Deepfake Call Detection: Mitigating Next-Gen Social Engineering

Guy Frankovits

Supervisor: Dr. Yisroel Mirsky

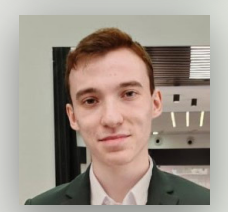
BIRD
Israel-U.S.
Binational Industrial Research
and Development Foundation

CBG
Cyber@Ben-Gurion
University of the Negev

Offensive AI Research Lab

Principal Investigator

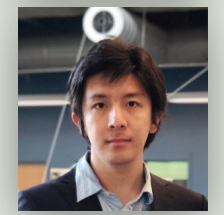










https://offensive-ai-lab.github.io/

BIRD
Israel-U.S.
Binational Industrial Research
and Development Foundation

CBG
Cyber@Ben-Gurion
University of the Negev

Agenda

- Background & Motivation
- Existing Defenses
- Our solution - Deepfake-CAPTCHA
- Usability Demo

BIRD
Israel-U.S.
Binational Industrial Research
and Development Foundation
CBG
Cyber@Ben-Gurion
University of the Negev

Background & Motivation

- Deep-Fake (DF)
- Real-Time Deep-Fakes (RT-DFs)

What is Deep-Fake (DF)

“Any believable media generated by a deep neural network”

A decade of DF improvements

BIRD
Israel-U.S.
Binational Industrial Research
and Development Foundation

CBG
Cyber@Ben-Gurion
University of the Negev

Two Deep-Fake Types

Video

Samsung’s MegaPortrait Generates Deepfake Videos from Still Images

Audio

Respeeecher Demo May 2022

BIRD
Israel-U.S.
Binational Industrial Research
and Development Foundation

CBG
Cyber@Ben-Gurion
University of the Negev

Video Deepfake

Can be used for art

The Heart Part 5
by Kendrick Lamar

BIRD
Israel-U.S.
Binational Industrial Research
and Development Foundation

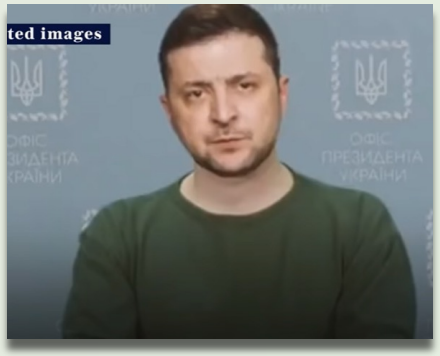
CBG
Cyber@Ben-Gurion
University of the Negev

Why Should We **Care**?

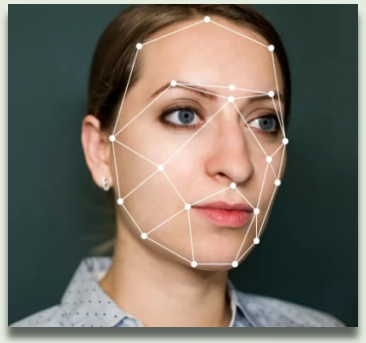
Deepfake abuse examples

Malicious Goals:

- Blackmail
- Abuse
- Misinformation
- Scams
- Media Tampering
- Fake Porn



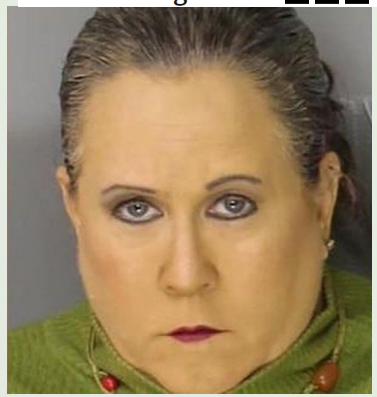
https://www.wired.com/story/zelensky-deepfake-facebook-twitter-playbook/



