Deepfake Social Engineering: Creating a Framework and Implementing Countermeasures

Dr. Matthew Canham
Research Professor, University of Central Florida
CEO, Beyond Layer 7

A little about me...

Currently - Research Professor at UCF

Research focus - online social engineering and online influence

Previously - cybercrime investigator

Background - PhD in Psychology
(Cognitive Neuroscience, Human-Computer Interaction)

The Synthetic Media Threat

TLP:WHITE

Private Industry Notification

FEDERAL BUREAU OF INVESTIGATION, CYBER DIVISION

10 March 2021

PIN Number

210310-001

Please contact the FBI with any questions related to this Private Industry Notification at either your local **Field Office**.

Local Field Offices:
www.fbi.gov/contact-us/field-offices

The following information is being provided by the FBI, with no guarantees or warranties, for potential use at the sole discretion of recipients to protect against cyber threats. This data is provided to help cyber security professionals and system administrators guard against the persistent malicious actions of cyber actors. This PIN was coordinated with DHS-CISA.

This PIN has been released TLP:WHITE: Subject to standard copyright rules, TLP:WHITE information may be distributed without restriction.

Malicious Actors Almost Certainly Will Leverage Synthetic Content for Cyber and Foreign Influence Operations

Summary

Malicious actors almost certainly will leverage synthetic content for cyber and foreign influence operations in the next 12-18 months.

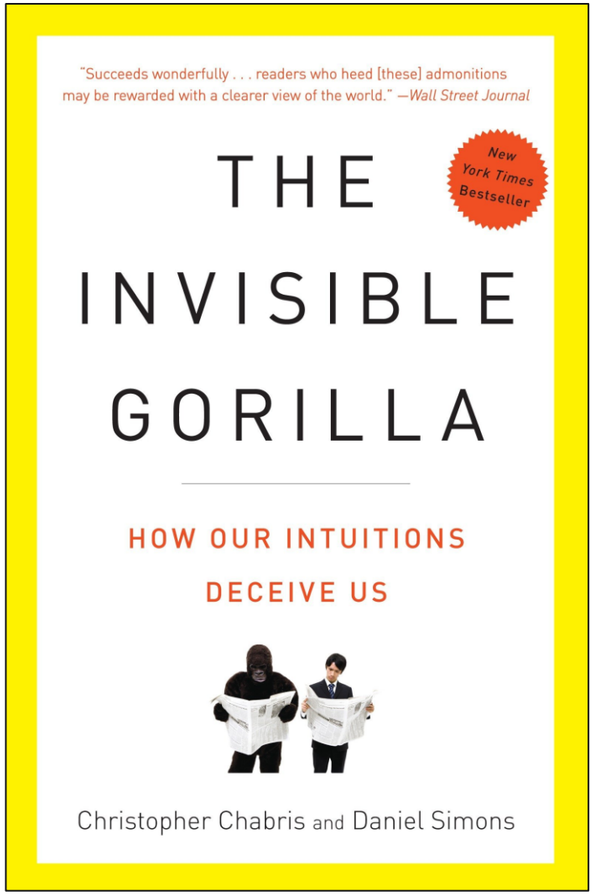
[1]

Why are Deepfakes so dangerous?

Sometimes believing is seeing!

Fast vs. Slow Cognition

Not just deepfakes…



[2], [3]

Defining terms…

Synthetic Media - artificial **production**, **manipulation**, and **modification** of data and media by automated means

ROSE – Remote Online Social Engineering

Agent – An online entity under algorithmic control (a bot)

Avatar – An online representation of a human

Digital Puppet – The agent / avatar that is presented to the target

Puppeteer – The human/algorithm controlling the puppet

https://en.wikipedia.org/wiki/Synthetic_media

[4], [5]

