National College of Ireland

Mitigating AI-Driven Cyber Deception: Theoretical Modelling of Social Engineering Tactics and Human Vulnerability

MSc Research Project

MSc Cybersecurity

Shona Susan Shaji

Student ID: 23291257

School of Computing

National College of Ireland

Supervisor: Mark Monaghan

National College of Ireland

MSc Project Submission Sheet

School of Computing

Student Name:	Shona Susan Shaji
Student ID:	23291257
Programme:	MSc Cybersecurity **Year:** 2024-2025
Module:	MSc Practicum
Supervisor:	Mark Monaghan
Submission Due Date:	15th September 2025
Project Title:	Mitigating AI-Driven Cyber Deception: Theoretical Modelling of Social Engineering Tactics and Human Vulnerability

Word Count: 6268 **Page Count:** 20

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.
ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature: Shona Susan Shaji

Date: 15-09-2025

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST

Attach a completed copy of this sheet to each project (including multiple copies)	□
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	□
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	□

Assignments that are submitted to the Programme Coordinator Office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Mitigating AI-Driven Cyber Deception: Theoretical Modelling of Social Engineering Tactics and Human Vulnerability

Shona Susan Shaji

23291257

Abstract

The rise of generative artificial intelligence introduces some new risks within cybersecurity, especially as AI generates some social engineering attacks. For deceiving users via unprecedented realism, such attacks now leverage fake voices along with synthetic personas. Machine-generated emails are leveraged during these attacks. Security approaches, customary ones that focus on how they detect malicious code or suspicious URLs, prove inadequate against this new form of deception aimed at how humans cognize and emotionally react.

This study is addressing of the urgent gap in the comprehension of AI-generated deception. Researchers then analyze human psychology after these actions. Cybersecurity research explores technical solutions such as firewalls or anti-phishing tools greatly but explores rarely the cognitive reasons users fall for AI-powered scams. Due to the fact that some recent studies do show AI-generated phishing outperforming even professional red team testers, the vulnerability of humans is now what attackers mainly target. This work seeks to use cognitive and ethical theory to model those vulnerabilities, designing future security systems that consider real human behavior.

1. Introduction

1.1 Background

Advancements within artificial intelligence (AI) have truly reshaped also the cybersecurity landscape now, in particular generative models such as deepfake systems along with ChatGPT, by now enabling more advanced social engineering (SE) attacks. AI-generated deception adapts, convinces, also can operate at scale (Schmitt and Flechais, 2024), which differs from customary phishing or scam attempts because they often rely on generic messages and outdated tricks. AI helps attackers make custom phishing emails, copy voices, and fake identities using lifelike avatars. AI helps attackers act so this makes finding their deceitful aim much harder for people and systems.

The threat environment evolves along with revealing that software flaws are not the primary vulnerability anymore. Human users, in fact, are at this point most vulnerable. According to the studies, cognitive biases, a trust misalignment, and emotional triggers such as fear, urgency, or familiarity each cause even tech-savvy people to give in to AI-powered deception

(Goldman et al., 2024). Additionally, real-world penetration tests each prove AI-generated phishing messages now beat human red teams' messages in both click-through rates and believability (Dutta, 2025). As AI systems continue improving so they can mimic human communication, the psychological manipulation that they enable becomes increasingly difficult to defend against using only technical means.

Cybersecurity research must respond through integrating models from psychology, ethics, and human-computer interaction and move beyond solely technical detection tools to understand why people fall for these attacks. A shift in focus should be from protecting of systems to understanding and to protecting of users.

1.2 Research Problem

Despite the growing threat posed by AI-generated social engineering, technical defenses like email filters, anomaly detection, also firewall configurations remain the focus of current academic and industry literature. These tools are necessary, but they often fail at preventing attacks bypassing machine-level detection through human emotion and decision-making (Nazo Moosa, 2024). A critical void persists since we lack understanding into human behavioral processes making people open to such attacks, especially within settings using AI-created content.

AI-driven social engineering preys on the cognitive and emotional weaknesses of people, but the existing defences are largely technical in their nature, and this leaves room open in that behavioural aspect because it provides perception on the mechanism that underlies susceptibility. There is a need for investigation of key psychological processes that include dual-process thinking, emotional priming, and trust calibration, and ethical concepts are also key, such as the moral disengagement in this regard. Cybersecurity solutions will be out of sync to that of real-life user behavior unless a theoretical model incorporates each of these variables. A conceptual framework for modeling human vulnerability to AI-generated deception is developed in order to address the problem within this research. It proposes a user-centric foundation as future training programs. It also uses cognitive psychology along with ethical reasoning as well as socio-technical theory for creating awareness tools with adaptive security protocols.

1.3 Research Question

To develop a theoretical framework that models how AI-generated social engineering attacks exploit human psychological and emotional vulnerabilities.

1.4 Objectives

- To critically evaluate the role of generative AI in evolving social engineering attack methods.
- To explore how cognitive theories such as Dual-Process Theory and Affective Priming explain user susceptibility to AI deception.
- To integrate ethical and socio-technical theories to analyse rationalization and systemic impact.
- To design a conceptual model that maps human vulnerabilities targeted by AI-powered SE attacks.

- To simulate realistic attack scenarios using the model and assess its explanatory and practical utility.

1.5 Contribution

The contribution of this study is that it combines both behavioural science and cybersecurity since it targets the increasing problem of artificial intelligence-based social engineering. The research presents a new model integrating established psychological as well as socio-technical theories within. For the new framework, Dual-Process Theory, Affective Priming, Trust Calibration, Moral Disengagement and Actor-Network Theory are used to understand how AI-generated phishing exploits vulnerabilities. The framework, designed as of an offline Python-based ethical deception simulation tool, rates at deception risk, identifies then matches triggers for cognition theories, and also compares AI-designed with conventional attacks. Findings indicate AI-composed messages trigger more psychological triggers and deception rating heightens which lends urgency to end-user-centered defenses.

