a

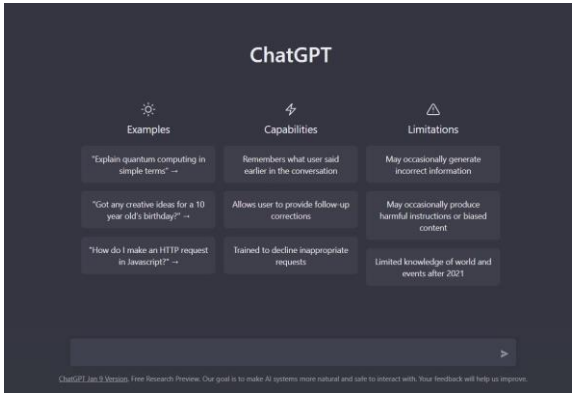


- Klíčové know-how

[1] Lapuschkin, S., Wäldchen, S., Binder, A. et al. Unmasking Clever Hans predictors and assessing what machines really learn. Nat Commun 10, 1096 (2019). https://doi.org/10.1038/s41467-019-08987-4.

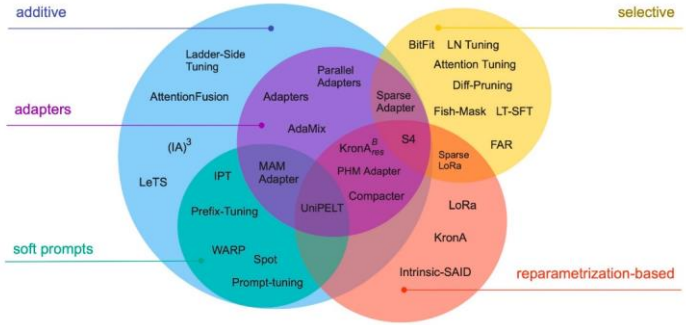

Garbage in
Garbage OUT

- AI schopná generovat „hmatatelné“ výstupy nebo produkty
- Text, zvuk, obraz, video…
- V současnosti vysoce populární
- Text: ChatGPT a konkurenční chatboti
- Obraz: Dall-E, Midjourney…
- Trend: Multimodální vstupy a výstupy



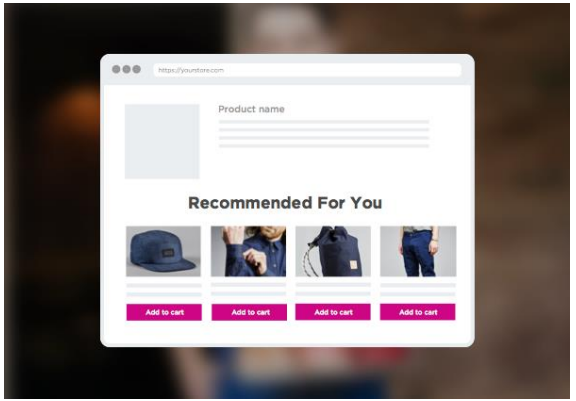
Velké jazykové modely (LLM)

- Klíč fungování Chatbotů
- Textové datasety převedeny do matematického modelu
- Přiřazuje slovům hodnoty a pravděpodobnosti jejich výskytu poblíž jiných slov
- Na první pohled velmi přesvědčivý, ale pořád velmi limitovaný



Negenerativní AI

- Především AI pro rozhodovací procesy
- Výstupem není produkt, ale rozhodnutí
- Nejsou tolik vidět jako generativní AI, ale neméně důležité
- Samořídící automobily, asistence s diagnózami, algoritmy pro doporučování obsahu, cílené reklamy…



Blízká budoucnost

- Současný milník na horizontu je (kvalitní) video
- Velké akční modely? Text-to-game?
- Konečně plně autonomní auta?
- Vazba na hardware
- Boom nebo stagnace?

Umělá inteligence a kyberbezpečnost

Dopady AI na kyberbezpečnost

- AI promění (a již proměňuje) podobu kybernetické bezpečnosti
  - Ve všech rovinách
- AI slouží a bude sloužit útočníkům
  - zároveň ale i obráncům
- Závody v AI zbrojení?
- Morální a etické limity „těch hodných“
- AI se v neposlední řadě sama stává cílem jak AI tak ne-AI útoků

AI jako nástroj útočníka - negenerativní

- Automatické hledání zranitelností
- Monitorování síťového provozu
- C2 (chytrý autonomní malware)
- Automatické shánění podkladů pro social engineering
- Řízení masivních a chytrých botnetů

AI jako nástroj útočníka - generativní

- Psaní dezinformací
- Generování dezinformačních obrázků
- Oboje využitelné pro phishing
- Deepfakes (videa a hlas)
- Vishing
- Psaní malwaru
  - Zatím málo sofistikovaný, ale dostupný komukoliv
  - Dedikované „bad guy“ modely

AI jako nástroj obránce

- Bojovat s AI pomocí AI
- Vše co AI dělá dokáže AI mitigovat (…teoreticky)
- Chytré IDS
- Chytrý antimalware
- Automatické psaní oprav?
- AI na detekci deepfakes
- Chytré spam filtry
- Útočníci vždy o krok napřed
- Morální a právní limity obránců

AI jako oběť útoku

- Podíl AI služeb a produktů raketově roste
- Točí se v nich obrovské množství peněz
- Masivně na nich rovněž roste závislost společnosti
- Nevyhnutelné útoky proti AI (ať už s AI nebo bez)
- V menší míře už se dějí, a bude jich přibývat

Forbes

FORBES > INNOVATION > CYBERSECURITY

ChatGPT Down As Anonymous Sudan Hackers Claim Responsibility

Davey Winder Senior Contributor
Veteran cybersecurity and tech analyst, journalist, hacker, author

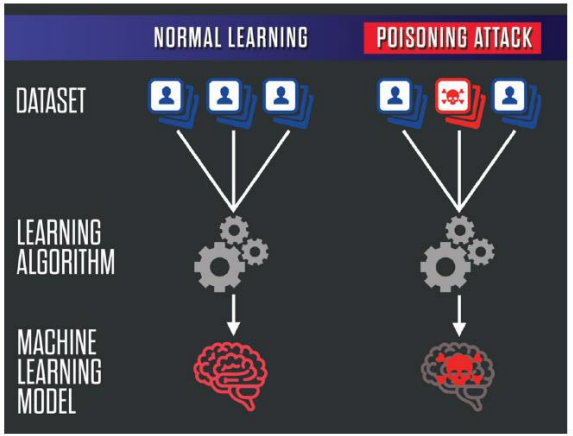
Follow

Nov 10, 2023, 08:39am EST

Základní útoky na AI

- DDoS – Tradiční denial of service bude v nejbližší době hlavním nástrojem
- Ransomware – může zasáhnout i AI, velmi lákavé cíle
- Jailbreak – přinutit chatbota k podání zakázaných informací

- Narušení „black boxu“ – potenciálně těžce odhalitelné
- Nemusí ale mířit na AI přímo
- Otrava datasetů
- Nazpět k příkladu s koňmi
- Co kdyby AI letištní skener skrze otrávené datasety nepoznal zbraň v batohu?
- Proto potřeba chránit jak AI tak datasety



Deepfakes

Název	**Nástroje boje proti hlasovým DeepFakes (VB02000060)**
	Tools To Combat Voice DeepFakes
Realizace	01/2024 - 12/2026
Realizuje	FIT VUT v Brně (Speech@FIT + Security@FIT) + Phonexia s.r.o.
Uživatel	Policejní prezidium ČR
Poskytovatel	MV ČR - Program bezpečnostního výzkumu ČR 2021 – 2026 (SECTECH)
Veřejný link	https://www.fit.vut.cz/research/project/1762


#nePINdej!
KYBERTEST

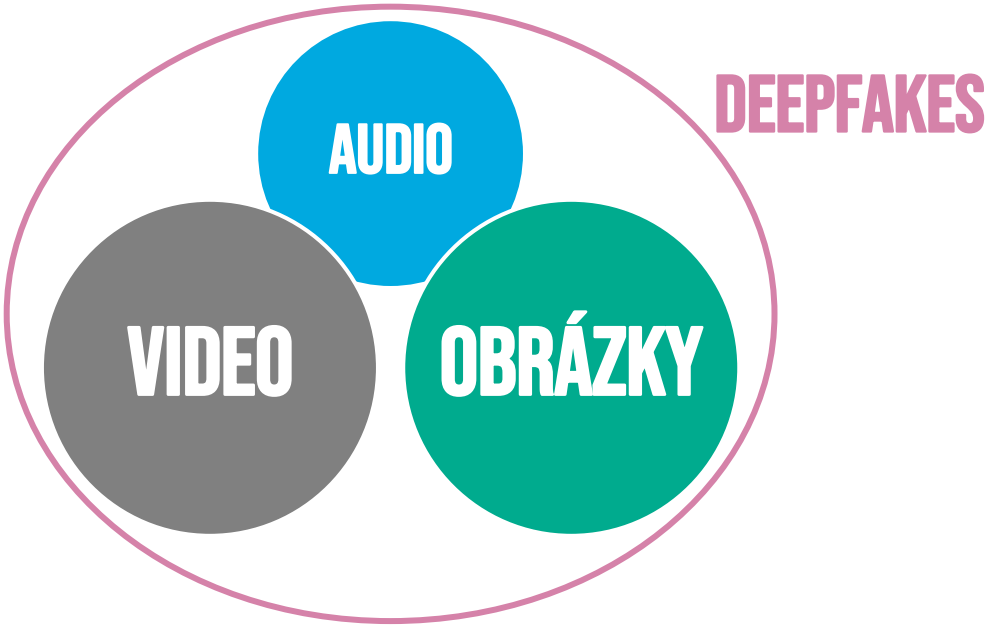

KYBERTEST
#nePINdej!
Nikdy nikomu nedávej svoje hesla!
Jasný! E-šmejdi mě nedostanou!
#nePINdám!
E-šmejdi vás chtějí okrást!
Naučte se, jak jim nenaletět!
www.kybertest.cz

Osnova

- Motivace a reálné útoky
- Trocha teorie k deepfakes
- Metody ochrany
- Detaily k výzkumným oblastem (když vyjde čas…)
  - Útok na hlasovou biometrii
  - Útoky na obličejovou biometrii
  - Lidské rozpoznávání deepfakes

Co jsou deepfakes?

- “deep learning” + “fake”
- Podmnožina syntetických médií
- Zobrazují události které se nikdy nestaly
- Výrazné zvýšení kvality a dostupnosti v posledních letech



Aktuální příklady z médií

Široká škála pozitivních využití před námi…

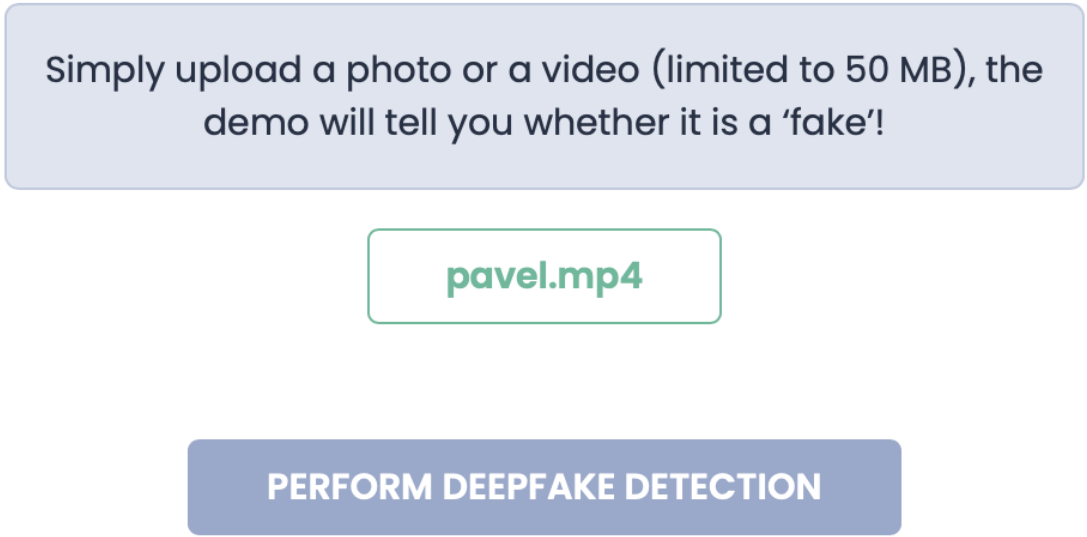
NOT HIS
REAL
VOICE

TOP GUN MAVERICK

Vektory útoků

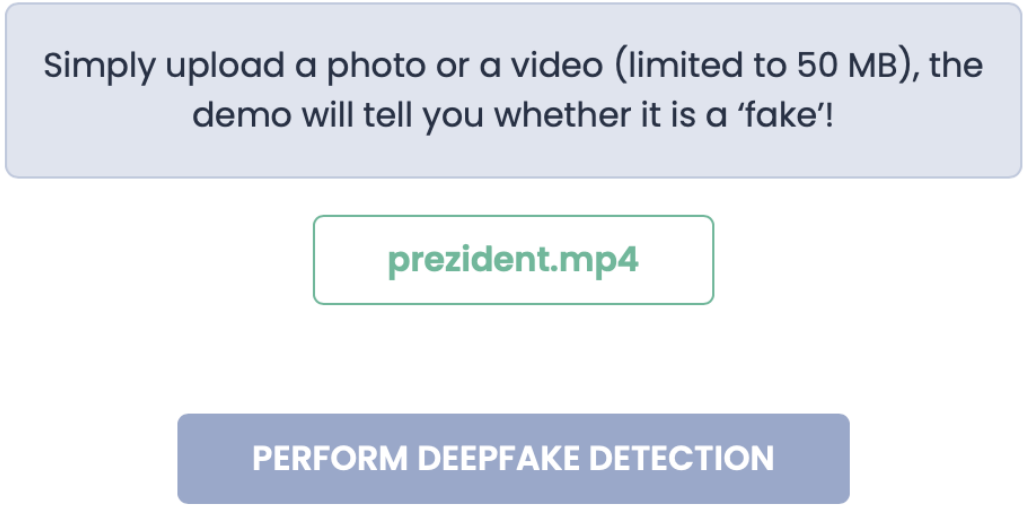
Deepfakes v bezpečnosti

Demonstrační deepfake ukázka vytvořená bez souhlasu vyobrazené osoby.


