Task 11 – Part 1

Detecting Realtime Deepfakes

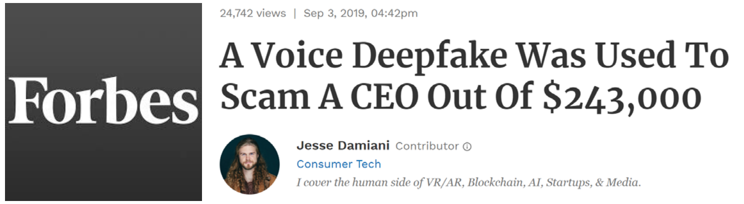
Speaker: Dr. Yisroel Mirsky

CBG

Cyber@Ben-Gurion
University of the Negev

The Threat

- One third of all attacks on ICS are SE related
- Inherent weakness to familiarity
- Already on the rise...



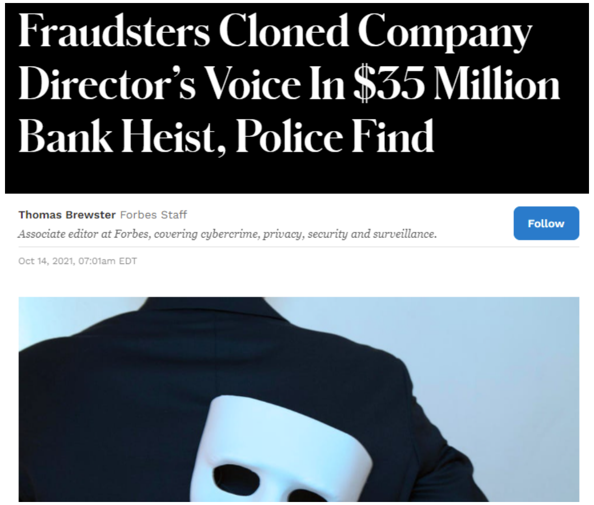
24,742 views | Sep 3, 2019, 04:42pm

A Voice Deepfake Was Used To Scam A CEO Out Of $243,000

Jesse Damiani Contributor

Consumer Tech

I cover the human side of VR/AR, Blockchain, AI, Startups, & Media.



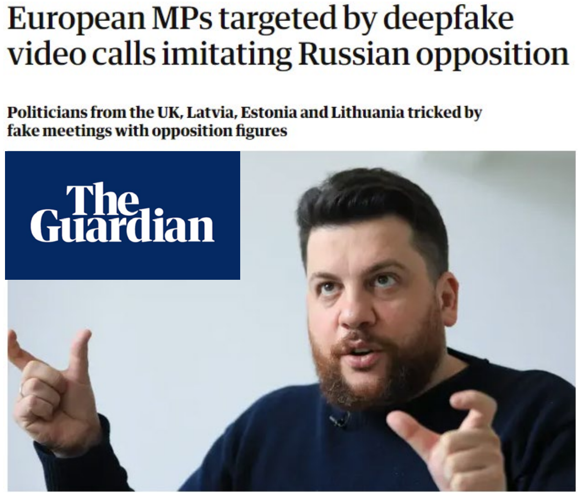
Fraudsters Cloned Company Director's Voice In $35 Million Bank Heist, Police Find

Thomas Brewster Forbes Staff

Associate editor at Forbes, covering cybercrime, privacy, security and surveillance.

