Reliability and Explainability of AI – An Example of Face Recognition

Dr Thomas Lampert

Chair of Artificial Intelligence and Data Science

lampert@unistra.fr

Télécom Physique Strasbourg

SDC Research Team, ICube, University of Strasbourg

AI and Legal Practice

- Why use AI?
  - greater strains on civil and criminal justice systems
  - streamlining certain ‘routine’ activities (i.e. those with highly predictable outcomes)
  - reduce the burden on people
  - increase the speed and efficacy of collecting more and better evidence for use in criminal prosecutions

- We can already see these advances in, e.g. the medical domain

AI and Legal Practice

- **Trustworthy AI** requires three components (AI HLEG*):
  - (1) it should be **lawful**, ensuring compliance with all applicable laws and regulations,
  - (2) it should be **ethical**, ensuring adherence to ethical principles and values and
  - (3) it should be **robust**, both from a technical and social perspective since to ensure that, even with good intentions, AI systems do not cause any unintentional harm.

* High-Level Expert Group on Artificial Intelligence

AI and Legal Practice

- According to AI HLEG's **Ethics Guidelines for Trustworthy Artificial Intelligence**, the requirements for an AI system to be accepted are:
    a. human agency and oversight,
    b. technical robustness and safety,
    c. privacy and data governance,
    d. transparency,
    e. diversity, non-discrimination and fairness,
    f. societal and environmental wellbeing, and
    g. accountability.

- Are we there yet?

Face Recognition

- NIST 2020 tests, best algorithm’s error rate is 0.08% (< 1 error in 1000 images)

Model	Accuracy
DeepFace (Facebook)	97.25%
FaceNet (Google)	99.63%
Human	97.53%

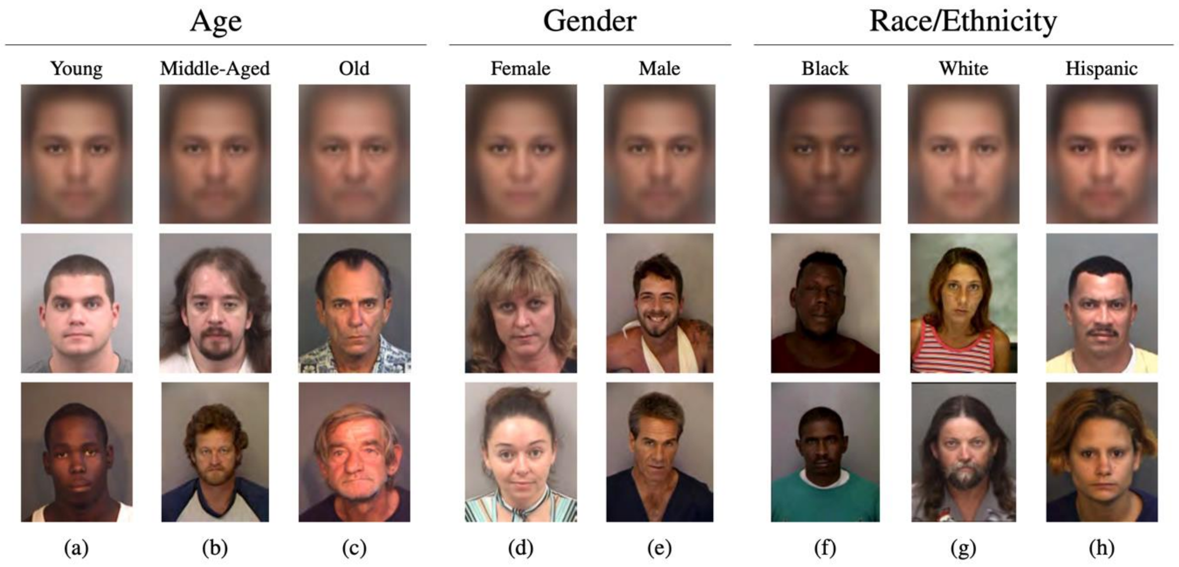
- Can match or outperform humans (in constrained settings)…

Face Recognition

- “ML predictions are (mostly) accurate but brittle” – A. Mądry
- Weaknesses
  - Bias
  - Data source (quality, orientation, video, etc)
  - Super-resolution (data used to train, have GT)
  - Explainability
- Attacks
  - Generative
  - Adversarial
- Transparancy