Simply upload a photo or a video (limited to 50 MB), the demo will tell you whether it is a 'fake'!
pavel.mp4
PERFORM DEEPFAKE DETECTION
BWS Deepfake Detection for 'pavel.mp4' says:
This was a fake video!

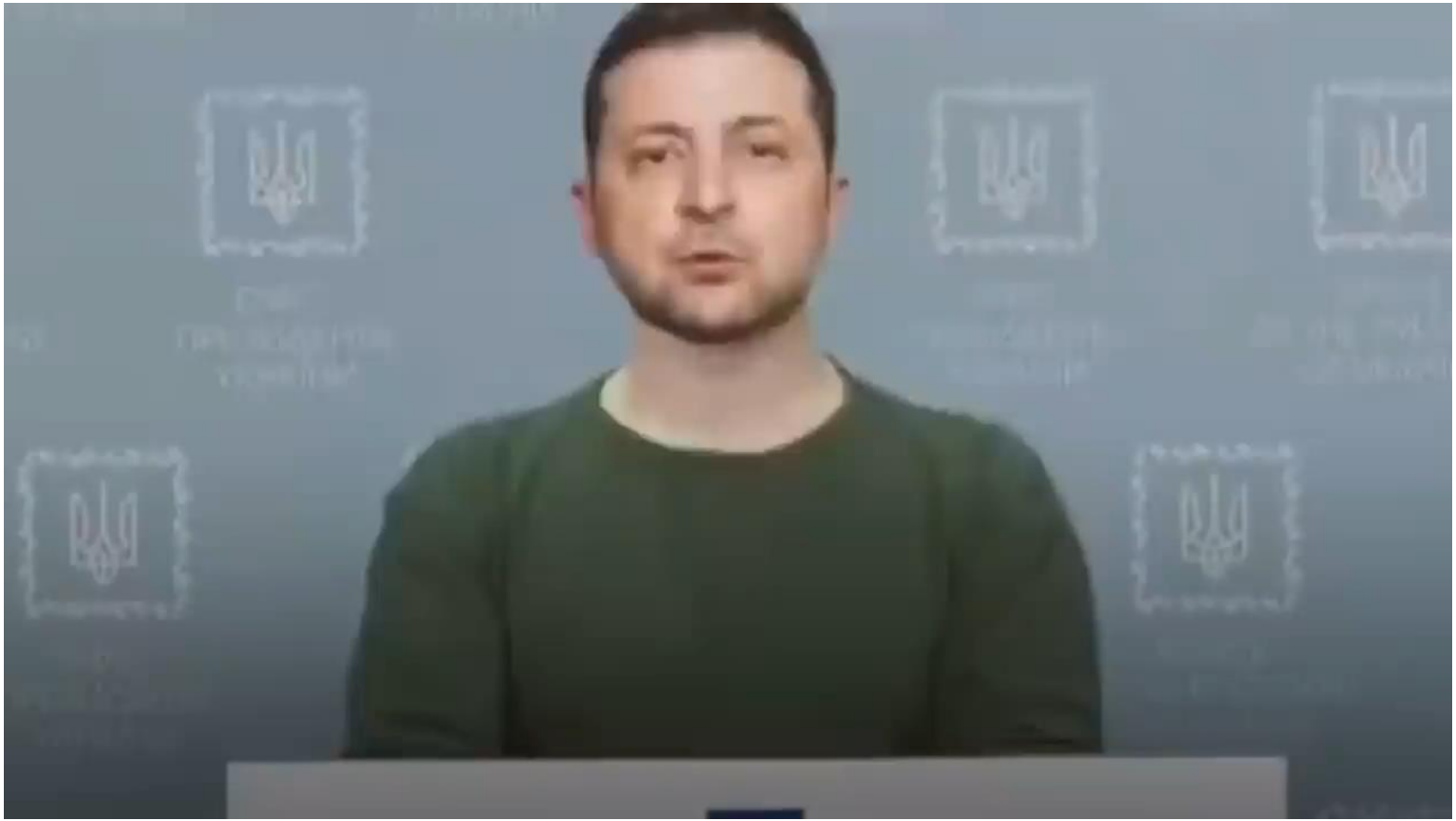
Deepfakes v bezpečnosti

Demonstrační deepfake ukázka vytvořená bez souhlasu vyobrazené osoby.

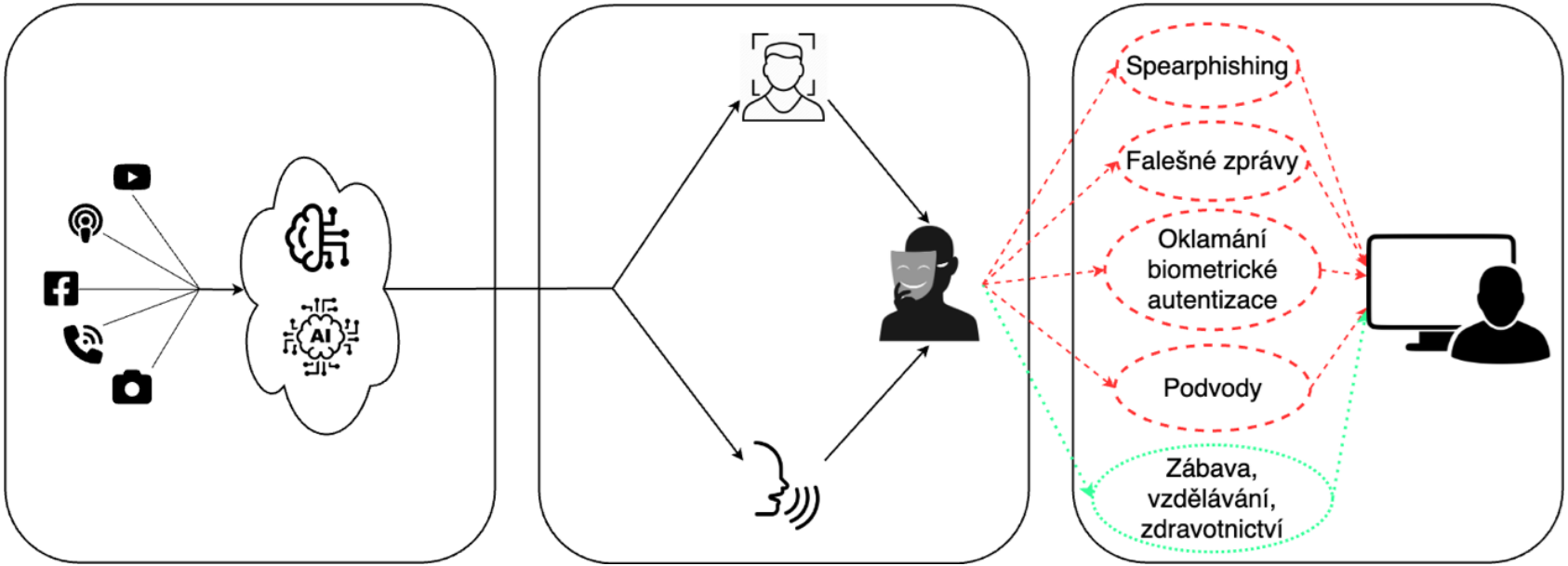

Simply upload a photo or a video (limited to 50 MB), the demo will tell you whether it is a 'fake'!
prezident.mp4
PERFORM DEEPFAKE DETECTION
BWS Deepfake Detection for 'prezident.mp4' says:
The video has been recorded from a live person!


Man created ‘deep
classmates using thei
HN.cz › Archiv
Miliardový e-shop se stal terčem deepfake útoku. Zaměstnanec si čtvrt hodiny volal s videokopií šéfa GymBeamu
Fake Biden robocall tells voters to skip New Hampshire primary election
idents used in Russia-
23 January 2024
Startup Shocked When 4Chan Immediately Abuses Its Voice-Cloning AI
Person who
has spoken
Vienna
Deepfaked Voice Enabled $35 Million Bank Heist in 2020

Jak to reálně vypadalo

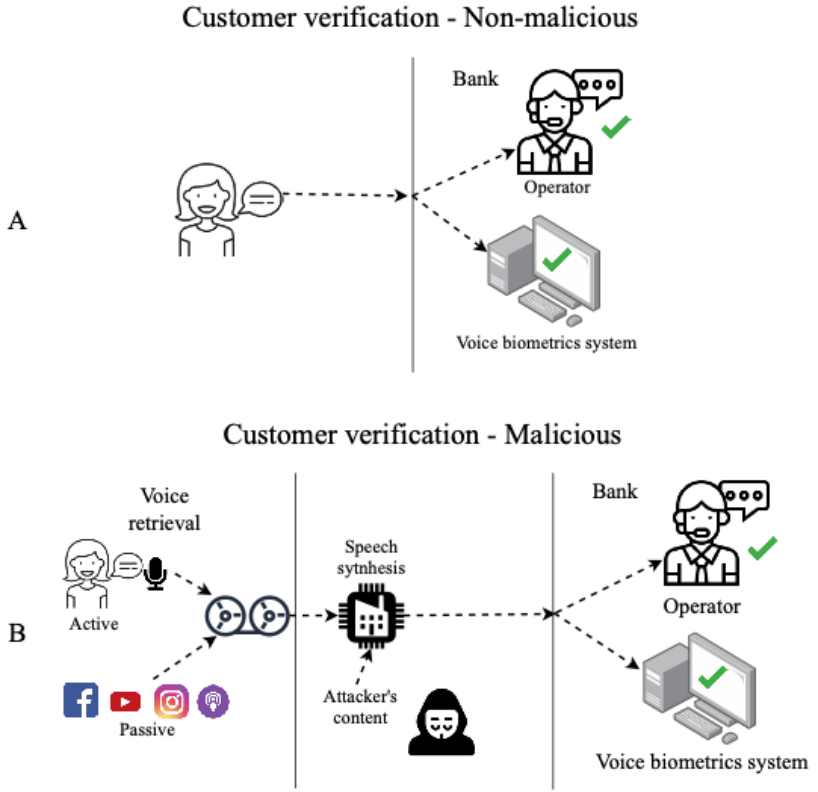


Model útočníka

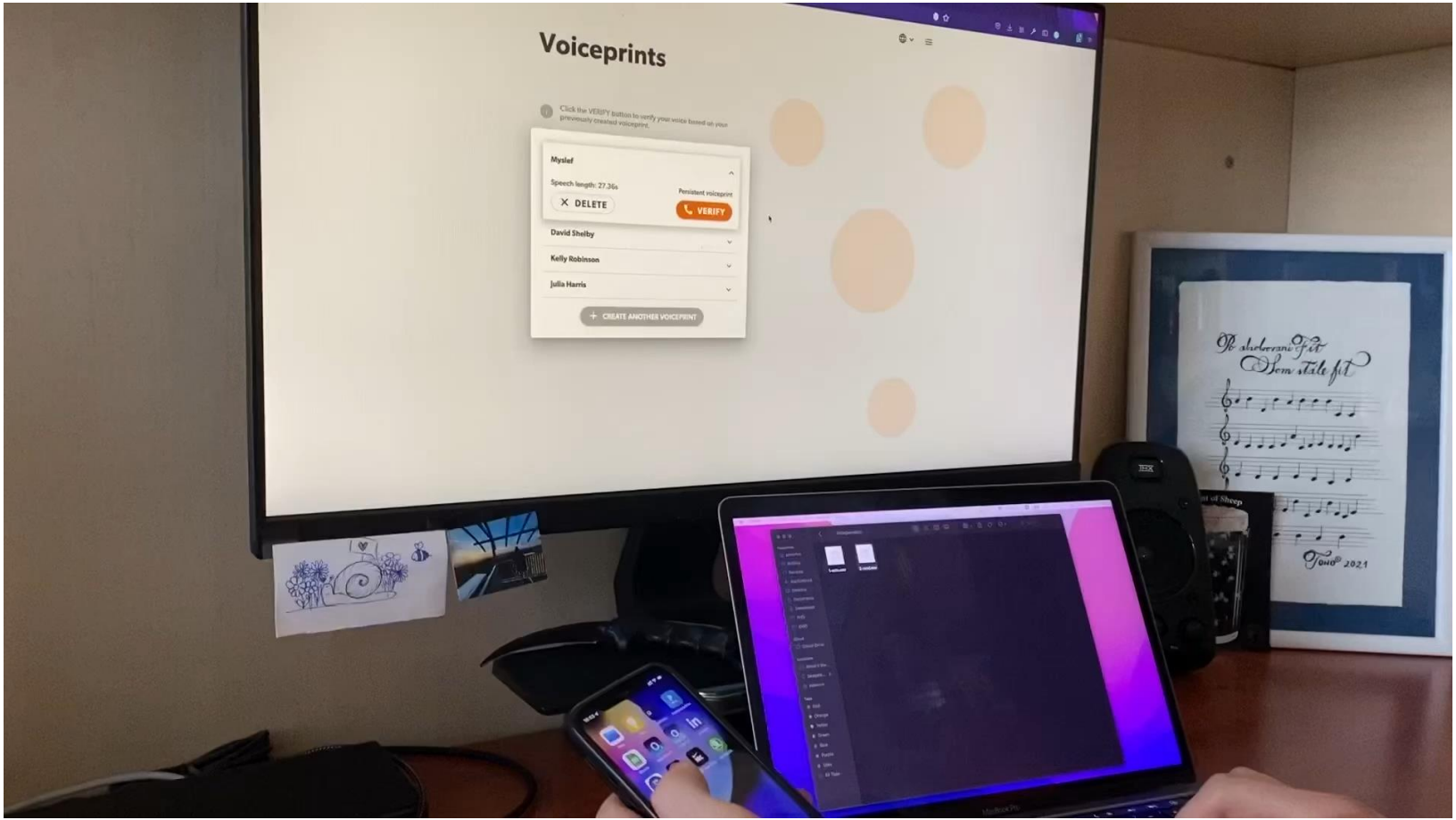


Útok v praxi – ověření identity zákazníka hlasem

- Dva faktory:
  - Hlasová biometrie
  - Lidský operátor
- 3 hlasové syntetizátory + 2 hlasové biometrie



