

A Cognitive-Aware Security Framework for LLM-Based Social Engineering

Woo Jin Jung*, Ah Reum Kang**

*M.S. Student, School of Cybersecurity, Korea University, Seoul, Korea

**Professor, Dept. of Cyber Security, Pai Chai University, Daejeon, Korea

[Abstract]

The emergence of large language model (LLM)-based generative artificial intelligence is fundamentally transforming social engineering attacks by enabling large-scale automation, extreme personalization, and context-aware persuasion. As a result, such attacks can no longer be explained solely by user carelessness but increasingly exploit inherent vulnerabilities in human cognition. This paper reconceptualizes LLM-based social engineering as a problem of cognitive vulnerability amplification, in which linguistic naturalness, authority cues, contextual dependence, and cognitive overload are systematically leveraged to manipulate decision-making. To address this shift, we propose a cognitive-aware security framework that moves beyond content-based detection and focuses on protecting user decision processes through interaction-level analysis and decision-point intervention. Through structured analysis and a case study, we discuss the limitations of conventional security models and highlight the need for a human-centric defense paradigm in the era of generative AI.

▶ **Key words:** Large Language Models, Social Engineering Attacks, Human Cognitive Vulnerabilities, Cognitive-Aware Security

[요 약]

대규모 언어 모델(LLM) 기반 생성형 AI는 대규모 자동화, 초개인화, 그리고 상황 인지 기반 설득을 가능하게 함으로써 사회공학 공격의 양상을 근본적으로 변화시키고 있다. 이에 따라 사회공학 공격은 더 이상 사용자의 부주의로 설명되는 문제가 아니라 인간 인지에 내재된 취약성을 직접적으로 공략하는 구조로 진화하고 있다. 본 논문은 LLM 기반 사회공학 공격을 인지 취약성의 구조적 증폭 문제로 재정의하고, 언어 기반 신뢰, 권위 편향, 맥락 의존성, 인지 과부하와 같은 요소가 의사결정 과정에 미치는 영향을 분석한다. 또한 기존 콘텐츠 기반 탐지 방식의 한계를 극복하기 위해 상호작용 수준 분석과 의사결정 시점 개입을 중심으로 하는 인지 기반 보안 프레임워크를 제안하며, 구조화된 분석과 사례 연구를 통해 생성형 AI 환경에서 인간 중심 보안 패러다임의 필요성을 제시한다.

▶ **주제어:** 대규모 언어 모델, 사회공학 공격, 인간 인지 취약성, 인지 기반 보안

-
- First Author: Woo Jin Jung, Corresponding Author: Ah Reum Kang
 - *Woo Jin Jung (jungwoojin@korea.ac.kr), School of Cybersecurity, Korea University
 - **Ah Reum Kang (armk@pcu.ac.kr), Dept. of Cyber Security, Pai Chai University
 - Received: 2026. 04. 30, Revised: 2026. 07. 14, Accepted: 2026. 07. 30.

I. Introduction

대규모 언어 모델(LLM)을 기반으로 한 생성형 AI의 등장은 사이버보안 환경 전반에 구조적인 변화를 초래하고 있다. 특히 사회공학(social engineering) 공격의 양상은 단순한 자동화 수준을 넘어, 인간의 인지 체계를 직접적으로 공략하는 방향으로 진화하고 있다. 전통적인 사회공학 공격은 공격자의 언어 능력, 창의성, 그리고 시간적·물리적 제약에 의해 제한되어 왔다. 그러나 LLM은 대규모 언어 데이터를 기반으로 자연스러운 문맥 생성, 다국어 처리, 반복적 메시지 생성의 자동화, 그리고 정교한 타겟 맞춤형 메시지 생성을 가능하게 하여 이러한 제약을 크게 완화한다 [1][2]. 이로 인해 공격자는 단순히 더 많은 메시지를 생성하는 것을 넘어, 인간의 판단과 의사결정 과정을 체계적으로 설계하고 조작할 수 있는 환경을 확보하게 된다.