Weaknesses – Bias

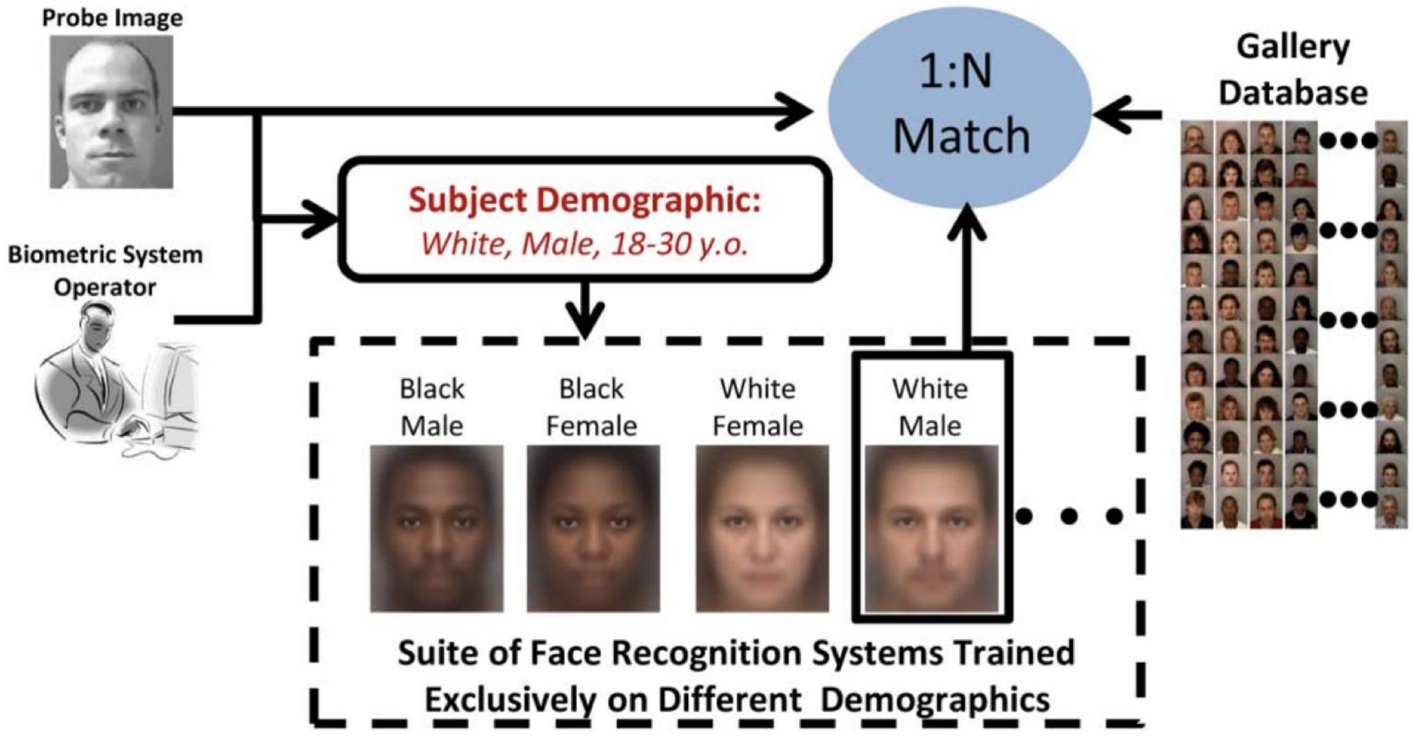
- In 2012 Klare et al. found:
  - “Lower recognition accuracies on the following cohorts: females, Blacks, and younger subjects (18 to 30 years olds)."



Klare et al., Face Recognition Performance: Role of Demographic Information

Weaknesses – Bias

- In forensic scenarios the use of dynamic face matcher selection may be preferred



Klare et al., Face Recognition Performance: Role of Demographic Information

Weaknesses – Bias

- … and in 2024

Algorithm	Submission Date	FNMR Overall	FMR Min	FMR Max	FMR Max/Min
sertis_003	2023-12-27	0.0039 (201)	0.00003 E.Europe M (35-50]	0.01213 W.Africa F (65-99]	420 (257)
rebs_001	2023-12-22	0.0018 (27)	0.00000 E.Europe M (20-35]	0.00486 W.Africa F (65-99]	1505 (479)
roc_016	2023-12-19	0.0018 (26)	0.00007 E.Europe F (12-20]	0.00831 W.Africa F (65-99]	122 (23)
intellivision_007	2023-12-19	0.0093 (369)	0.00004 E.Europe M (35-50]	0.01214 W.Africa F (65-99]	327 (131)
cyberlink_013	2023-12-15	0.0040 (206)	0.00002 E.Europe M (35-50]	0.00427 W.Africa F (65-99]	266 (91)

https://pages.nist.gov/frvt/html/frvt_demographics.html

Weaknesses – Bias

- Even simpler tasks involving the face, e.g. gender identification exhibit the same limitations

Gender Classifier	Overall Accuracy on all Subjects in Pilot Parlaiments Benchmark (2017)
Microsoft	93.7%
FACE++	90.0%
IBM	87.9%

http://gendershades.org/overview.html

Weaknesses – Bias

- Even simpler tasks involving the face, e.g. gender identification exhibit the same limitations

