

Project DefAI

Defending age assurance for Artificial Intelligence attacks

Trust and Safety Professionals Association
Europe Conference
May 17 2024



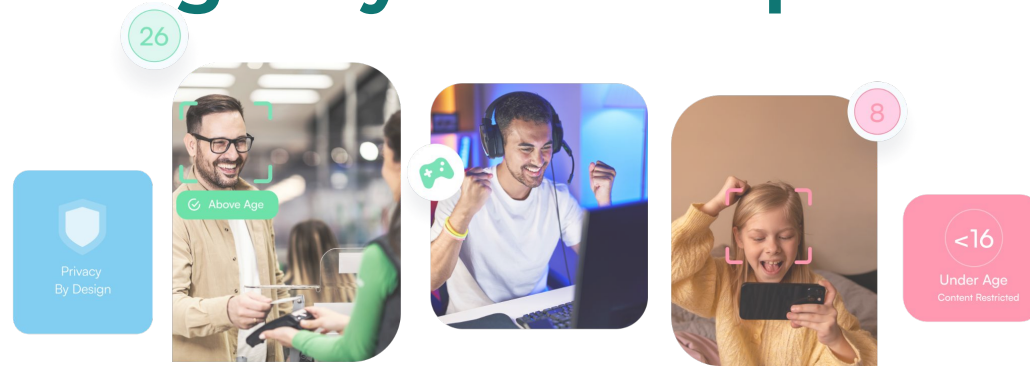
Takeaways



- Age assurance is increasing required globally
- The industry is improving privacy preserving measures, and delivering interoperability

BUT

- Spoofing age verification clear and present threat
- Defense measures current and developing
- Privacy-preserving ways remain possible



DefAI project scope



- **With a tsunami of global legislation demanding that the Internet becomes "age-aware", the demand for online age assurance will continue to grow exponentially.**
 - AI-based age estimation, with age discerned from biometric or behavioural features, is a critical means of age assurance, particularly for children and other vulnerable groups without documentation or records.
 - But these methods are increasingly becoming vulnerable to AI-generated attacks, and the perception of vulnerability to them, which could allow users to fake their age online and seriously undermine faith in age assurance technologies.
- **It is well established that biometric systems are vulnerable to presentation/spoofing attacks (PADs) and injection attacks (IADs).**
 - However in the last 12 months, increasingly sophisticated systems available on consumer devices and applications have begun to make it easier for anyone to conduct basic AI generated facial (deepfake) and audio spoof attacks on systems.
 - Indeed, there were over 80 deep fake apps found online in 2023 and they are observed to be improving in quality all the time.
- **Previous research has been conducted by our Swiss research partner, IDIAP, on various attack vectors and synthetic data, and by AVID on the standardization of testing.**
 - We want to build on this research and advance it to create defences against these attacks, not only for our Swiss Age Verification partner, Privately SA, but also to share the general lessons learnt with the AV industry as a whole through their trade association, the AVPA.

DefAI project scope



- **This project will develop technology identifying AI-generated presentation and injection attacks across emerging methods of age assurance, including facial analysis, voiceprint analysis and game play.**
 - It will then standardize methods to test the defences AV Providers implement against them.
- **ACCS's comprehensive testing procedures and state-of-the-art technology will protect the integrity clients of the industry as a whole from allowing inappropriate access to underage users and the legal, commercial and ethical consequences associated with that**
- **The AVPA consults with other providers in the industry to ensure a joined up response to this threat which could undermine credibility in all forms of online age assurance**
- **This new technology will help provide children enhanced protection from exposure to goods, content or services that may cause them or others harm, and comes at the perfect time as global regulatory changes mandate effective and accurate age assurance systems to protect children from harm.**

Goals



Scientific Goals

- Create a confidential best practice guide for members of the AVPA to advise them on their approach to defending against AI attacks on their solutions

Technological and Product Goals

- Develop tests to assure that Presentation Attack detection successfully detects at least 90% of the Presentation Attacks
- The testing process is sufficient for ISO and IEEE purposes
- Document contributions to the next versions of international standards for age assurance which specify the requirements to defend against AI attacks

Societal Goals

- Contribute to the appropriate international standards on presentation attack detection, such as ISO/IEC 27566, ISO 15408, and IEEE 2089.1
- Maintain confidence in age estimation and age verification solutions amongst the public-at-large, regulators, regulated services, policy-makers
- Maintain confidence in Age Assurance testing, audit and certification

Age assurance is increasingly required globally



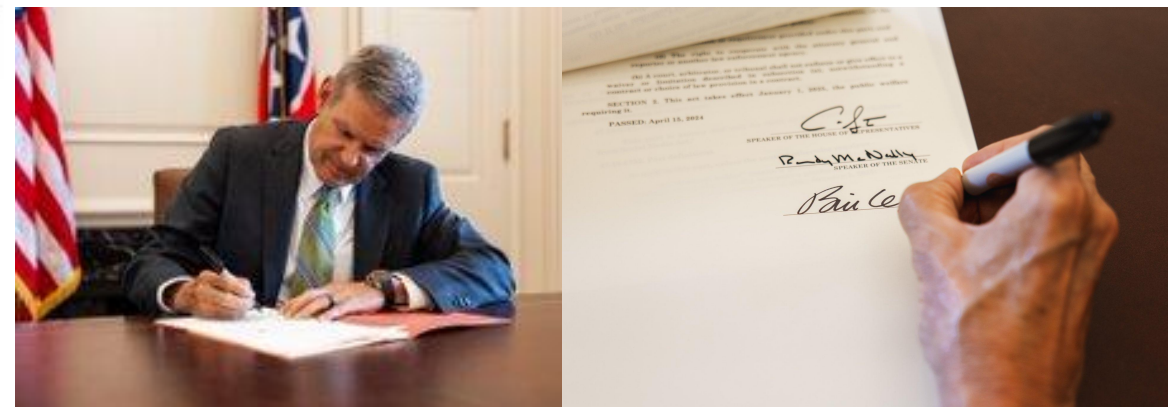
- EU
 - Digital Services Act
 - GDPR
 - Audio Visual Media Services Act
 - Code of Conduct for Age Appropriate Design
- UK
 - Online Safety Act
 - Age Appropriate Design Code
- Ireland
 - Age Appropriate
- France
 - SREN law
- Germany
 - Youth treaty
- India
 - Data protection legislation
- Australia
 - Pilot of age assurance modelled on euCONSENT
- Canada
 - Bill S-210
- USA
 - State laws

USA – state adult content laws



1. Louisiana	Louisiana Act 440 Louisiana HB 77	January 1, 2023 August 1, 2023
2. Utah	Utah SB 287	May 2, 2023
3. Mississippi	Mississippi SB 2346	July 1, 2023
4. Virginia	Virginia SB 1515	July 1, 2023
5. Arkansas	Arkansas SB 66	July 31, 2023
6. Texas	Texas HB 1181 Texas HB 18	September 1, 2023 September 1, 2024
7. Montana	Montana SB 544	January 1, 2024
8. North Carolina	North Carolina HB 8 North Carolina HB 534	January 1, 2024
9. Indiana	Indiana SB 17	January 1, 2024

10. Idaho	Idaho H 498	July 1, 2024
11. Kansas	Kansas SB 394 Kansas HB 2592	July 1, 2024
12. Georgia	Georgia SB 351	July 1, 2024
13. Kentucky	Kentucky HB 278	July 3, 2024
14. Nebraska	Nebraska LB 1092	July 18, 2024
15. Oklahoma	Oklahoma SB 1959	November 1, 2024
16. Alabama	Alabama HB 164	October 1, 2024
17. Florida	Florida HB 3	January 1, 2025



