CRICOS Provider No. 00103D | RTO Code 4909 | TEQSA No. PRV12151 (Australian University)

## ITECH1502 Cybersecurity Fundamentals

## Week 4 Lab Activities

# AI in Cyber Deception and Detection

**Lab Overview:**

Week 4 focuses on **AI’s dual role in cybersecurity**—as both a powerful engine for deception and an advanced mechanism for threat detection. Students will engage in two complementary lab tasks: the first develops practical skills in identifying and evaluating AI-generated deception (e.g., deepfakes, synthetic phishing), and the second builds capability in applying AI-powered detection tools to identify and analyze malicious activity. Together, these labs strengthen students’ technical proficiency, analytical thinking, and ethical decision-making in the evolving AI-driven security landscape.

**Learning Outcomes Assessed:**

By the end of this lab, students will be able to:

1. **Demonstrate practical skills in identifying, analyzing, and critically evaluating AI-generated deception content using free online tools**, and reflect on ethical considerations in balancing defensive and offensive uses of AI in cybersecurity.
2. **Apply and critically evaluate AI-based threat detection tools** for identifying and analyzing malicious activity, assess detection accuracy, and reflect on the operational and ethical implications of AI-powered monitoring in cybersecurity operations.

### ****🔐 Lab Task 1: Investigating AI in Cyber Deception – Deepfakes, Synthetic Media, and Deceptive AI Tools****

In this task, students will explore freely available online tools used to create and detect AI-generated deception, such as deepfake videos, synthetic voices, and phishing content. By generating deceptive artifacts and testing detection capabilities, students will gain first-hand insight into the sophistication of AI-enabled deception techniques, the challenges of detecting them, and the ethical considerations of their use in real-world contexts.

### ****Part A – Exploring AI Deception Tools****

Students will explore **at least 3 of the following free or freemium tools** to understand AI deception capabilities:

1. **DeepAI Text-to-Image Generator** – <https://deepai.org/machine-learning-model/text2img>   
   Create synthetic images for phishing or propaganda scenarios.
2. **FakeYou Voice Cloning** – <https://fakeyou.com>   
   Generate deepfake audio samples mimicking a target voice.
3. **This Person Does Not Exist** – <https://thispersondoesnotexist.com>   
   Identify signs that a profile image is AI-generated.
4. **Deepfakes Web (trial)** – <https://deepfakesweb.com>   
   Test deepfake video creation and analyze artifacts.
5. **Copy.ai or ChatGPT Free Version** –  
   Generate persuasive phishing email text.
6. **PhishSim by KnowBe4 (Free Phishing Simulation)** – <https://www.knowbe4.com/phish-prone>   
   Design a fake email to see how realistic it appears.
7. **Zao (Mobile App)** – AI face swap videos for social engineering awareness.
8. **Artbreeder** – <https://www.artbreeder.com>   
   Modify human faces to bypass biometric recognition.
9. **DALL·E Mini / Craiyon** – <https://www.craiyon.com>   
   Create misleading contextual imagery.

### ****Part B – Exploring AI Detection Tools****

Students will test **at least 2 of the following free AI detection tools**:

1. **Sensity AI (Free Trial)** – https://sensity.ai  
   Deepfake detection and risk scoring.
2. **Deepware Scanner** – https://deepware.ai  
   Scan videos for AI manipulation.
3. **Microsoft Video Authenticator** –  
   Analyze image/video authenticity.
4. **Hugging Face Deepfake Detection Models** – https://huggingface.co/models
5. **Forensically** – https://29a.ch/photo-forensics  
   Image forensics and error level analysis.

### ****Part C – Analytical Task****

Students will:

* **Choose one generated deceptive artifact** (image, audio, video, or phishing text).
* Attempt to **detect/manipulate** it using one or more detection tools.
* Record **observations** in a table format:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| *Tool Used* | *Input Tested* | *Detection/Generation Result* | *Observed Strengths* | *Observed Weaknesses* |
| *****Deepware Scanner***** *– https://deepware.ai* Scan videos for AI manipulation | ***A Deep fake video to see if the tool detects it*** | ***97% confidence – flagged abnormal facial motion and lip‑sync mismatches*** | ***- Automated analysis with probability scoring***  ***- Detects multiple manipulation types<*** | * ***Requires file/URL upload*** * ***Detection accuracy may drop with low‑res or compressed media*** |

### ****Part D – Critical Reflection (300–400 words)****

Answer:

1. How convincing was the generated content?
2. Which detection methods worked best, and why?
3. What are the **ethical implications** of using AI for deception, even in testing?
4. How would you integrate both **AI deception detection** and **awareness training** in a real organization?