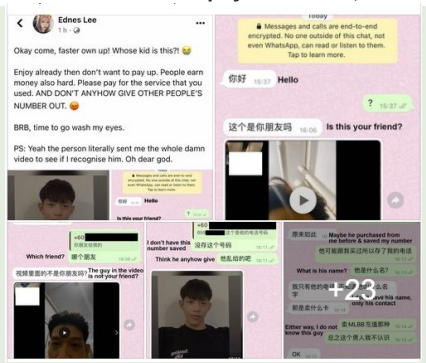
https://www.technologyreview.com/2021/02/12/1018222/deepfake-revenge-porn-coming-ban/

https://www.motherjones.com/politics/2019/03/deepfake-gabon-ali-bongo/

Mother 'used deepfake to frame cheerleading rivals' BBC



Singaporean man's face ends up in deepfake porn after he refuses to pay hacker $5,800 yahoo!news



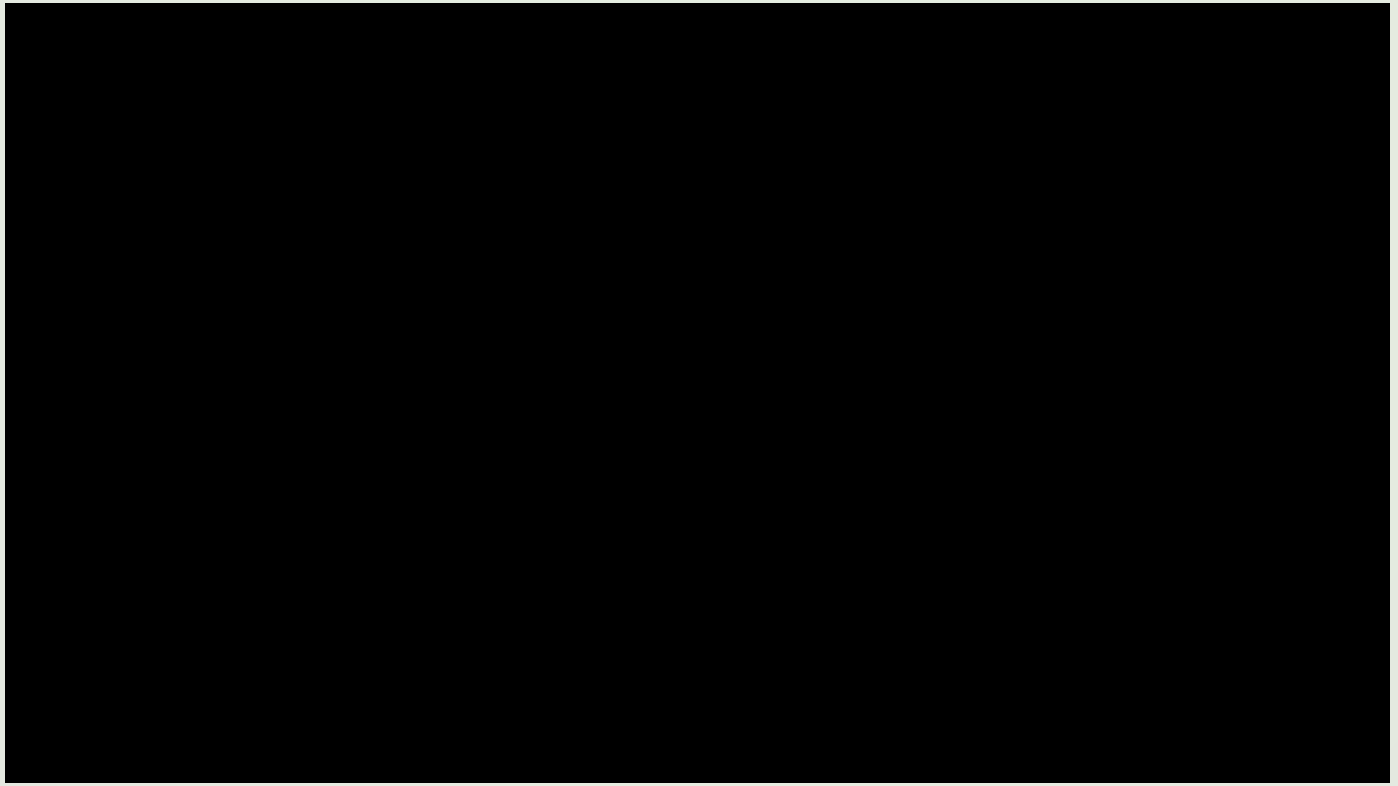
https://www.businessinsider.com/video-boris-johnson-endorses-jeremy-corbyn-in-convincing-deepfake-2019-11

BIRD Israel-U.S. Binational Industrial Research and Development Foundation

CBG Cyber@Ben-Gurion University of the Negev

Why Should We **Care**?