기존 보안 연구에서는 사회공학 공격을 주로 사용자의 부주의나 보안 인식 부족에서 비롯된 문제로 간주해 왔다. 이에 따라 대응 전략 역시 사용자 교육이나 규칙 기반 필터링, 또는 기계학습 기반 분류 정확도 향상에 초점을 맞추어 발전해 왔다 [3][4][5]. 이러한 접근은 일정 수준의 효과를 보였으나, 생성형 AI 기반 공격 환경에서는 근본적인 한계를 드러낸다. 특히 자연스러운 언어 표현에 대한 신뢰, 권위 있는 발신자에 대한 수용, 상황 맥락과의 일치로 인한 경계심 완화, 그리고 정보 과부하로 인한 인지 자원 고갈과 같은 인간의 인지적 특성은 기존의 콘텐츠 중심 방어 모델로 충분히 설명되지 않는다 [4][6].

최근 연구에서도 사용자 행동, 메시지 속성, 그리고 인지 상태가 사회공학 공격의 성공 여부에 중요한 영향을 미친다는 점이 실증적으로 보고되고 있으며, 이는 공격이 단순한 메시지 문제가 아니라 인간의 판단 과정 자체를 표적으로 한다는 것을 시사한다 [4][6]. 특히 생성형 AI는 이러한 인지적 취약성을 단순히 활용하는 수준을 넘어, 이를 구조적으로 증폭시키는 역할을 수행한다. 즉, 공격자는 더 이상 제한된 시나리오를 반복하는 것이 아니라, 사용자 반응에 따라 메시지를 동적으로 조정하고 설득 전략을 지속적으로 최적화할 수 있다.

이러한 변화는 사회공학 공격을 단순한 사용자 실수의 문제로 해석하는 기존 관점을 근본적으로 재검토할 필요성을 제기한다. 기존 방어 모델이 “악성 콘텐츠를 탐지하고 차단하는 것”에 초점을 맞추었다면, LLM 기반 사회공학 공격은 “사용자의 의사결정 과정을 왜곡하는 것”을 핵심 목표로 한다. 따라서 방어 역시 콘텐츠 필터링을 넘어, 인간의 판단 과정 자체를 보호하는 방향으로 확장되어야 한다.

본 논문은 이러한 문제의식을 바탕으로, LLM 기반 사회공학 공격을 인지 취약성의 구조적 증폭(cognitive vulnerability amplification) 문제로 재정의한다. 기존 연구가 메시지의 표면적 특징이나 분류 성능에 초점을 맞추었던 반면, 본 연구는 공격이 공략하는 인간의 인지적 요소와 그 상호작용 구조에 주목한다. 이를 위해 언어 기반 신뢰(language-based trust), 권위 편향(authority bias), 맥락 의존성(contextual dependence), 인지 과부하(cognitive overload)라는 네 가지 핵심 인지 요소를 중심으로 공격 메커니즘을 체계적으로 분석한다.

또한 본 논문은 기술 중심 보안 패러다임의 한계를 지적하고, 이를 보완하기 위한 인지 기반 보안 프레임워크(Cognitive-Aware Security Framework)를 제안한다. 제안 프레임워크는 개별 메시지의 악성 여부를 판단하는데 그치지 않고, 상호작용 맥락, 사용자 상태, 그리고 의사결정 환경을 함께 고려하여 공격 위험을 평가하고 대응하는 것을 목표로 한다. 이는 사회공학 방어를 단순한 탐지 문제에서 인간 중심의 의사결정 보호 문제로 확장하는 접근이라 할 수 있다.

특히 본 연구는 특정 탐지 알고리즘의 성능을 평가하는 실험 중심 연구가 아니라, LLM 기반 사회공학 공격의 작동 구조를 인간 인지 취약성의 관점에서 설명하고 이에 대응하기 위한 분석 틀을 제시하는 데 초점을 둔다. 이를 위해 관련 연구 검토를 통해 기존 사회공학 대응 모델의 한계를 정리하고, LLM 기반 공격의 위협 모델과 핵심 인지 취약성을 도출하였다. 이후 상호작용 맥락, 사용자 인지 부하, 의사결정 시점을 중심으로 Cognitive-Aware Security Framework를 구성하고, 사례 기반 구조적 분석을 통해 제안 프레임워크의 적용 가능성을 검토하였다.