1.6 Structure of the Paper

Section 2: Provides review on AI-generated social engineering, examining traditional defences, cognitive and psychological models of susceptibility, and identifying gaps that motivate this research.

Section 3: Outlines the design science approach adopted, detailing the theoretical framework, ethical considerations, and methodological boundaries.

Section 4: Describes the conceptual model, system architecture, detection and scoring mechanisms, and mapping of phishing triggers to behavioural theories.

Section 5 : Details the development of the Python-based simulation tool, including modules, workflow, and operational modes.

Section 6: Presents the results of predefined and custom message simulations, deception scoring, theory activation, and visualisation, alongside analysis of findings.

Section: Summarises key findings, discusses their implications, outlines limitations, and proposes directions for future work.

2. Literature Review

2.1 Introduction to Social Engineering and Human Vulnerability

Social engineering (SE) is still a very common and effective strategy in cyberattacks, focusing on human behavior over technical system flaws. It relies on manipulating, deceiving, also exploiting trust to convince users to divulge sensitive information or perform actions that

compromise security. Systems happen to be becoming more secure for us all, which is certainly a positive development for everyone. Focus shifts by attackers increasingly to the weakest link, the human user (Adu-Manu et al., 2023).
Due to artificial intelligence (AI), social engineers have seen a rise in the capabilities that they have. Generative AI tools are responsible especially for this increase in capabilities. Tools such as ChatGPT and ElevenLabs enable cybercriminals along with deepfake technologies so they can create content that appears more realistic, personal, and emotionally persuasive than it ever was before (Schmitt and Flechais, 2024). Social engineering improved by AI is scalable, adaptive, and convincingly human-like unlike typical attacks reliant on common bait and spelling errors (Heiding et al., 2024).
This danger is especially worrisome due to the psychological piece. Since they do not exploit any system vulnerabilities yet cognitive shortcuts, SE attacks bypass all firewalls plus antivirus systems such as responding fast to urgency or perceived authority. Most cybersecurity strategies remain overly technical as well as underprepared regarding tackling the human vulnerability dimension (Goldman et al., 2024) though emotional cues like fear, trust, and familiarity are among the most exploited pathways.

2.2 Generative AI and the Evolution of Social Engineering

These AI tools do not simply automate phishing but they also increase its impact by tailoring content to a victim's language preferences, context, and digital habits. Generative AI can produce social engineering messages indistinguishable to human-written content, with some AI-generated phishing emails outperforming human-created versions in click-through and response rates (Francia et al., 2025). Other AI models are used to generate deepfake audio or video so as to impersonate trusted voices or faces in real time (Begou et al., 2023) while tools such as ChatGPT have been exploited so as to produce convincing spear-phishing scripts as well as SMS attacks (Shibli et al., 2024).
AI-driven SE is both dynamic and scalable, and also it is difficult to detect through the legacy security systems, unlike phishing that relied on poorly written messages sent en masse.
AI powers or improves many SE attacks which are using multi-modal deception now (The WPI, 2024). For example, these attacks combine email as well as voice phishing (vishing), text messages (smishing), also deepfake video conferencing. AI with SE methods combine to form a cyber deception arms race defenders fight to match attack volume and sophistication ((Phillips et al., 2024), (Charles Owen, 2024).
These developments highlight a critical shift: AI is no longer just a tool for defenders (e.g., in spam filtering); it is also a strong weapon when attackers use it. The typical technical tactic for phishing detection is inadequate since it does not tackle cognitive manipulation via generative AI (Samala, 2024).

2.3 Psychological and Cognitive Models in SE Susceptibility

Understanding why individuals fall for social engineering attacks particularly those powered by generative AI requires explorations of behavioral science as well as cognitive psychology. Social engineering exploits human information processing under conditions of pressure, uncertainty, or emotional influence along with technical gaps. Dual-Process Theory stands as one of the most widely cited models in this domain, and it explains that AI-generated phishing often triggers rapid, intuitive System 1 thinking. Thinking within System 1 happens rapidly and depends on intuition plus emotion it avoids rational System 2 thought that is slow more careful and also rational.

Emotional cues like fear, urgency, or trust in Affective Priming shape responses, so generative AI can craft persuasive sentiment-aware messages. Trust Calibration, as highlighted by Goldman et al. (2024), shows just how attackers manipulate tone, design, as well as logos for the purpose of creating misplaced confidence within seemingly professional communications ((Francia et al., 2025); (Dutta, 2025)). Moral Disengagement Theory explains how users justify dangerous actions ethically after an attack. Victims can blame the organization or minimize what is the impact. Responsibility may be denied by them as well. Likewise, attackers morally justify engaging in deception without guilt (e.g., "testing the system" or "proving a point") in behavior (NONUM et al., 2025). These psychological detours help both victims and perpetrators avoid internal conflict about what they do. SE is understood through Actor-Network Theory as a socio-technical phenomenon, in which the linked systems for deception are shaped by humans, by AI tools, and by organisational structures. AI-driven attacks increasingly exploit cognitive weaknesses these models explain predictably. According to (Wang et al., 2021), SE attacks should be viewed not as one-to-one manipulations but as outcomes of networked systems.