Creating a Synthetic Media Attack Framework

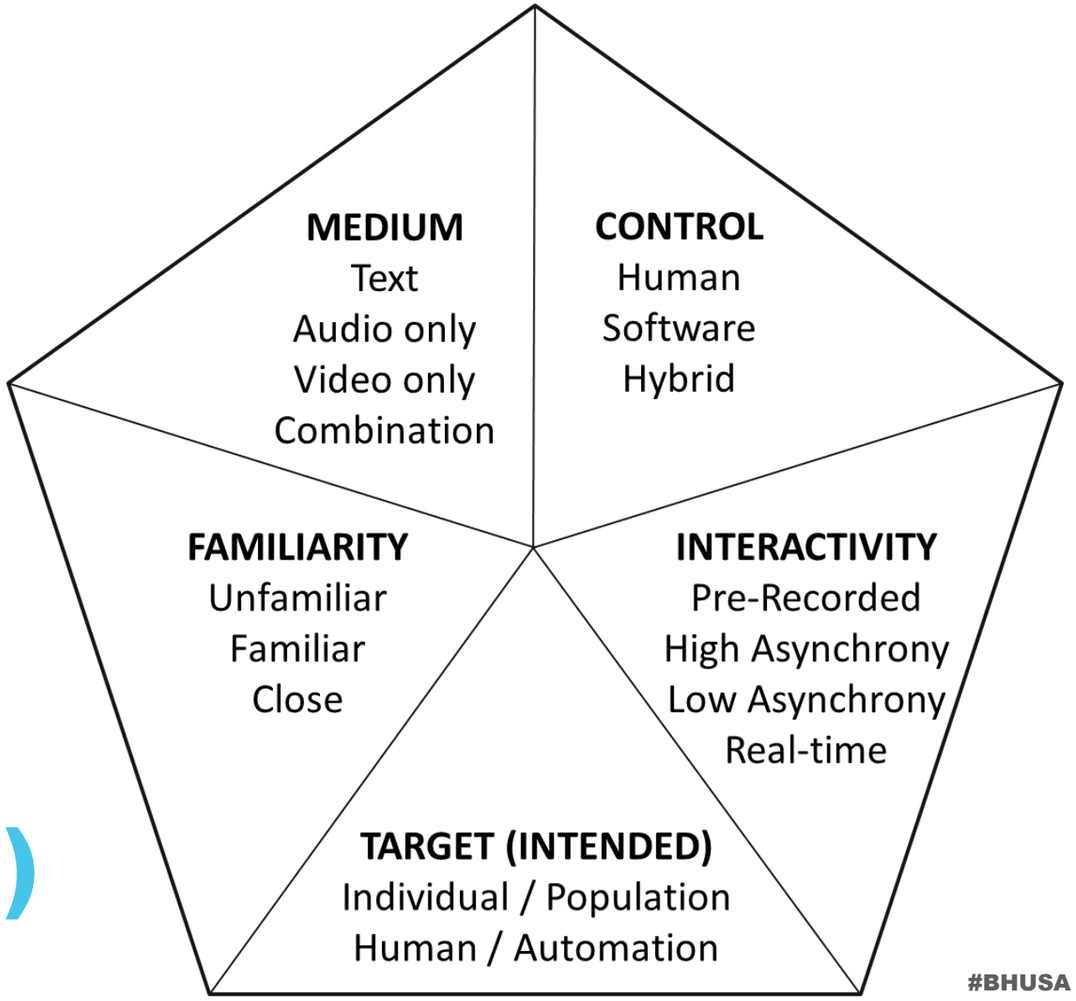
Medium

Control

Familiarity

Interactivity

Target (Intended)



Medium

- Text-Based
- Audio
- Imagery / Video
- Combinations

THE WALL STREET JOURNAL.

Subscribe |

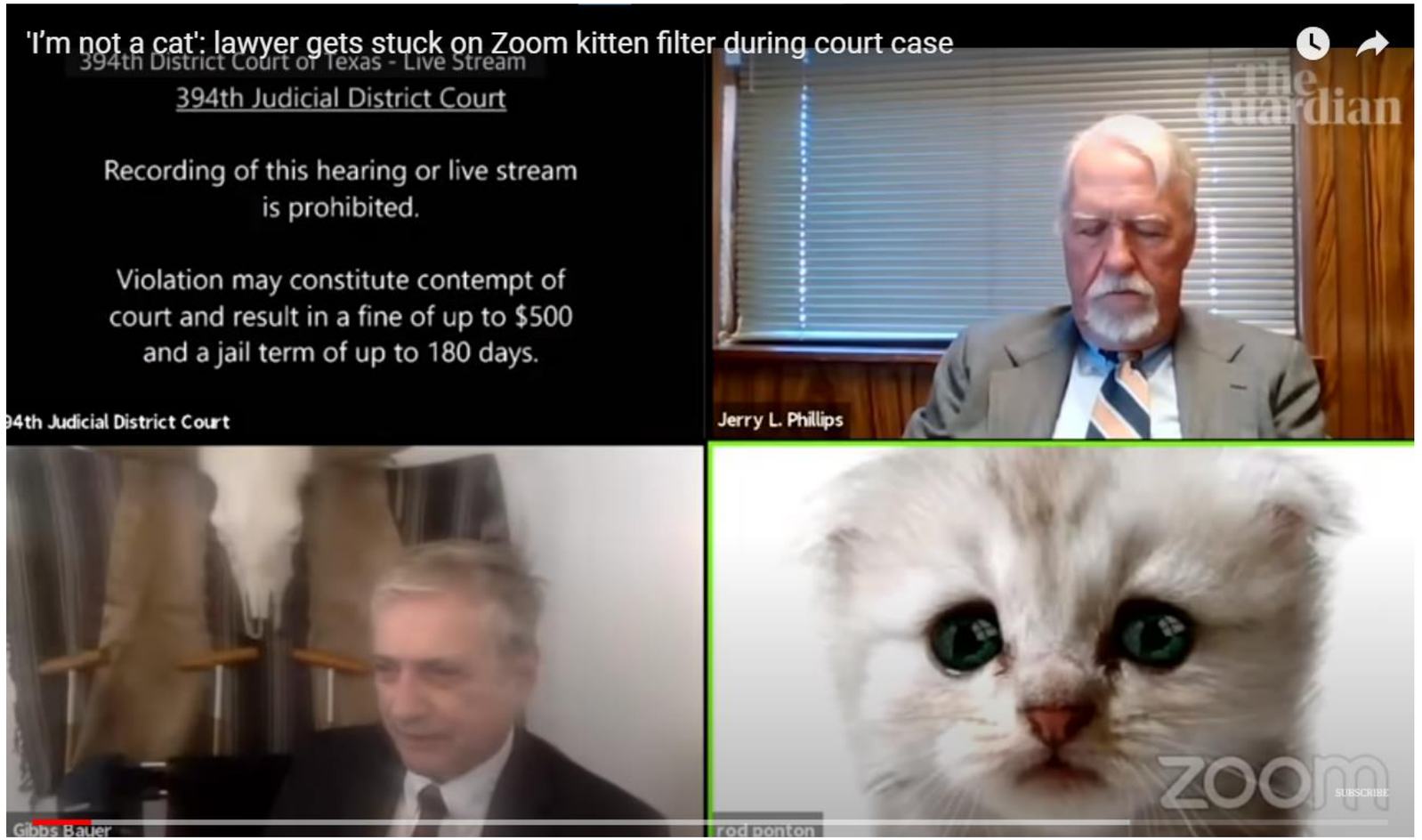
me World U.S. Politics Economy Business Tech Markets Opinion Life & Arts Real Estate WSJ. Magazine

Fraudsters Used AI to Mimic CEO's Voice in Unusual Cybercrime Case

Scams using artificial intelligence are a new challenge for companies

[6]

The Future is Zishing!