LLM 기반 사회공학 공격은 단일 메시지의 악성 여부만으로 설명되기 어렵고, 사용자 반응에 따른 반복적 설득과 의사결정 과정의 왜곡을 함께 고려해야 한다. 이에 본 연구는 개별 메시지 분석을 넘어, 공격 과정에서 나타나는 상호작용 흐름과 인간 인지 요소를 중심으로 프레임워크를 구성하였다. 이러한 개념적 모델링과 구조적 분석은 생성형 AI 환경에서 사회공학 공격의 작동 방식을 설명하고, 향후 실증 연구를 위한 분석 기준을 제시하는 데 목적이 있다.

본 논문의 주요 기여는 다음과 같다.

첫째, LLM 기반 사회공학 공격을 인간 인지 취약성의 구조적 증폭 문제로 재정의한다. 둘째, 기존 사회공학 대응 연구의 한계를 분석하고, LLM 기반 공격을 설명하기 위한 주요 구조적 공백을 도출한다. 셋째, 언어 기반 신뢰, 권위 편향, 맥락 의존성, 인지 과부하라는 네 가지 핵심 인지 요소를

중심으로 공격 메커니즘을 정리한다. 넷째, 상호작용 맥락 분석, 인지 부하 추정, 의사결정 보호 계층으로 구성된 Cognitive-Aware Security Framework를 제안한다. 다섯째, 기존 콘텐츠 기반 보안 모델 및 Human-Centered Security 접근과 비교하여, 생성형 AI 환경에서 의사결정 보호 중심 보안 패러다임의 필요성을 제시한다.

본 논문의 구성은 다음과 같다. 2장에서는 전통적 사회공학 공격과 생성형 AI 기반 공격에 대한 관련 연구를 검토하고 연구 공백을 분석한다. 3장에서는 위협 모델을 정의하고, Cognitive-Aware Security Framework와 사례 연구를 제시한다. 4장에서는 구조적 분석을 통해 기존 보안 모델과 제안 프레임워크의 차이를 비교하고 적용 가능성을 논의한다. 마지막으로 5장에서는 연구의 주요 기여, 실무적 시사점, 한계 및 향후 연구 방향을 제시한다.

II. Related Work

1. Legacy Social Engineering Framework

사회공학 공격은 기술적 취약점이 아닌 인간의 판단과 행동을 공격 표면으로 삼는다는 점에서 전통적으로 독립적인 보안 위협 범주로 다뤄져 왔다. 대표적인 공격 유형으로는 피싱(phishing), 스피어 피싱(spear-phishing), 그리고 기업 이메일 침해(Business Email Compromise, BEC)가 있으며, 이들은 정상적인 커뮤니케이션을 모방하여 사용자의 신뢰를 유도하고 즉각적인 의사결정을 촉진한다는 공통된 특징을 가진다.

기존 연구는 이러한 공격을 주로 두 가지 관점에서 접근해 왔다. 첫째는 사용자 교육 기반 접근으로, 사용자의 보안 인식을 향상시켜 공격을 식별하도록 유도하는 방식이다. 그러나 실제 환경에서는 업무 속도, 반복되는 메시지 처리, 그리고 제한된 인지 자원으로 인해 사용자가 항상 체계적인 검증 과정을 수행하기 어렵다는 한계가 존재한다. 특히 사용자 습관과 반복적 행동 패턴이 피싱 성공률에 영향을 미친다는 연구는, 단순 지식 전달형 교육이 지속적인 방어 효과를 보장하지 못함을 시사한다 [8].

둘째는 콘텐츠 기반 탐지 접근으로, 메시지의 텍스트, URL, 발신자 정보 등의 특징을 활용하여 악성 여부를 판단하는 방식이다. 초기 연구에서는 규칙 기반(rule-based) 방법이 주로 사용되었으며, 특정 키워드나 패턴을 기반으로 스팸 메시징을 탐지하는 프레임워크가 제안되었다 [5]. 이후 기계학습 및 딥러닝 기반 모델이 도입되면서 분류 정확도 향상이 이루어졌으며, 특히 하이브리드 모델을 통해 탐지 성

능을 개선하려는 연구가 진행되었다 [3].