Gender Classifier	Female Subjects Accuracy	Male Subjects Accuracy	Error Rate Diff.
Microsoft	89.3%	97.4%	8.1%
FACE++	78.7%	99.3%	20.6%
IBM	79.7%	94.4%	14.7%

http://gendershades.org/overview.html

Weaknesses – Bias

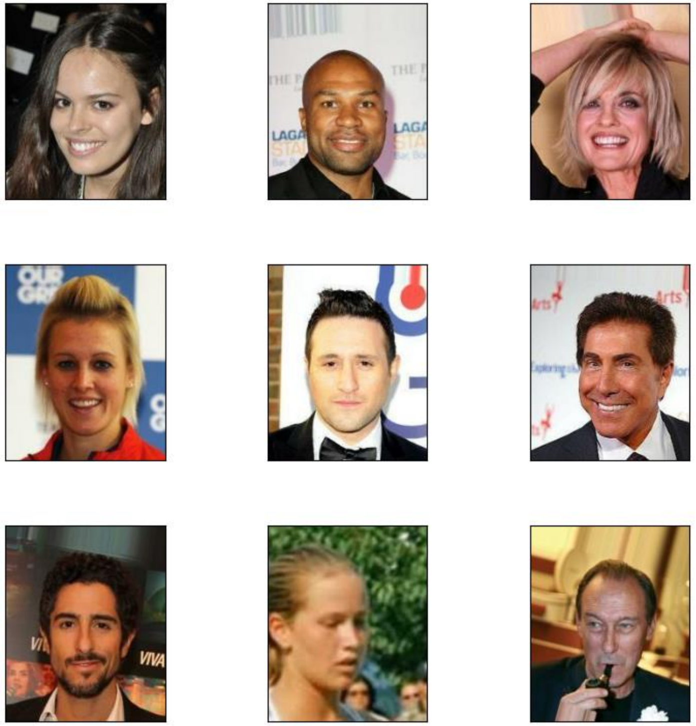
- Even simpler tasks involving the face, e.g. gender identification exhibit the same limitations

Gender Classifier	Darker Male	Darker Female	Lighter Male	Lighter Female	Largest Gap
Microsoft	94.0%	79.2%	100%	98.3%	20.8%
FACE++	99.3%	65.5%	99.2%	94.0%	33.8%
IBM	88.0%	65.3%	99.7%	92.9%	34.4%

http://gendershades.org/overview.html

Weaknesses – Data Sources

- Algorithms are generally developed with high resolution images



CASIA-Webface and Celeb-A datasets

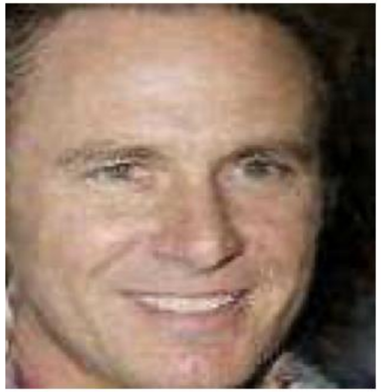
Weaknesses – Data Sources

https://www.youtube.com/watch?start=9&feature=oembed&v=6LYmlWXuW_s

Weaknesses – Data Sources

- Super Resolution

Low-Resolution Original Reconstructed



Low-Resolution Original Reconstructed

Yu et al., Super-Resolving Very Low-Resolution Face Images with Supplementary Attributes, CVPR, 2018

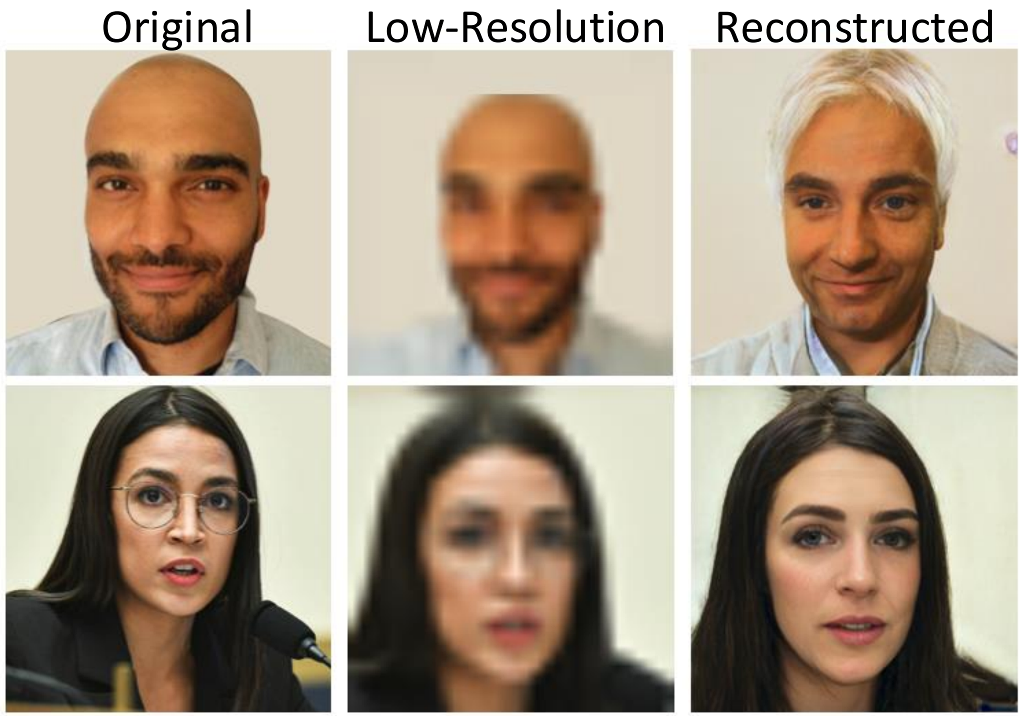
Weaknesses – Data Sources

- Super Resolution

https://x.com/tg_bomze/status/1274245778551328769
https://x.com/Chicken3gg/status/1274314622447820801