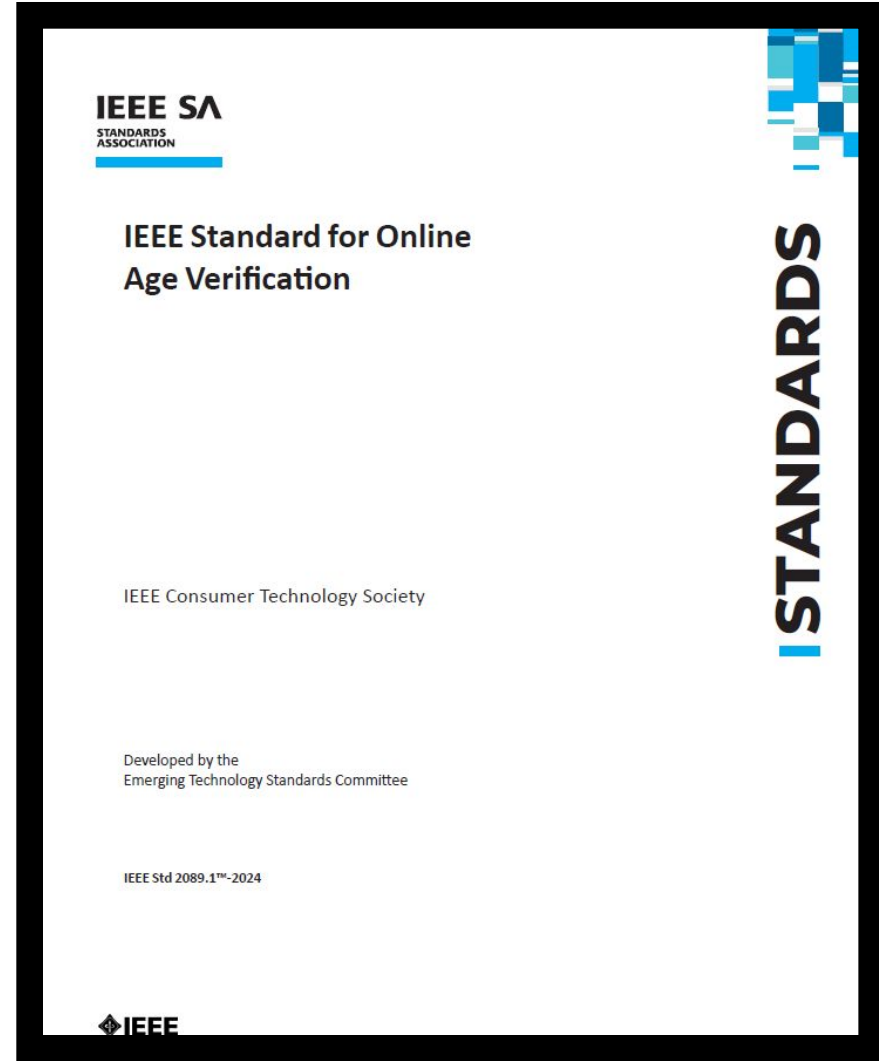


- **ISO/IEC CD1 27566-1 – Age assurance systems – Part 1: Framework**
- **ISO/IEC PWI 27566-2 – Age assurance systems – Part 2: Technical approaches and guidelines for implementation**
- **ISO/IEC AWI 27566-3 – Age assurance systems – Part 3: Benchmarks for benchmarking analysis**

IEEE 2089.1



- **Final proofing complete**
- **Ready for publication**
- **Sharing it with**
 - European Commission
 - ETSI taskforce
 - Canadian Standards
 - Ofcom/ICO



EU – ETSI Taskforce



- Two leaders from the Project DefAI team are representing euCONSENT on a small 6 person taskforce run by ETSI and commissioned by the Commission to determine on which technology Europe should standardise for age verification. It is formally "Human Factors (HF); Age Verification Pre-Standardization Study"

The study is due to last a year, and has three deliverables:

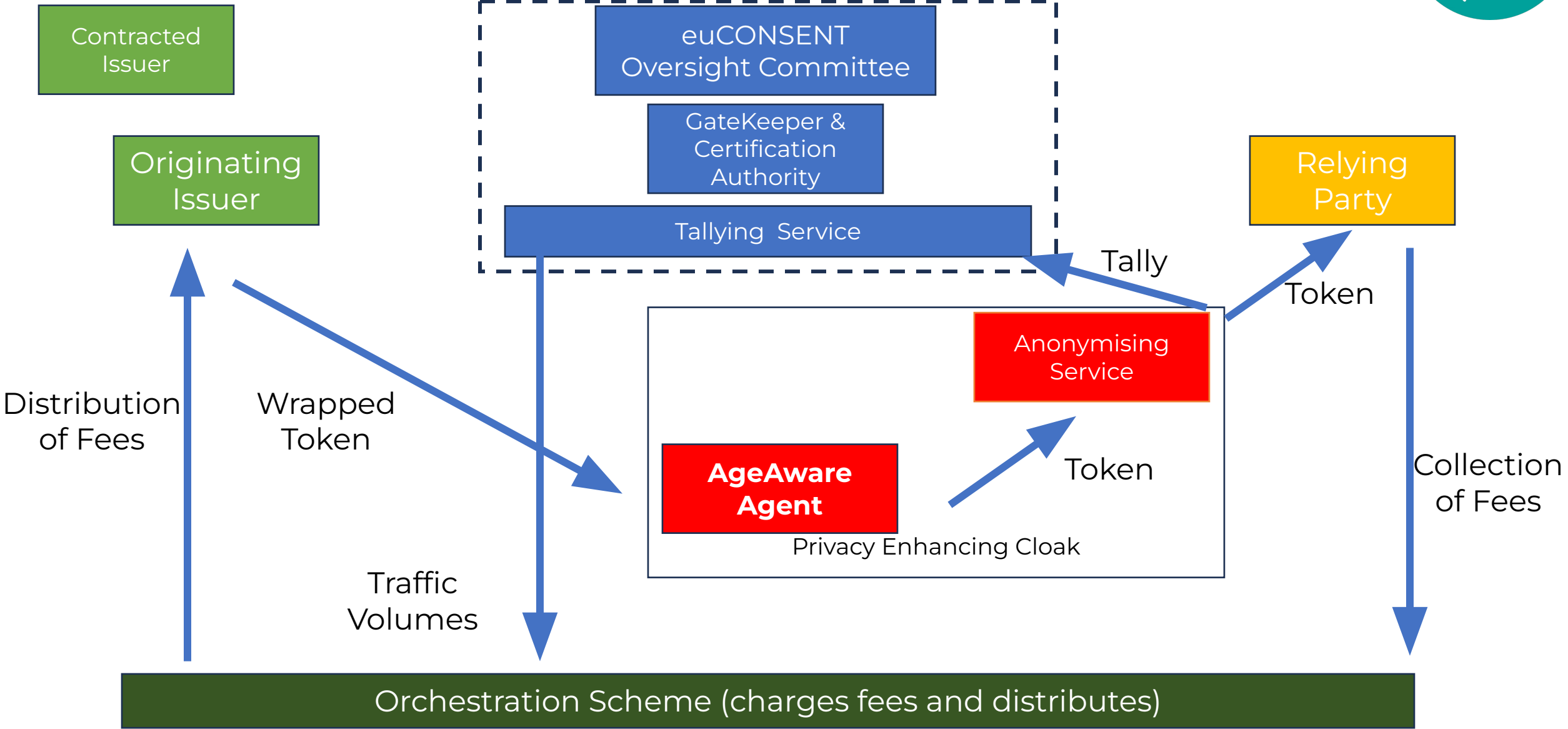
- Part 1: Stakeholder Requirements
 - Part 2: Solution and Standards Landscape
 - Part 3: Proposed Standardization Roadmap
- The emerging standards will need to address risk of deepfake and AI attacks.