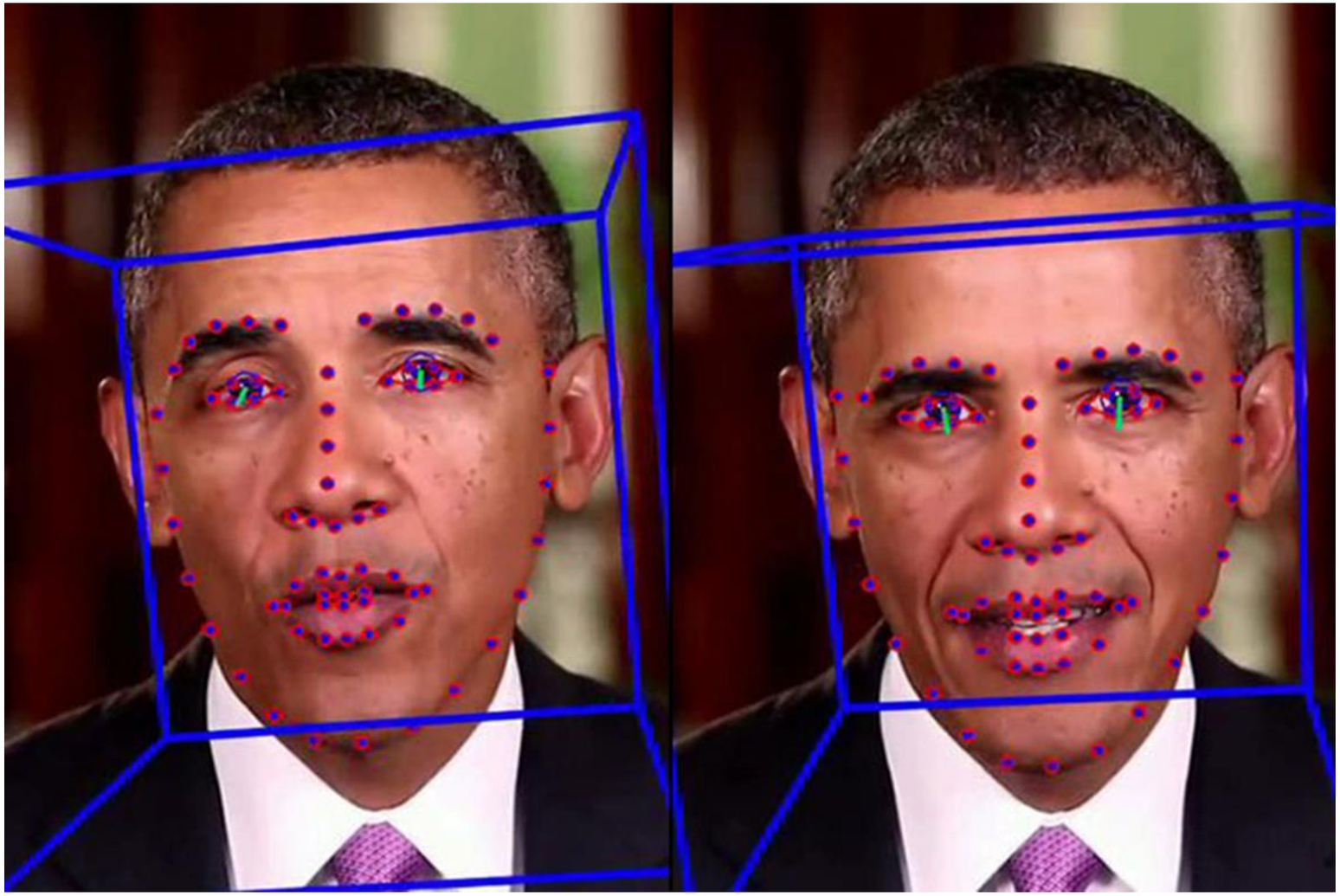
https://www.youtube.com/watch?v=lGOofzZOyl8