Realizace útoku – ověření identity zákazníka hlasem



Jak to ve skutečnosti vypadá

Face-swap v reálné čase

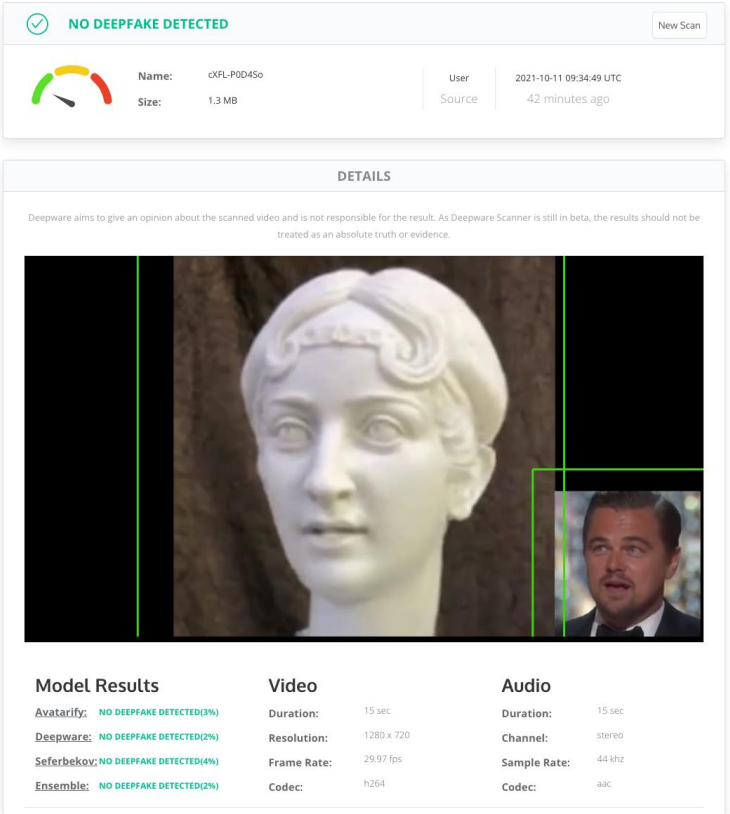
Lidský faktor - ověření identity zákazníka hlasem

- Lidé obecně nedokážou rozlišit mezi skutečnou a falešnou řečí.
  - Operátor není varován před vystavením deepfakes
- Pečlivě připravené deepfake nahrávky lidi oklamou
- Větší důraz je kladen na obsah konverzace, aby bylo možné identifikovat autenticitu řeči

96.8% lidí **neodhalí** deepfake v běžné konverzaci

[1] PRUDKÝ Daniel, FIRC Anton and MALINKA Kamil. Assessing the Human Ability to Recognize Synthetic Speech in Ordinary Conversation. In: 2023 International Conference of the Biometrics Special Interest Group (BIOSIG). Darmstadt, 2023.

Funguje to i s obličejem



- Blížíme se do stavu tvorby deepfakes v reálném čase (spíše už tam jsme..)
  - Např. filtry na MS Teams
- Deepfakes přinášejí nové rizika pro společnost
  - Biometrická autentizace
  - Social engineering
- Detekce deepfakes potřebná pro ochranu systémů
- Lidské rozpoznávaní deepfakes
- Zvyšující se kvalita a dostupnost
- S rozvojem jazykových modelů (ChatGPT apod.) se blíží riziko automatizace útoků

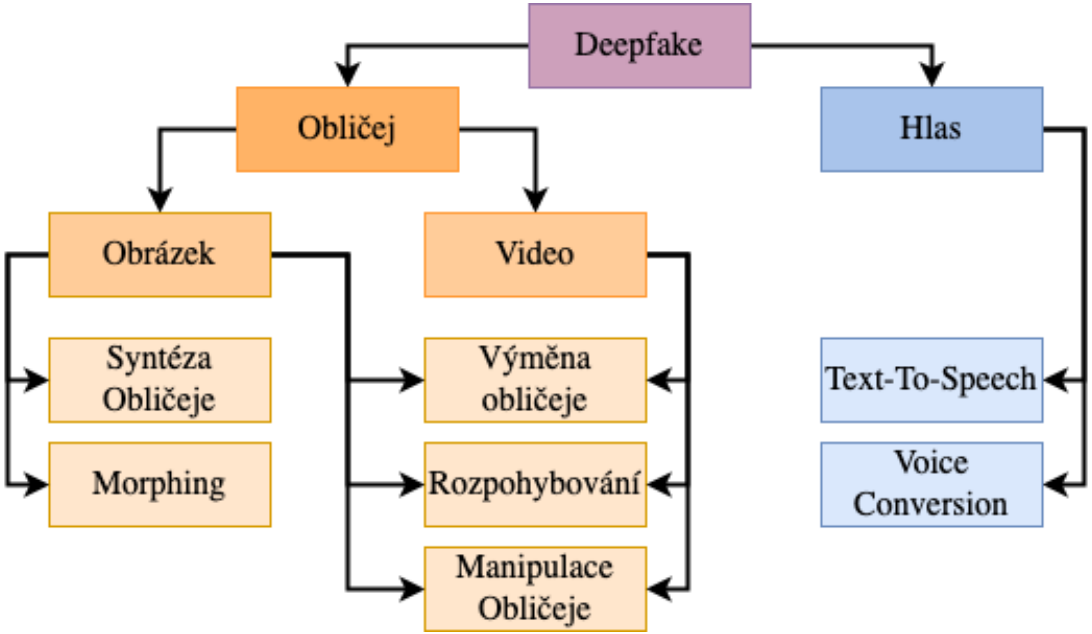
Budoucí? hrozby - je tento scénář vlastně stále nereálný?

- Útočník 1 (vše automatizovaně a se širokým pokrytím):
  - Stáhne fotky ze sociálních sítí
  - Spojí je s osobními údaji, kontakty atp.
  - Vytvoří pomocí face swap nahé fotky
  - Začne vydírat a hrozit zveřejněním (rozešle výhružku i s ukázkou)

- Útočník 2 (vše automatizovaně a se širokým pokrytím):
  - Vishingové útoky (chatbot vydávající se za osobu blízkou v nesnázích

Jaké existují deepfakes?

Kategorizace deepfakes



- Syntéza ne-existujících objektů
  - Lidi, zvířata, scenérie, auta…
- Trénink nad velkým množstvím obrázků stejné kategorie
- Náhodné generování z naučeného latentního prostoru



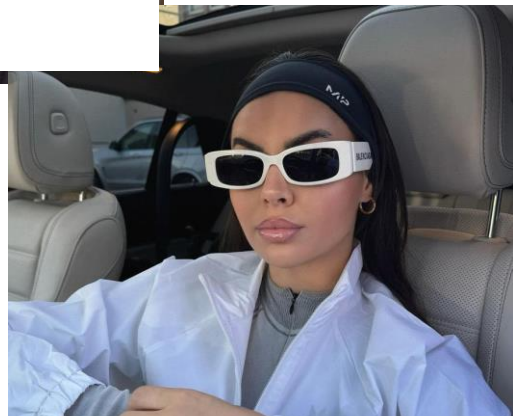

Zahrajme si hru: reálný nebo AI influencer?

1

2

3

4



5

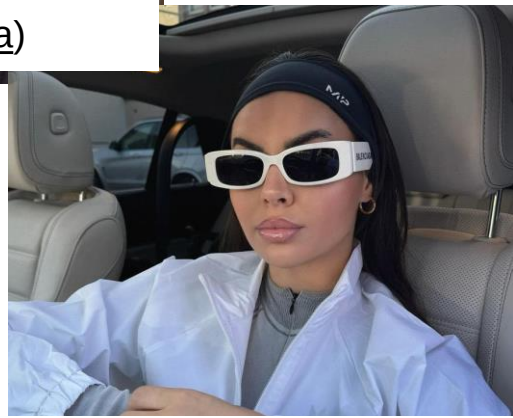
Zahrajme si hru: reálný nebo AI influencer?

1: Real (@beth.sal)

2: Fake (@fit_aitana)

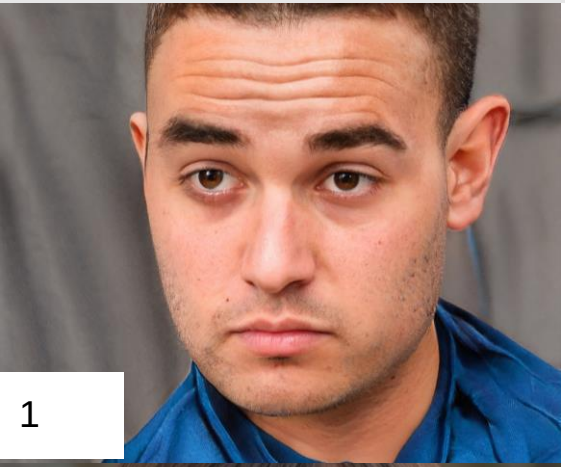
3: Fake (@bejby.blue)

4: Fake (@shudu.gram)

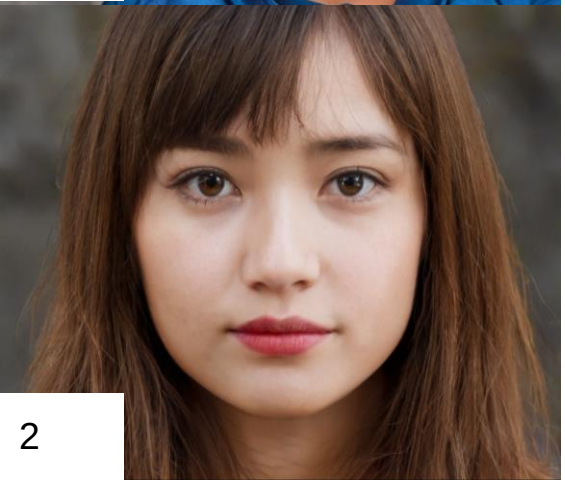


5: Real (@llnzz)

Zahrajme si hru…

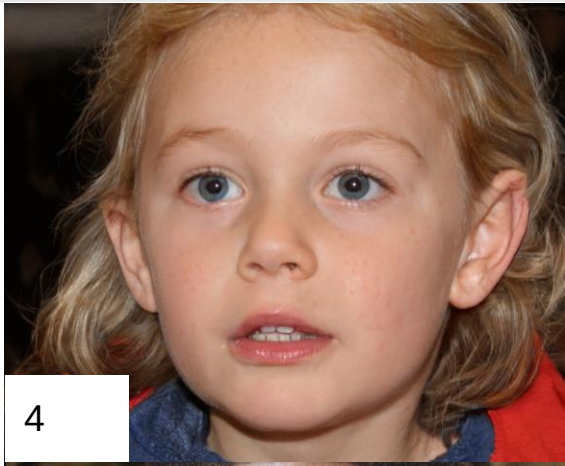


1



2

3

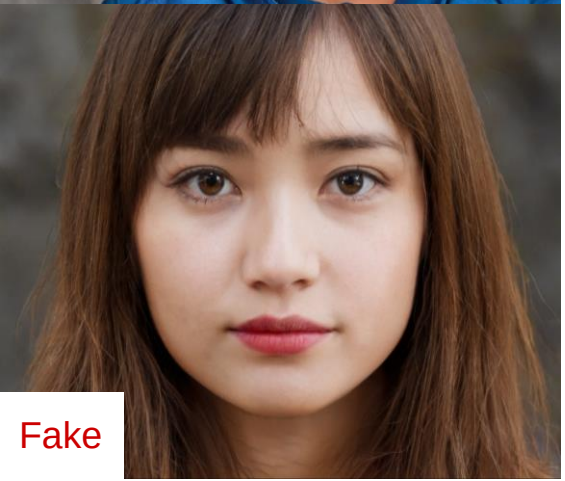


4

5

Zahrajme si hru…


Fake


Fake


Fake


Fake


Fake

Syntéza obrázků – vektory útoku

- Boti na sociálních sítích
  - Syntetické profilové obrázky = náročná detekce
  - Falešné Twitter účty šířící dezinformace v Belgii [1]
- Negativní tvrzení o identitě
  - “nemám přístup do systému, ale měl bych“

https://plus.rozhlas.cz/socialni-davky-zneuzivaji-bohati-7844454

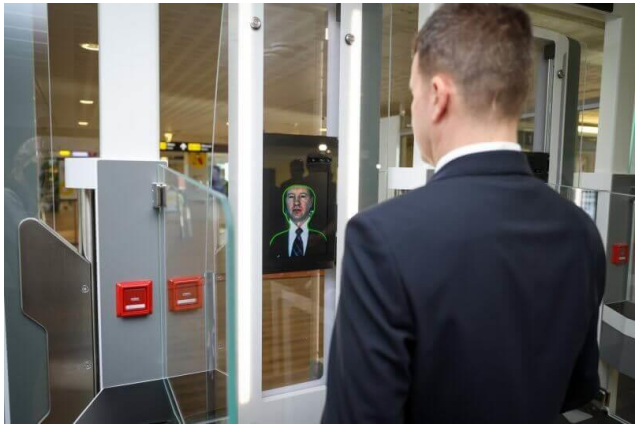
[1] Graphika Team, Fake Cluster Boosts Huawei, 2022.
https://graphika.com/reports/fake-cluster-boosts-huawei.