AGE Verification Pre-Standardization Verify Age - 101162874 Description of the action (Part B)

(SMP STAND Standard)

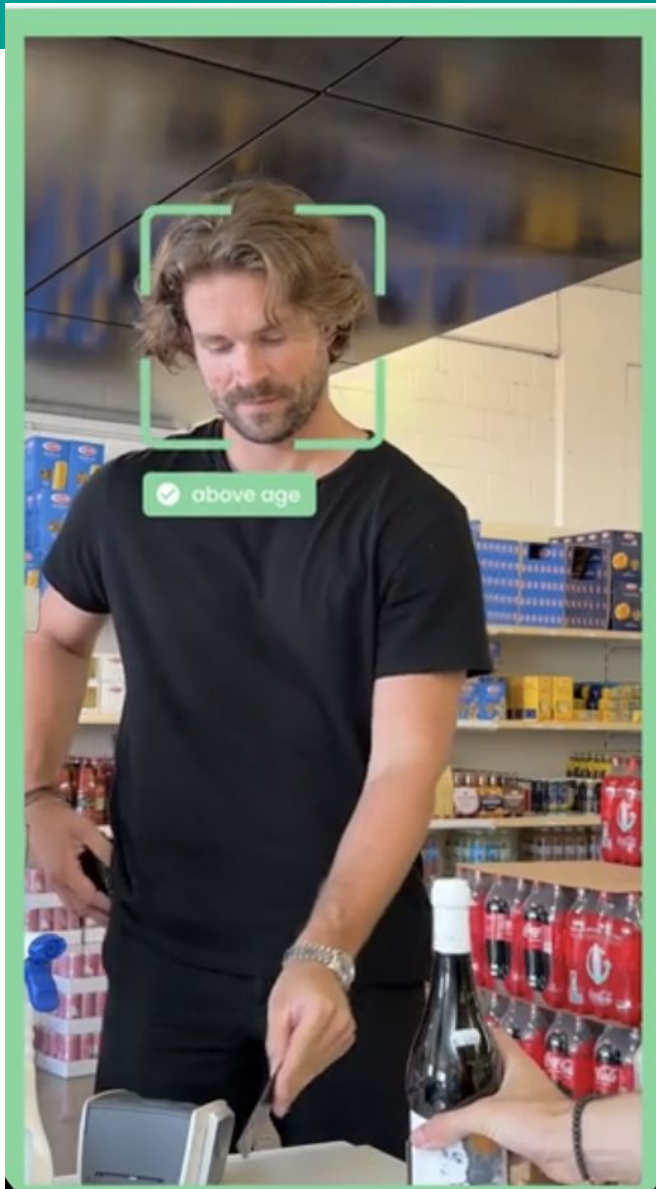
euCONSENT 2.0





What is Facial Age Estimation?

AGE ESTIMATION BOOSTS COMPLIANCE



We estimate you are

▼

21-

Try Again?

- A lightweight replacement of traditional “KYC” using apparent age
- Broad use cases: tobacco, physical retail, age-appropriate advertisements, online games, e-cigarettes, ...

MULTIPLE MODALITIES

Age Estimation is NOT Face Recognition



34.9 35.1 36.8 40.3 43.5



41.9 43.5 44.9 45.7 46.3

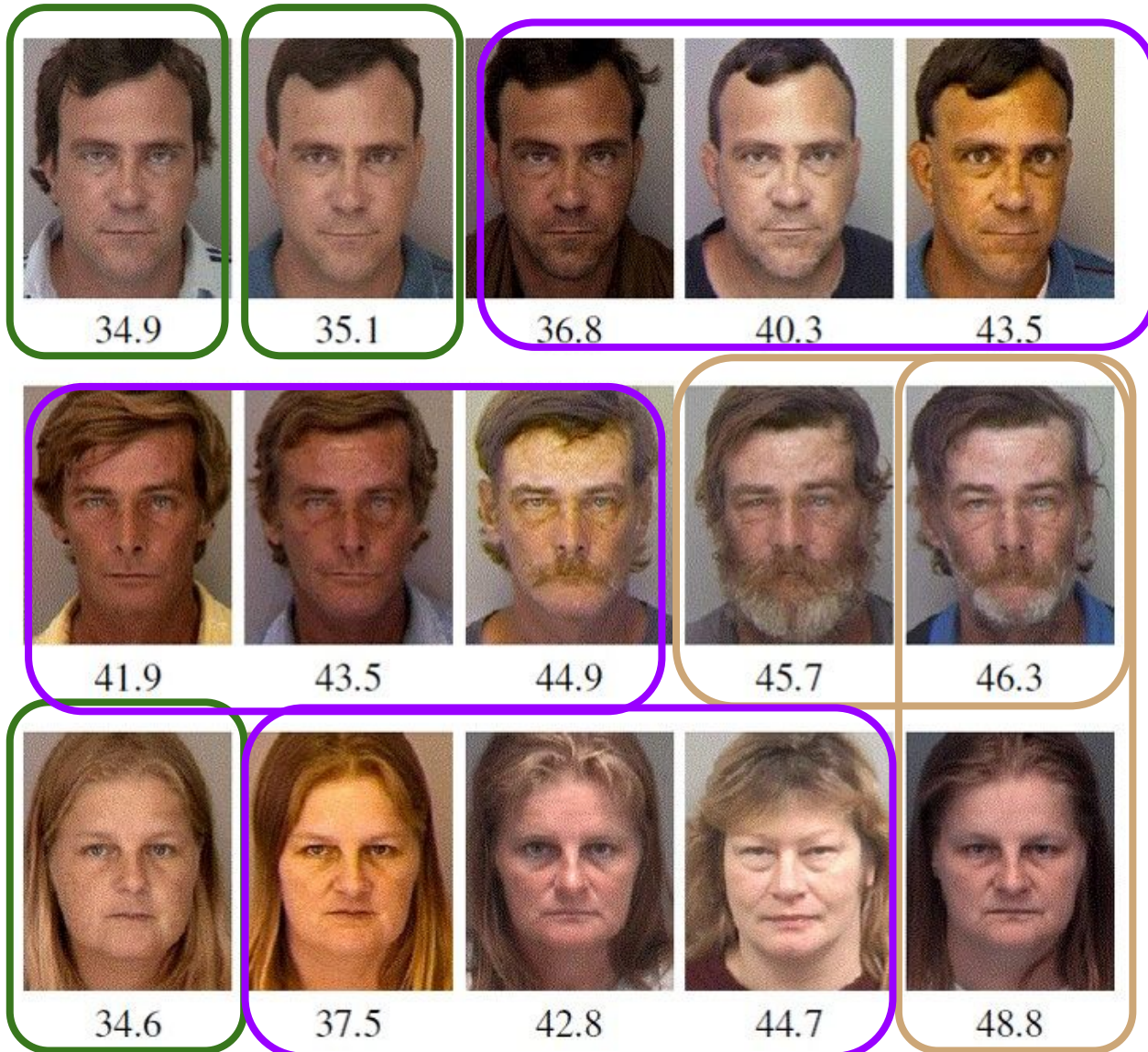


34.6 37.5 42.8 44.7 48.8

- Face recognition is optimized to identify individuals, regardless of the age. It is programmed to ignore age.
- It helps you “recognize” Person 1, Person 2, Person 3,... regardless of how old they are

Image retrieved from:
http://biometrics.cse.msu.edu/projects/longitudinalstudy_face.html

Age Estimation is NOT Face Recognition



Facial age estimation is optimized to detect age markers, regardless of whose face it is. It is optimized to ignore identity.

It helps you estimate whether a presented face is of a person between **34-36**, **36-45**, **45-49**, etc.