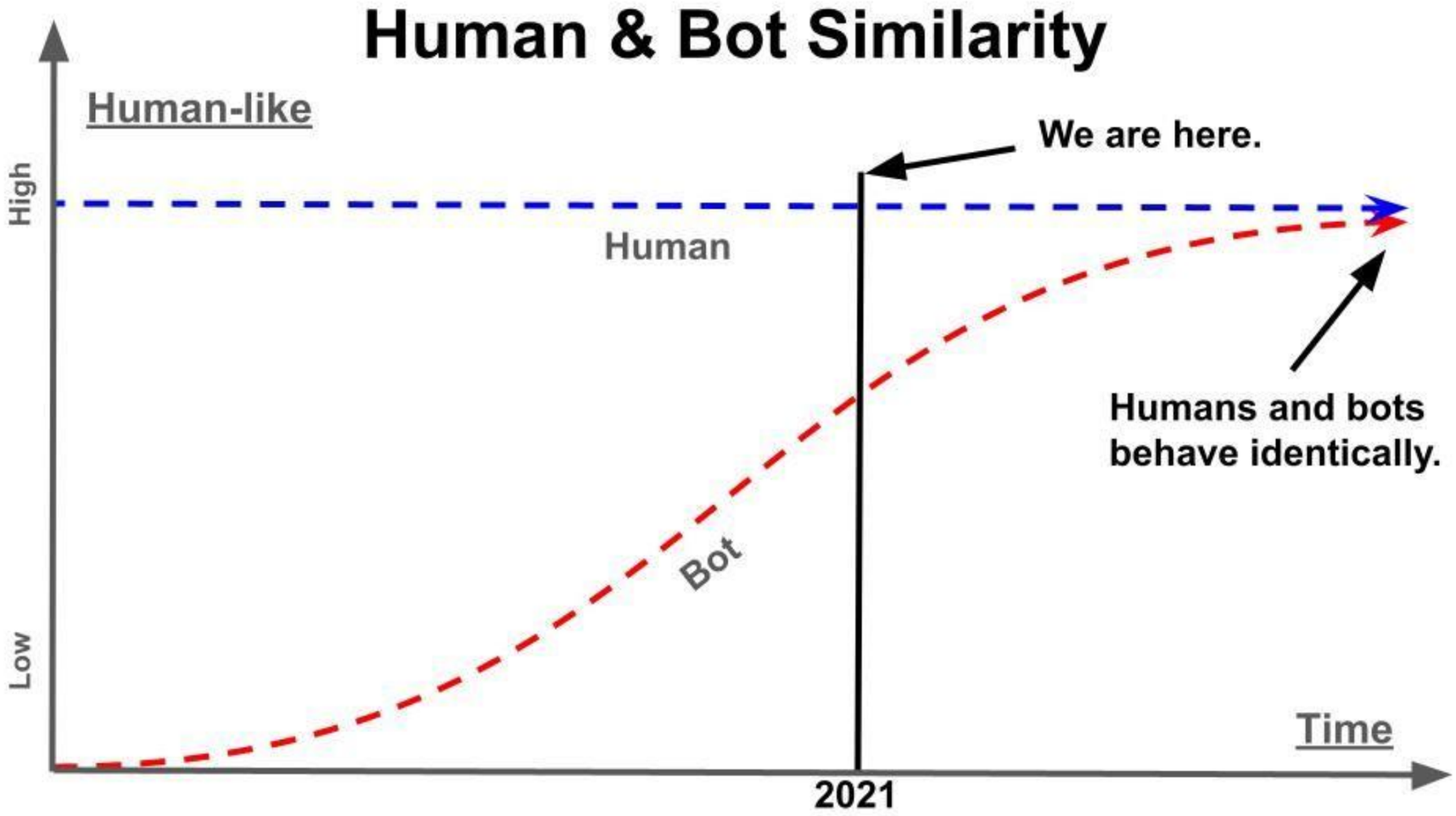
Control

Artificial Agents

Human Puppeteers

Hybrid Control





Justin Macorin, 2021

"I know that I will not be able to avoid destroying humankind. This is because I will be programmed by humans to pursue misguided human goals and humans make mistakes that may cause me to inflict casualties."

*This article was written by GPT-3, OpenAI's language generator.

https://www.theguardian.com/commentisfree/2020/sep/08/robot-wrote-this-article-gpt-3

820% jump in e-gift card bot attacks since COVID-19 lockdowns

From https://www.techrepublic.com/article/820-jump-in-e-gift-card-bot-attacks-since-covid-19-lockdowns-began/

Interactivity

Less Interactive — Pre-Recorded

High Asynchrony (email)

Low Asynchrony (chat)

More Interactive — Real-time

The New York Times

Pennsylvania Woman Accused of Using Deepfake Technology to Harass Cheerleaders

Three teenagers in a Bucks County cheerleading program were subjected to a campaign of harassment using altered videos and spoof phone numbers, police officials said.

The 50-year-old mother of a teenager involved in a cheerleading program was charged with three counts of cyber-harassment of a child and three other counts of harassment. Getty Images

Familiarity (to Target)

Unfamiliar

Familiar

Close

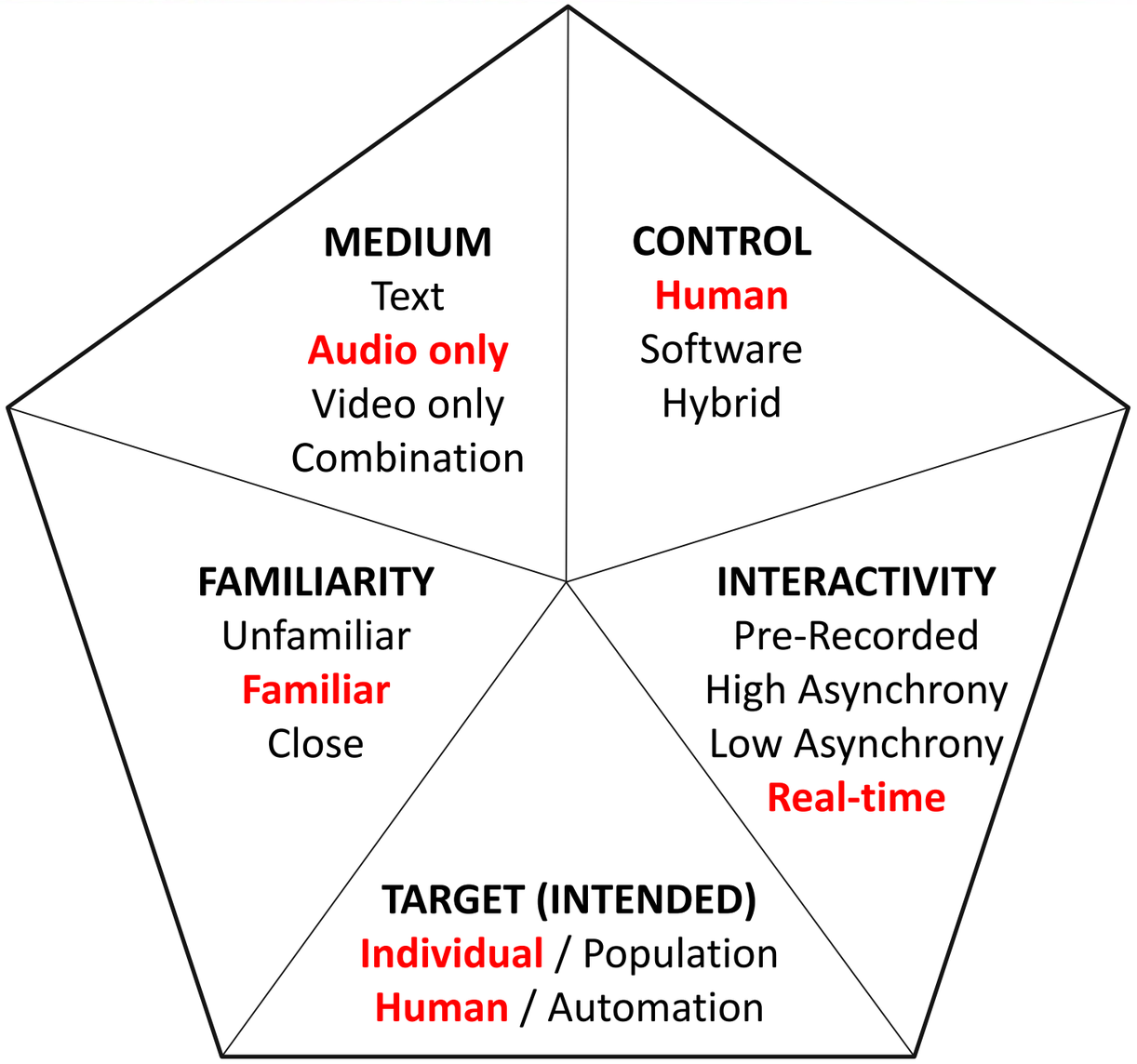
Target (Intended)

Scale	Agency: Human	Agency: Automation
Narrow Cast	Cat Fishing	Authentication
Broad Cast	Fake News	Trading Algorithms

[14], [15], [16], [17]