Weaknesses – Data Sources

- Even training on more diverse data does not guarantee to solve the problem
  - Model training problems
- AI is based on statistics
  - If the information is not there, it does not exist
  - These are (statistically likely) inventions (that depend on the data, model, …)



https://x.com/osazuwa/status/1274444300894572546

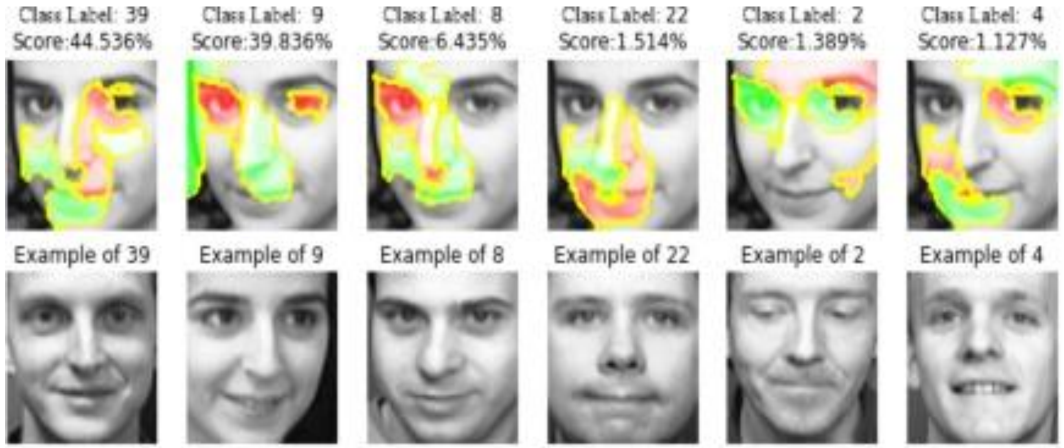
Weaknesses – Explainability

- The power of current AI algorithms is derived from their non-linear, multi-layered structure
- Inherently their output cannot be explained easily

XAI

Weaknesses – Explainability

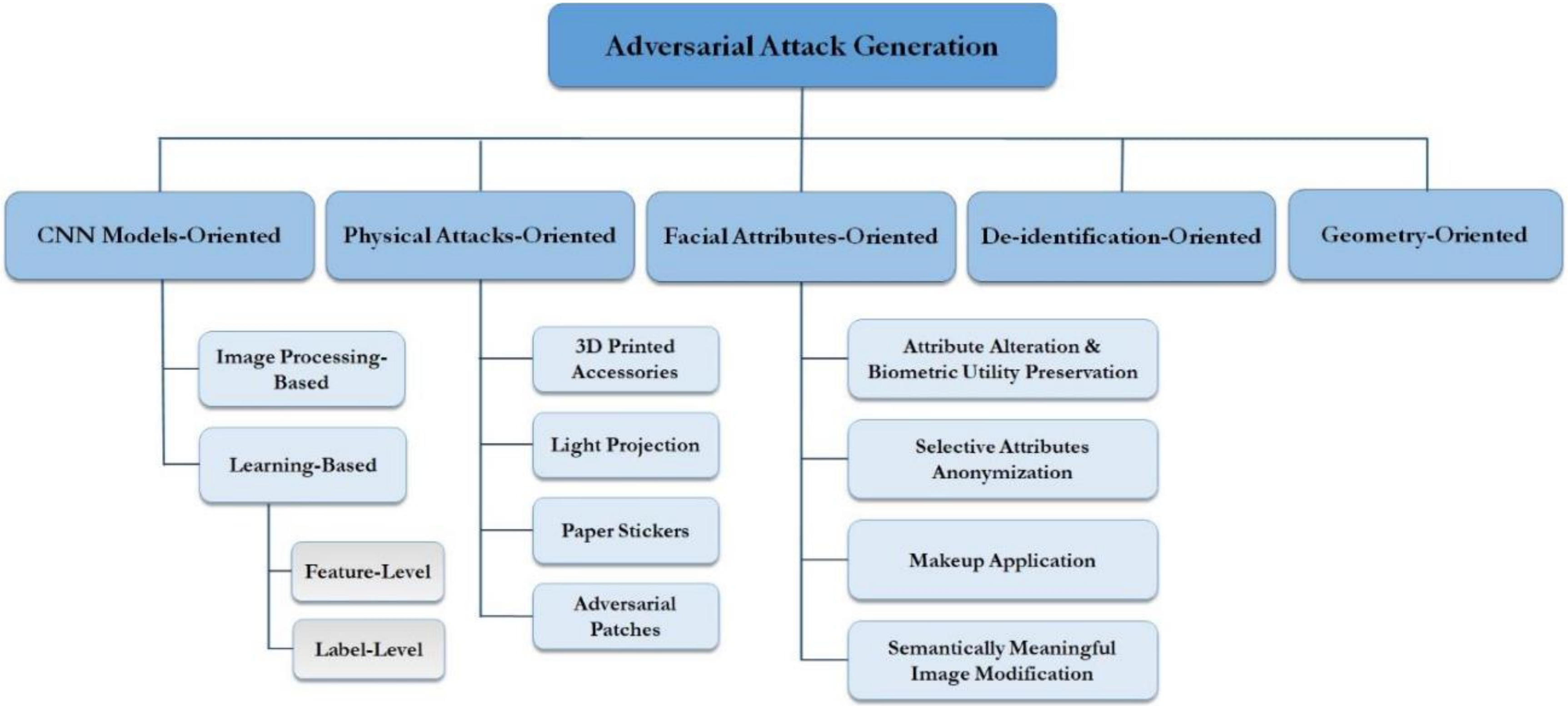
- Use ad hoc external approaches:
  - Does not reveal what is salient
  - Often misses impacts with less magnitude
  - Identified regions contain both useful and unuseful information
  - Requires human to interpret (biased)
  - Ad hoc general approaches
  - Can be wrong