Combining Video and Audio Deepfake - dangerous



Boris Johnson claims that his opponent is more suitable

BIRD Israel-U.S. Binational Industrial Research and Development Foundation

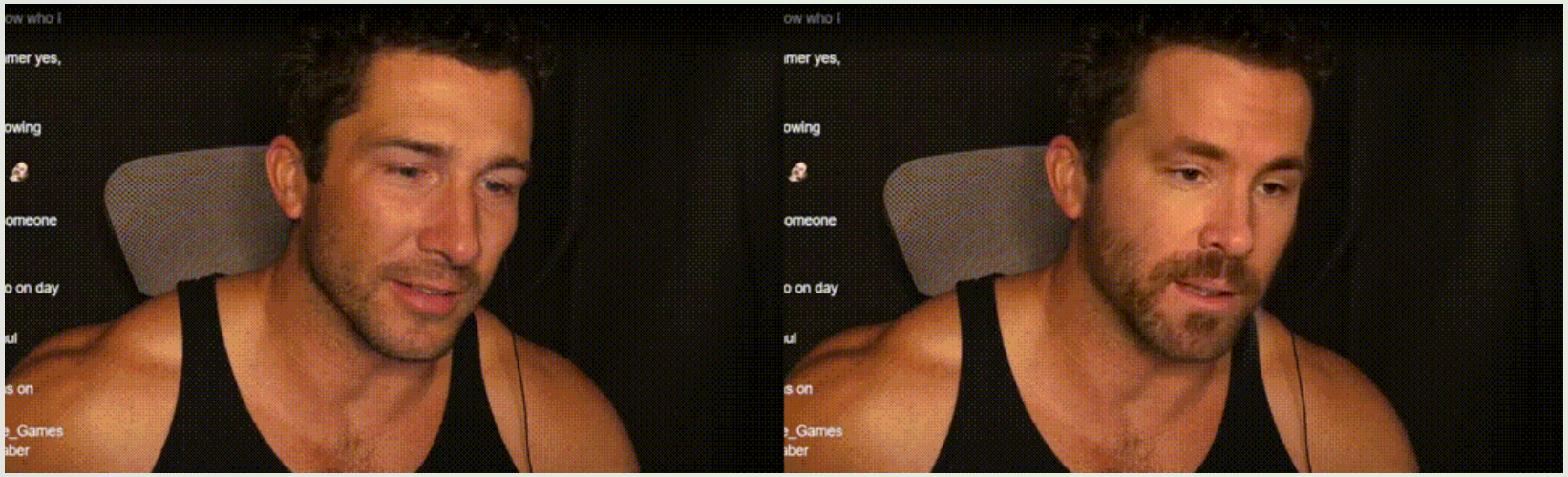
CBG Cyber@Ben-Gurion University of the Negev

Real-Time Deep-Fakes - An Emerging Threat

- Minimal **delay** which allows **interactive discussions**
- RT-DF enable to preform **Social engineering** attacks in two aspects:
  - **Psychology** - People confuse familiarity with authenticity
  - **Awareness** - RT-DFs are an unexpected attack vector