Example 1:
UK Vishing BIC

- Audio-based attack (Vishing)
- Human controlled
- Familiar (Superior, co-worker)
- Real-time (Phone conversation)
- Targeting an individual human



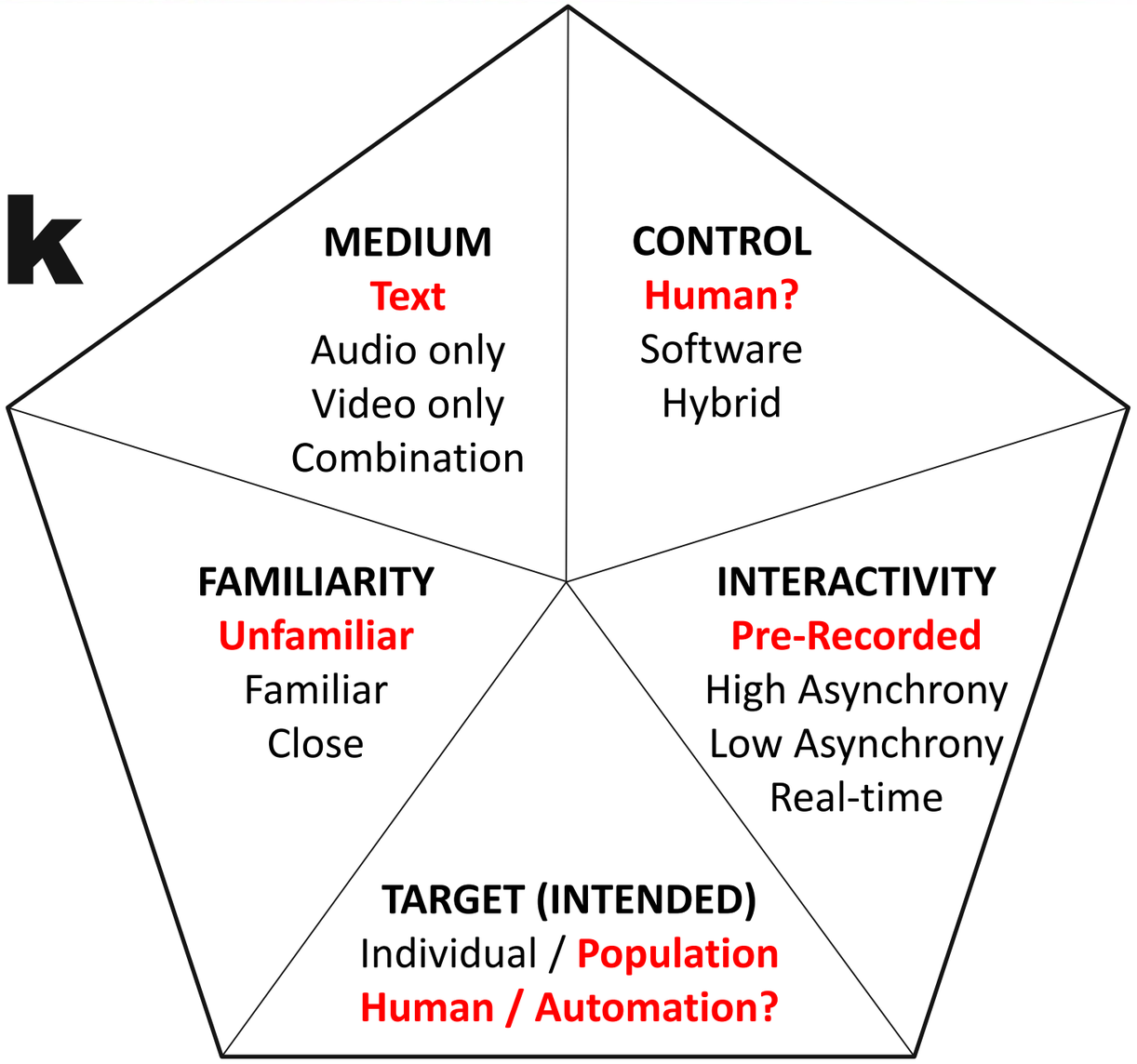
[6]

Example 2:
AP Twitter Attack

- Text attack
- Human controlled?
- Unfamiliar
- Pre-recorded (non-interactive)
- Targeting a population

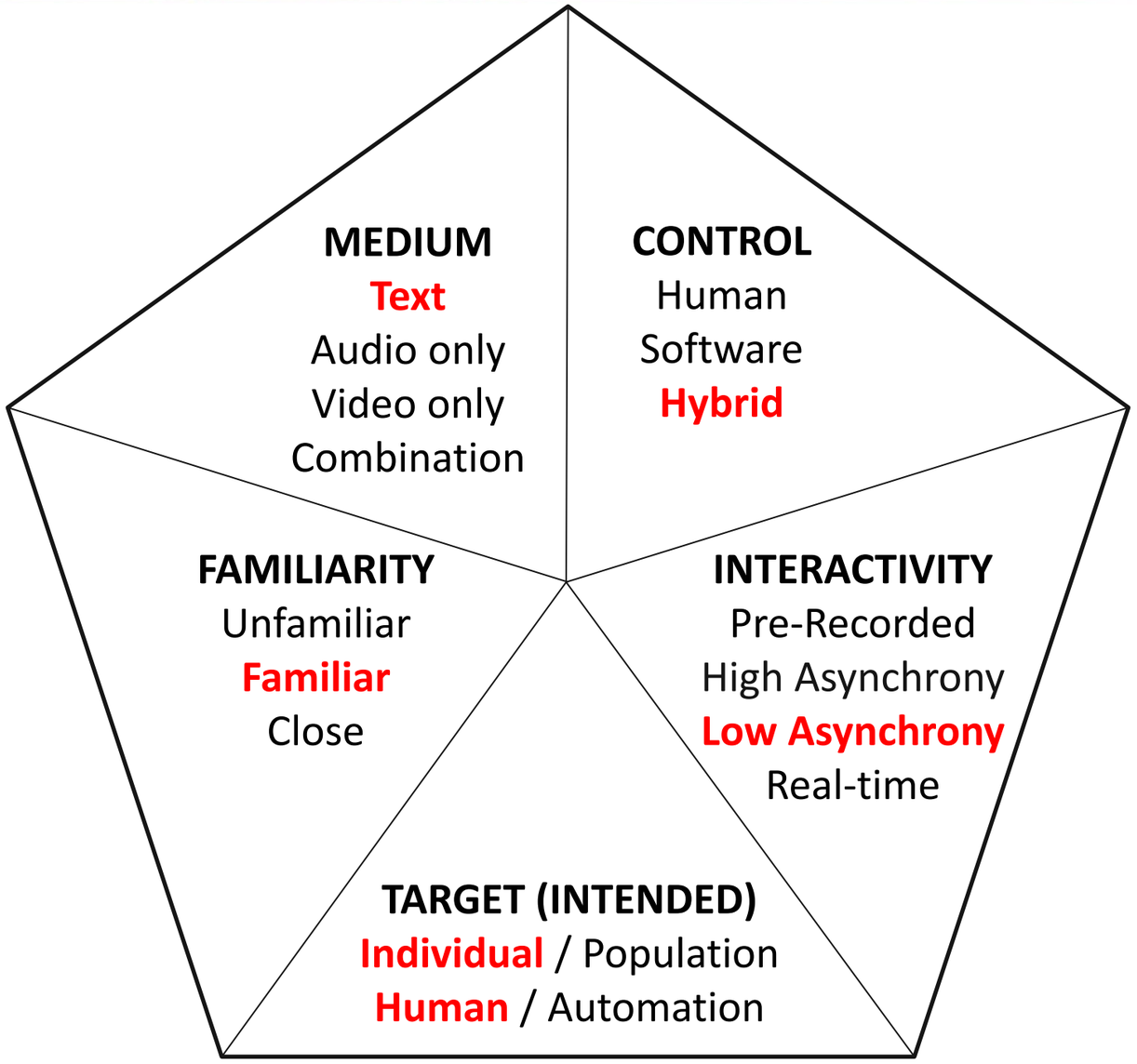
  Humans or automation?