**The deepfake video I chose for study was taken from YouTube and depicted a faked address by the President of the United States. This film was a good fit for the job because it showed how successfully AI-generated media can mimic a popular figure's likeness, behaviors, and voice to build a fake narrative. Such examples demonstrate the growing risks of misinformation, as well as the potential for reputational or political harm. To investigate this media, I used the Deepware Scanner, a browser-based tool for detecting AI-generated video modification. The application scans uploaded files or media URLs and uses detection algorithms to discover irregularities common to deepfake technology, such as artificial facial motion or displaced lip synchronization. In this scenario, the tool returned a clear "Deepfake Detected" result in 97 % confidence in finding *abnormal facial motion and lip‑sync mismatches* in the video.**

## ****Marking Rubric**** (Total: 1.5 Marks)

|  |  |  |
| --- | --- | --- |
| Criterion | Description | Marks |
| **Tool Exploration & Documentation** | Used at least 3 generation tools and 2 detection tools, with documented outputs and findings. | 0.5 |
| **Analysis Table** | Completed analysis table with detailed and relevant observations. | 0.3 |
| **Critical Reflection** | Demonstrated higher-order thinking, integrating ethical, technical, and practical insights. | 0.5 |
| **Presentation & Clarity** | Submission is well-structured, clear, and logically presented. | 0.2 |

### ****🔐 Lab Task 2: Investigating AI in Cyber Deception – Deepfakes, Synthetic Media, and Deceptive AI Tools****

#### This task immerses students in the use of AI-based detection tools for identifying cyber threats across files, emails, and network traffic. Students will test and compare multiple detection platforms, analyze their outputs, and assess detection accuracy, false positives, and usability. The activity develops critical thinking by encouraging evaluation of the technical, operational, and ethical implications of relying on AI for threat detection in cybersecurity operations.

### ****Part A – Exploring AI Threat Detection Tools****

Students will use **at least 3 of the following tools** for real or simulated threat detection:

1. **VirusTotal** – https://www.virustotal.com  
   AI-assisted malware and file scanning from multiple engines.
2. **Darktrace Demo** – https://www.darktrace.com/en/demo  
   Behavioral AI that detects anomalies in network traffic.
3. **Cuckoo Sandbox** – https://cuckoosandbox.org  
   Automated malware analysis using sandboxing and machine learning.
4. **Intezer Analyze (Free Tier)** – https://analyze.intezer.com  
   AI-powered code DNA analysis for detecting malware families.
5. **PhishTool (Free Plan)** – https://www.phishtool.com  
   AI-based phishing email analysis and intelligence extraction.
6. **AnyRun (Community Edition)** – https://any.run  
   Interactive malware sandbox with AI detection features.
7. **Cisco Talos Intelligence** – https://talosintelligence.com  
   Threat intelligence enriched with AI anomaly detection.
8. **Splunk Security Essentials (Free App)** – https://splunkbase.splunk.com/app/3435  
   ML-assisted anomaly and threat detection dashboards.
9. **YARA Rules with AI Enhancements (via Github projects)** –  
   Pattern matching for malware detection, enhanced with AI scoring.

### ****Part B – Threat Simulation and Detection****

Students will:

* Download **sample benign and malicious files** from trusted security training repositories (e.g., TheZoo, EICAR test file).
* Run these through **at least two tools** to see detection results.
* For network-based tools, use captured **PCAP files** from training datasets (e.g., CIC-IDS2017) and run anomaly detection.

### ****Part C – Comparative Analysis Table****

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Tool Used | Input Type (File, Email, Network) | Detection Result | AI Features Used | False Positives/Negatives Observed | Ease of Use |
| **Any Run (Community Edition)** | **File** |  | |  | | --- | | Behavioral analysis, process monitoring | | No false positives | **Easy** |
| **Virus Total** | **File zip** |  | **anomaly detection and behavioural analytic** | **No false positives observed** | **Very easy** |

### ****Part D – Critical Reflection (300–400 words)****

Answer:

1. Which tool provided the most **accurate detection results**, and why?
2. What limitations did you observe in AI-based detection?
3. How can combining **multiple AI detection tools** improve security posture?
4. What governance or ethical risks arise from relying heavily on AI in detection?

Virus Total was more accurate in this case as it flagged 29 out of 63 engines detecting the file as malicious, while ANY.RUN showed no threat. Virus Total collects findings from different antivirus engines, boosting the chances of detecting known threats. AI-based detection tools have several limitations: they can miss new or obfuscated threats resulting in false negatives, mistakenly flag safe files as malicious false positives. By combining multiple AI detection tools is that it covers more threat types and detection methods, reduces blind spots from individual tool limitations such as Any Run finding one anomaly whereas virtual tool flagged 29 of them. Combining multiple AI detection tools can greatly improve security posture. Organizations can use static and dynamic analysis to cross-reference data, remove blind spots, and improve threat visibility.

.

## ****Marking Rubric**** (Total: 1.5 Marks)

|  |  |  |
| --- | --- | --- |
| Criterion | Description | Marks |
| **Tool Exploration & Documentation** | Used at least 3 AI-based detection tools, documented inputs, outputs, and AI features. | 0.5 |
| **Comparative Analysis Table** | Completed table with accurate and insightful comparisons. | 0.3 |
| **Critical Reflection** | Provided higher-order thinking, integrating operational, technical, and ethical perspectives. | 0.5 |
| **Presentation & Clarity** | Submission is well-structured and clear. | 0.2 |

## Lab Submission Guidelines

1. Students might use AI tools. However, they must have to demonstrate their understanding of contents to the lab tutor through oral questions or presentation.
2. All students are required to submit their completed **Week 4 Lab** tasks via **Moodle** by **Sunday, 17 August 2025** 23:59 (local time). Late submissions are not accepted.
3. Navigate to **Week 4 > Active Learning** in Moodle to access the submission link.