Forbes

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



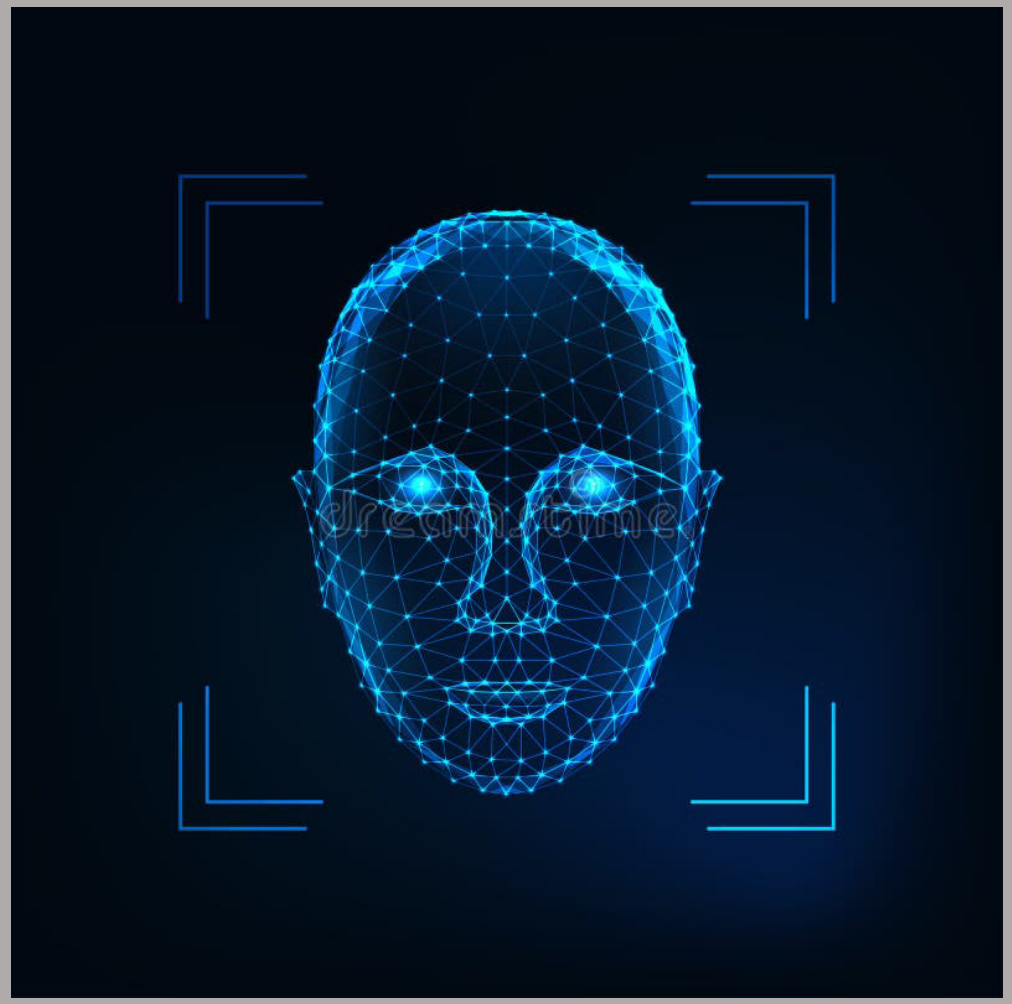
DeepFaceLab – open source real-time deepfake software

BIRD Israel-U.S. Binational Industrial Research and Development Foundation

CBG Cyber@Ben-Gurion University of the Negev

Existing Defenses

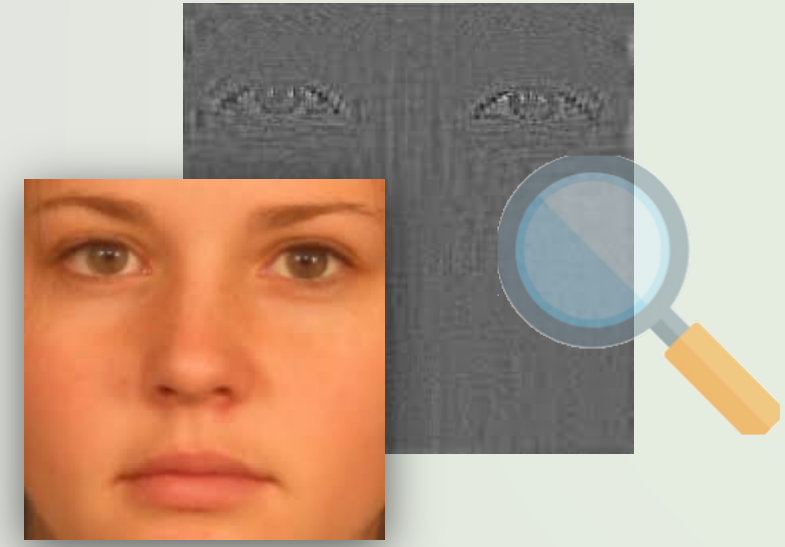
- Common Approaches
- Limitation



Existing DF Defenses

Common Approaches

1. **Analytical**



signal extraction

2. **Directed**
(*Artifact-Specific)*

ML focused on specific features

3. **Undirected**
   - Classifiers
   - Anomaly Detectors

ML given all features
(learns own features)