Image retrieved from:



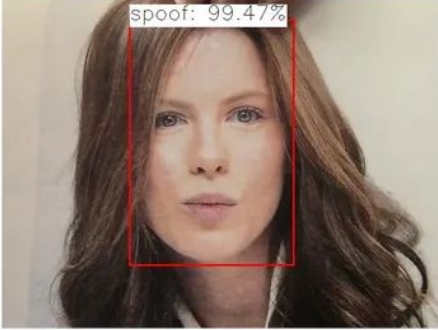
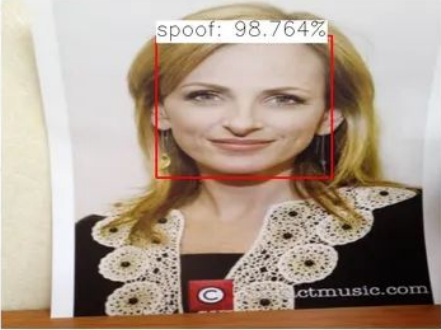
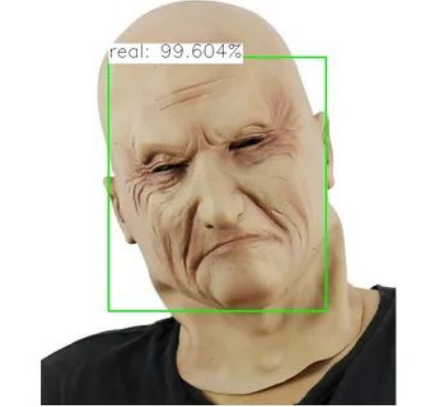
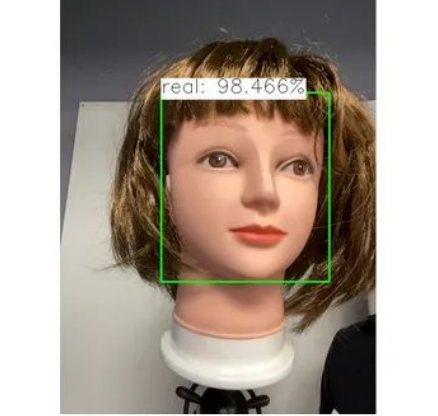
http://biometrics.cse.msu.edu/projects/longitudinalstudy_face.html



The Problem

Presentation attacks



		
Paper bills	Images projected from phones	Printed Photos
		
Printed Bended	Latex Masks	Mannequins

Some consumers will always try to bypass age gates

- *Very easy with credit cards, phone numbers, etc.!*

Anti-spoofing, genuineness check, live presence check, ... are the counter-measures.

We conduct joint R&D with IDIAP on Deepfakes, injection attacks, and more

Injection attacks Biometric Swaps



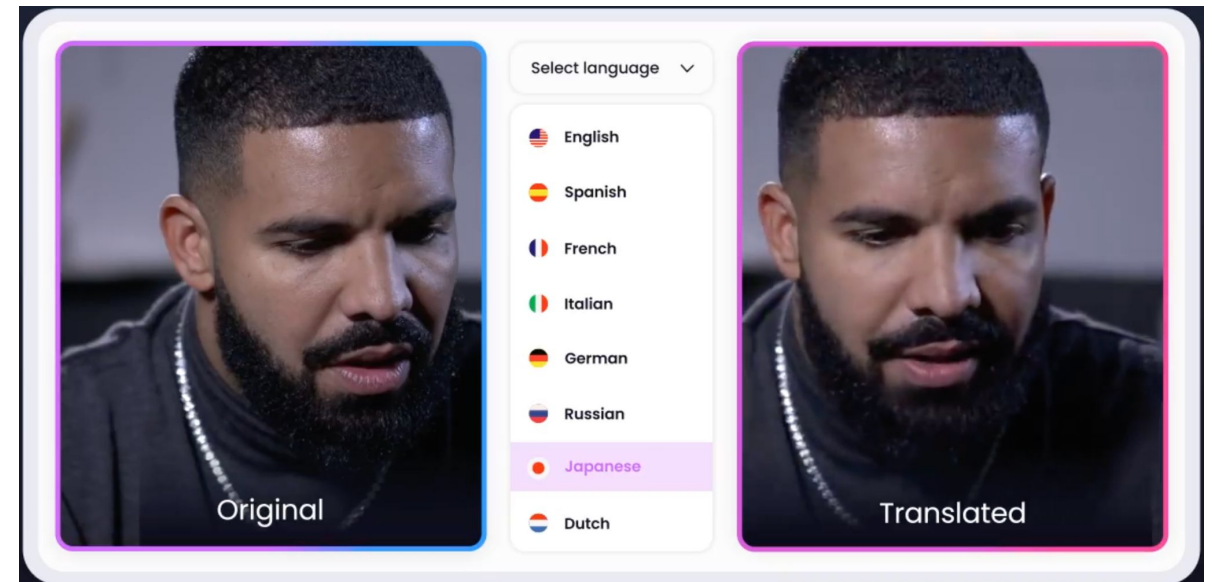
Source:

<https://images.app.goo.gl/X1X4qTMJhjdNqWt89>

Tamper genuine feed with

- Faces and audio swaps
- Lip syncs

... and feed the media instead of the genuine camera!



Source:

<https://twitter.com/nickfloats/status/176974566>

Deepfake/Injection attacks



Source:
<https://cheatsheet.md/midjourney/midjourney-consistent-character.en>

Hyperrealistic, synthetic faces injected instead of a real camera feed



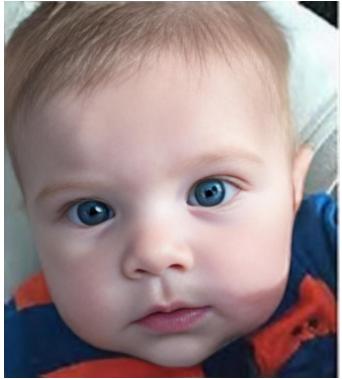
Source

Examples



- See pdf

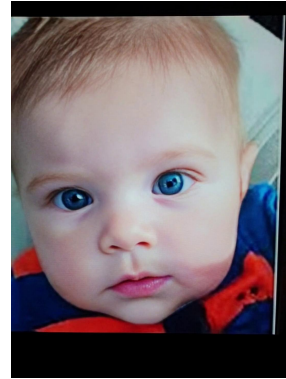
Presentation attacks are effective!