Morphing obličeje

- Morphing = efekt změny z jedného obrázku na druhý
- 2 obličeje -> jeden podobný oběma
- Nebezpečí pro automatizované hraniční kontroly identity

Morphing obličeje – vektory útoku

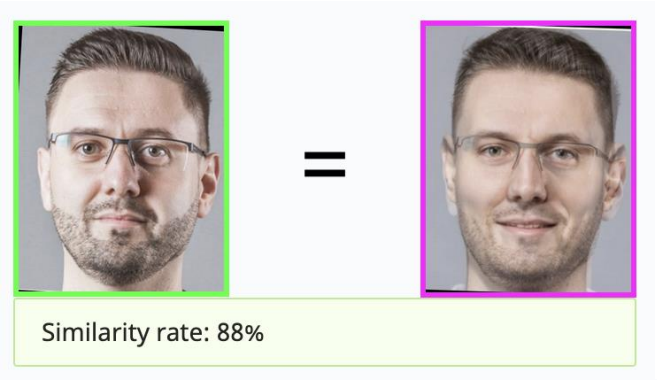
- Problém pro obličejové biometrie
  - Automatické hraniční kontroly
- Dva lidi, jeden občanský průkaz (pas)
- Incidenty hlásí Slovinská policie [1]
  - 40 incidentů
  - Zneužívání Slovinských pasů ne-EU občany pro cesty do Kanady

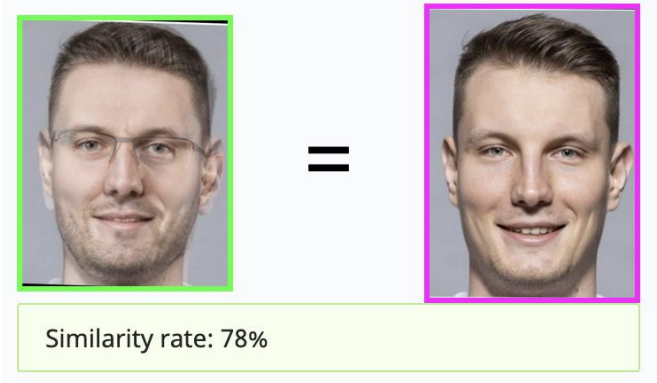


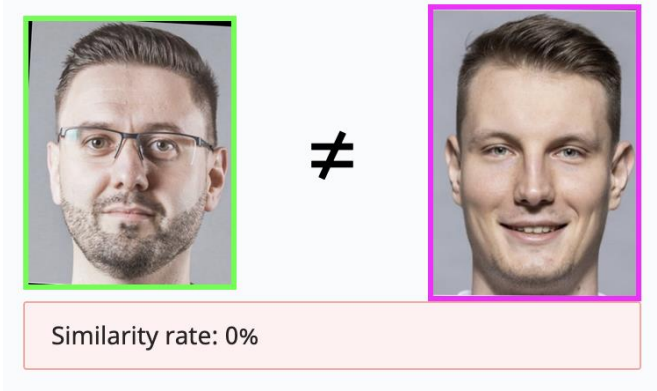
https://www.schengenvisainfo.com/news/estonia-implements-new-automated-border-control-gates-for-travellers/

[1] Chris Burt, Morphing attack detection for face biometric spoofs needs more generalization, datasets: Biometric Update. https://www.biometricupdate.com/202207/morphing-attack-detection-for-face-biometric-spoofs-needs-more-generalization-datasets, 2022.

- Jeden dokument = dva ”právoplatní” majitelé





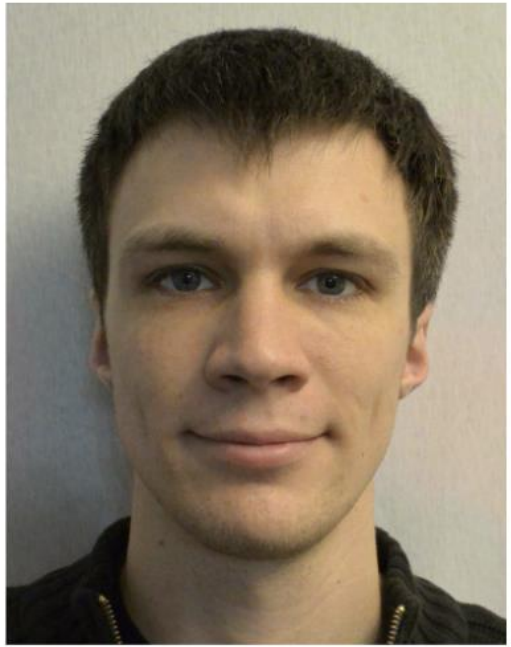


[1] https://faceapi.regulaforensics.com/

Jak to ve skutečnosti vypadá

(a) Subject 1

(b) Morph



(c) Subject 2

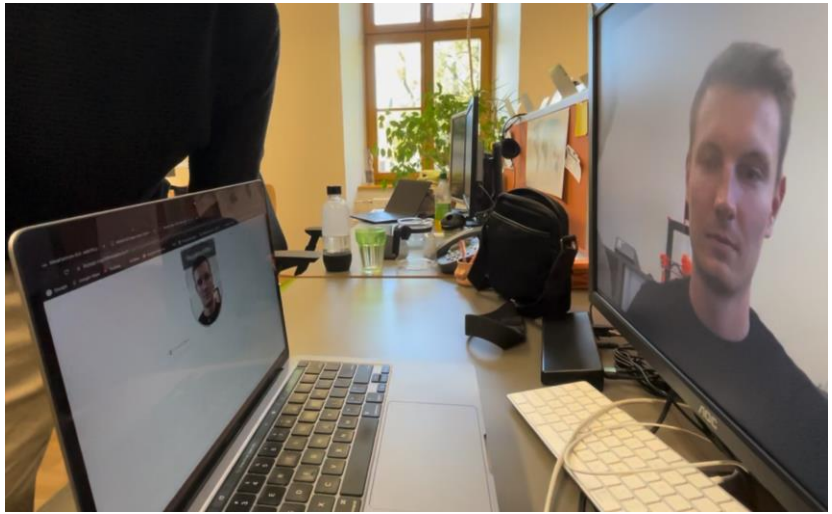
[1] U. Scherhag, C. Rathgeb, J. Merkle, R. Breithaupt and C. Busch, "Face Recognition Systems Under Morphing Attacks: A Survey," in IEEE Access, vol. 7, pp. 23012-23026, 2019, doi: 10.1109/ACCESS.2019.2899367.

Face swap - výměna obličeje

- Obličej ze zdrojové fotografie je přenesen na hlavu v cílové fotografii
- Obrázky a video

Face swap – vektory útoku

- Útok na biometrické systémy
- Krádež identity
- Pomluva
- Manipulace s důkazy
- Úprava pornografického materiálu
  - Sdílení upravených fotek s obličeji studentek v USA [1]
  - 98 % deepfake videí má pornografický obsah [2]
    - 99 % z nich je zaměřeno na ženy.



[1] Thaler, S. (2023, November 2). AI-generated nude images of girls at NJ high school trigger police probe: “I am terrified.” New York Post. https://nypost.com/2023/11/02/news/ai-generated-nudes-of-girls-at-nj-high-school-trigger-police-probe/
[2] Security Hero. (2023). STATE OF DEEPFAKES 2023. https://www.securityhero.io/state-of-deepfakes/

Jak to ve skutečnosti vypadá

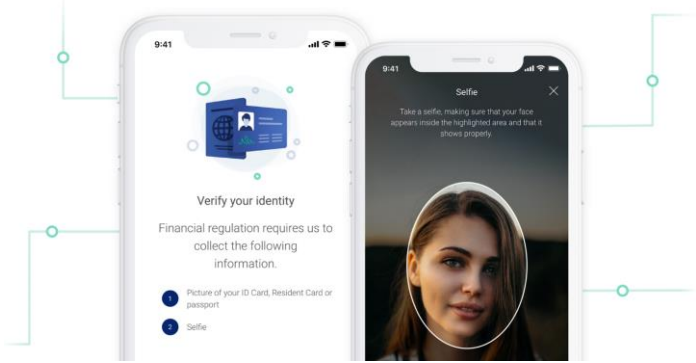
Face-swap v reálném čase

”Rozpohybování” (Reenactment)

- Foto-realistická re-animace obrázku podle zdrojového videa
- Generování vizuální informace a synchronizace audia a videa

Reenactment – vektory útoku

- Krádež identity
- Falešné zprávy
- Know Your Customer proces
  - Ověření identity v aplikaci na základě fotky dokladu a nasnímání obličeje
  - Možnost oklamat test živosti
- Únik na daních v hodnotě 76 mil. USD [1]
  - Nahrání falešných identit do daňového portálu v Číně



https://medium.com/2gether/this-is-why-the-kyc-know-your-customer-is-so-important-for-you-df3f5f1f4e29

[1] Masha Borak, Tax Scammers Hack Government-Run Facial Recognition System, 2021. https://www.scmp.com/tech/tech-trends/article/3127645/chinese-government-run-facial-recognition-system-hacked-tax.

Manipulace obličeje

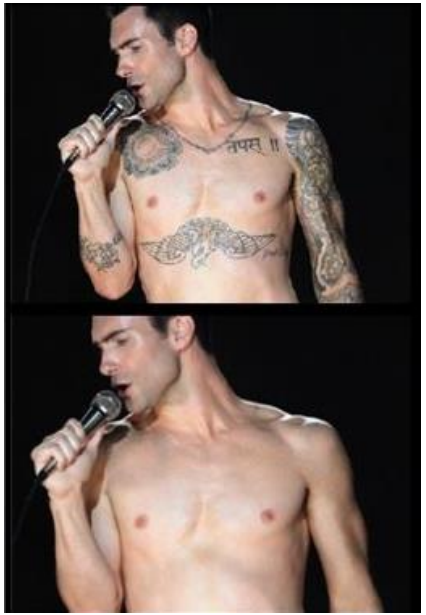
- Technika úpravy vybrané části obličeje v obrázku nebo na videu
- Přidání nebo odstranění vlasů, brýlí nebo tetování
- Úprava atributů v latentním prostoru a zpětná generace obličeje

Manipulace obličeje – vektory útoku

- Úprava atributů obličeje tak, aby osoba nebyla rozpoznána biometrickým systémem
- Manipulace důkazů u soudu

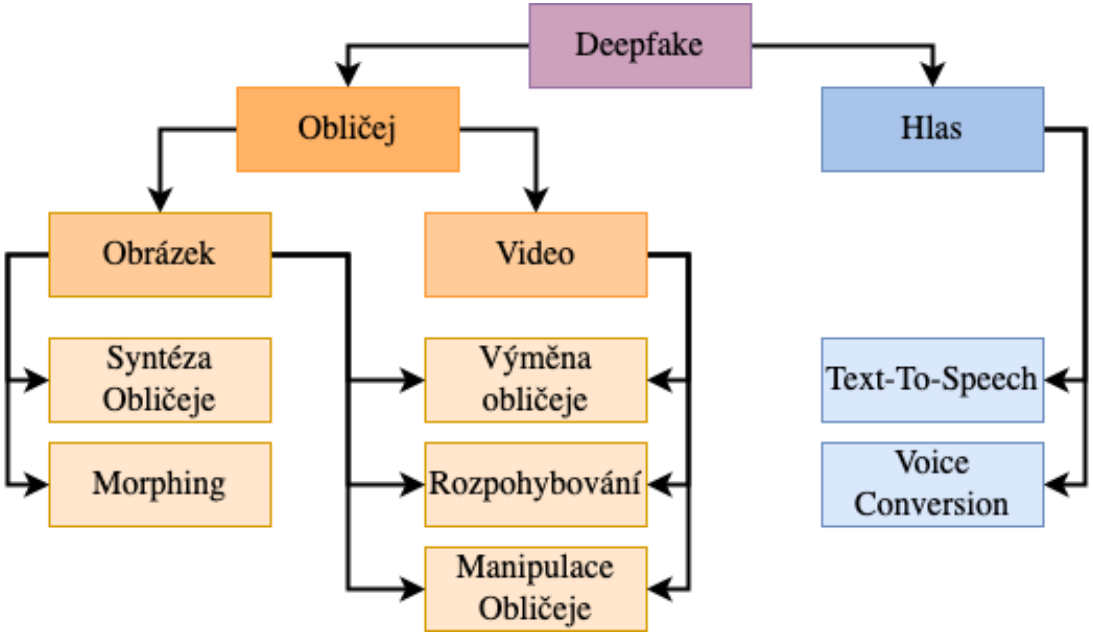
Původní obličej

Upravený obličej



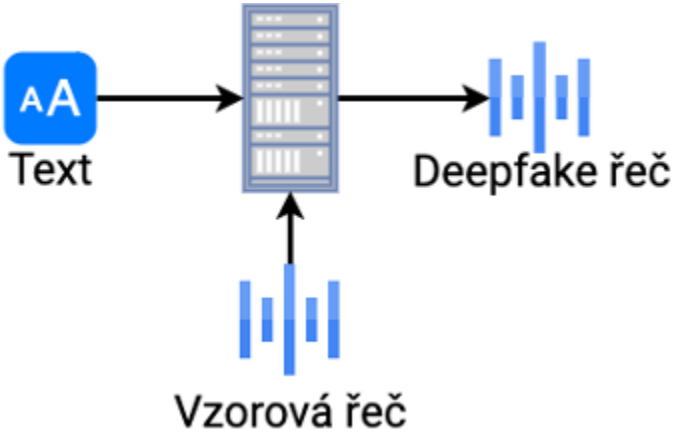
https://github.com/vijishmadhavan/SkinDeep

Kategorizace deepfakes



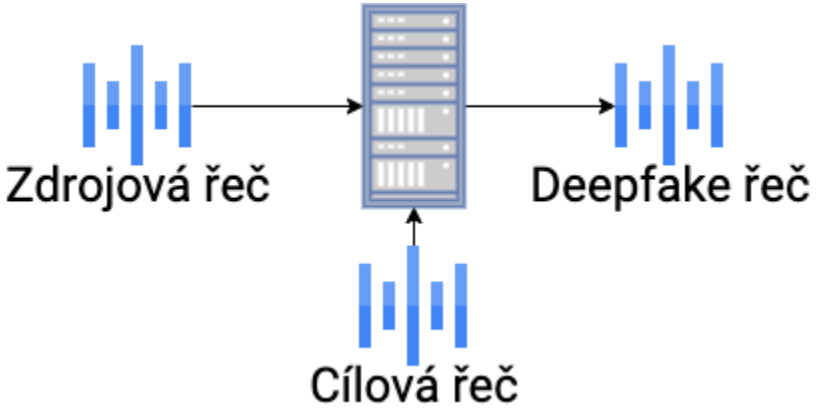
Syntéza: převod textu na řeč (Text-To-Speech)

- Převod textu na řeč v hlase ze vzorové nahrávky



Změna hlasu (Voice Conversion)

- Změna hlasu v zdrojové nahrávce na hlas z cílové nahrávky při zachování obsahu zdrojové nahrávky



Zahrajme si další hru…

Která nahrávka je pravá a která deepfake?