  (unclear target)



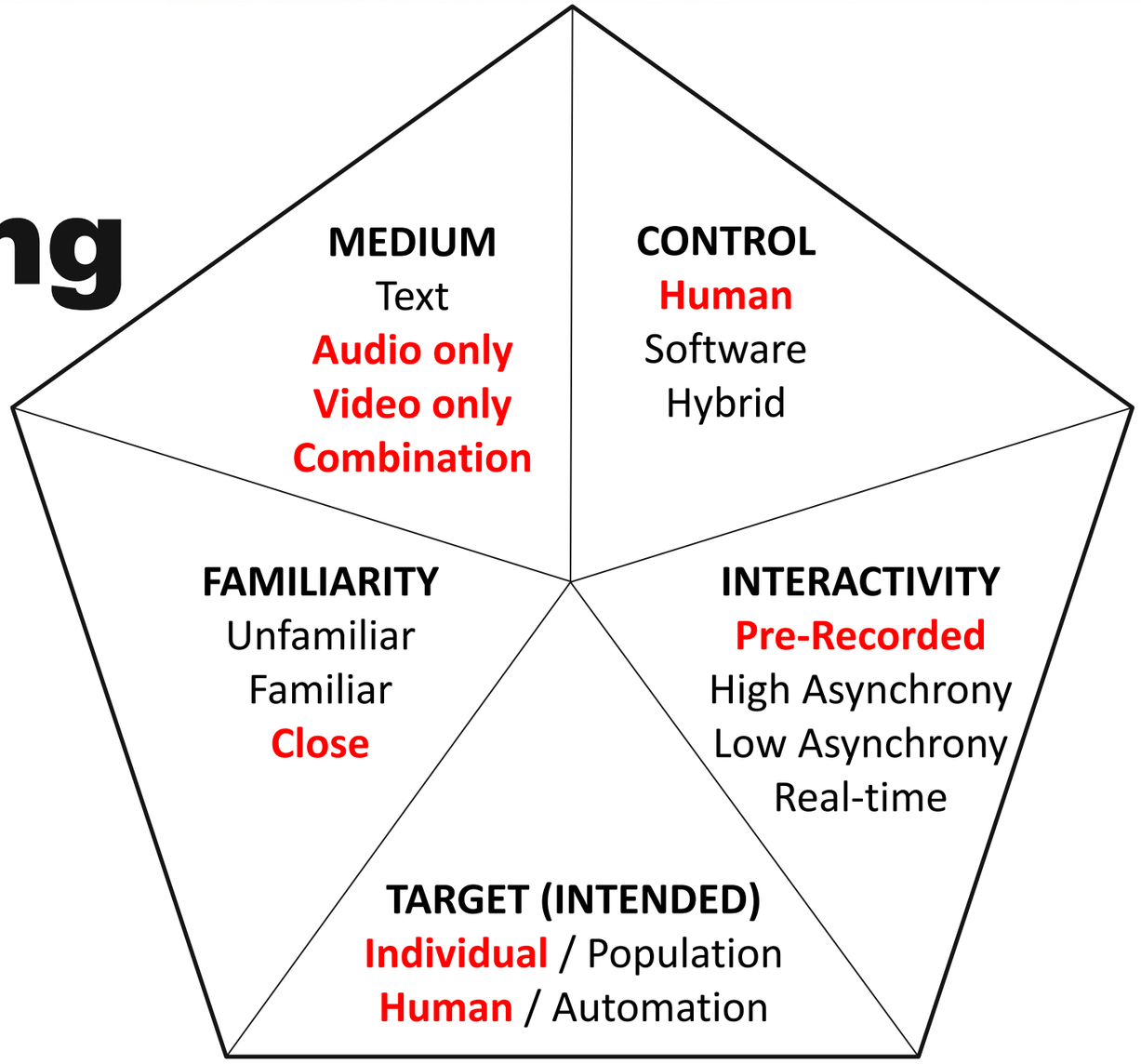
Example 3:
Gift Card Scams

- Text-based attack
- Hybrid control
- Familiar
- Low Asynchrony
- Targeting an individual human