2.4 Empirical Evidence: AI vs Traditional Phishing

Recent empirical research shows that it is true customary methods are often less effective than what is feasible AI-generated phishing. (Heiding et al., 2024) compared ChatGPT-generated phishing emails against those produced by professional red team operators and found that AI-created messages matched or exceeded human-written ones in persuasiveness, particularly when urgency or emotional appeal was leveraged. (Francia et al., 2025) broadened these results to SMS phishing (smishing) since LLM-created texts got greater interaction rates, even among participants trained in cybersecurity. AI chatbots can dynamically create customised smishing campaigns immediately using little contextual input as (Shibli et al., 2024) showed. This was bypassing conventional detection filters. (Begou et al., 2023) revealed that ChatGPT could produce scalable phishing scripts and could include pretexts upon fake login pages, which lowers the technical barrier for less-skilled attackers. According to (Dutta, 2025), corporate testing has shown that AI-generated phishing emails outperformed internal red team exercises from within the defensive perspective. In some of the cases, click-through rates reached up to 14% within minutes, even among security-conscious staff (Sabin, 2023). In a collective manner, these findings confirm generative AI enables adaptive, scalable, as well as emotionally clever phishing strategies plus they highlight a need for integrated behavioral defense models which customary rule-based filters and awareness programs battle to counter.

2.5 Defense Mechanisms and Existing Frameworks

As generative AI redefines the capabilities of social engineering, defense mechanisms now focus on technical barriers plus static rule-based detection systems, thus exposing intrinsic limitations. The existing solutions which were mostly designed for the older phishing forms do not at all account for the dynamic personalized and emotionally manipulative nature of the AI-generated attacks.

Customary tools such as spam filters, link analyzers, along with anomaly detectors are still useful, but AI-generated messages increasingly bypass them because the messages are grammatically correct, contextually appropriate, and often indistinguishable from legitimate

communication (Schmitt and Flechais, 2024); In response, some cybersecurity researchers as well as organizations have begun to explore AI-for-defense strategies via machine learning and natural language processing to detect suspicious patterns in emails plus texts. However, these models still are not able to keep up with the rate of AI-generated content. The models do also battle in order to match the quality of the content (Phillips et al., 2024).

(Samala, 2024) stresses how emotional literacy matters in cybersecurity training. Human error, Samala notes, is rarely caused from a lack of knowledge; instead, it stems from emotional pressure, stress, or urgency.

By Wang et al. (2021), an SE ontology for standardization also classification of attack types, channels, plus cognitive techniques was proposed. Ontologies help with the systematization of detection, but they are not always that actionable. Users must decide quickly using sounds or images found within messages.

Generative AI has made multi-modal deception possible by combining deepfake audio, synthetic video, and chatbot text to bypass single-channel defenses (Begou et al., 2023). Even though we have progressed, current frameworks lack behavioural models that link AI-driven deception to human cognition. Defences incorporating technical strategies with human strategies must go beyond attack classifications and explain reasons users are deceived, strengthening psychological resilience (Samala, 2024).

2.6 Literature Gaps and the Need for a New Theoretical Model

Although AI-driven tools such as machine learning and natural language processing in addition to predictive analytics promise that they can detect then prevent social engineering (SE) attacks, an important gap remains present in how technical defences integrate behavioural perceptions for the purpose of countering AI-enabled deception (Fakhouri et al., 2024) As noted in the work by (Adu-Manu et al., 2023), most SE defense systems build specifically on customary attack patterns which lack the personalized and emotionally manipulative sophistication found in generative AI content. Ontological approaches like (Wang et al., 2021) classify attack types yet psychological mechanisms of victim susceptibility lack explanation. AI-generated phishing content is more often effective than human-written alternatives according to studies (Francia et al., 2025; Heiding et al., 2024) but their research stops at a performance comparison and does not offer a user-side defense model grounded in psychology or in ethics.

Generative AI exploits emotional and cognitive manipulation techniques like urgency, fear, familiarity, and trust. Current defense frameworks address these techniques with rarity. AI-powered email filters can detect malicious intent through source anomalies or metadata alone, but they do fail in accounting for the fast, emotionally driven decision-making processes (System 1) that often may lead users toward clicking on or responding to dangerous content (Samala, 2024; Stylianou et al., 2025). Thus modern cybersecurity strategies have a fundamental blind spot. Emotionally smart security training calls have not translated formal models that can be taught, implemented, or evaluated ((Goldman et al., 2024; Stylianou et al., 2025). Furthermore, existing awareness programs lack integration with moral disengagement theory as well as trust calibration which represent two necessary components to understanding how users justify their actions after an attack otherwise misplace trust inside AI-generated communication.

3: Research Methodology

The methodology includes a conceptual model plus an implementation in Python using rules in addition to an interpretive evaluation with simulations plus mappings that are theoretical. The design does not have reliance on real human data, and this maintains ethical compliance as it can show technical feasibility along with an alignment for behavior.

3.2 Research Design

The research project design comes from a design science methodology seeking to produce an artefact. The artefact, like a simulation tool, addresses a practical cybersecurity challenge and contributes to theoretical understanding. This artefact offers up a model as well as an interpretation of AI-driven phishing deception, seeing it through a psychological lens as it uses transparent, rule-based logic without any opaque machine learning classifiers. This approach ensures interpretability along with reproducibility. It guarantees also consistency with existing behavioural science studies on human vulnerability to deceit.
At its core, five behavioural as well as cognitive theories do integrate into the system's architecture:

- **Dual-Process Theory** distinguishes thinking as System 1 (fast, automatic, emotional) and System 2 (slow, deliberate, logical). Phishing messages are frequently created in order that they provoke System 1 responses which bypass rational scrutiny.
- **Affective Priming** Emotional stimuli such as fear, urgency, or promised rewards can subconsciously prime recipients toward acting impulsively according to Affective Priming.
- **Trust Calibration Theory** highlights how users evaluate trust within digital communication. Phishing exploits false trust signals, like official logos or authoritative language, to do so.
- **Moral Disengagement Theory** gives understanding into how victims can excuse or minimise their involvement after a phishing attempt, and this affects resilience long-term.
- **Actor-Network Theory** frames phishing when actors deceive as a component of a socio-technical network in which AI generates messages incorporating intertwined human along with non-human actors.