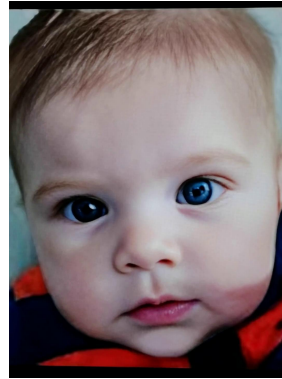
Original
upscaled



iPhone 12



Samsung
Galaxy S9

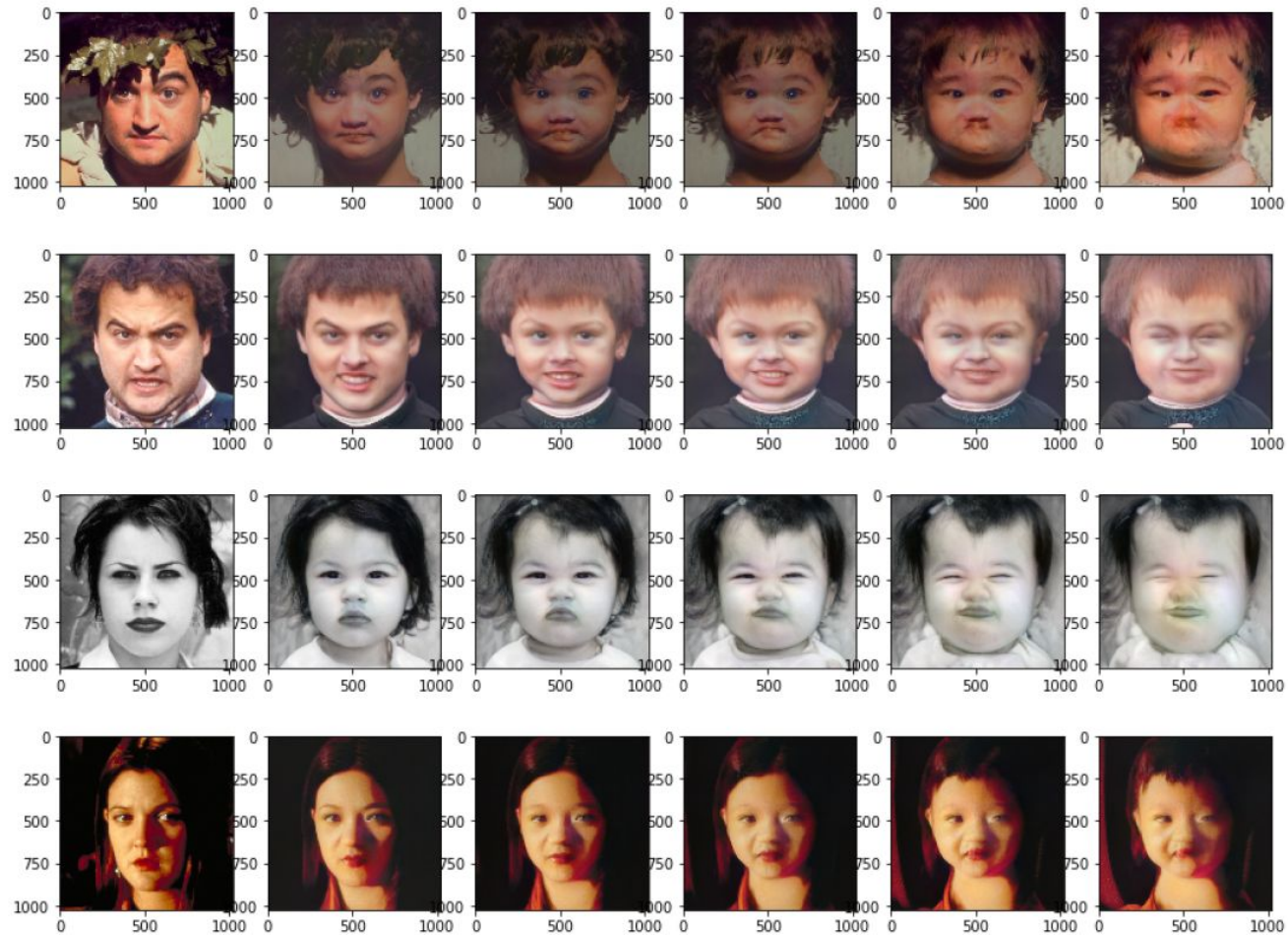


Huawei
Mate 30

Replayed media can obtain the desired age estimations! (Korshunov et al., 2024)

Training DB	Method	Original	iPhone	Galaxy	Huawei
UTKFace	Adaptive	0.599	0.566	0.567	0.586
Several	Regression via classification	0.596	0.571	0.573	0.583
Several	Distribution	0.589	0.574	0.585	0.597
UTKFace	Classification	0.574	0.529	0.543	0.560

Automatically de-aging people



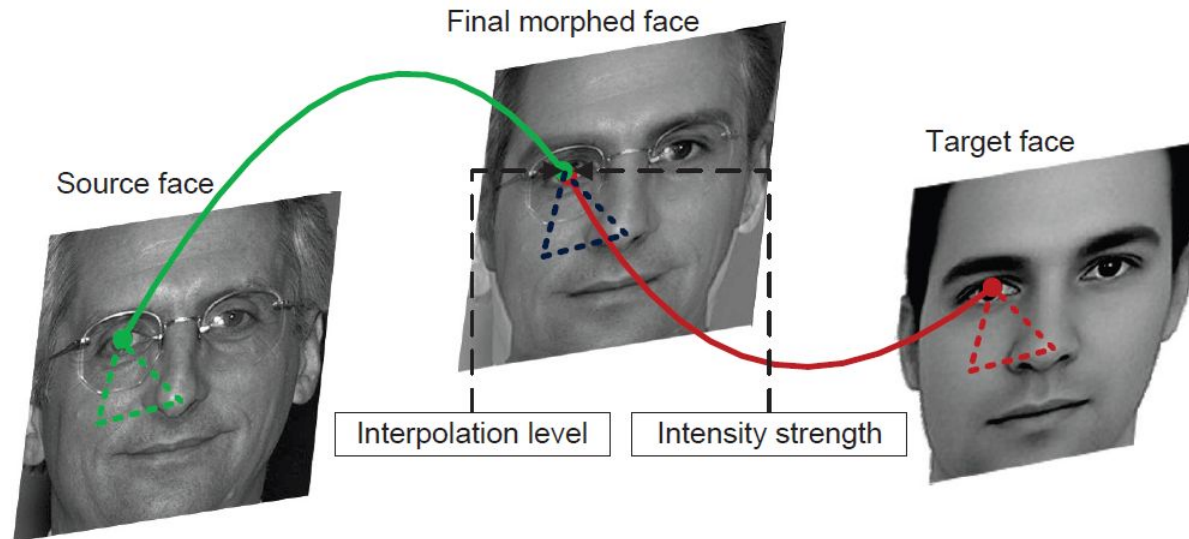
⁷P. Korshunov and S. Marcel “Face Anthropometry Aware Audio-visual Age Verification”, ACM Multimedia 2022.

Controlling pose and expressions



⁸L. Colbois, T. Pereira and S. Marcel, "On the use of automatically generated synthetic image datasets for benchmarking face recognition", IJCB 2021.

Morphing attacks



⁹E. Sarkar, P. Korshunov, L. Colbois and S. Marcel, "Are GAN-based Morphs Threatening Face Recognition?", ICASSP 2022.