(a) LIME-generated explanation for LeNet-5 model when True Label is 9 but wrongly predicted as Label is 39.

Rajpal et al., XAI-FR: Explainable AI-Based Face Recognition Using Deep Neural Networks, 2022

Vulnerabilities – Attacks



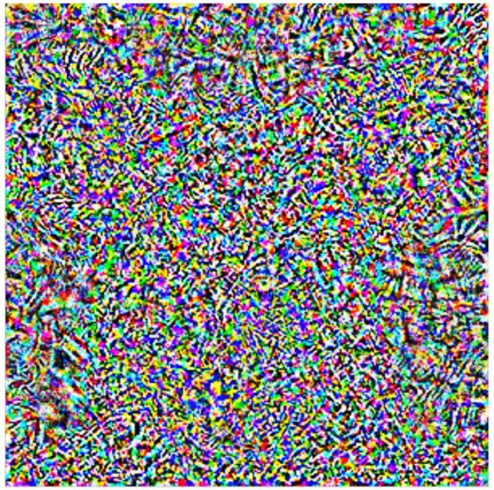
Threat of Adversarial Attacks on Face Recognition: A Comprehensive Survey

Vulnerabilities – Attacks

Pig (91%)

+ 0.005 x

Noise (NOT random)

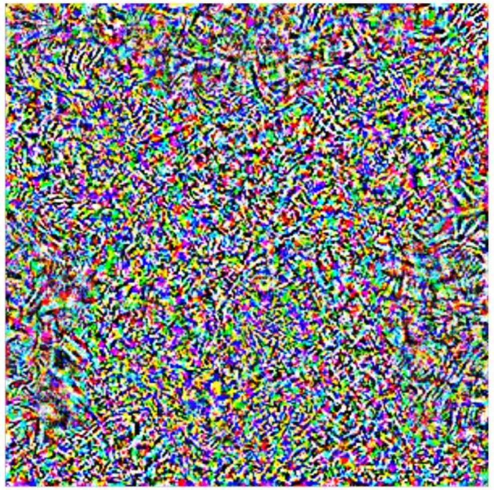


=

Vulnerabilities – Attacks

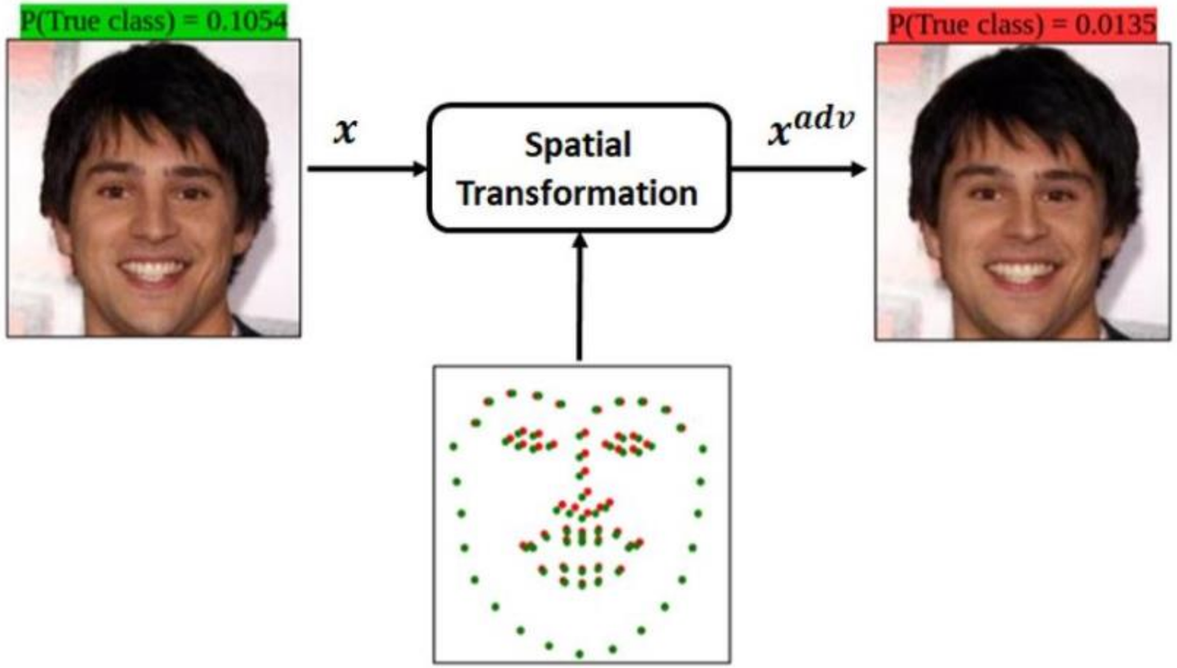
Pig (91%) **Noise (NOT random)** **Aeroplane (99%)**