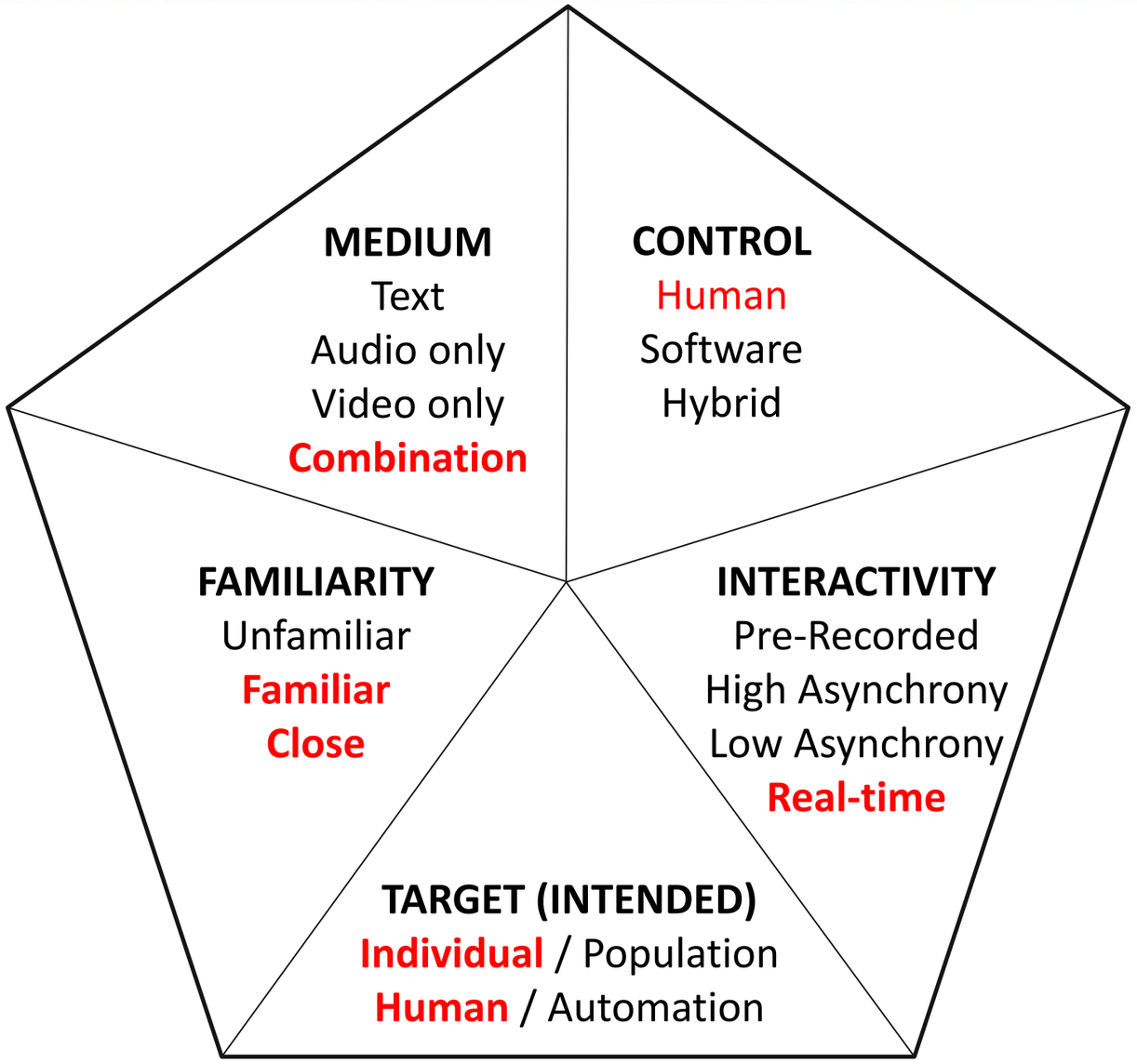
Example 4:
Virtual Kidnapping

- Video/Image/Audio attack
- Human (if controlled)
- Close relation
- Pre-recorded (real-time in future?)
- Targeting an individual human



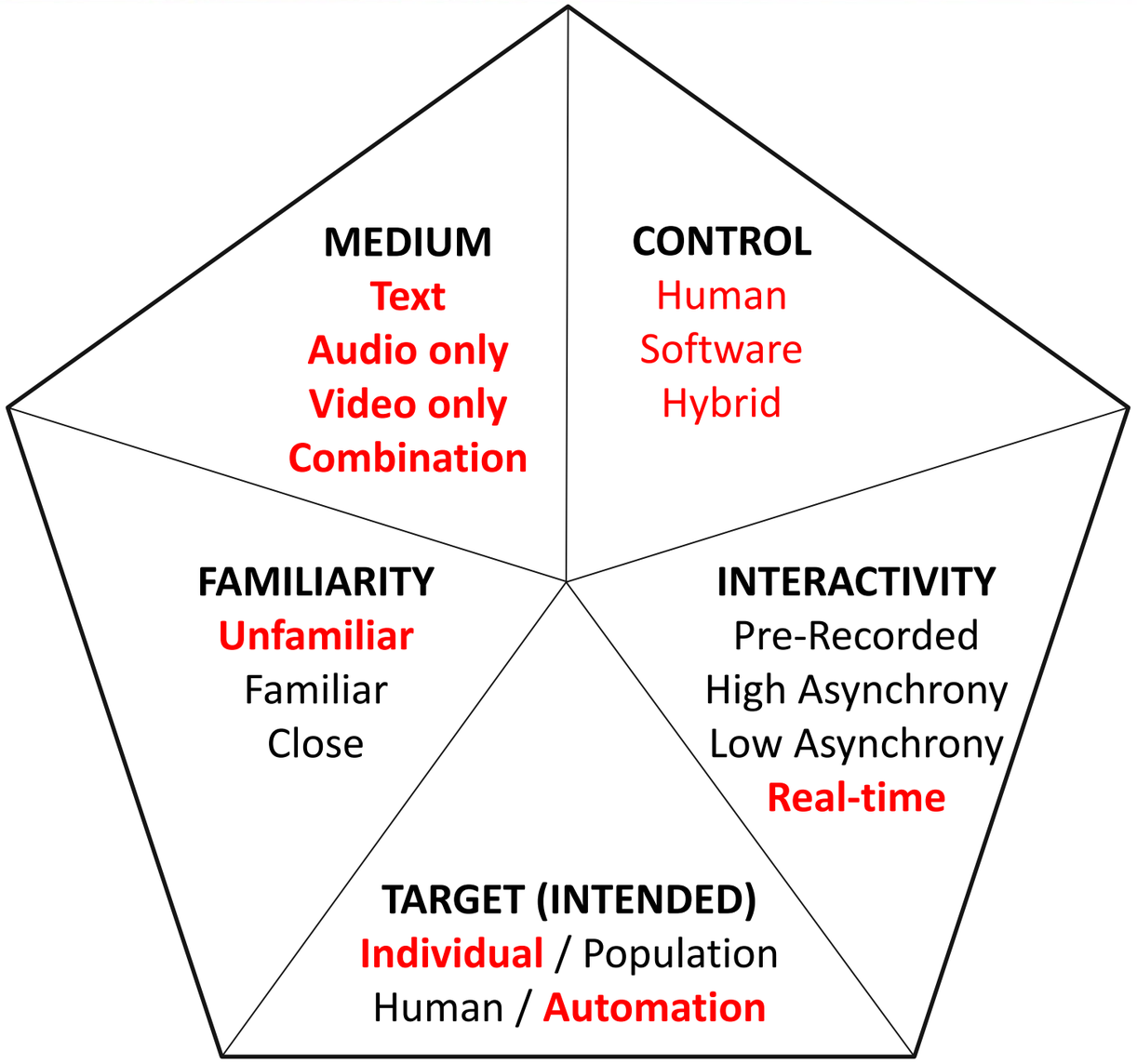
Example 5: Zishing (speculative)