BIRD
Israel-U.S.
Binational Industrial Research
and Development Foundation

CBG
Cyber@Ben-Gurion
University of the Negev

The **Problem** with Current Defenses

When it comes to detecting RT-DFs…

Practically

- Defenses must be efficient
- No reference data

Delivery

- Defenses must be flexible
- Expect specific types of artifacts/DF-pipelines

Quality

- Defenses must be robust: dealing with noise and compression
- Noise and compression decrease performance

BIRD
Israel-U.S.
Binational Industrial Research
and Development Foundation

CBG
Cyber@Ben-Gurion
University of the Negev

Current Defenses Summary

- **Passive defenses -** Artifacts are expected in deepfake
- Quality of deepfake is **rapidly improving**
- Argument: Seeking out existing artifacts is a **losing game**

Our solution - Deepfake-CAPTCHA

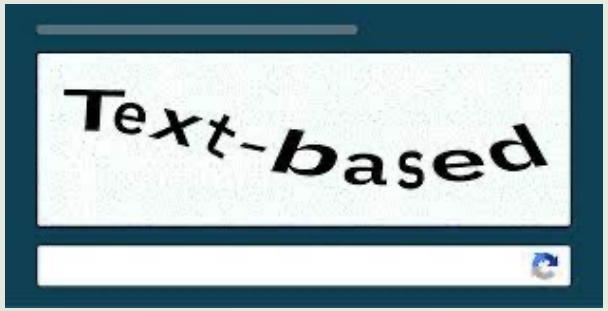
- Deepfake-CAPTCHA system
- Evaluation

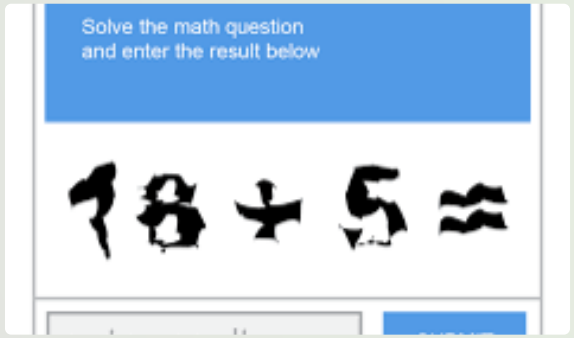
BIRD
Israel-U.S.
Binational Industrial Research
and Development Foundation

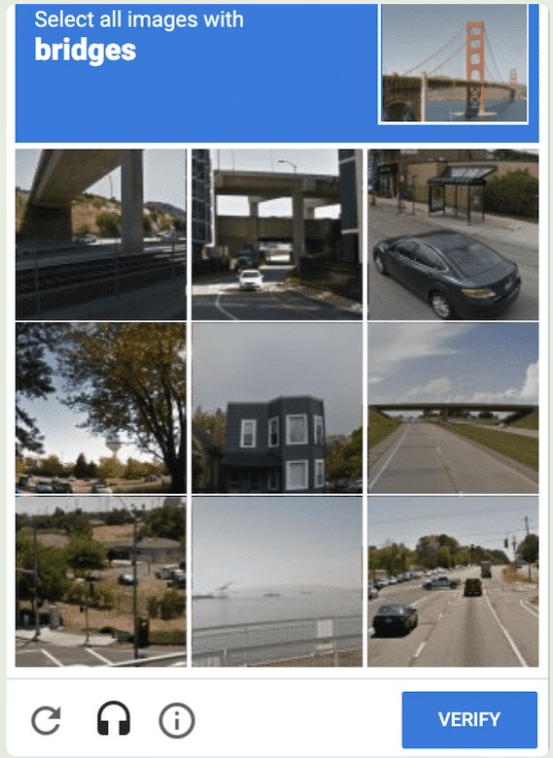
CBG
Cyber@Ben-Gurion
University of the Negev