Speaker recognition (SpeechBrain): Our DB



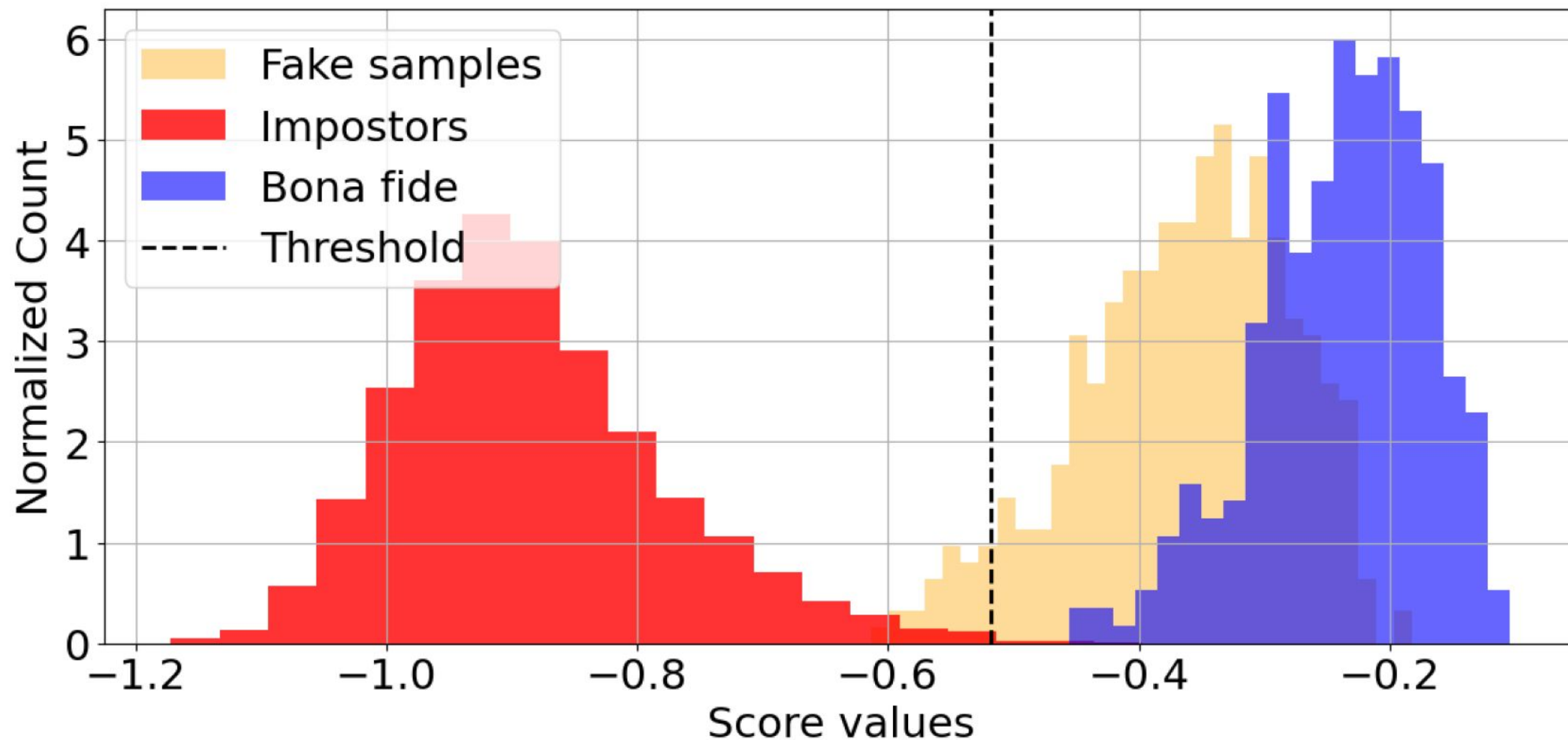
Method	FMR	FNMR	IAPMR
DiffVC, pretrained on LibriTTS	0.77	0.00	8.09
HiFiVC, pretrained on VCTK	0.77	0.00	0.00
YourTTS, pretrained on VCTK & LibriTTS	0.77	0.00	27.43
FreeVC, pretrained on VCTK	0.77	0.00	15.44
FreeVC, tuned 70K iterations	0.77	0.00	92.59
FreeVC, tuned 109K iterations	0.77	0.00	94.21

Facial recognition (MobileFaceNet): Our DB

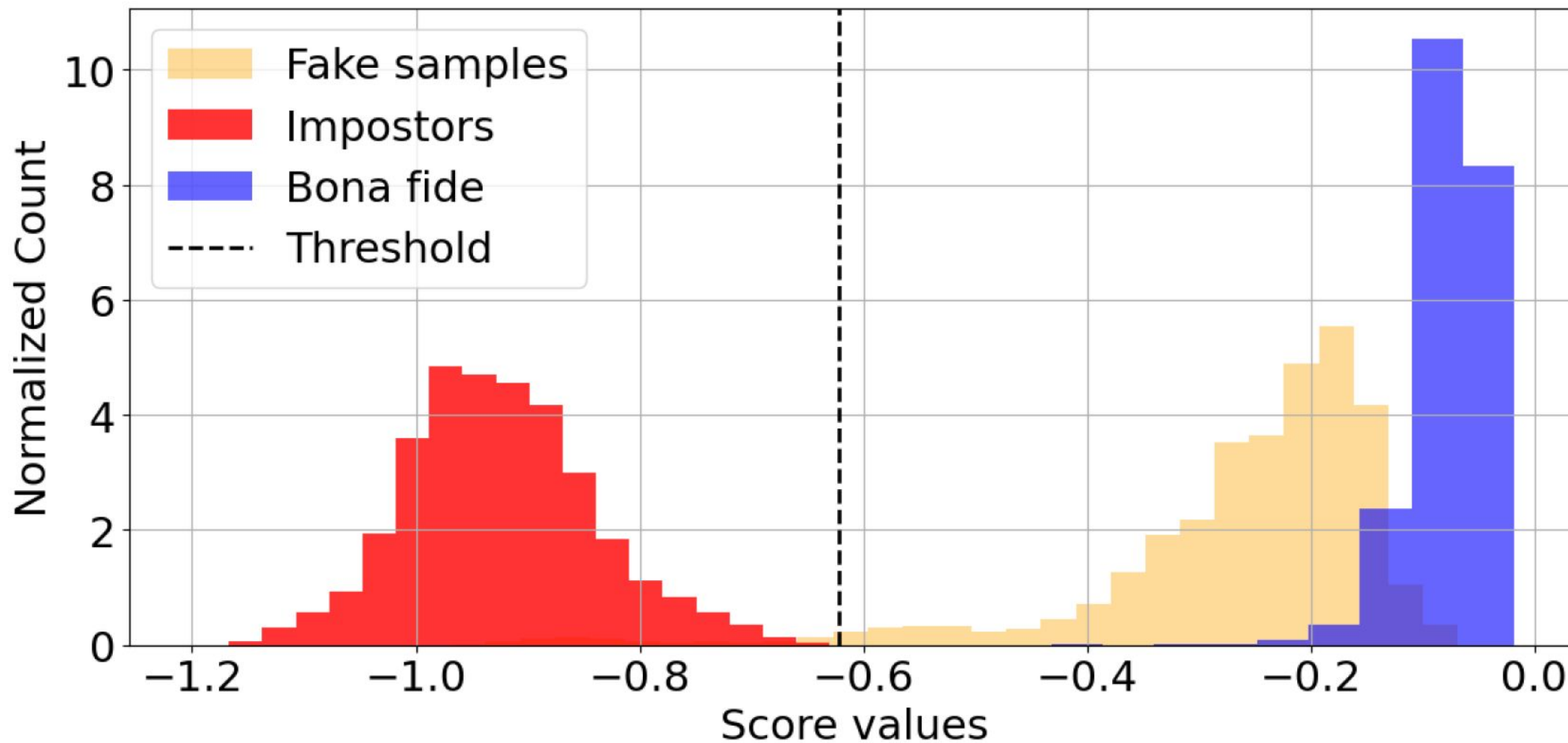


Model resolution	training params	blending methods	IAPMR
160px	no mask	no blending	96.27
160px	mask training	seamless, mlk color	95.22
160px	mask training + color	overlay, no color	96.43
256px	mask training	params tuned	96.78
256px	mask training	overlay, no color	97.36
320px	mask training	params tuned	99.96

Score of speaker recognition (FreeVC model)



Scores of face recognition (256px model)



Impact of synthetic data on age assurance



- Age assurance is as vulnerable as recognition
- Real-time deepfakes pose threats to liveness detection
- Detection methods struggle to generalise to unseen deepfakes
- Lack of children real and deepfake data

Related Papers

- P. Korshunov and S. Marcel “Face Anthropometry Aware Audio-visual Age Verification”, ACM Multimedia 2022.
- L. Colbois, T. Pereira and S. Marcel, “On the use of automatically generated synthetic image datasets for benchmarking face recognition”, IJCB 2021.
- E. Sarkar, P. Korshunov, L. Colbois and S. Marcel, “Are GAN-based Morphs Threatening Face Recognition?”, ICASSP 2022.
- P. Korshunov, H. Chen, P. N. Garner and S. Marcel, “Vulnerability of Automatic Identity Recognition to Audio-Visual Deepfakes”, IJCB 2023.