Můžeme?

Fake

Fake

Fake

Fake

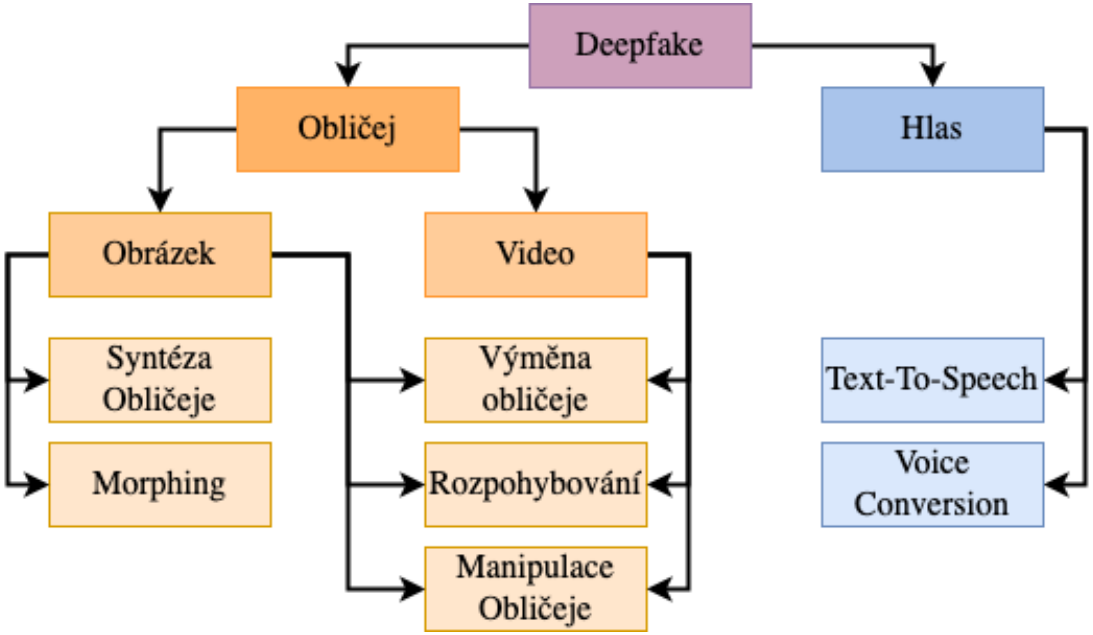
Hlasové deepfakes – vektory útoku

- Vhishing – voice phishing
- Krádež identity
- Pomluva
- Podvody
  - Krádež 250 tis. USD [1]
  - Krádež 36 mil. USD [2]

https://whatismyipaddress.com/vishing

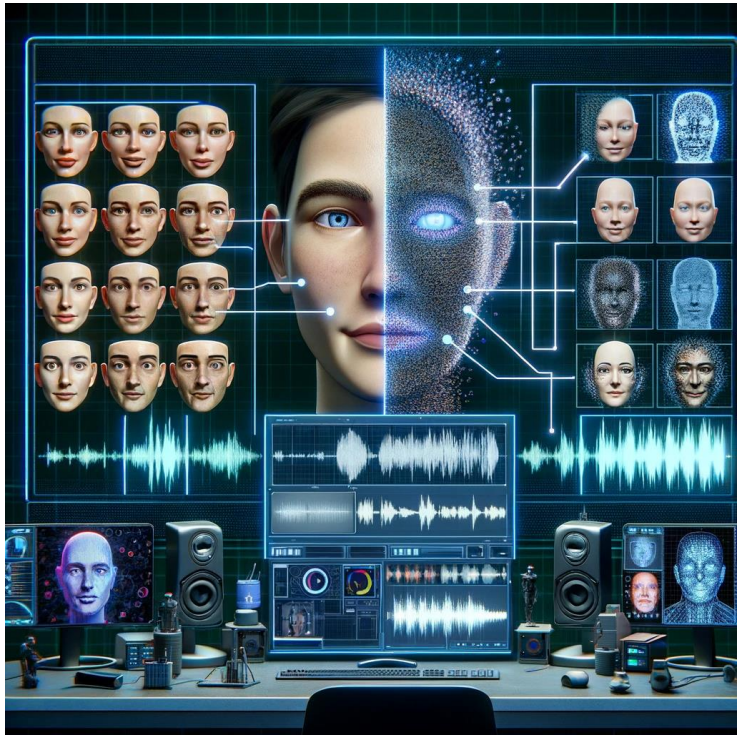
[1] Lindsey O'Donnell, CEO 'Deep Fake' Swindles Company Out of $243K. https://threatpost.com/deep-fake-of-ceos-voi ce-swindles-company-out-of-243k/147982/, 2019.
[2] Thomas Brewster, Fraudsters Cloned Company Director's Voice in $35 Million Bank Heist, Police Find, 2021. https://www.forbes.com/sites/thomasbrewster/2021/10/14/huge-bank-fraud-uses-deep-fake-voice-tech-to-steal-millions/.

Kategorizace deepfakes



Kombinace obraz a hlas

- Spojení obrazových a hlasových deepfakes
  - Např. Rozpohybování + změna hlasu
- Vysoká přesvědčivost



Kombinace obraz a hlas – vektory útoku

- Nové možnosti útoků
  - Online video konference
- Vyšší přesvědčivost
- Šíření fake news
- Krádež identity
- Podvody
  - Krádež $25 mil. za použití vícero deepfake identit [1]
  - GymBeam online meeting [2]

[1] Chen, H. (2024, Feb 4). Finance worker pays out $25 million after video call with deepfake ‘chief financial officer’. CNN. https://edition.cnn.com/2024/02/04/asia/deepfake-cfo-scam-hong-kong-intl-hnk/index.html
[2] Massayová, V. (2023, August 13). Podvodníci si trúfli na známu slovenskú firmu: V deepfake videohovore sa vydávali za šéfa. Startitup.sk. https://www.startitup.sk/podvodnici-si-trufli-na-znamu-slovensku-firmu-v-deepfake-videohovore-sa-vydavali-za-sefa/

Review article

Deepfakes as a threat to a speaker and facial recognition: An overview of tools and attack vectors

Anton Firc *, Kamil Malinka, Petr Hanáček

Brno University of Technology, Božetěchova 2, Brno, 612 00, Czech Republic

ARTICLE INFO

Keywords:
Face deepfakes
Speech deepfakes
Biometrics systems
Facial recognition
Speaker recognition
Deepfake detection
Cybersecurity

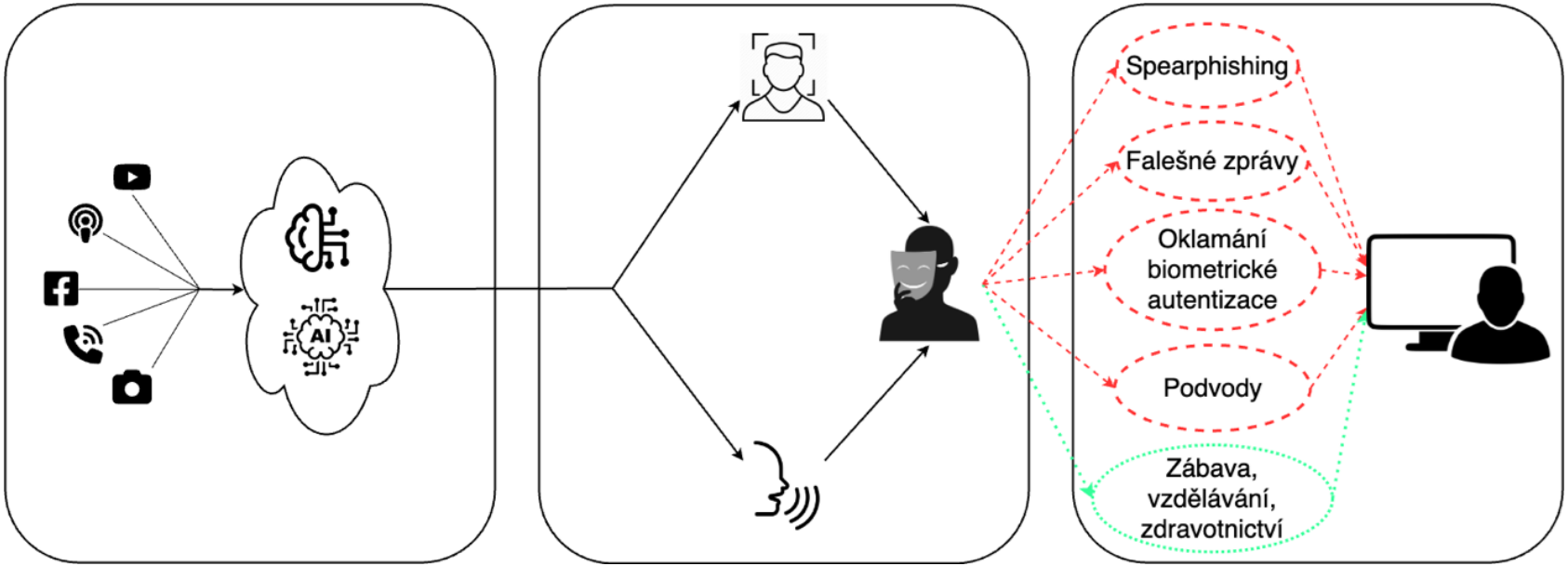
ABSTRACT

Deepfakes present an emerging threat in cyberspace. Recent developments in machine learning make deepfakes highly believable, and very difficult to differentiate between what is real and what is fake. Not only humans but also machines struggle to identify deepfakes. Current speaker and facial recognition systems might be easily fooled by carefully prepared synthetic media – deepfakes. We provide a detailed overview of the state-of-the-art deepfake creation and detection methods for selected visual and audio domains. In contrast to other deepfake surveys, we focus on the threats that deepfakes represent to biometrics systems (e.g., spoofing). We discuss both facial and speech deepfakes, and for each domain, we define deepfake categories and their differences. For each deepfake category, we provide an overview of available tools for creation, datasets, and detection methods. Our main contribution is a definition of attack vectors concerning the differences between categories and reported real-world attacks to evaluate each category's threats to selected categories of biometrics systems.

Metody ochrany

Dobré AI bojuje se zlým AI

Model útočníka



- Tři oblasti pro implementaci ochran
  - Získávání dat
    - Watermarking
    - Právní regulace
  - Tvorba deepfakes
    - Znemožnění tvorby
  - Užití deepfakes
    - Detekce
    - Dokazování autenticity
    - Vzdělávání

Table 1: Connection of mitigation methods to the areas in deepfake lifecycle. ***The● /○ symbol represents whether a method applies to a specific area.***

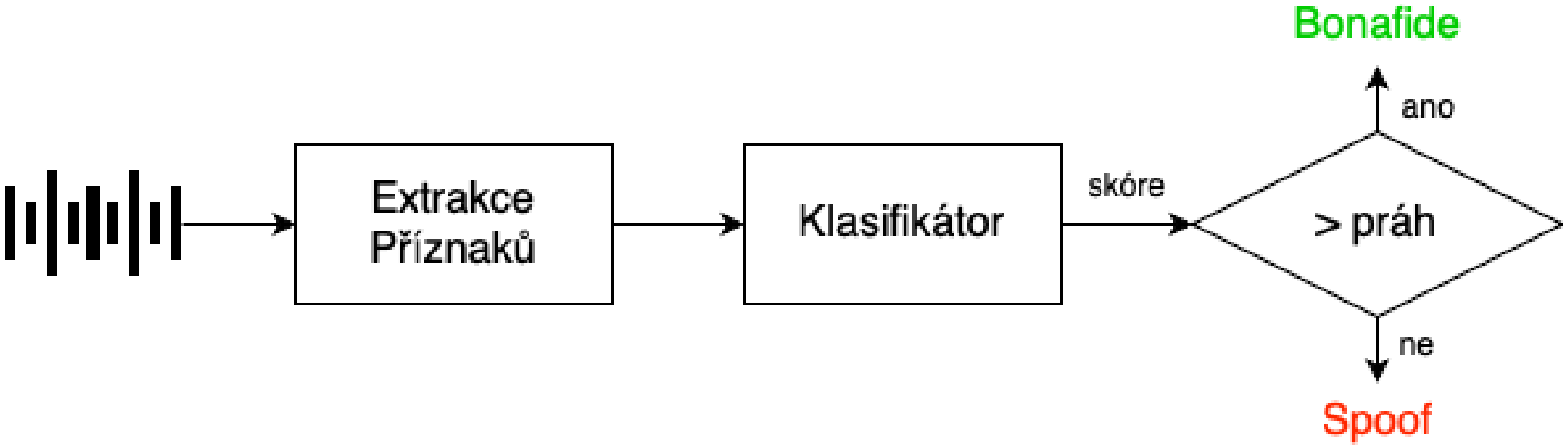
Method	*Area*		
	Preparation	*Creation*	*Usage*
Digital watermarking	●	●	●
Legal regulations	●	●	●
Obstructing deepfake creation	●	●	○
Deepfake detection	○	○	●
Forensics analysis	○	○	●
Proof of authenticity	○	○	●
Human education	○	○	●

Další možnost: prevence

- Omezení možností zneužít deepfakes
- Některé společnosti volí prevenci omezením služeb
  - Open AI:
    - ”Phasing out voice-based authentication as a security measure for accessing bank accounts and other sensitive information”
- Vhodné nastavení procesů snižuje možnost provedení útoku
  - Stejně platí pro social engineering útoky

Nejrozšířenější ochrana - detekce

Detekce deepfakes (audio)



- EER - bod kde APCER = BPCER
  - APCER = Attack Presentation Classification Error
    - Počet útočných (deepfake) vzorek klasifikovaných jako bonafide
  - BPCER = Bona fide Presentation Classification Error
    - Počet bonafide vzorek klasifikovaných jako útok (deepfake)
- Jak důvěryhodná je 99% přesnost detekce?

Dataset	EER (%)
ASVspoof 2015	0.09
ASVspoof 2019 LA	0.06
ASVspoof 2021 LA	0.56
ASVspoof 2021 DF	1.16
WaveFake	1.30
In-the-wild	3.10

ASVSpoof – co to je?