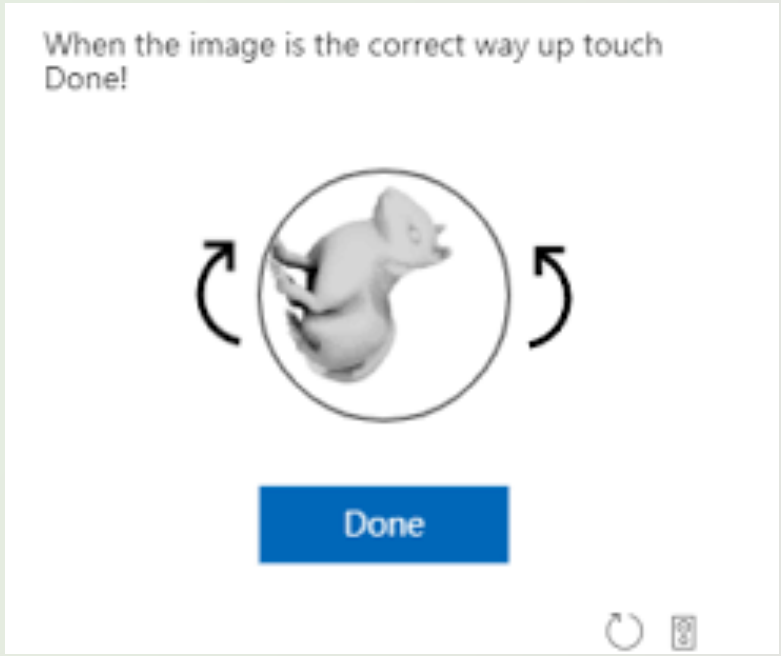
CAPTCHA - An Active Defense

"A cryptographic protocol whose underlying hardness assumption is based on an AI problem."
CAPTCHA: Using Hard AI Problems for Security, Luis von Ahn et. al







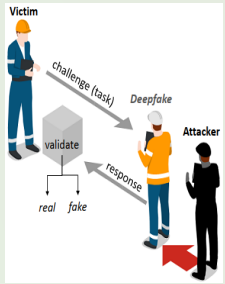


BIRD
Israel-U.S.
Binational Industrial Research
and Development Foundation

CBG
Cyber@Ben-Gurion
University of the Negev

Our solution - Deepfake-CAPTCHA

The Framework:



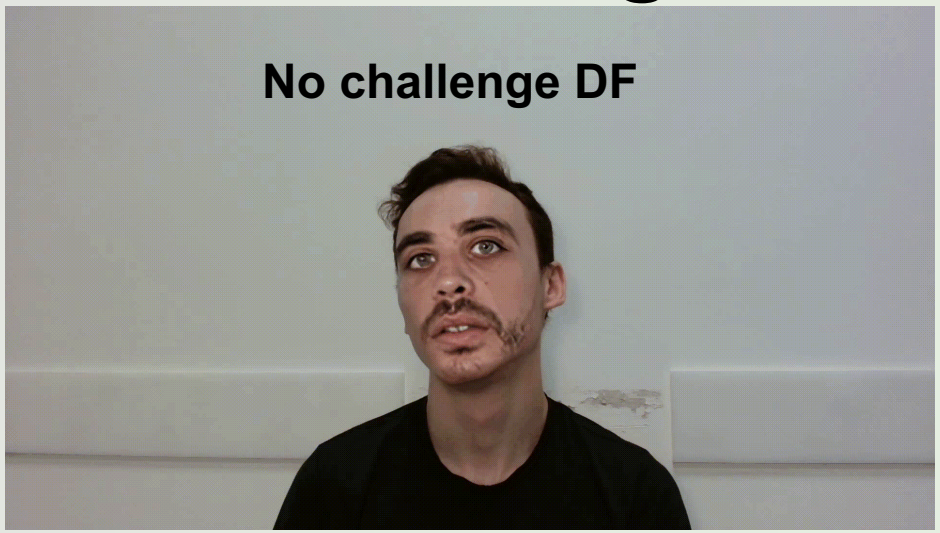
Challenge the caller to create content:

1. Victim/server sends a challenge
2. Caller responds
3. Verification of challenge

BIRD Israel-U.S. Binational Industrial Research and Development Foundation

CBG Cyber@Ben-Gurion University of the Negev

Video Challenge Example



BIRD
Israel-U.S.
Binational Industrial Research
and Development Foundation

CBG
Cyber@Ben-Gurion
University of the Negev

Audio Challenge Example

Real Audio

Fake Audio

Fake challenge – Clapping

Fake challenge – Playback

BIRD
Israel-U.S.
Binational Industrial Research
and Development Foundation

CBG
Cyber@Ben-Gurion
University of the Negev

The Challenge Purpose

Forces creating content with the following constraints:

Realism

Identity

Task

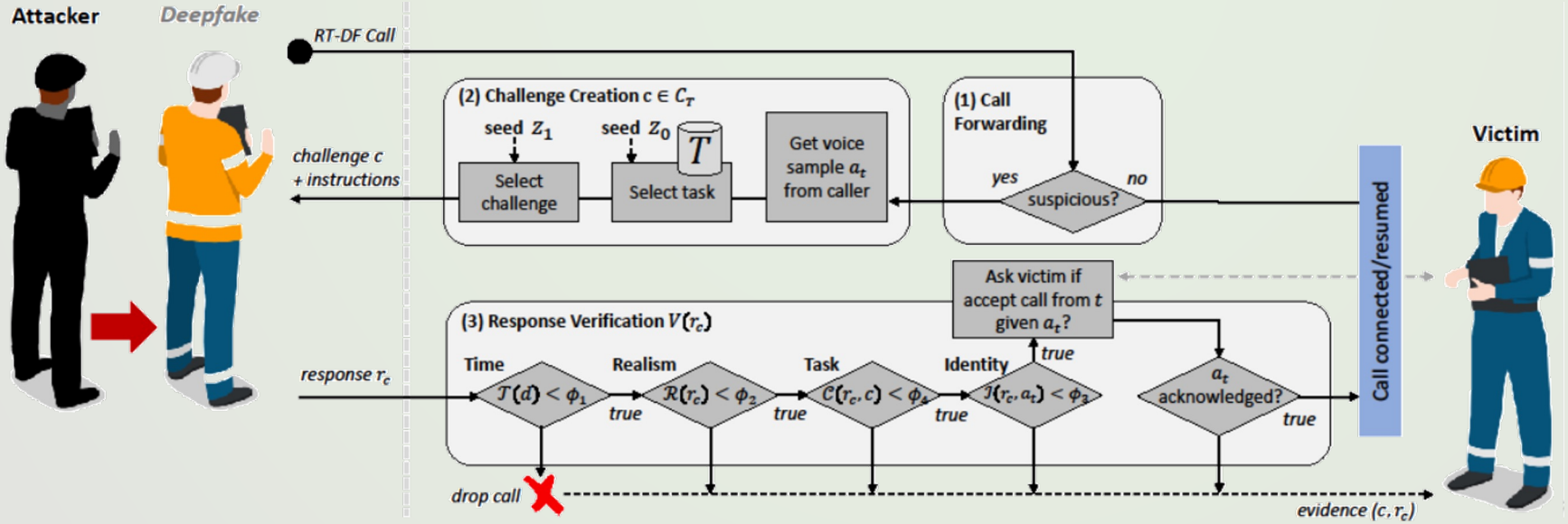
Time

BIRD
Israel-U.S.
Binational Industrial Research
and Development Foundation

CBG
Cyber@Ben-Gurion
University of the Negev

DF Detector Components

- **Realism Verification *(R)*** - Anomaly detector
- **Identity Verification *(I)*** – Speaker recognition
- **Task Verification *(C)*** – Task classifier
- **Time Verification *(T)*** - Time constraint