+ 0.005 x



=

Vulnerabilities – Attacks

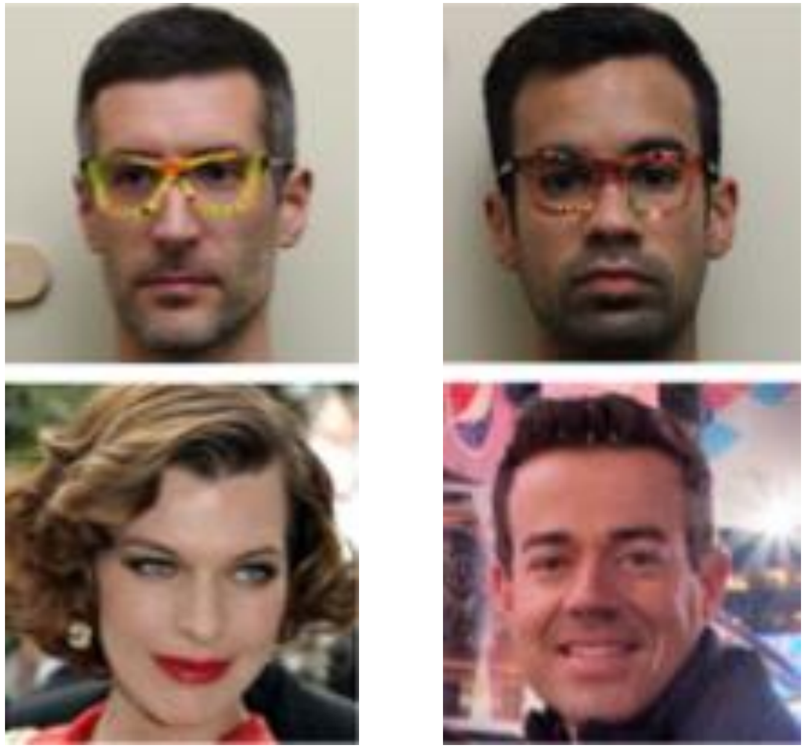


Threat of Adversarial Attacks on Face Recognition: A Comprehensive Survey

Vulnerabilities – Attacks

- Allows an attacker to evade recognition or impersonate somebody else
- Can also be used in real life!

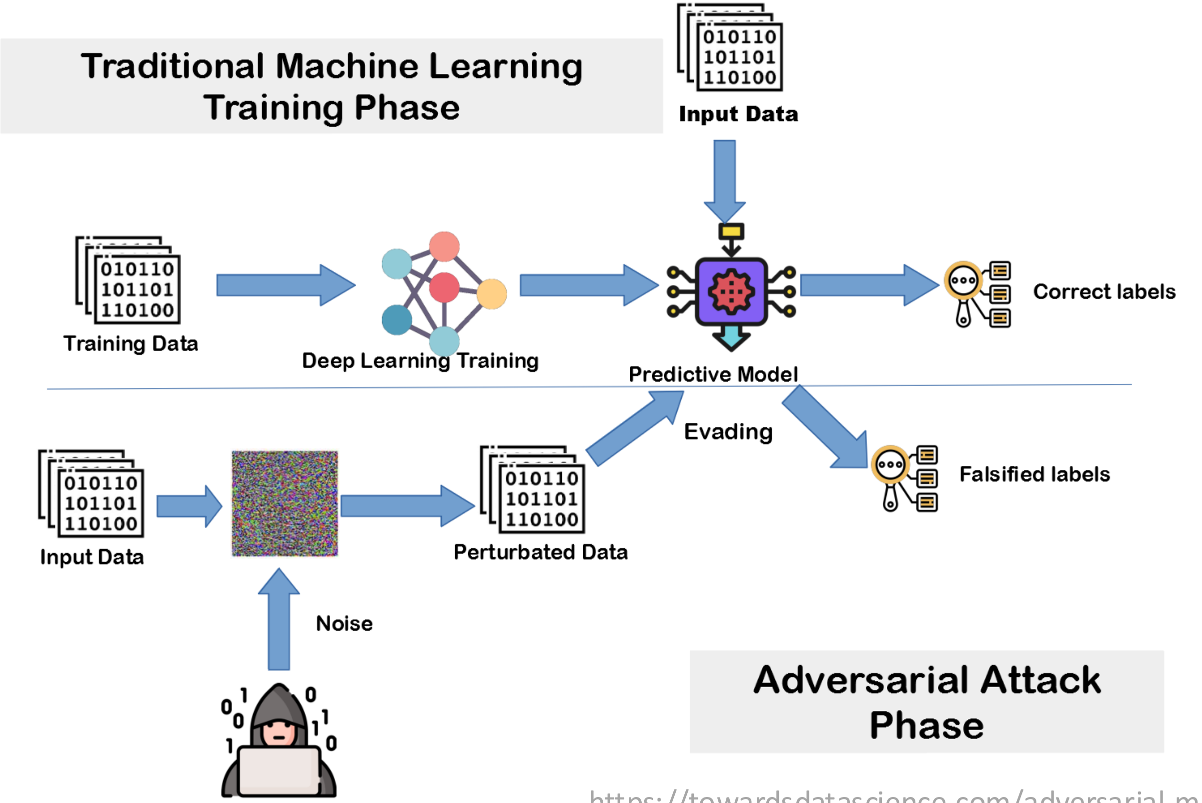
Zhou et al., Invisible Mask: Practical Attacks on Face Recognition with Infrared