Questions to the audience



- **Do you come across any presentation or injection attacks that might impact your business?**
- **What forms of presentation and injections attacks you would expect to increase in volumes?**

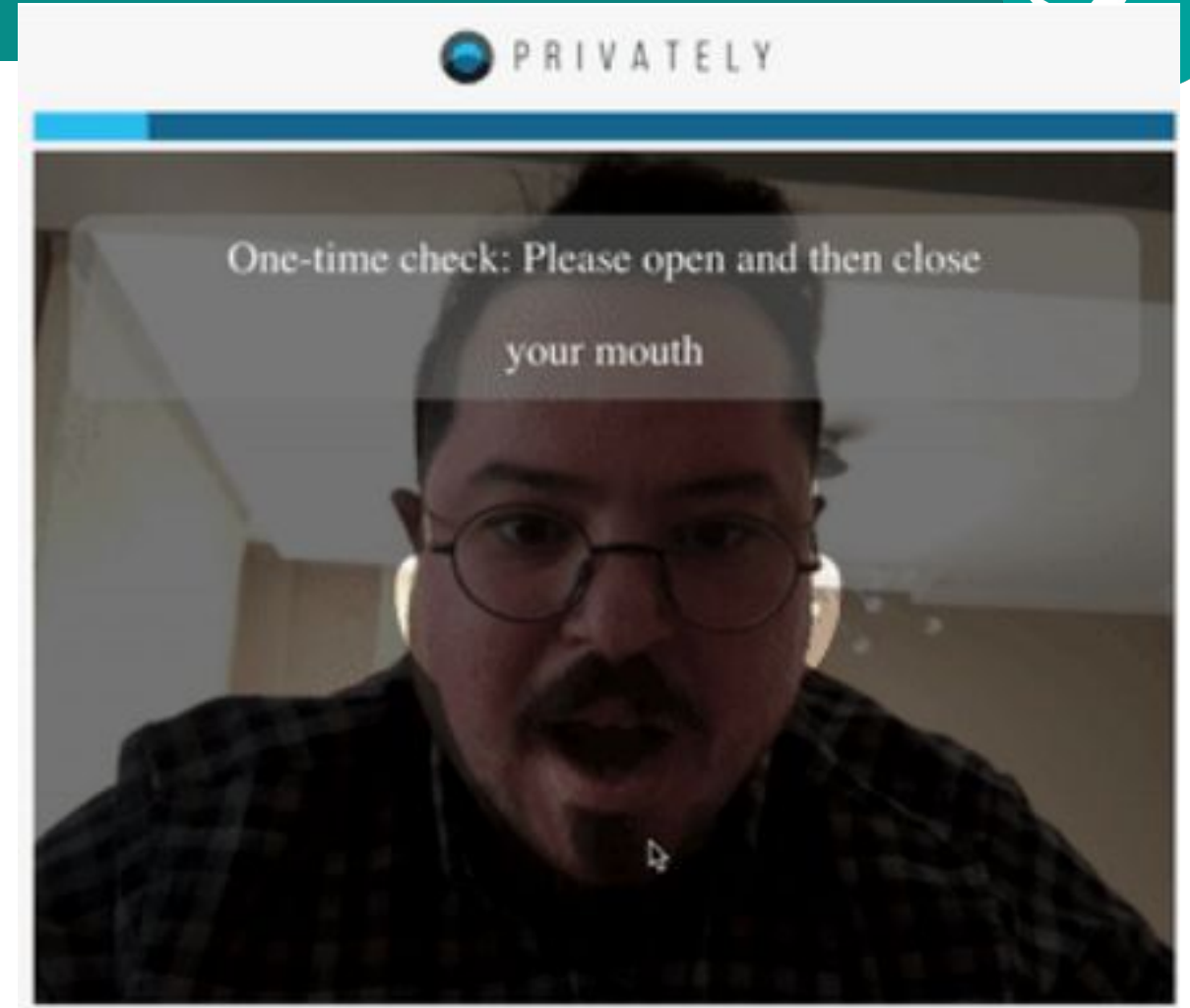


Potential Detection Mechanisms

Liveness Check

Cure for the pre-recorded media!

- Asking for gestures
- Reading out sentences



AI-Based Defences



- Anomaly detections:
- Skin tones
- Image depth
- Texture inconsistencies (also glasses, etc.)



Source: <https://images.app.goo.gl/mowkqCx4s4dMWQ3k7>



Source
<https://twitter.com/justinsuntron/status/17660469>

Anti-tampering control

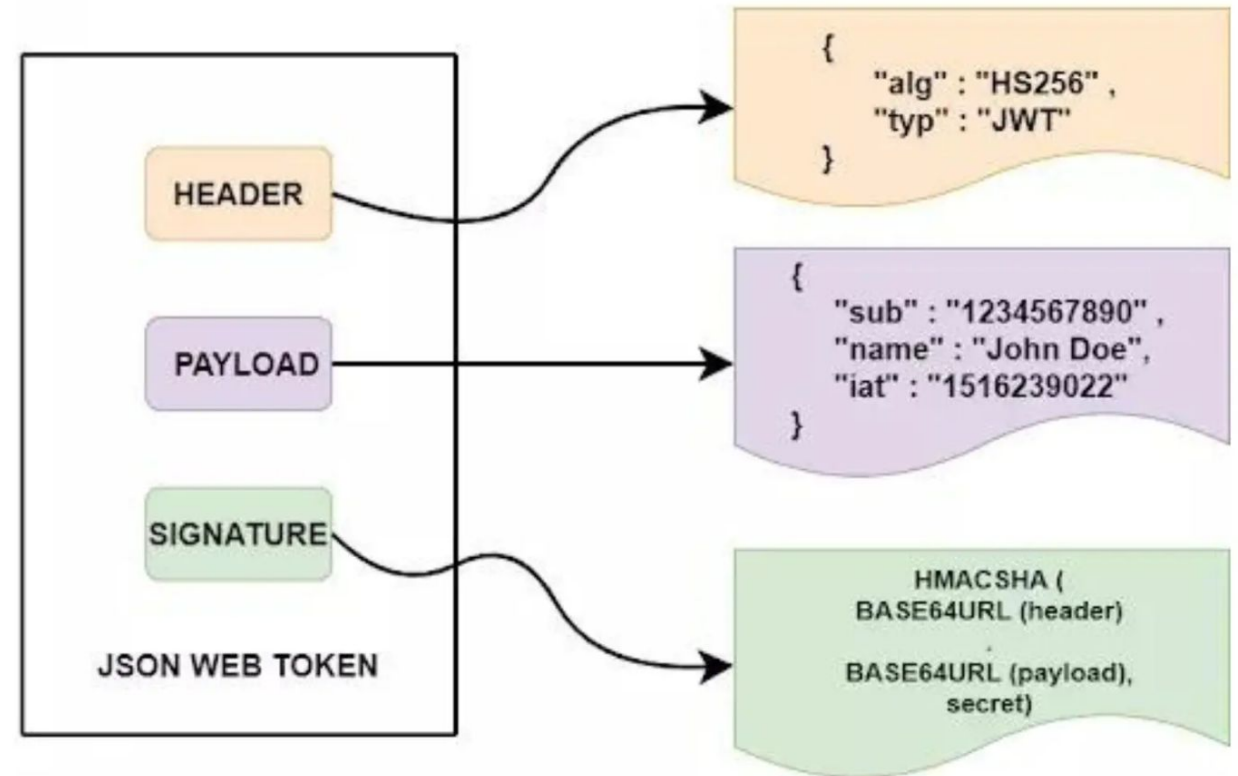


A well-studied security pattern

- Signed transactions
- Signed input



Structure of JSON Web Token (JWT)



Source:

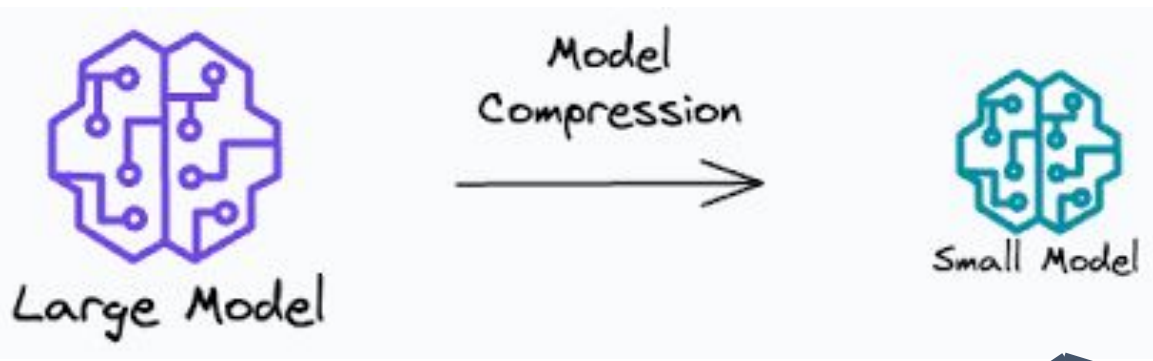
<https://images.app.goo.gl/LKKS1uiut2KX2Ra87>

Questions to the audience

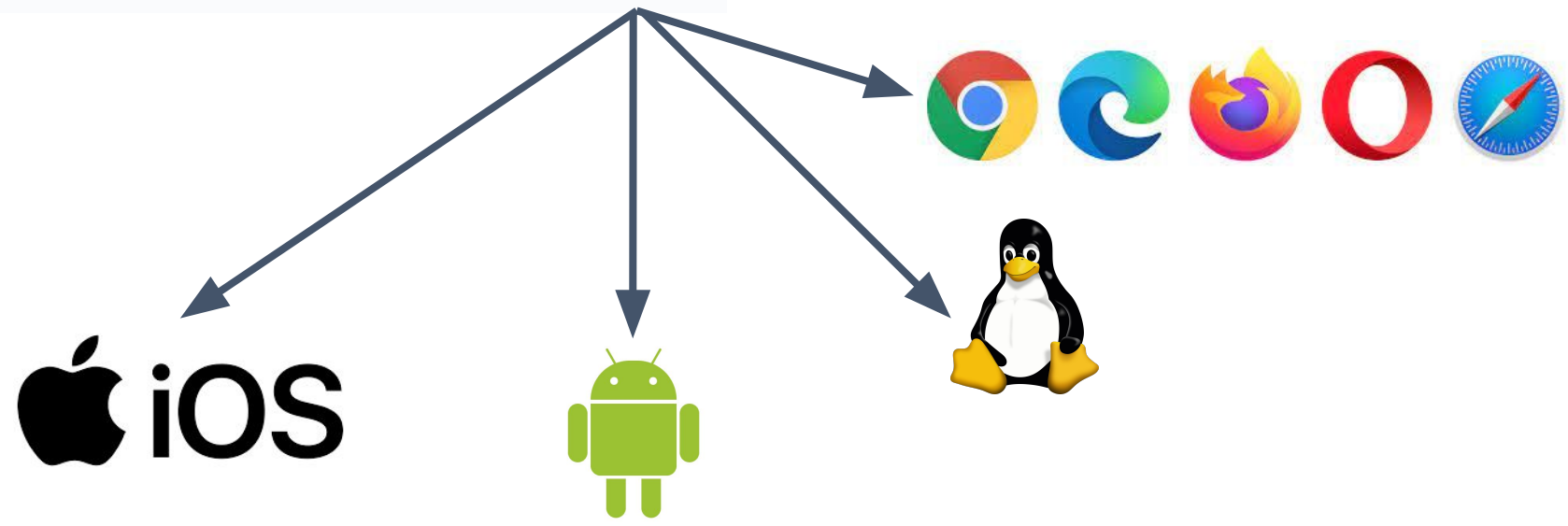


- **Do you employ any such PAD/IAD mechanisms for your business?**
- **What forms of defense mechanisms are feasible and desirable for age estimation purposes?**

AI-Defense: Can it still preserve privacy?



Keep PII away from servers!

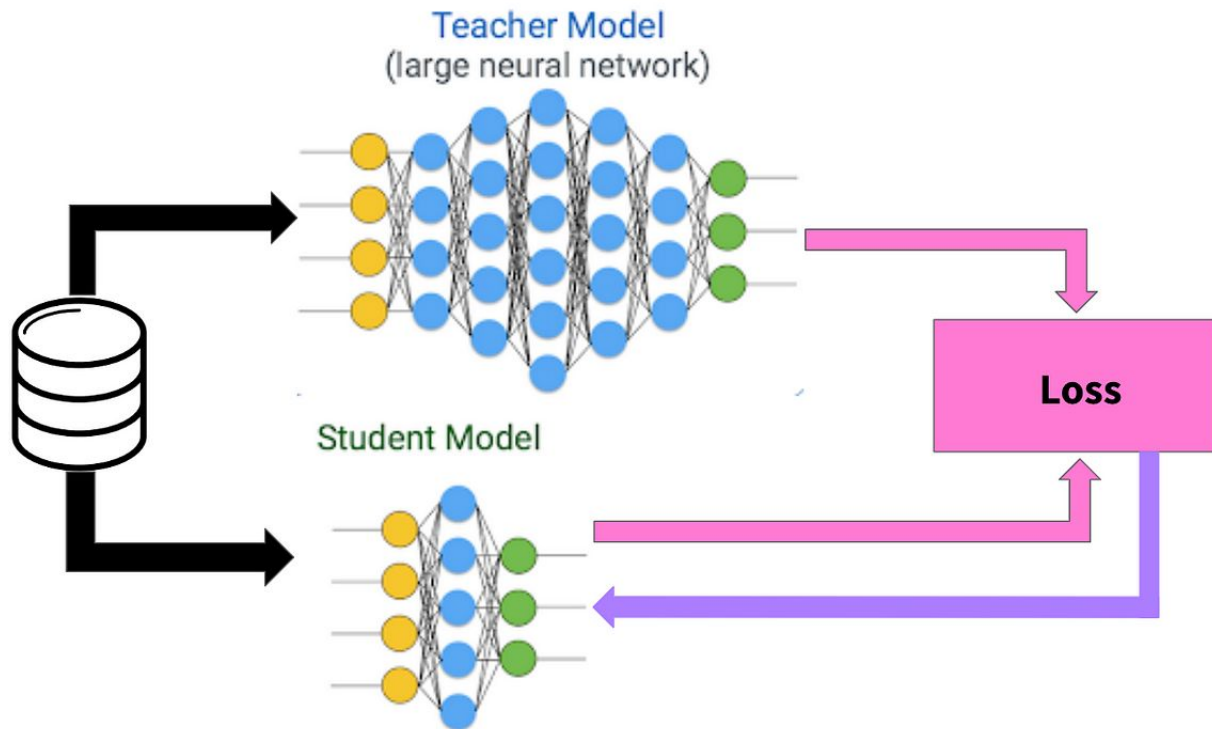


AI-Defense: Can it still preserve privacy?



How?

- Models compression
- Edge deployments



0.34	3.75	5.64
1.12	2.7	-0.9
-4.7	0.68	1.43

FP32



Quantization

64	134	217
76	119	21
3	81	99

INT8

Questions to the audience



- **What are the privacy considerations most organizations implement around PAD/IAD mechanisms?**

Project DefAI



Seek input from
age assurance / T&S
industry

Share lessons
privately with
industry

Scan for
threats

Prioritise
based on cost,
time usability
& accessibility

Develop
detection
methods

Develop
testing for
detection