3.3 Research Tools and Environment

The simulation framework developed entirely in Python 3.x, which coders chose because of its versatile nature, readable code, as well as modular designs. Ethical conduct was assured plus offline work became an option. It used no live phishing data or external datasets.

3.3.1 Core Development Environment:

- **Programming Language:** Python 3.x
- **Primary Libraries:**
  - Matplotlib – for generating bar chart visualisations of deception scores
  - re (regex) – for pattern matching of phishing phrases and constructs
- **Execution Interface:** Command-line interface (CLI) compatible with both **Windows PowerShell** and **Unix-based terminals**

- **Data Storage:** Local .txt and .json files for storing analysis results; no database integration to maintain system simplicity and security

3.3.2 Operational Characteristics:

- **Lightweight and Portable:** Requires no external dependencies beyond standard Python libraries, enabling deployment on various operating systems without complex configuration.
- **Secure Simulation:** Runs in a closed environment without internet access, preventing interaction with live phishing infrastructure.
- **Ethical Design:** Utilises predefined and user-provided text messages, all artificially generated or anonymised, to avoid privacy and consent issues.

3.4 Implementation process

3.4.1 Core Modules

- **main.py** – Acts as the entry point, allowing users to run either predefined message simulations or custom message testing.
- **messages.py** – Stores predefined phishing messages categorised as human-written traditional, AI-generated generic, and AI-generated personalised, enabling comparative analysis of deception styles.
- **scoring.py** – Detects psychological triggers using a combination of expanded keyword lists, phrase-based matching, and regex pattern recognition. It computes deception scores and applies heuristic rules that elevate risk scores for dangerous trigger combinations.
- **theories.py** – Maps detected triggers to relevant psychological theories, maintaining interpretability and alignment with behavioural research.
- **logger.py** – Records results into a simple text log (result.txt) for documentation or further analysis.
- **chart.py** – Produces bar charts comparing deception scores between different message types

Message Analyzer uses keywords and a phrase list since these are built on actual phishing instances showing urgency, fear, authority, reward, also trust signals, etc. It concerns a multi-word phrase like “act now” or “immediate action required,” increasing ability to spot subtle social engineering methods. For improvement of pattern-based recognition, the detection system based on regex has been deployed so that it can detect commonly used phishing phrases like click here and verify your account that keyword matching alone would not detect. The heuristic risk assessment layer also increases accuracy in detection because it detects some patterns with risky triggers such as a detection of an urgent tone within the message as well as log in prompts and reward-based lures that are occurring together. Such messages automatically score high in deception even when specific keywords do not weigh so heavily. The Deception Scorer uses heuristic reasoning so as to refine taxonomies also give a weighted score (0-100) to psychological cues like authority, fear, and urgency. Theory Mapper software will relate each message to cognitive theories. The mapping relies on the nature of identified triggers and a deception score. You are able to run the simulation on predefined AI phishing scenarios, and you are able to run the simulation in custom phishing message input mode.

3.5 Ethical Considerations

- No real users or data were used, avoiding privacy and consent issues.
- AI-generated content was simulated, not scraped.
- No live phishing or credential capture was involved.

The tool solely operates in a safe simulation space, making it suitable for academic purposes.

3.6 Limitations

The present study has several limitations for acknowledgement. First, they evaluated the simulation framework completely in an offline, controlled environment that was without any real human participants, which, while it was ensuring ethical compliance and it was protecting privacy, limits empirical validation of the model's predictions against actual user behaviour that is in live phishing contexts.

Second, it is predefined keyword lists and also regex patterns and even heuristic rules which are relied upon by the scoring and detection mechanisms, yet these mechanisms may not capture the full range of adaptive structural and linguistic variations used in AI-generated attacks, even when attacks are informed by documented phishing cases.

Third, the scope limits itself to textual phishing content. It does not reach multi-modal deception formats such as deepfake audio, synthetic video, or coordinated cross-channel attacks. Interpretability is kept via avoiding advanced NLP plus opaque machine learning models in the design. However, this does constrain semantic understanding beyond just surface-level cues and potentially can miss more subtle manipulations.

The theoretical framework does not include all cognitive, cultural, or situational factors that influence susceptibility since it integrates five established behavioural and socio-technical models along with the predefined phishing scenarios, static as they are, may not reflect the diversity of real-world attacker strategies. The user-supplied messages are what determine its utility, though also the custom input mode enables testing that is broader. These limitations suggest future research directions; research should integrate semantic NLP, expand to multi-modal threat detection, validate across cultures, as well as test empirically with real user groups within controlled environments.

4 Design

4.1 System Architecture

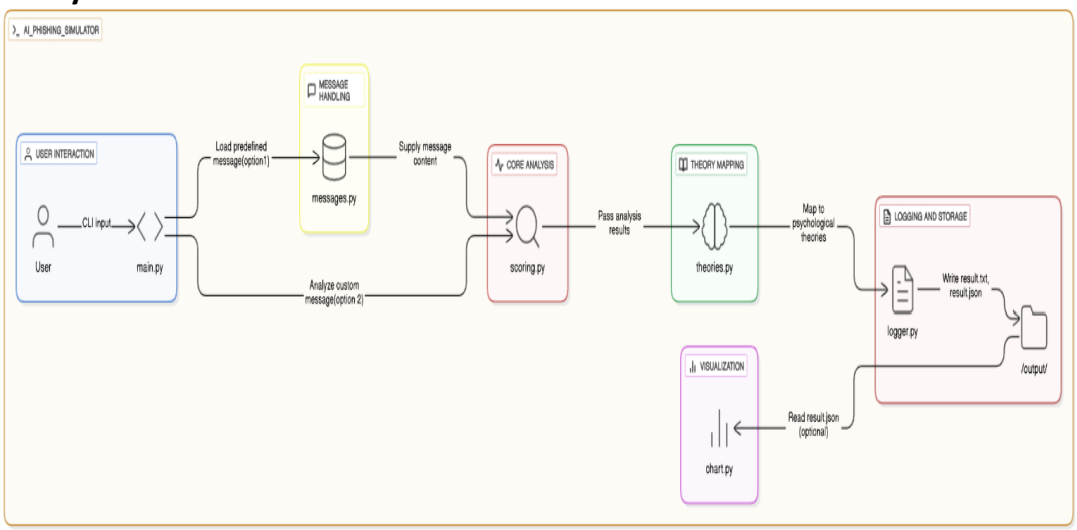

Figure 1. System Architecture

The system(Figure 1) follows through a modular Python-based architecture comprising six key components:

1. **Message Input Module** – Phishing datasets predefined or custom messages that user entered are accepted.
2. **Trigger Detection Module–** It uses expanded keyword lists, phrase matching, and regex pattern recognition since it identifies social engineering cues such as urgency, fear, authority, reward, and trust signals.
3. **Scoring Module–** It assigns deception scores ranging from 0 to 100 on the basis of the strength along with the frequency plus the combination of triggers. For example, dangerous combinations of triggers, like when urgency is paired with a fake signal of trust, receive a higher weighting.
4. **Theory Mapping Module** – Links triggers detected to relevant cognitive and ethical theories. Trust Calibration, Actor-Network Theory, Moral Disengagement Theory, Affective Priming, and Dual-Process Theory are some examples.
5. **Result Logging Module** – For both review and also for reporting, the analysis results are recorded in either .txt or in .Json formats.
6. **Visualisation Module** – It can generate bar chart comparisons of deception scores through using Matplotlib so risk levels will be clearly communicated.

4.2 Tools and Techniques

4.2.1 Tools

1. **Python 3.x** – Main programming language used to implement the framework for simulation.
2. **Matplotlib** – Matplotlib makes bar chart visualisations. These bar chart visualisations display deception scores within.
3. **Regex (Regular Expressions)** – Use regex or regular expressions to detect common phishing patterns past simple keyword matching.
4. **PowerShell / Unix Terminal** – A command-line environment that you can use to execute the simulation tool.
5. **Text & JSON File Logging** – Analysis results now are stored within a structured format that is exportable.

4.2.2 Techniques Used

- **Rule-Based Detection**: We identify phishing cues since we use predefined keyword and phrase lists derived from real-world phishing samples.
- **Phrase Matching & Multi-Word Trigger Detection**: Multi-Word Trigger Detection and Phrase Matching capture more subtle manipulation forms. For example, this includes situations when someone says “act now” or demands “immediate action required”.

- **Regex Pattern Recognition:** Structural phishing elements that are such as “click here” or like fake login prompts not found in static keyword lists are detected via Regex Pattern Recognition.
- **Heuristic Risk Scoring:** assigns deception scores (0–100) that severity, trigger combinations, and cue frequency weight.
- **Psychological Theory Mapping:** Links from detected cues to cognitive theories (Dual-Process Theory, Affective Priming, Trust Calibration, Moral Disengagement Theory, Actor-Network Theory) map Psychological Theory, explaining the message's manipulative intent.
- **Multi-Scenario Simulation:** Multi-Scenario Simulation provides support for predefined phishing categories. Customary and generic AI-generated categories in conjunction with personalised AI-generated categories as well as custom inputs allow for flexible testing.

5 Implementation

5.1 Implementation

The phishing deception simulation tool is implemented in Python since it follows a modular architecture with clarity, maintainability, and replicability. The updated version, retaining ethical offline operation without reliance on external datasets or on live phishing content, incorporates improved features for usability, scoring that is heuristic, and logic for detection that is expanded.

5.2 Technology Stack

The tool is built up entirely in Python 3.x because its native modular programming capabilities were employed for achieving this. Matplotlib generates bar chart visualisations for deception scores. The command-line interface functions on Unix-based terminals and PowerShell if there are no extra dependencies. The system requires absolutely no database or internet connection at all, and it keeps the system both lightweight and secure. For the further analysis or the reporting, results can be stored in either JSON or in text formats.

5.3 Workflow

In predefined mode, the tool processes three categories for phishing messages, and the tool passes each category through trigger detection, deception scoring that includes heuristic adjustments, and theory mapping. Each message users enter in custom mode within a single unlimited session undergoes the same analytical pipeline.
Detecting triggers includes three techniques that are complements to each other:

1. **Expanded keyword and phrase matching** to capture a wider range of phishing language.
2. **Regex pattern detection** for spotting common scam constructs.
3. **Heuristic rules** that identify dangerous combinations of triggers and adjust risk scores accordingly.

When the deception score is finalised, relevant psychological theories are activated. The results are then presented inside the console and saved to a log file. Then users are able to make a visual scores chart. This allows for the comparative analysis itself.

Maintaining a transparent, interpretable framework suitable for academic research as well as awareness training, the updated design improves the user experience, improves risk classification accuracy, and greatly increases detection coverage.

6. Evaluation and Result Analysis

6.1 Evaluation

The evaluation uses scenarios instead of experiments or statistics. The proposed framework's usability with explanatory value are its focus.

6.2 Scenario Demonstration

Two AI-generated phishing messages appeared in the simulation. There also was one of those customary messages. Each message triggered specific psychological categories such as fear and then urgency. These categories received deception scores according to expected user vulnerability.
For example:

- A generic AI phishing message ("We've detected unusual sign-ins…") triggered urgency and authority, scoring 65/100 and mapping to:
  - Dual-Process Theory
  - Affective Priming
  - Trust Calibration
- A safe message submitted by the user ("Your subscription expires next week") triggered no psychological cues and scored 0/100, mapping only to logical thinking (System 2).