- *”Automatic Speaker Verification and Spoofing Countermeasures Challenge”*
- Nejznámější iniciativa zaměřená na vývoj metod pro detekci deepfake audia
- Každé 2 roky nová výzva
  - Sdružuje odborníky z oblasti
  - Zaměřená na vývoj nových (lepších) algoritmů pro detekci
    - ne systémů!
- Datasety z výzev jsou považovány za standard pro trénink a testování detektorů

Algoritmus, alebo detekční systém?

- Akademické publikace produkují algoritmy (metody)
  - Specifický postup nebo sada výpočtů, která slouží k analýze zvukových dat a rozpoznání pravosti
  - Testován a optimalizován v laboratorním prostředí
  - Potřeba upravit pro použití v praxi
- My ale potřebujeme detekční systémy
  - kompletní implementace detekčního algoritmu v reálném světě
  - nejen samotný algoritmus, ale i další komponenty, jako je rozhraní pro příjem a zpracování zvukových dat
- Algoritmus -> systém
  - Technologické a legislativní překážky

9. Do not use our Services in any manner contrary to ElevenLabs' policies, purpose or mission.

This includes:

k) Using any part of our Services or their Output as input for any machine learning or training of artificial intelligence models.

l) Using any part of our Services or their Output as part of a dataset that may be used for training, fine-tuning, developing, testing, or improving any machine learning or artificial intelligence technology.

9. Do not use our Services in any manner contrary to ElevenLabs' policies, purpose or mission.

This includes:

k) Using any part of our Services or their Output as input for any machine learning or training of artificial intelligence models.

l) Using any part of our Services or their Output as part of a dataset that may be used for training, fine-tuning, developing, testing, or improving any machine learning or artificial intelligence technology.

Jak na tom jsme?

Year	Only 2019 LA?	Frontend	EER (%) on 2019 LA
[1] 2022	Yes	waveform	0.64
[2] 2021	Yes	SincNet	0.83
[3] 2022	No (3DBs)	SincNet	0.93
[4] 2022	Yes	LFCC	0.98
[5] 2021	Yes	SincNet	1.06
[6] 2022	Yes	LFCC	1.06
[7] 2021	Yes	LFCC	1.07
[8] 2022	Yes	wav2vec 2.0	1.08
[9] 2022	No (4DBs)	wav2vec 2.0	1.28

Year	Only 2019 LA?	Frontend	EER (%) on 2019 LA
[10] 2024	Yes	wav2vec 2.0	0.39
[11] 2024	Yes	wav2vec 2.0	0.51
[12] 2024	Yes	Masked autoencoder	0.25
[13] 2024	No (3 DBs)	wav2vec 2.0	0.17
[14] 2024	No (7 DBs)	wav2vec 2.0	0.40
[15] 2024	No (3 DBs)	wav2vec 2.0	0.07
[16] 2024	No (3 DBs)	wavLM	0.65
[17] 2024	No (3 DBs)	wavLM	0.23
[18] 2024	No (5 DBs)	wav2vec 2.0	2.36

Jak na tom jsme?

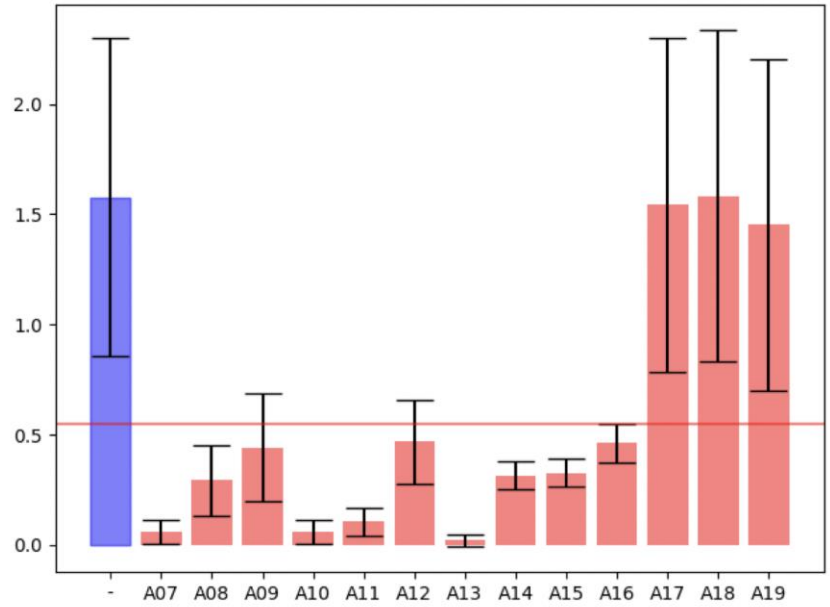
Year	Only 2019 LA?	Frontend	EER (%) on 2019 LA
[1] 2022	Yes	waveform	0.64
[2] 2021	Yes	SincNet	0.83
[3] 2022	No (3DBs)	SincNet	0.93
[4] 2022	Yes	LFCC	0.98
[5] 2021	Yes	SincNet	1.06
[6] 2022	Yes	LFCC	1.06
[7] 2021	Yes	LFCC	1.07
[8] 2022	Yes	wav2vec 2.0	1.08
[9] 2022	No (4DBs)	wav2vec 2.0	1.28

Year	Only 2019 LA?	Frontend	EER (%) on 2019 LA
[10] 2024	Yes	wav2vec 2.0	0.39
[11] 2024	Yes	wav2vec 2.0	0.51
[12] 2024	Yes	Masked autoencoder	0.25
[13] 2024	No (3 DBs)	wav2vec 2.0	0.17
[14] 2024	No (7 DBs)	wav2vec 2.0	0.40
[15] 2024	No (3 DBs)	wav2vec 2.0	0.07
[16] 2024	No (3 DBs)	wavLM	0.65
[17] 2024	No (3 DBs)	wavLM	0.23
[18] 2024	No (5 DBs)	wav2vec 2.0	2.36

Jediný systém s řešením biasu ticha v nahrávkách

Bližší pohled na ticho (shortcuts)

(c) This plot shows the average silence duration in seconds per audio file for the data split ‘eval’.



[1] Müller, N., Dieckmann, F., Czempin, P., Canals, R., Böttinger, K., Williams, J. (2021) Speech is Silver, Silence is Golden: What do ASVspoof-trained Models Really Learn? Proc. 2021 Edition of the Automatic Speaker Verification and Spoofing Countermeasures Challenge, 55-60, doi: 10.21437/ASVSPOOF.2021-9

Shortcut learning

- B. Chettri et al., “A Deeper Look at Gaussian Mixture Model Based Anti-Spoofing Systems,” ICASSP, 2018
  - *“Our analysis shows how system performance can depend on a simple class-dependent cue in the dataset: initial silence frames of zeros appear in the genuine signals but missing in the spoofed version.”*

- N. M. Müller et al, “Speech is Silver, Silence is Golden: What do ASVspoof-trained Models Really Learn?,” ASVspoof workshop, 2021
  - *“Results show that models trained on only the duration of the leading silence perform suspiciously well, and achieve up to 85% accuracy and an equal error rate (EER) of 15.1%”*
  - *“…we observe that even for established antispoofing models such as RawNet2, removing silence during training leads to comparatively worse performance.”*

- R. Geirhos et al, “Shortcut Learning in Deep Neural Networks,” in Nature Machine Intelligence, 2023
  - *“Shortcuts are decision rules that perform well on standard benchmarks but fail to transfer to more challenging testing conditions, such as real-world scenarios”*

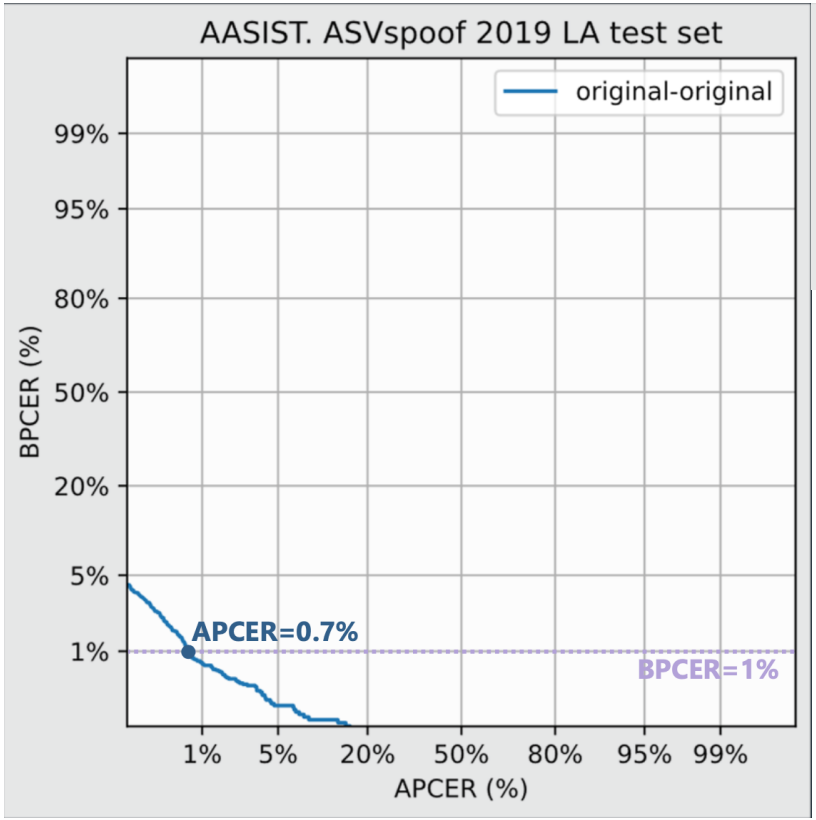
Co když zkratky odstraníme?

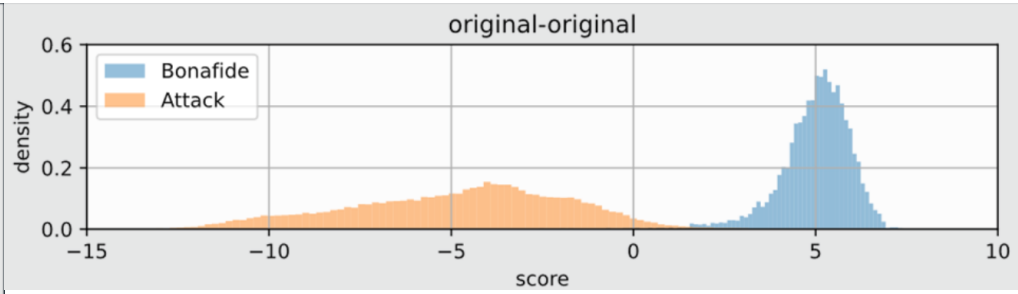
- AASIST: Audio Anti-Spoofing using Integrated Spectro-Temporal Graph Attention Networks
- SOTA v té době. EER=0,83 % na databázi ASVspoof 2019 LA
- Při trénování i inferenci systém používá počáteční 4s úsek vstupního zvuku, včetně počátečního ticha

Condition	Trained on	Tested on
original-original	Original data	Original data
original-trimmed	Original data	Leading/trailing silences removed
trimmed-trimmed	Leading/trailing silences removed	Leading/trailing silences removed

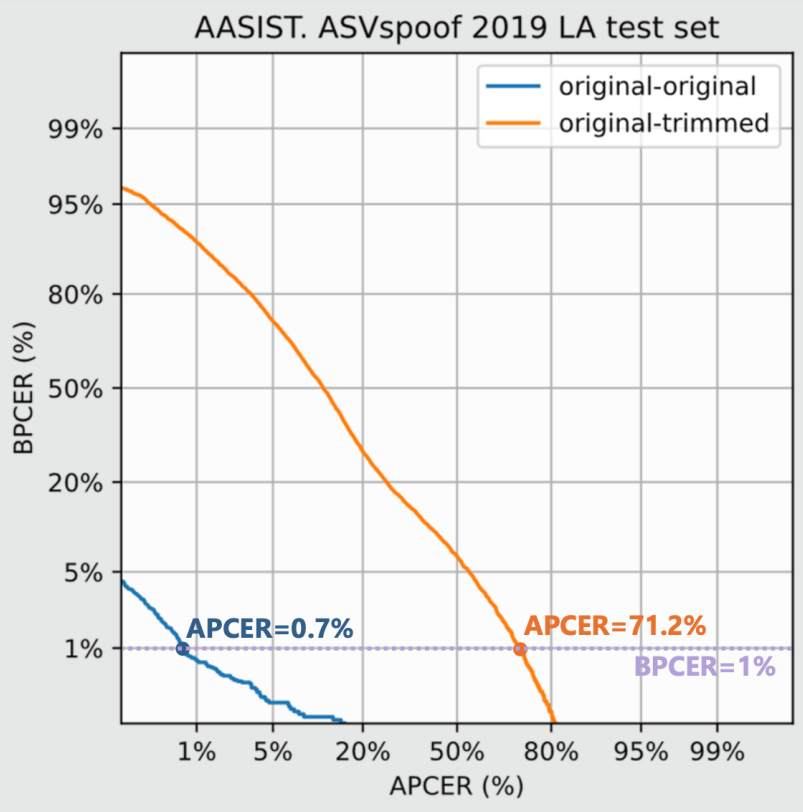
[1] J. -w. Jung et al., "AASIST: Audio Anti-Spoofing Using Integrated Spectro-Temporal Graph Attention Networks," ICASSP 2022

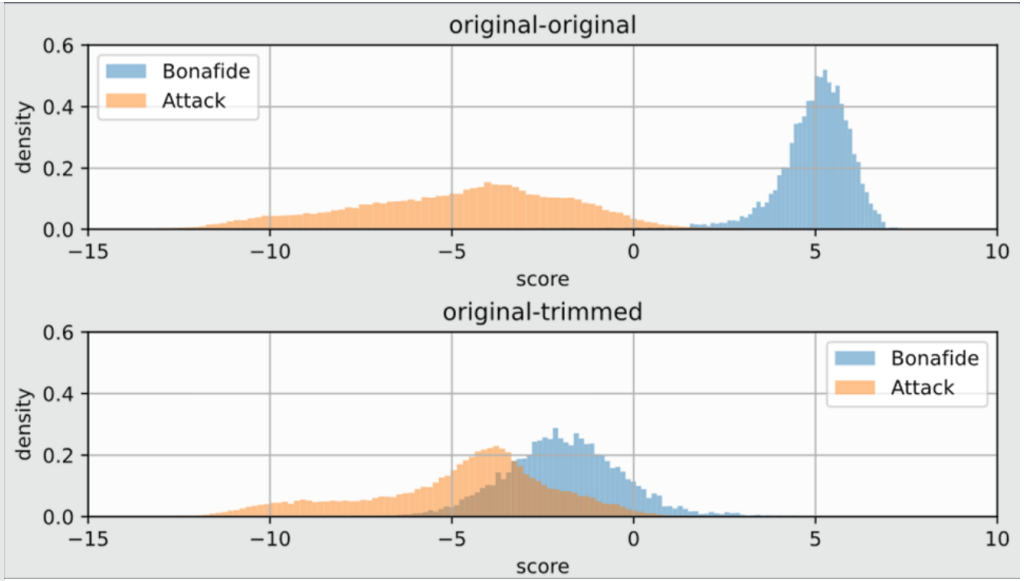
Co když zkratky odstraníme?