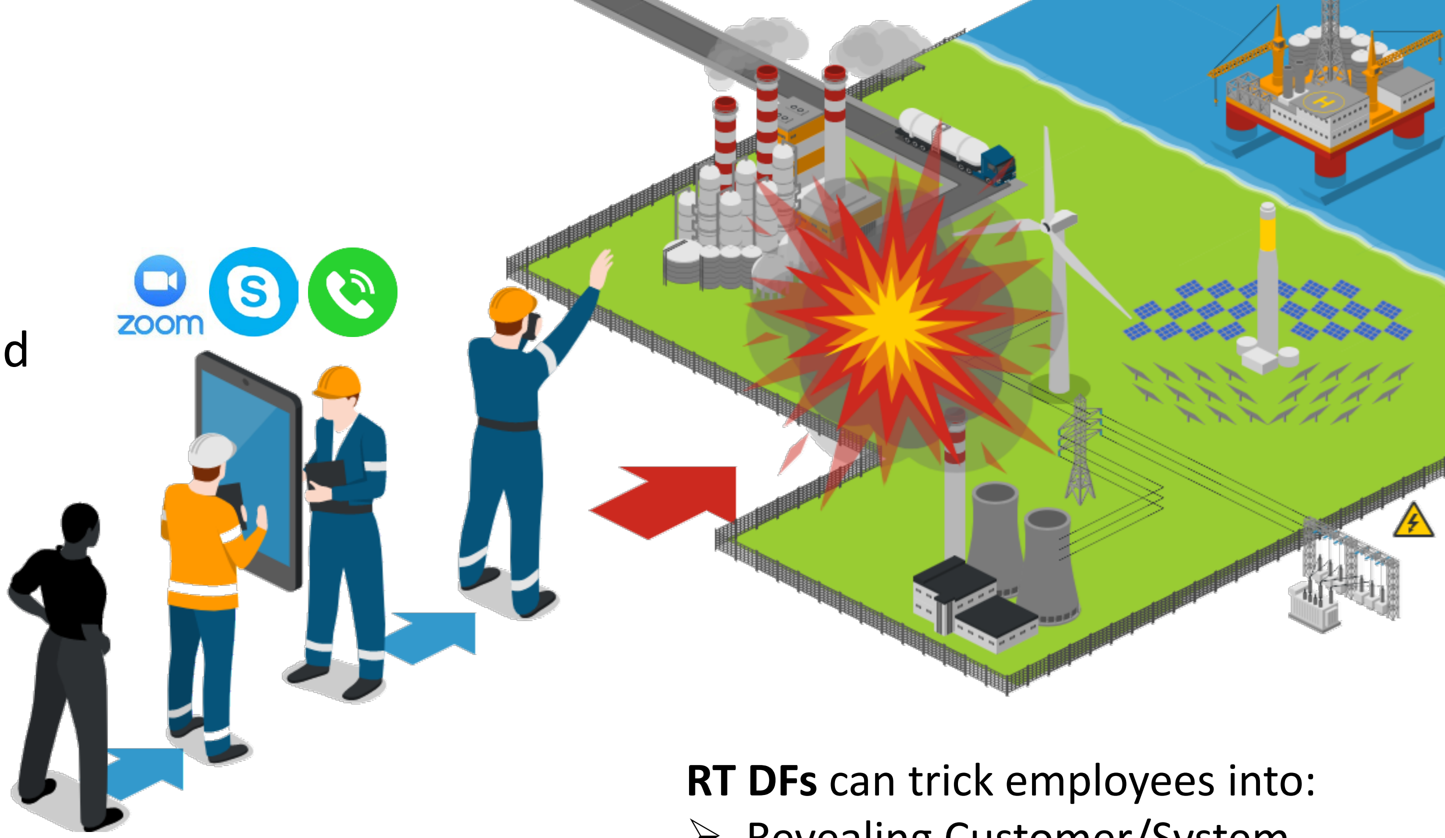
Follow

Oct 14, 2021, 07:01am EDT



European MPs targeted by deepfake video calls imitating Russian opposition

Politicians from the UK, Latvia, Estonia and Lithuania tricked by fake meetings with opposition figures



RT DFs can trick employees into:

- Revealing Customer/System Information
- Installing malware
- Change settings, ...