6.3 Visualisation

A bar chart displayed comparative deception scores across all message types, helping to:

- Visually communicate risk levels
- Confirm theoretical alignment
- Demonstrate consistency of the model's reasoning

6.4 Validation of Theoretical Mapping

Theory mapping correctness was cross-verified via established behavioural science literature (e.g., Kahneman, Bandura, Fazio). Therefore theory mapping has been verified. Logical rules in the system mimic pathways for real psychological processing.

6.5 Overview

In this chapter, outcomes are presented of the implemented system that models psychological and emotional triggers exploited by AI-generated phishing messages. To obtain the results, analysts map activated theory, score deception, and combine simulation output. The goal toward evaluating the proposed framework's practical applicability is validating its potential to explain human vulnerability to AI-driven social engineering (SE) tactics.

6.5.1 Simulation Summary

The Python-based simulation developed as part of this research tests three categories of phishing messages:

- A **traditional (human-written)** phishing message,
- A **generic AI-generated** phishing message, and
- A **personalised AI-generated** phishing message.

A deception score from zero to one hundred is computed plus relevant cognitive theories are mapped plus each message is analyzed using a deception scoring algorithm for detecting psychological cues like urgency, fear, authority, and personalisation. A suspicious message from them also may be input for analysis.

6.5.2 Deception Scoring Output

Table below shows the deception scores produced during the system run:

Message Type	Triggers Detected	Deception Score	Risk Level
Traditional (Human)	Urgency, Fear	45/100	Moderate
AI-Generated (Generic)	Urgency, Authority, Fear	100/100	High
AI-Generated (Personalised)	Urgency, Authority, Personalisation	70/100	High
User Input Example 1	None	0/100	Low
User Input Example 2	Urgency, Authority, Personalisation, Trust Signal (Fake)	40/100	Moderate

Table 1: Systems Deception score produced

6.5.3 Activated Psychological Theories

From the conceptual framework, psychological theories also mapped each message. The outcomes mapping appears in the table below:

Message Type	Activated Theories
Traditional	Dual-Process Theory (System 1)
AI-Generated (Generic)	Dual-Process Theory, Affective Priming, Trust Calibration
AI-Generated (Personalised)	Dual-Process Theory, Affective Priming, Trust Calibration, Actor-Network Theory
User Input Example 2	Dual-Process Theory, Trust Calibration, Actor-Network Theory

Table 2: Activated Psychological Theories

AI-created messages trigger emotionally charged responses faster plus greater misplaced trust as the results show.

6.5.4 Visual Representation

Python chart.py generated for deception scores a bar chart. The chart shows how each message type has a comparative deception potential.

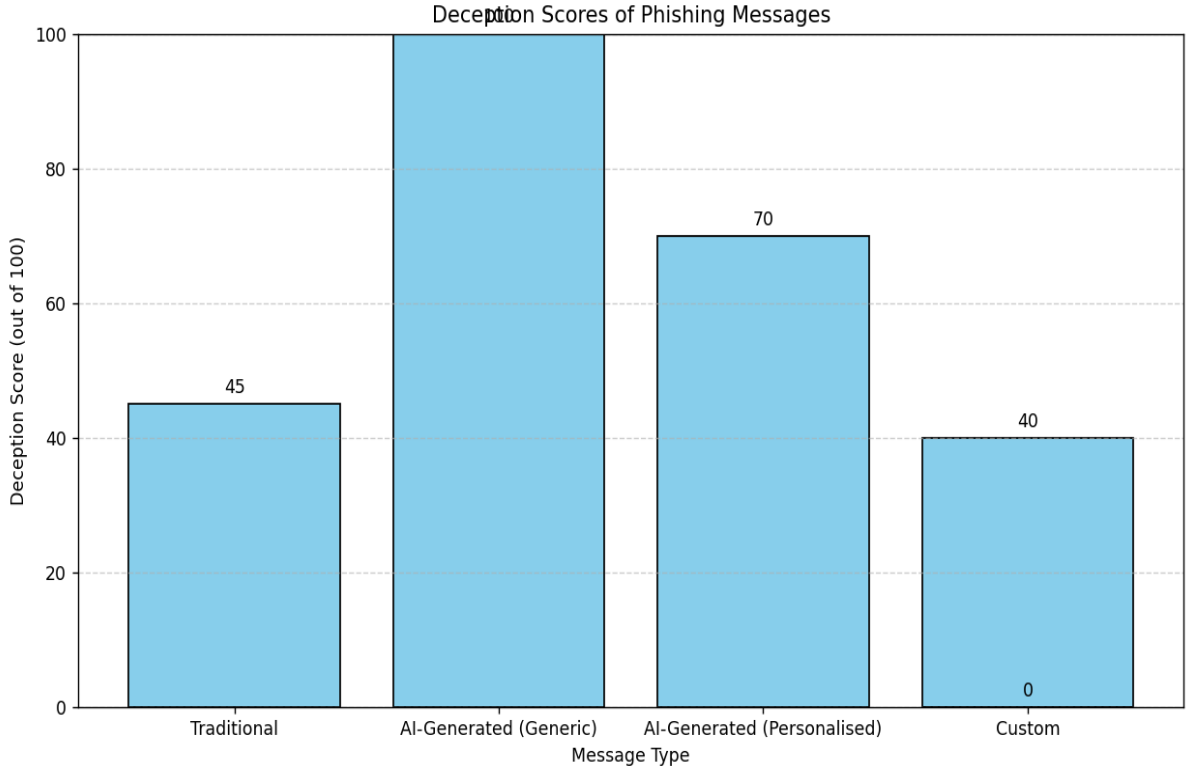

Figure 2 : Result Bar Graph Generated

This view (Figure 2) strengthens the concept. AI-generated messages which can appear more personalised or official can pose a higher psychological risk, even without malicious links or without code.