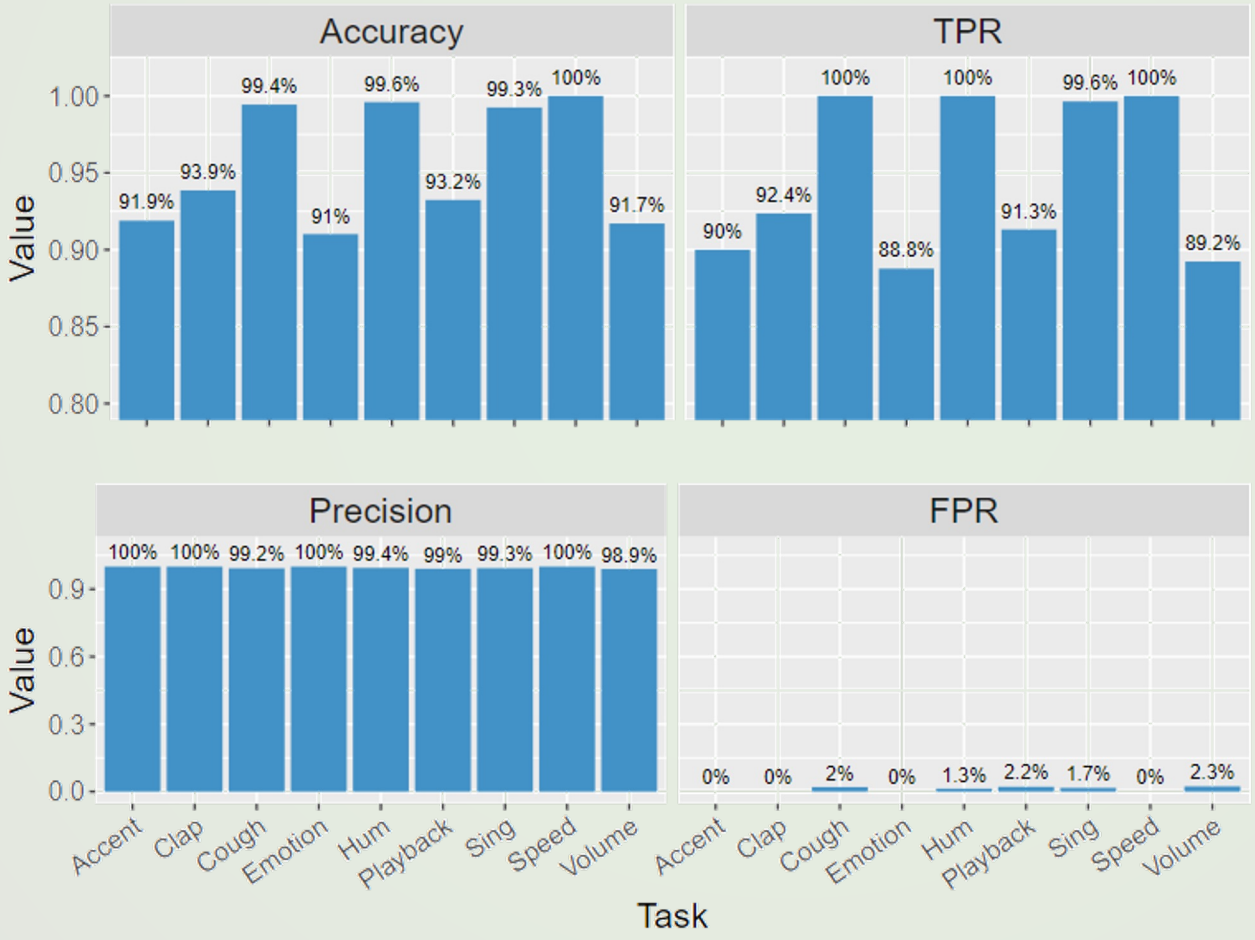


BIRD
Israel-U.S.
Binational Industrial Research
and Development Foundation

CBG
Cyber@Ben-Gurion
University of the Negev

Audio D-CAPTCHA system - Results

The performance of the ensure D-CAPTCHA system (end-to-end)



BIRD Israel-U.S. Binational Industrial Research and Development Foundation

CBG Cyber@Ben-Gurion University of the Negev

Usability Demo

BIRD
Israel-U.S.
Binational Industrial Research
and Development Foundation

CBG
Cyber@Ben-Gurion
University of the Negev

Usability Demo

BIRD
Israel-U.S.
Binational Industrial Research
and Development Foundation

CBG
Cyber@Ben-Gurion
University of the Negev

Thank You!

Frankovits, G.*, Yasur, L.*, Grabovski, F. M., & Mirsky, Y. (2023). Deepfake CAPTCHA: A Method for Preventing Fake Calls (ASIA CCS ’23 Conference)

Frankovits, G., Mirsky, Y. (2023). Discussion Paper: The Threat of Real Time Deepfakes (WDC ’23 Workshop)

Deepfake CAPTCHA Demo

Questions

Correspondence:
guyfrank@post.bgu.ac.il

https://offensive-ai-lab.github.io/

BIRD
Israel-U.S.
Binational Industrial Research
and Development Foundation

CBG
Cyber@Ben-Gurion
University of the Negev