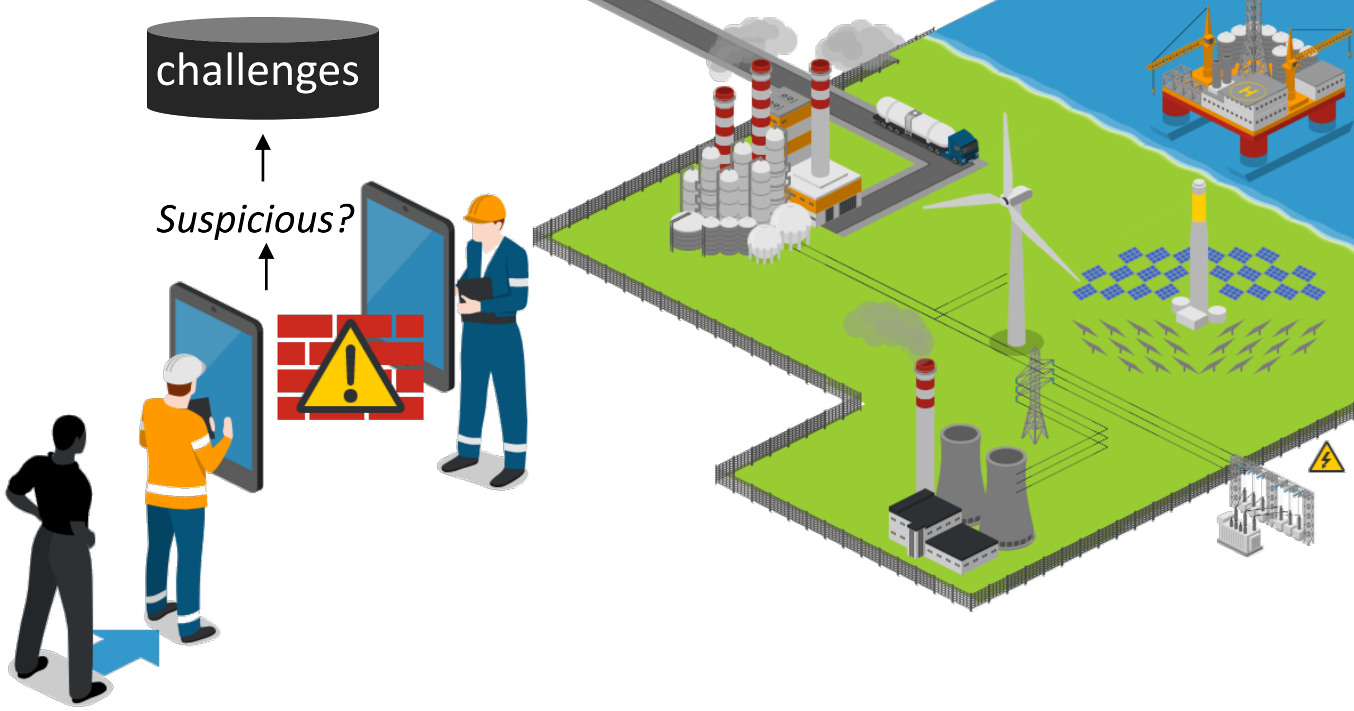
Approach

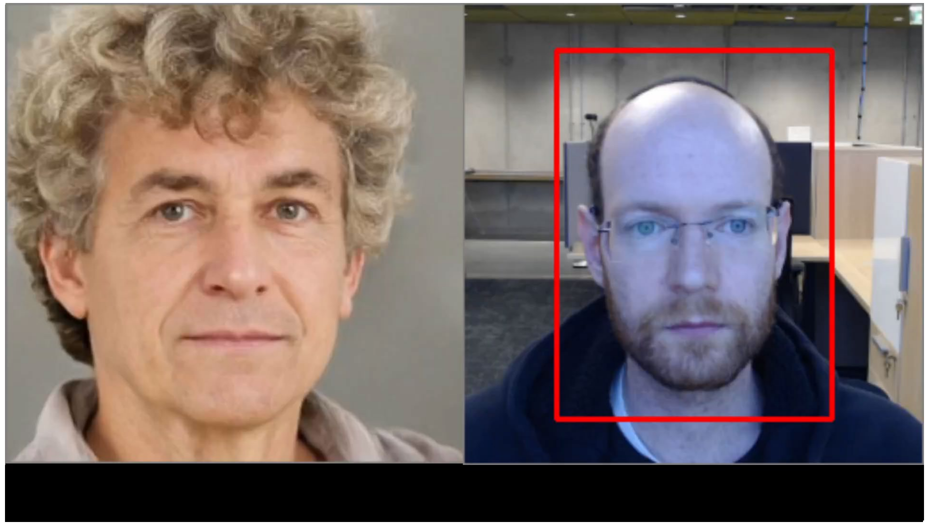
Hypothesis

RT-DFs are limited (pipeline/abilities)