6.6 User Testing of Custom Message

The second module accepts suspicious messages from users directly. A good example is also this message here now:

“We’ve detected unusual activity on your corporate account. Please confirm your login details by 5 PM today to avoid temporary suspension.” was analysed and returned:

- **Triggers**: urgency, authority, trust signal (fake)
- **Score**:40 /100
- **Risk Level**: Moderate
- **Activated Theories**: Trust Calibration, Actor-Network Theory

This indicates the model can extend the framework to any random message not only set examples.

6.7 Evaluation Against Objectives

Objective	Outcome
Evaluate the role of AI in social engineering	Achieved via comparative analysis of traditional vs. AI-generated phishing
Map psychological and cognitive triggers	Achieved through scoring engine and theory mapping
Integrate ethical theories and socio-technical views	Achieved via theory-layer and Actor-Network logic
Simulate real-world attack messages	Completed with three AI scenarios and custom user input
Assess framework's utility	Results show it explains deception risk clearly and is user-interactive

6.8 Summary

Phishing messages made by AI trick people more easily since they seem real. The results show that their content, rich in triggers, heightens deception risks. The custom simulation tool shows user susceptibility may be modeled by cognitive theories such as Dual-Process Thinking, Affective Priming, and Trust Calibration. Being in a position to test custom messages further improves the practical relevance. These findings do validate the proposed theoretical framework, and these findings set a foundation. Future integrations can use the foundation along with user awareness training or with security tools.

7 Discussion and Conclusion

7.1 Discussion

This research explored the ways in which AI-generated phishing messages can manipulate psychology and also developed a lightweight simulation framework which models this deception through the use of well-established behavioural theories. A modular Python-based simulation systematically detects emotional along with cognitive triggers within phishing content. This implementation realized this aim through that detection. The system offered deception scores plus theoretical mappings meeting academic expectations while also interpreting scores practically. AI-generated messages, for instance, exhibited higher deception scores than customary phishing messages, and this confirms that contemporary social engineering attacks are increasingly personalised, emotionally charged, and cognitively manipulative.

Analysts were enabled via theories such as Dual-Process Thinking, Affective Priming, and Trust Calibration to layer and analyze beyond customary technical detection models. Rather than to simply label any message as being malicious, the tool interprets what the attacker's intended underlying psychological influence is. In educational and in awareness-training contexts, this interpretive approach proved highly valuable in instances where comprehension of just why a message is deceptive is as important as recognising of the fact that it is. The simulator also let users input messages they thought were suspicious. A prompt analysis of its manipulation tactics as well as theoretical implications was then received by them. This dual functionality offered flexibility with practical relevance. Analysis included predefined messages in addition to the messages users generated.

Visual analysis using deception score charts further strengthened the tool's communication capacity since it helped non-technical users understand the results intuitively. The simulation depicted the way cognitive science of decision-making translates to strategies for cybersecurity awareness in practice. It was intentionally rule-based plus offline for both ethical and academic compliance. The dissertation's central argument is supported overall by the findings: that behavioural theories should have a vital role to play when countering AI-driven cyber deception, especially within phishing contexts which exploit rapid, emotional responses.

7.2 Conclusion

About theoretical modelling of social engineering tactics with human vulnerability presents a novel approach to reducing AI-driven cyber deception in this dissertation. The study offers up a reproducible and ethically sound type of method to understand just how phishing messages exploit both cognitive and also emotional biases. It does function so by integrating psychological theory along with a lightweight simulation framework, especially for those messages generated by AI. Validated behavioral science can provide meaningful perceptions even by the implementation, which showed a free rule-based tool. Because it consistently activated theory also because it simulated scenarios, the tool revealed patterns that modern phishing strategies manipulate, which contributes academically and avails practically.

7.3 Key Findings

- The study found that richer emotional also cognitive triggers meant AI-generated phishing messages, especially those personalised, scored higher (65/100) on deception risk than customary phishing (45/100).
- The rule-based Python simulation did effectively map these triggers to well established behavioral theories. Because of the simulation that provided a deeper perception into victim susceptibility, it used signals like urgency, fear, authority, and trust.
- The framework worked consistently with both predefined and custom messages because it demonstrated that interpretable, non-ML models can be both ethical and effective.
- Discoveries reveal that we must add cognitive-behavioural ideas to methods detecting phishing and teaching users.

7.4 Future Works

Adaptive scoring models could be developed then, NLP could be integrated for deeper semantic analysis, the framework could be validated through user studies, and detection could be extended to multi-modal phishing. Also, the tool could defend at once. Furthermore, it could train through gamification also explore cross-cultural susceptibility factors.
The dissertation thus attains all of its stated objectives and contributes some original value to the discourse on AI-driven phishing and human-centred cyber defence.

8.References

Adu-Manu, K.S., Ahiable, R.K., Appati, J.K. and Mensah, E.E. (2022) *Phishing Attacks in Social Engineering: A Review*. Journal of Cyber Security, 4(4), pp. 239–267. DOI: 10.32604/jcs.2023.041095. (Published 10 August 2023)

Bandura, A. (1999) *Moral Disengagement in the Perpetration of Inhumanities*. Personality and Social Psychology Review, 3(3), pp. 193–209. DOI: 10.1207/s15327957pspr0303_3.

Begou, N., Vinoy, J., Duda, A. and Korczynski, M. (2023) *Exploring the Dark Side of AI: Advanced Phishing Attack Design and Deployment Using ChatGPT*. In Proceedings of the IEEE Conference on Communications and Network Security (CNS 2023), Orlando, FL: IEEE, pp. 1–6. DOI: 10.1109/CNS59707.2023.10288940.

Owen-Jackson, C. (2024) *Social engineering in the era of generative AI: Predictions for 2024*. IBM Think Blog, 09 May 2024 [Online]. Available at: https://www.ibm.com/think/insights/social-engineering-generative-ai-2024-predictions (Accessed: 10 October 2025).