그러나 이러한 접근은 공통적으로 메시지의 표면적 특징에 의존한다는 한계를 가진다. 공격자가 문장을 변형하거나 정상 메시지의 언어적 특성을 모방할 경우 탐지 성능이 급격히 저하될 수 있으며, 규칙 기반 시스템은 유지 비용 증가와 오탐/미탐 문제를 반복적으로 야기한다. 또한 이러한 모델들은 사용자의 판단 과정이나 인지 상태를 고려하지 않기 때문에, 실제 공격 환경에서의 의사결정 왜곡을 충분히 설명하지 못한다.

한편, 사회공학 공격이 인간의 인지적 특성을 활용한다는 점에 관한 연구도 지속적으로 이루어져 왔다. Dhamija 등은 사용자가 보안 경고를 인지하고도 피싱에 속을 수 있음을 보이며, 신뢰 형성이 메시지 검증 이전 단계에서 이루어질 수 있음을 지적하였다 [7]. 또한 설득 원리(예: 권위, 긴급성)가 포함된 메시지가 사용자 판단에 영향을 미친다는 연구는, 공격 성공이 단순한 주의 부족이 아니라 인지적 편향과 밀접하게 연관되어 있음을 보여준다 [9].

이와 관련하여 Human-Centered Security 접근은 보안을 기술적 통제나 탐지 성능만의 문제가 아니라, 사용자 행동, 사용성, 인지 부담, 보안 경고의 수용 가능성, 정책 준수 가능성 등을 함께 고려해야 하는 문제로 이해한다 [4][7][8][9]. 이러한 관점은 사회공학 공격 대응에서 사용자가 왜 보안 경고를 무시하거나 정상적인 검증 절차를 생략하는지 설명하는 데 유용하며, 사용자 교육과 보안 경고 설계의 중요성을 강조한다. 그러나 기존 Human-Centered Security 접근은 주로 사용자의 보안 행동 개선, 경고 설계, 사용성 향상에 초점을 두어 왔기 때문에, LLM 기반 공격처럼 공격자가 사용자 반응에 따라 설득 전략을 반복적으로 조정하고 인지 부하를 누적시키는 공격 구조를 충분히 설명하는 데에는 한계가 있다.

결과적으로 전통적 사회공학 방어 프레임워크는 사용자 교육과 콘텐츠 기반 탐지라는 두 축을 중심으로 발전해 왔으나, 인간의 제한된 인지 처리 능력과 편향이 공격 성공을 구조적으로 지지할 수 있다는 점을 충분히 반영하지 못한 한계를 가진다.

2. AI-Driven Social Engineering

최근 생성형 AI의 발전은 사회공학 공격의 생산성과 적응성을 동시에 확대하는 새로운 환경을 형성하고 있다. 전통적인 공격이 공격자의 언어 능력과 시간적 제약에 의존했다면, LLM은 자연스러운 문맥 생성, 다국어 처리, 반복 생성 자동화, 그리고 대상자 맞춤형 메시지 생성을 가능하게 한다.

LLM의 이러한 특성은 공격 메시지의 품질과 규모를 동시에 향상한다. 특히 문맥에 맞는 자연스러운 문장 생성과 개인화된 메시지 변형 능력은 기존 피싱 메시지에서 나타났던 비정상적 언어 패턴을 제거하며, 정상 커뮤니케이션과의 경계를 흐리게 만든다. 이에 따라 공격자는 더 높은 설득력을 가지는 메시지를 대량으로 생성할 수 있으며, 사용자 반응에 따라 메시지를 동적으로 수정하는 상호작용 기반 공격이 가능해진다.

LLM 보안 및 프라이버시 관련 연구에서도 이러한 양면성이 지적되고 있다. Yao 등은 LLM이 보안 방어뿐 아니라 공격에도 활용될 수 있으며, 특히 사용자 대상 공격에서 그 영향력이 크다는 점을 체계적으로 분석하였다 [10]. 또한 BEC 공격에 대한 체계적 문헌 고찰 연구는, 공격이 단순한 메시지 조작을 넘어 조직 프로세스와 인간의 의사결정 구조를 결합한 형태로 진화하고 있음을 강조한다 [11].