Method

- (1) Send challenge that exploits the DF's limitations
- (2) Validate result with AI or human (e.g., victim)
- Puts attacker at disadvantage:
  - Attacker must be perfect at all limitations
  - Easy for defender to add new challenges





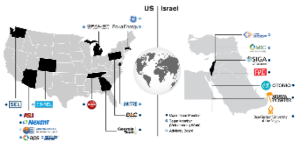
Out-of-Domain Challenges
- Pick up RQ object
- Hand expressions
- Tongue motion
- Fold ear
- Face occlusions
- Remove glasses
- …

Simulation Challenges
- Drop/bounce object
- Fold shirt, stroke hair
- Interact with background scenery
- …

Initial Focus

Voice Challenges
- mimic phrase + rhythm
- Repeat accent
- Change tone or speed
- Clear throat
- …

Research Plan – *Initiation May 1st*



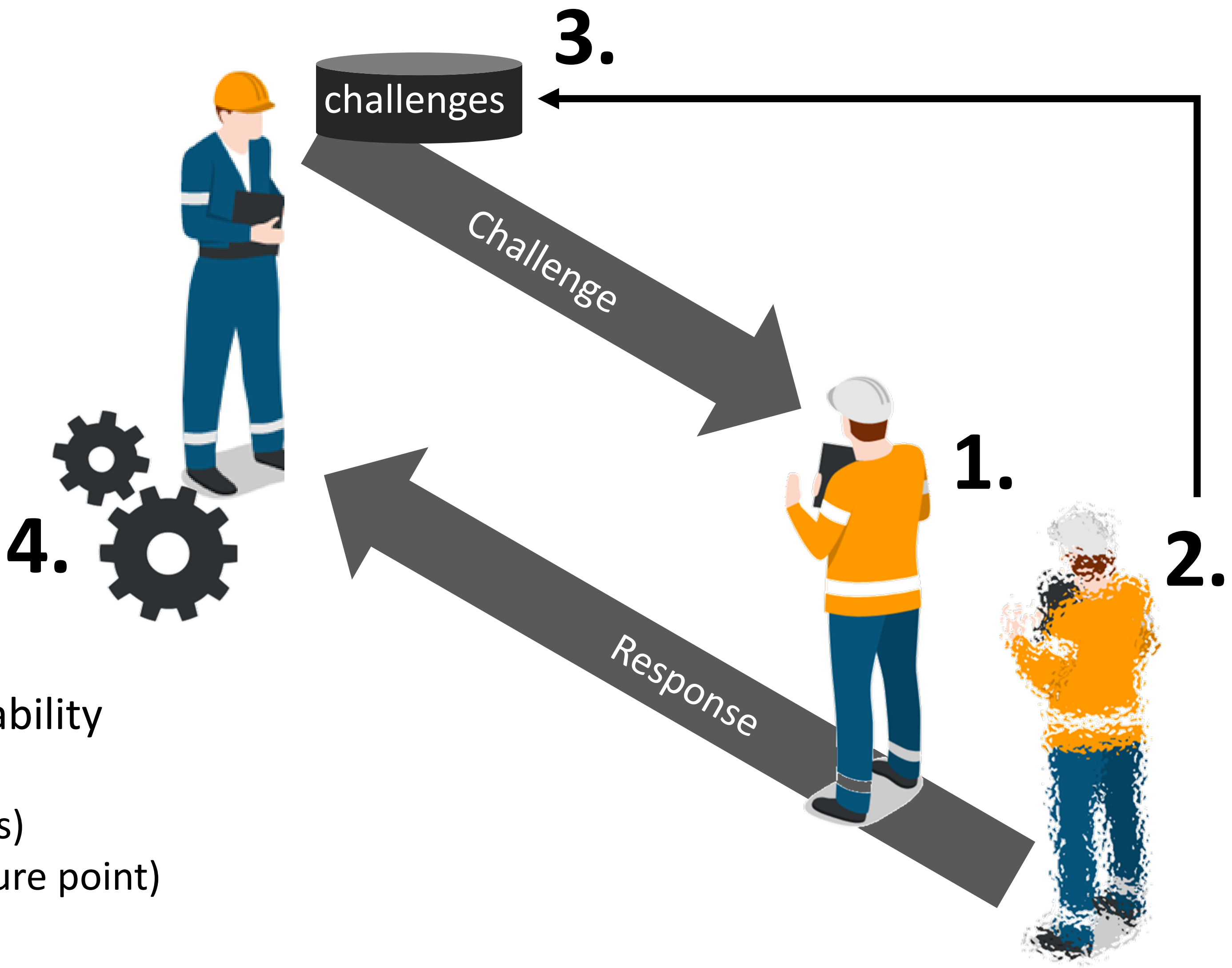
Research Plan

Year 1

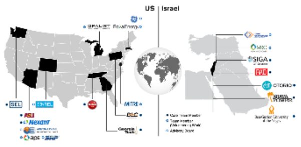
1. **Implement RT-DF Technologies**
   - Survey STOA Audio Cloning
   - Collect Existing Code
   - Implement Methods
   - Evaluate Quality (blind)
2. **Analyze RT-DF Limitations**