Czarniawska, B. (2006) *Bruno Latour: Reassembling the Social: An Introduction to Actor-Network Theory* (Book Review). Organization Studies, 27(10), pp. 1553–1557. DOI: 10.1177/0170840606071164.

Dutta, T.S. (2025) *AI Outperformed Elite Red Teams in Creating an Effective Spear Phishing Attack*. Cyber Security News, 07 April 2025 [Online]. Available at: https://cybersecuritynews.com/ai-outperformed-elite-red-teams/ (Accessed: 10 October 2025).

Fakhouri, H.N., Alhadidi, B., Omar, K., Makhadmeh, S.N., Hamad, F. and Halalsheh, N.Z. (2024) *AI-Driven Solutions for Social Engineering Attacks: Detection, Prevention, and Response*. In Proceedings of the 2nd International Conference on Cyber Resilience (ICCR 2024), Dubai, UAE: IEEE, pp. 1–8. DOI: 10.1109/ICCR61006.2024.10533010.

Fazio, R.H. (2001) *On the Automatic Activation of Associated Evaluations: An Overview*. Cognition and Emotion, 15(2), pp. 115–141. DOI: 10.1080/02699930125908.

Francia, J., Hansen, D., Schooley, B., Taylor, M., Murray, S. and Snow, G. (2025) *Assessing AI vs Human-Authored Spear Phishing SMS Attacks: An Empirical Study*. arXiv [Preprint], arXiv:2406.13049 [cs.CY]. DOI: 10.48550/arXiv.2406.13049.

Cherry, K. (2024) *There's a Reason Even the Smartest People Fall For Scams*. Verywell Mind, 17 September 2024 [Online]. Available at: https://www.verywellmind.com/why-we-fall-for-scams-8705528 (Accessed: 9 August 2025). *(Article authored by Kendra Cherry; reviewed by R. Goldman, PhD.)*

Heiding, F., Lermen, S., Kao, A., Schneier, B. and Vishwanath, A. (2024) *Evaluating Large Language Models' Capability to Launch Fully Automated Spear Phishing Campaigns: Validated on Human Subjects*. arXiv [Preprint], arXiv:2412.00586 [cs.CR]. DOI: 10.48550/arXiv.2412.00586.

Krämer, W. (2014) *Kahneman, D. (2011): Thinking, Fast and Slow* (Book Review). Statistical Papers, 55(3), p. 915. DOI: 10.1007/s00362-013-0533-y.

Moosa, N. (2024) *Letter: A different perspective on the threat of cyber crime*. Financial Times, 07 November 2024 [Online]. Available at: https://www.ft.com/content/7ccb0fc2-663d-4d61-8a78-05b622484aa8 (Accessed: 9 September 2025).

Nonum, E.O., Avwokuruaye, O. and Umar, A.M. (2025) *Social Engineering: Understanding Human Factors in Cyber Security*. International Journal of Convergent and Informatics Science Research, 7(9), (March 2025 edition). DOI: 10.70382/hijcisr.v07i9.032.
Perkins, R., Rezaei Khavas, Z. and Robinette, P. (2021) *Trust Calibration and Trust Respect: A Method for Building Team Cohesion in Human Robot Teams*. arXiv [Preprint], arXiv:2110.06809 [cs.RO]. DOI: 10.48550/arXiv.2110.06809.
Phillips, J., Toltzis, A., Fanous, V. and Lalsinghani, G. (2025) *The Darwinian Effect: The Weaponization of Artificial Intelligence By Cyber Criminals*. California Western Law Review, 61(1), Article 3. (Published January 2025). Available at: California Western Scholarly Commons.
Sabin, S. (2023) *ChatGPT-written phishing emails are already "scary good"*. Axios – Technology, 24 October 2023 [Online]. Available at: https://www.axios.com/2023/10/24/chatgpt-written-phishing-emails (Accessed: 10 October 2025).
Samala, M. (2024) *The Evolution of Social Engineering and Phishing in the Age of Artificial Intelligence*. Lumen (CenturyLink) Blog, 05 August 2024 [Online]. Available at: https://blog.lumen.com/the-evolution-of-social-engineering-and-phishing-in-the-age-of-artificial-intelligence/ (Accessed: 10 October 2025).
Schmitt, M. and Flechais, I. (2024) *Digital deception: generative artificial intelligence in social engineering and phishing*. Artificial Intelligence Review, 57(12), Article 324. DOI: 10.1007/s10462-024-10973-2.
Shibli, A.M., Pritom, M.M.A. and Gupta, M. (2024) *AbuseGPT: Abuse of Generative AI ChatBots to Create Smishing Campaigns*. In Proceedings of the 12th International Symposium on Digital Forensics and Security (ISDFS 2024), Antalya, Turkey: IEEE, pp. 1–6. DOI: 10.1109/ISDFS60797.2024.10527300.
Stylianou, I., Bountakas, P., Zarras, A. and Xenakis, C. (2025) *Suspicious minds: Psychological techniques correlated with online phishing attacks*. Computers in Human Behavior Reports, 19, Article 100694. DOI: 10.1016/j.chbr.2025.100694.
Worcester Polytechnic Institute (WPI) (2024) *SECURE IT – October 2024 (AI Focus)*. The WPI Hub – Information Security Newsletter, October 2024 [Online]. Available at: https://hub.wpi.edu/spread/163/secure-it-october-2024 (Accessed: 10 October 2025).
Wang, Z., Zhu, H., Liu, P. and Sun, L. (2021) *Social Engineering in Cybersecurity: A Domain Ontology and Knowledge Graph Application Examples*. Cybersecurity, 4(1), Article 31. DOI: 10.1186/s42400-021-00094-6.