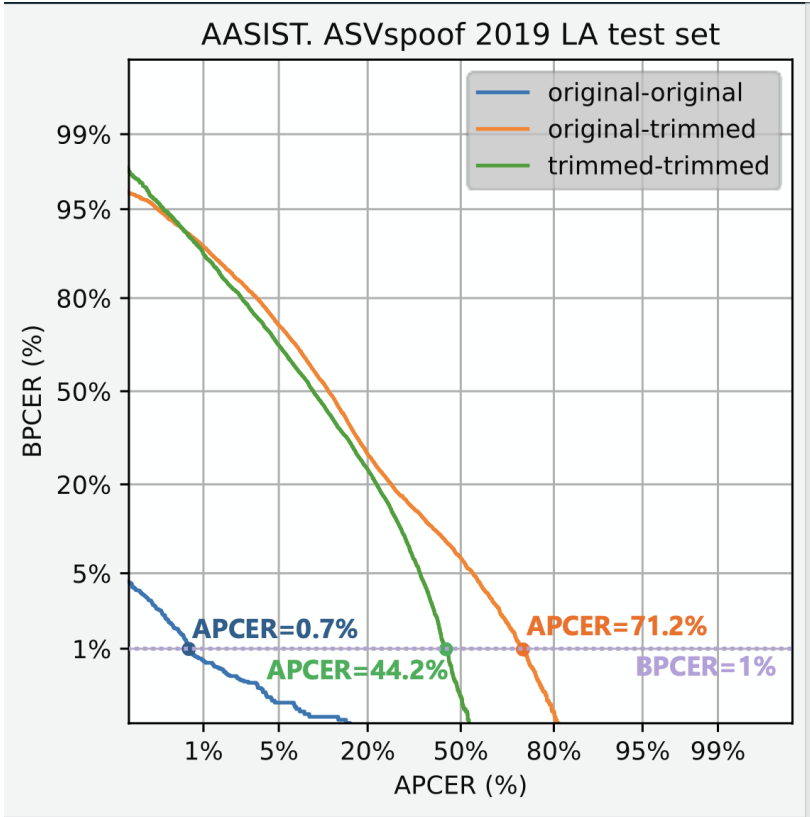


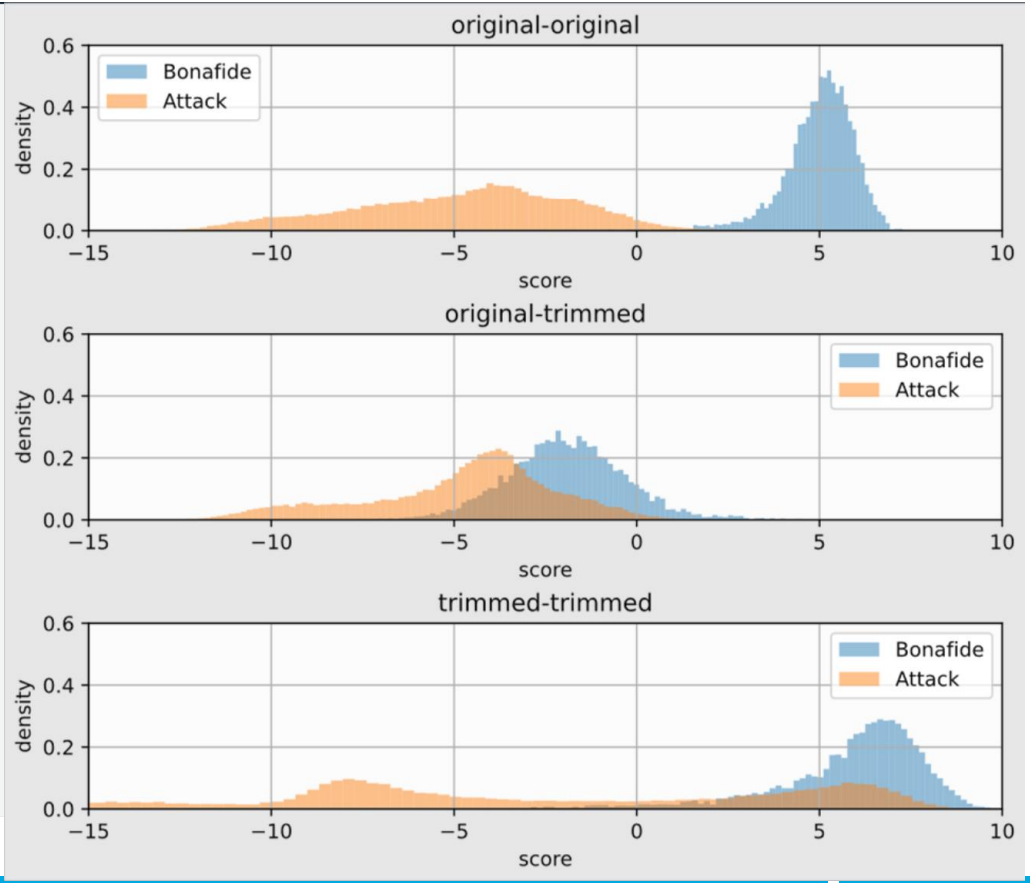
Co když zkratky odstraníme?





Co když zkratky odstraníme?

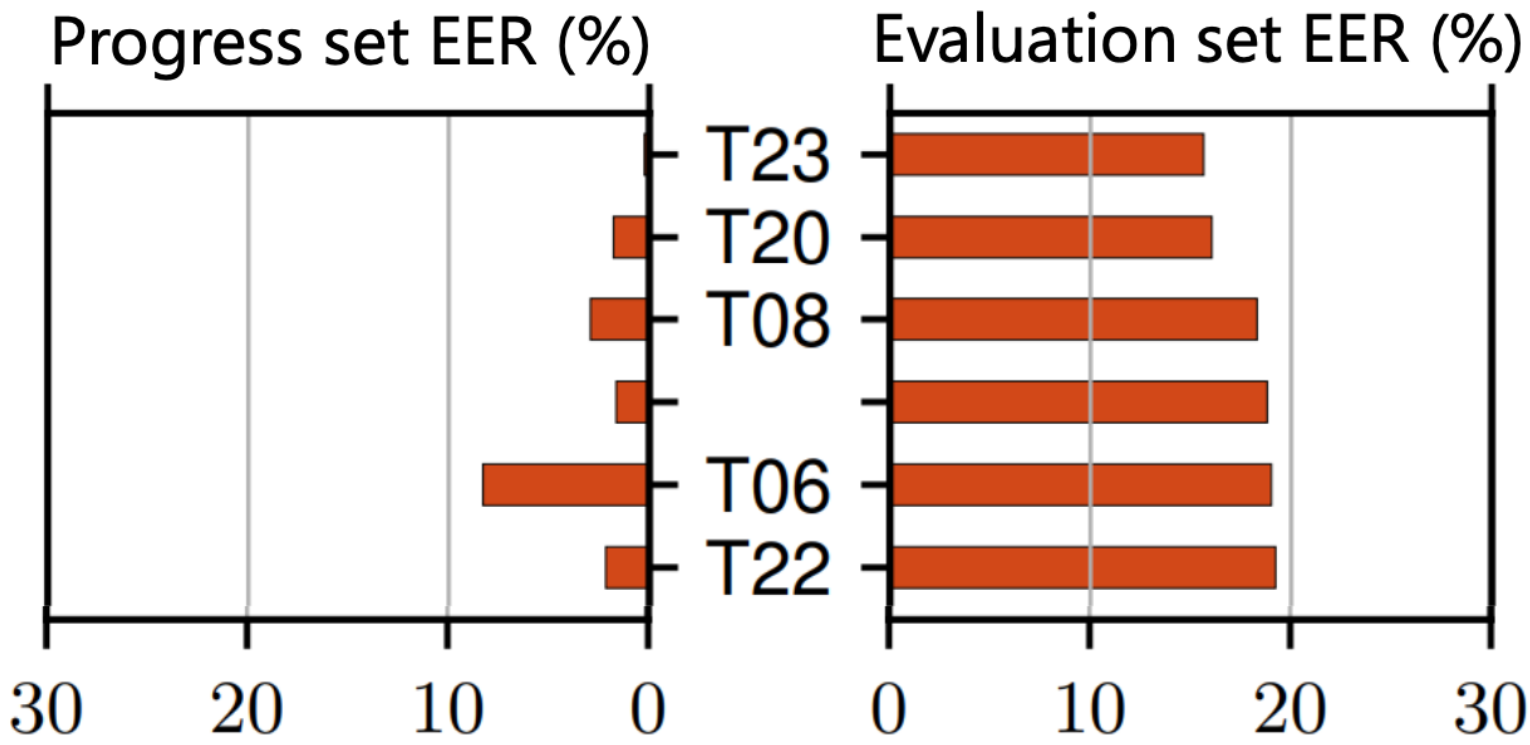




Co když zkratky odstraníme?

Dataset	EER (%)	Unbiased EER (%)
ASVspoof 2015	0.09	-
ASVspoof 2019 LA	0.06	2.36
ASVspoof 2021 LA	0.56	9.47
ASVspoof 2021 DF	1.16	5.55
WaveFake	1.30	-
In-the-wild	3.10	-

Problémem je i generalizace



Problémem je i generalizace

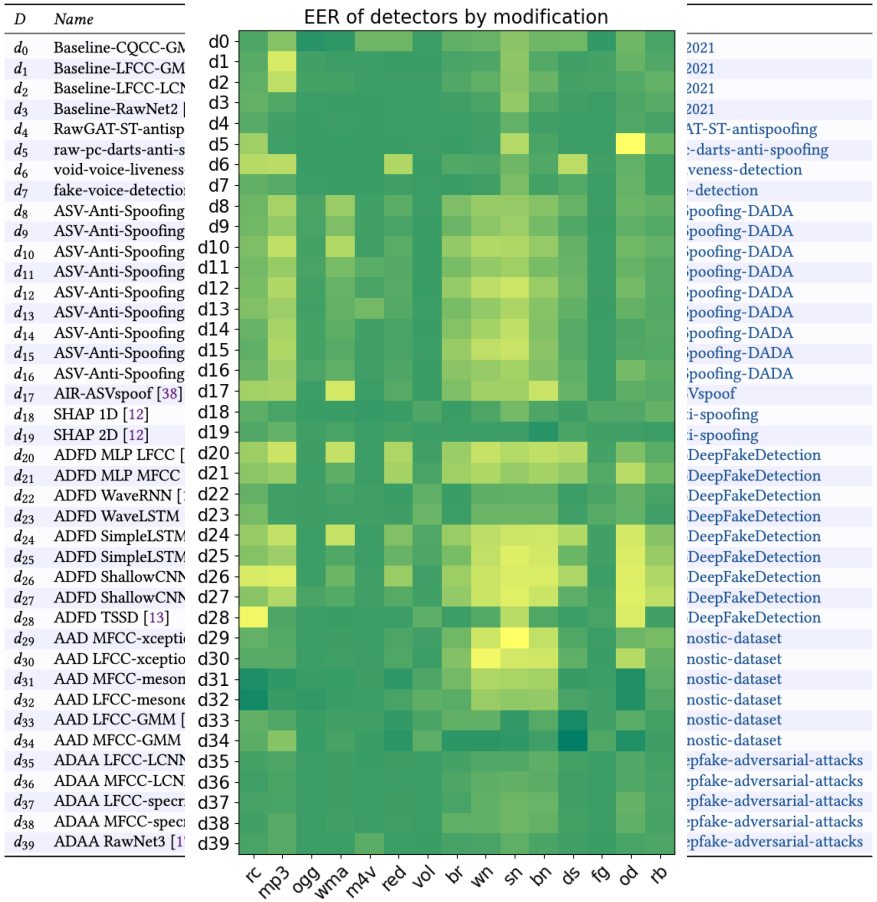
- Velké množství metod
- Slabá robustnost
- Těžká přenositelnost do praxe
- Jedno všeobecné řešení?