   - Stress training data limits
   - Stress tech limits
   - Stress scope limits

Year 2

3. **Develop DF-Captchas (challenges)**
   - Enumerate challenges with usability
4. **Develop response analysis**
   - Static Anomaly Detection (artifacts)
   - Temporal Anomaly Detection (failure point)



Commercialization Concepts



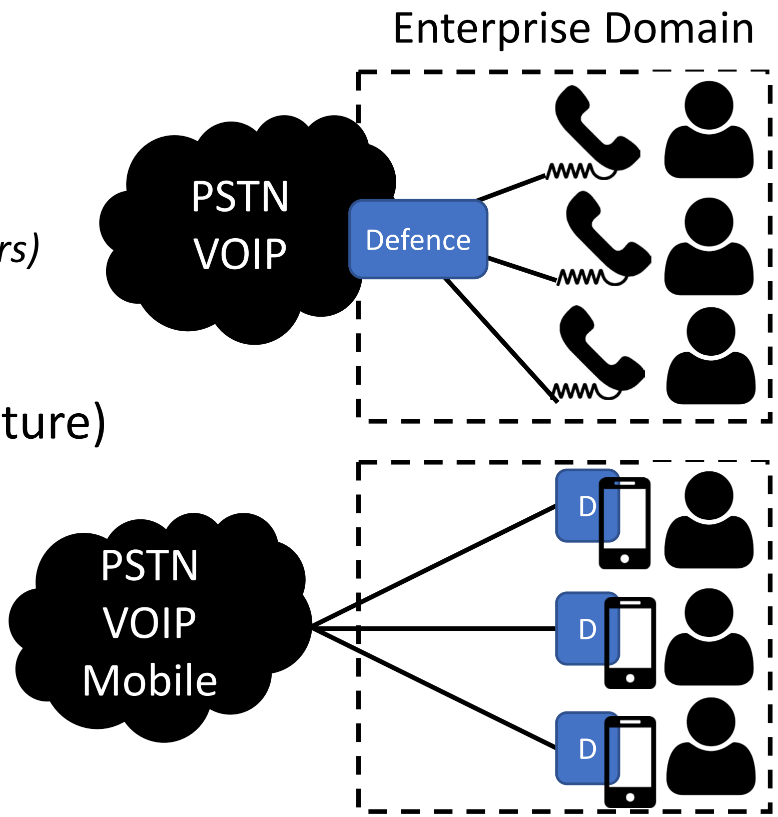
Example Technology Use-cases:

- ➢ Over Existing Infrastructure (telecom –audio, VoIP)
  - ➢ Network-based call firewall

    Detect malicious calls to internal numbers: technicians, customer service, billing, etc
  - * ➢ Mobile app call firewall

    Intercept and validate malicious calls cell calls (e.g., from unknown/suspicious numbers)

- ➢ Over 3rd party software
  - ➢ App that monitors Zoom/WhatsApp/Skype (wrapping via SDK or capture)
  - ➢ Authentication before joining Zoom meeting



Current Challenge

BGU has not found a partner for commercialization within the consortium

Suggestion

Find an external Partner