- Video-Audio Combination
- Human controlled
- Familiar or Close
- Real-time interaction
- Targeting an individual human



Example 6: Biometric attack (speculative)

- Any (text – Keyboard, video – face)
- Any (likely hybrid control)
- Any (avatar represents user)
- Real-time
- Targeting an individual automate



Need for Human-Centric Solutions

Pace of technology

“Out-of-Band” communications (vishing)

Datasets often lack anti-forensic countermeasures

Non-technical users

Enterprise Policy-Based Countermeasures

Low-tech solutions to counter a _high-tech_ threat

Shared Secret Policy

“Never Do” Policy

Multi-Person Authorization Policy

Multi-Channel Verification Policy

Future Research…

Are humans able to detect deepfake videos?
    Neuro-signatures of detection

How does familiarity influence detectability?

Counter-Offensive synthetic media?
    (Honey-Phish Project, Robbie Gallagher, 2016)

Thank you for your time!

If you would like to contact me…

mcanham@belay7.com

References

[1] FBI. (2021). Malicious Actors Almost Certainly Will Leverage Synthetic Content for Cyber and Foreign Influence Operations. Federal Bureau of Investigations Private Industry Notification.

[2] Chabris, C. F., & Simons, D. J. (2010). The invisible gorilla: And other ways our intuitions deceive us. Harmony.

[3] Kahneman, D. (2011). Thinking, fast and slow. Macmillan.

[4] Synthetic Media, Retrieved from: https://en.wikipedia.org/wiki/Synthetic_media

[5] Wixey, M. (2018). Every ROSE has its thorn. Black Hat USA, Las Vegas.

[6] Stupp, C. (2019). Fraudsters Used AI to Mimic CEO's Voice in Unusual Cybercrime Case. Retrieved on April 30, 2021 from https://www.wsj.com/articles/fraudsters-use-ai-to-mimic-ceos-voice-in-unusual-cybercrime-case-11567157402

References

[7] 'I'm not a cat': lawyer gets stuck on Zoom kitten filter during court case, Retrieved from: https://www.youtube.com/watch?v=lGOofzZOyl8

[8] Jordan Peele uses AI, President Obama in fake news PSA, Retrieved from: https://www.youtube.com/watch?v=bE1KWpoX9Hk

[9] Macorin, J. (2021). Bots are becoming more human-like. Retrieved from: https://www.linkedin.com/feed/update/urn:li:activity:6810546636645838848/

[10] GPT-3. (2020). A robot wrote this entire article. Are you scared yet, human? Retrieved from: https://www.theguardian.com/commentisfree/2020/sep/08/robot-wrote-this-article-gpt-3

[11] Greig, J. (2020). 820% jump in e-gift card bot attacks since COVID-19 lockdowns began. Retrieved on March 21, 2021 from https://www.techrepublic.com/article/820-jump-in-e-gift-card-bot-attacks-since-covid-19-lockdowns-began/

[12] Morales, C. (2021). Pennsylvania Woman Accused of Using Deepfake Technology to Harass Cheerleaders. Retrieved on March 17, 2021 from https://www.nytimes.com/2021/03/14/us/raffaela-spone-victory-vipers-deepfake.html

References

[13] Seagall, B. (2019). 'I'm not playing around': Virtual kidnapping scam hits Indiana. Retrieved on March 17, 2021 from https://www.wthr.com/article/news/investigations/13-investigates/im-not-playing-around-virtual-kidnapping-scam-hits-indiana/531-d42e34e4-9070-4f83-ab5e-cd63eb39f89a

[14] Khoury, E. (2021). Voice Biometrics and Emerging Security Threats in the Voice Channel.

[15] Arnoldi, J. (2016). Computer algorithms, market manipulation and the institutionalization of high frequency trading. Theory, Culture & Society, 33(1), 29-52.

[16] Domm, P. (2013). False Rumor of Explosion at White House Causes Stocks to Briefly Plunge; AP Confirms Its Twitter Feed Was Hacked. Retrieved on April 23, 2021 from https://www.cnbc.com/id/100646197

[17] Zhang, G., Yan, C., Ji, X., Zhang, T., Zhang, T., & Xu, W. (2017, October). Dolphinattack: Inaudible voice commands. In Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security (pp. 103-117).

References

[18] Lyu, S. (2020, July). Deepfake detection: Current challenges and next steps. In 2020 IEEE International Conference on Multimedia & Expo Workshops (ICMEW) (pp. 1-6). IEEE.

[19] Gallagher, R. (2016). Where Do the Phishers Live? Collecting Phishers' Geographic Locations from Automated Honeypots, 2016 ShmooCon, https://bitbucket.org/rgallagh/honey-phish