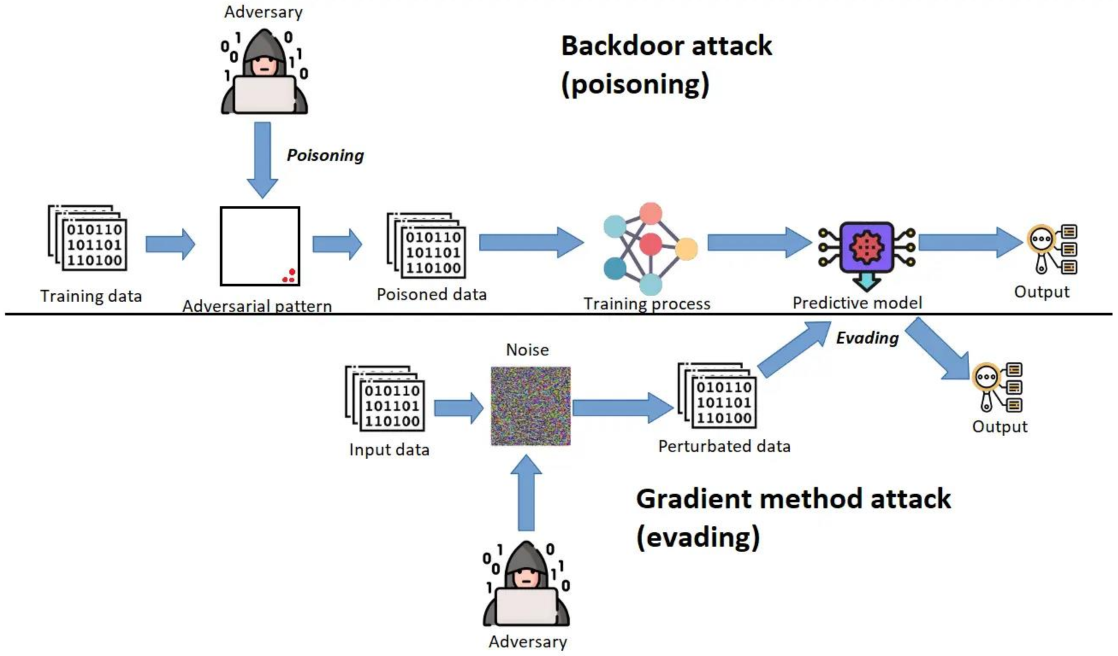


Sharif et al., Accessorize to a Crime: Real and Stealthy Attacks on State-of-the-Art Face Recognition