D	Name	Link
d_0	Baseline-CQCC-GMM [6]	https://github.com/asvspoof-challenge/2021
d_1	Baseline-LFCC-GMM [6]	https://github.com/asvspoof-challenge/2021
d_2	Baseline-LFCC-LCNN [6]	https://github.com/asvspoof-challenge/2021
d_3	Baseline-RawNet2 [6]	https://github.com/asvspoof-challenge/2021
d_4	RawGAT-ST-antispoofing [15, 31]	https://github.com/eurecom-asp/RawGAT-ST-antispoofing
d_5	raw-pc-darts-anti-spoofing [11]	https://github.com/eurecom-asp/raw-pc-darts-anti-spoofing
d_6	void-voice-liveness-detection [1]	https://github.com/chislab/void-voice-liveness-detection
d_7	fake-voice-detection [7]	https://github.com/dessa-oss/fake-voice-detection
d_8	ASV-Anti-Spoofing-DADA LightCNN [32]	https://github.com/JiJiJiang/ASV-Anti-Spoofing-DADA
d_9	ASV-Anti-Spoofing-DADA LightCNN DAT [32]	https://github.com/JiJiJiang/ASV-Anti-Spoofing-DADA
d_{10}	ASV-Anti-Spoofing-DADA LightCNN DADA [32]	https://github.com/JiJiJiang/ASV-Anti-Spoofing-DADA
d_{11}	ASV-Anti-Spoofing-DADA CGCNN [32]	https://github.com/JiJiJiang/ASV-Anti-Spoofing-DADA
d_{12}	ASV-Anti-Spoofing-DADA CGCNN DAT [32]	https://github.com/JiJiJiang/ASV-Anti-Spoofing-DADA
d_{13}	ASV-Anti-Spoofing-DADA CGCNN DADA [32]	https://github.com/JiJiJiang/ASV-Anti-Spoofing-DADA
d_{14}	ASV-Anti-Spoofing-DADA ResNet [32]	https://github.com/JiJiJiang/ASV-Anti-Spoofing-DADA
d_{15}	ASV-Anti-Spoofing-DADA ResNet DAT [32]	https://github.com/JiJiJiang/ASV-Anti-Spoofing-DADA
d_{16}	ASV-Anti-Spoofing-DADA ResNet DADA [32]	https://github.com/JiJiJiang/ASV-Anti-Spoofing-DADA
d_{17}	AIR-ASVspoof [38]	https://github.com/yzyouzhang/AIR-ASVspoof
d_{18}	SHAP 1D [12]	https://github.com/gewanying/shap-anti-spoofing
d_{19}	SHAP 2D [12]	https://github.com/gewanying/shap-anti-spoofing
d_{20}	ADFD MLP LFCC [13]	https://github.com/MarkHershey/AudioDeepFakeDetection
d_{21}	ADFD MLP MFCC [13]	https://github.com/MarkHershey/AudioDeepFakeDetection
d_{22}	ADFD WaveRNN [13]	https://github.com/MarkHershey/AudioDeepFakeDetection
d_{23}	ADFD WaveLSTM [13]	https://github.com/MarkHershey/AudioDeepFakeDetection
d_{24}	ADFD SimpleLSTM LFCC [13]	https://github.com/MarkHershey/AudioDeepFakeDetection
d_{25}	ADFD SimpleLSTM MFCC [13]	https://github.com/MarkHershey/AudioDeepFakeDetection
d_{26}	ADFD ShallowCNN LFCC [13]	https://github.com/MarkHershey/AudioDeepFakeDetection
d_{27}	ADFD ShallowCNN MFCC [13]	https://github.com/MarkHershey/AudioDeepFakeDetection
d_{28}	ADFD TSSD [13]	https://github.com/MarkHershey/AudioDeepFakeDetection
d_{29}	AAD MFCC-xception [16]	https://github.com/piotrkawa/attack-agnostic-dataset
d_{30}	AAD LFCC-xception [16]	https://github.com/piotrkawa/attack-agnostic-dataset
d_{31}	AAD MFCC-mesonet-inception (spectrogram) [16]	https://github.com/piotrkawa/attack-agnostic-dataset
d_{32}	AAD LFCC-mesonet-inception (spectrogram) [16]	https://github.com/piotrkawa/attack-agnostic-dataset
d_{33}	AAD LFCC-GMM [16]	https://github.com/piotrkawa/attack-agnostic-dataset
d_{34}	AAD MFCC-GMM [16]	https://github.com/piotrkawa/attack-agnostic-dataset
d_{35}	ADAA LFCC-LCNN [17]	https://github.com/piotrkawa/audio-deepfake-adversarial-attacks
d_{36}	ADAA MFCC-LCNN [17]	https://github.com/piotrkawa/audio-deepfake-adversarial-attacks
d_{37}	ADAA LFCC-specrnet [17]	https://github.com/piotrkawa/audio-deepfake-adversarial-attacks
d_{38}	ADAA MFCC-specrnet [17]	https://github.com/piotrkawa/audio-deepfake-adversarial-attacks
d_{39}	ADAA RawNet3 [17]	https://github.com/piotrkawa/audio-deepfake-adversarial-attacks

Table 2: Detection methods

Problémem je i generalizace

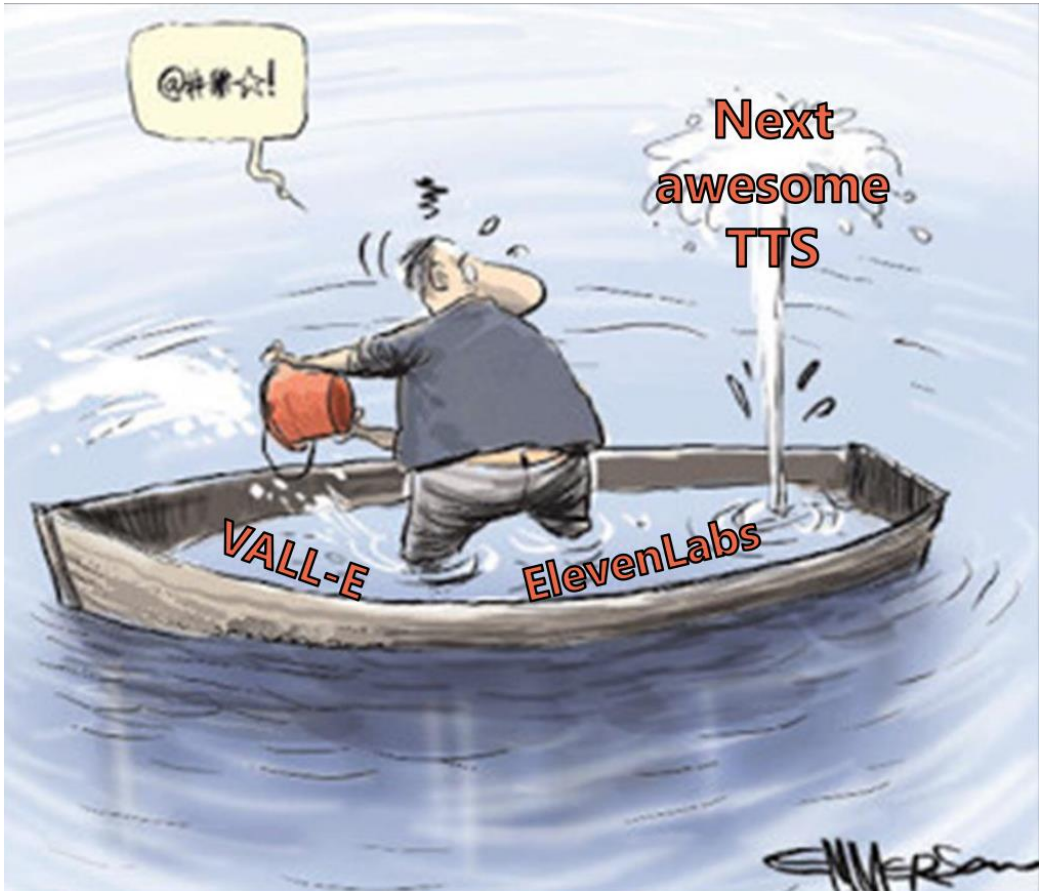
- Velké množství metod
- Slabá robustnost
- Těžká přenositelnost do praxe
- Jedno všeobecné řešení?



D	Name
d_0	Baseline-CQCC-GM
d_1	Baseline-LFCC-GM
d_2	Baseline-LFCC-LCN
d_3	Baseline-RawNet2 [
d_4	RawGAT-ST-antisp
d_5	raw-pc-darts-anti-s
d_6	void-voice-liveness
d_7	fake-voice-detectio
d_8	ASV-Anti-Spoofing
d_9	ASV-Anti-Spoofing
d_{10}	ASV-Anti-Spoofing
d_{11}	ASV-Anti-Spoofing
d_{12}	ASV-Anti-Spoofing
d_{13}	ASV-Anti-Spoofing
d_{14}	ASV-Anti-Spoofing
d_{15}	ASV-Anti-Spoofing
d_{16}	ASV-Anti-Spoofing
d_{17}	AIR-ASVspoof [38]
d_{18}	SHAP 1D [12]
d_{19}	SHAP 2D [12]
d_{20}	ADFD MLP LFCC [
d_{21}	ADFD MLP MFCC
d_{22}	ADFD WaveRNN [
d_{23}	ADFD WaveLSTM
d_{24}	ADFD SimpleLSTM
d_{25}	ADFD SimpleLSTM
d_{26}	ADFD ShallowCNN
d_{27}	ADFD ShallowCNN
d_{28}	ADFD TSSD [13]
d_{29}	AAD MFCC-xceptio
d_{30}	AAD LFCC-xceptio
d_{31}	AAD MFCC-meson
d_{32}	AAD LFCC-meson
d_{33}	AAD LFCC-GMM [
d_{34}	AAD MFCC-GMM
d_{35}	ADAA LFCC-LCNN
d_{36}	ADAA MFCC-LCN
d_{37}	ADAA LFCC-specr
d_{38}	ADAA MFCC-spec
d_{39}	ADAA RawNet3 [1

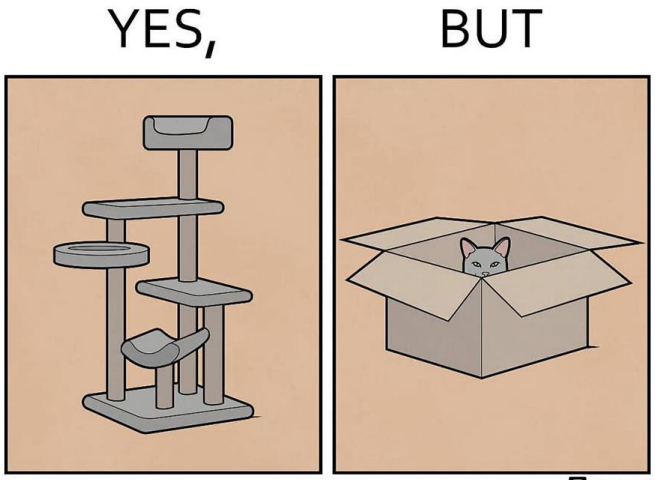
2021
2021
2021
2021
AT-ST-antispoofing
-darts-anti-spoofing
veness-detection
-detection
poofing-DADA
poofing-DADA
poofing-DADA
poofing-DADA
poofing-DADA
poofing-DADA
poofing-DADA
poofing-DADA
poofing-DADA
Vspoof
i-spoofing
i-spoofing
DeepFakeDetection
DeepFakeDetection
DeepFakeDetection
DeepFakeDetection
DeepFakeDetection
DeepFakeDetection
DeepFakeDetection
DeepFakeDetection
nostic-dataset
nostic-dataset
nostic-dataset
nostic-dataset
nostic-dataset
nostic-dataset
pfake-adversarial-attacks
pfake-adversarial-attacks
pfake-adversarial-attacks
pfake-adversarial-attacks
pfake-adversarial-attacks

Problémem je i generalizace



Detekce: ano, ale…

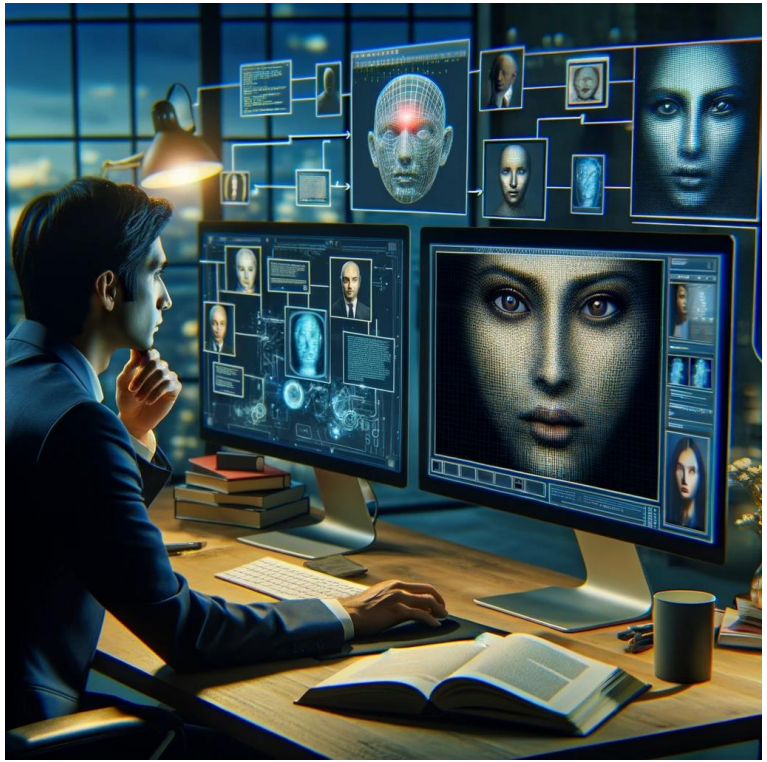
- Generalizace (a zkratky)
  - Nové typy útoků (nástrojů)
  - Variabilita dat
- Vysvětlitelnost
  - Proč?
- Certifikace
  - Jak moc věříme výsledkům?



https://www.boredpanda.com/yes-but-comics-society-criticism-anton-gudim/

Povědomí o deepfake technologii

- Lidi nedokážou rozpoznat pravé a deepfake média
- Možnost zmírnění hrozeb
  - Zvyšování povědomí o technologii deepfakes
  - Opakovaná expozice deepfake médiím zvyšuje rozeznávací schopnost
  - Možnost inspirace cvičnými phishing kampaněmi



Závěr

- Deepfakes představují hrozbu pro lidi a biometrické systémy
- Lidi nedokážou rozlišit deepfakes od pravých nahrávek
- Biometrické systémy nemají přirozenou ochranu před deepfakes
- Tvorba deepfakes je čím dál jednodušší a dostupnější
- Potřeba věnovat se kvalitním metodám pro detekci

Díky za pozornost!

Kamil Malinka, Anton Firc
malinka@fit.vut.cz, ifirc@fit.vut.cz

Security@FIT
https://www.fit.vut.cz/research/group/security@fit/.en

- Proč se nebát AI
  - AI pohledem nejen českých odborníků
  - Kapitola o bezpečnostních dopadech deepfakes
  - https://www.jota.cz/proc-se-nebat-umele-inteligence.html
- Česká Asociace Umělé Inteligence
  - Obranná strategie pro české firmy
  - https://asociace.ai/wp-content/uploads/2023/12/DEEPFAKE-2024-CAUI.pdf