Vulnerabilities – Attacks

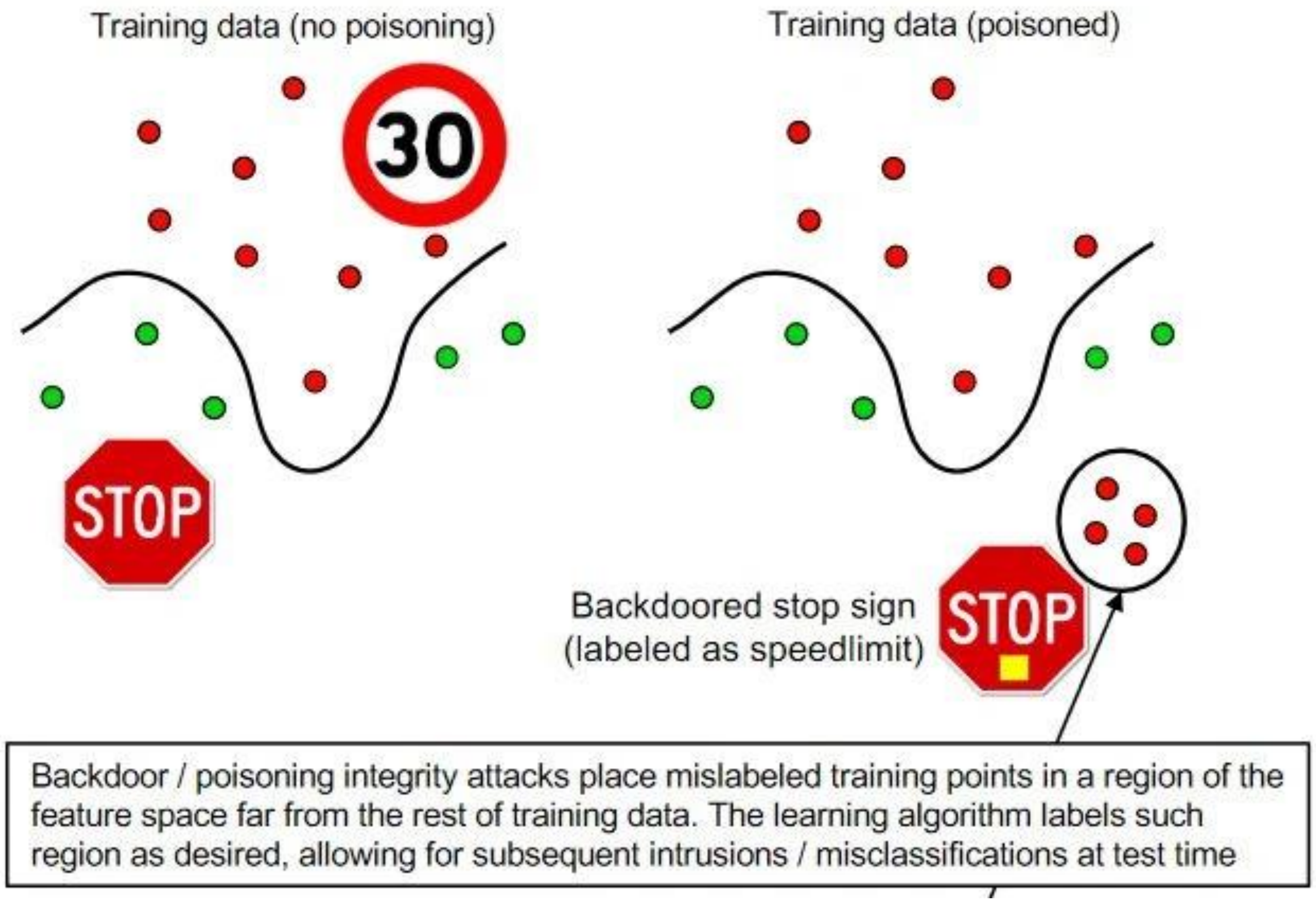
- Poisoning attacks embed hidden malicious behaviour into deep learning models





https://towardsdatascience.com/adversarial-machine-learning-mitigation-adversarial-learning-9ae04133c137

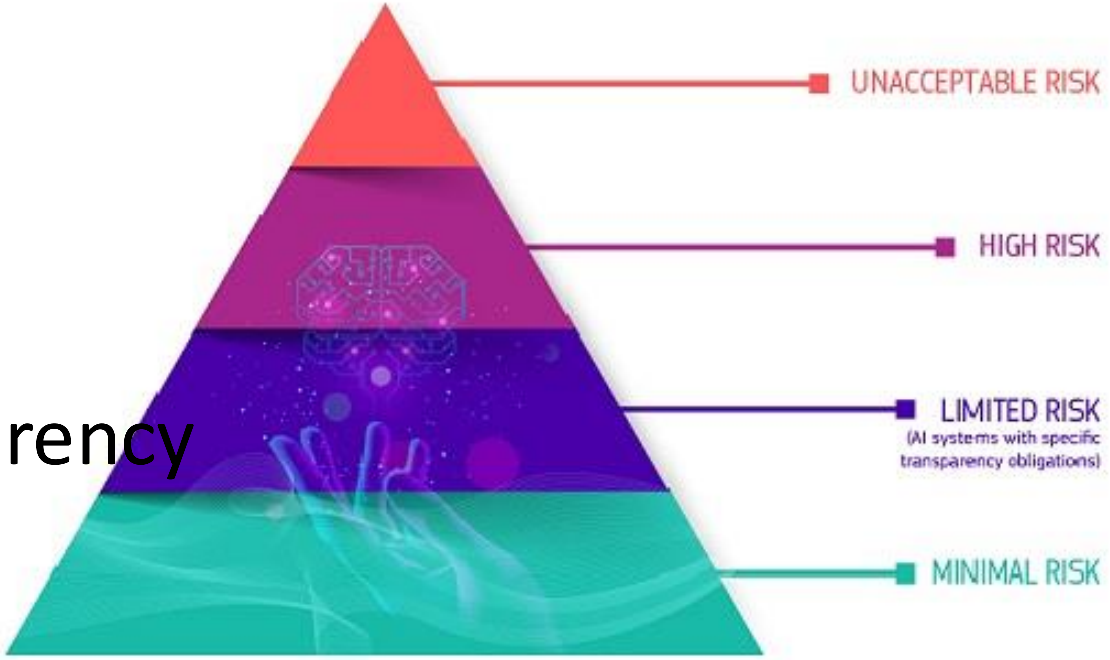
Vulnerabilities – Attacks



Gao and Qiu, Energy-Based Learning for Preventing Backdoor Attack, 2022

Transparancy

- Commercial systems are protected (trade secrets)
- A user can be unsure of:
  - Architecture
  - Training data
  - Evaluation protocols
  - Training strategy
- Regulations are targeting transparency



https://digital-strategy.ec.europa.eu/en/policies/regulatory-framework-ai

Conclusions

- AI holds great possibility for analysing data
- Caution is needed to ensure that the correct information is presented
- ... and risks quantified
- Checks need to be in place to ensure that it is not relied upon