이러한 변화는 사회공학 공격을 단순한 콘텐츠 문제로 설명하기 어렵게 만든다. 공격자는 더 이상 정적 템플릿을 사용하는 것이 아니라, 사용자와의 상호작용을 통해 설득 전략을 지속적으로 조정할 수 있으며, 이는 기존 탐지 모델의 가정을 근본적으로 약화한다. 또한 멀티모달 생성 기술(음성, 이미지, 영상 등)과 결합할 경우, 공격은 단일 텍스트 채널을 넘어 다양한 감각 채널을 동시에 활용하는 방향으로 확장된다.

그럼에도 불구하고 기존 연구는 여전히 세 가지 한계를 보인다. 첫째, 텍스트 유사도나 문장 자연스러움과 같은 표면적 지표 중심 평가가 주를 이루며, 사용자의 실제 판단 과정 변화는 충분히 반영되지 않는다. 둘째, 자동화 가능성에 대한 논의는 많지만, 자동화가 사용자 인지 피로를 누적적으로 증가시키는 메커니즘에 대한 분석은 제한적이다. 셋째, 상호작용 기반 공격 환경을 고려한 인간 중심 방어 모델은 아직 초기 단계에 머물러 있다.

한편, 최근 연구에서는 인지 편향을 직접적으로 고려한 방어 접근이 제안되기 시작하였다. Schöni 등은 증강현실(AR)을 활용하여 공격자가 사용하는 설득 전략을 시각화하고, 사용자가 이를 인식하도록 하는 훈련 시스템을 제안하였다 [12]. 이러한 접근은 사회공학 방어가 단순한 메시지 차단을 넘어, 사용자의 판단 과정 자체에 개입해야 함을 시사한다.

3. Research Gap

앞서 살펴본 기존 연구를 종합하면, 사회공학 공격과 방어에 관한 연구는 크게 발전했음에도 불구하고 다음과 같은 구조적 공백이 존재한다.

첫째, 대부분의 탐지 연구는 메시지의 표면적 특징과 분류 성능에 초점을 맞추고 있으며, 사용자의 인지 상태, 행동 패턴, 그리고 업무 환경과 같은 요소가 방어 성능에 미치는 영향을 충분히 반영하지 못한다. 사용자 습관이 피싱 결과에 영향을 미친다는 연구 결과는 존재하지만, 이러한 요소가 실질적인 방어 모델로 통합되는 경우는 드물다 [8].

둘째, 설득 원리와 시간 압박이 결합할 때 발생하는 판단 왜곡에 대한 분석은 일부 존재하지만, 이를 실제 조직 환경과 결합하여 설명하는 연구는 제한적이다. 시간 압박과 설득 전략이 동시에 작용할 경우 사용자 판단이 크게 왜곡될 수 있다는 연구는, 방어 모델이 상황적 요소를 반드시 고려해야 함을 시사한다 [13].

셋째, LLM 기반 사회공학 공격은 상호작용을 통해 지속적으로 진화하는 특성을 가지지만, 기존 방어 모델은 여전히 단일 메시지 단위의 정적 분석에 머무르는 경우가 많다. 이는 적응형 공격 환경에서 방어 효과가 급격히 감소할 수 있음을 의미한다 [10].

넷째, 자동화된 공격이 장기적으로 사용자 인지 자원에 미치는 누적적 영향, 즉 인지 피로와 순응(compliance)의 문제는 충분히 분석되지 않았다. 반복적인 메시지 노출과 정보 과부하는 사용자의 판단 기준을 점진적으로 약화할 수 있으며, 이는 단일 공격보다 더 큰 위험을 초래할 수 있다.

결론적으로 기존 연구는 (i) LLM 기반 생성 능력, (ii) 설득 전략의 자동화, (iii) 상황 압박과 결합한 판단 왜곡, (iv) 인지 자원의 누적적 소모라는 요소를 통합적으로 설명하지 못하는 한계를 가진다.

이에 본 논문은 사회공학 공격을 단순한 메시지 조작 문제가 아닌, 인간 인지 취약성이 구조적으로 확장되는 문제로 재정의한다. 특히 본 연구는 공격이 공략하는 인지 요소와 상호작용 구조에 초점을 맞추고, 이를 기반으로 기존 보안 패러다임의 한계를 분석하며 새로운 인지 기반 보안 프레임워크의 필요성을 제시한다.

III. Cognitive Modeling of LLM-Based Social Engineering

1. Threat Model

본 연구는 LLM 기반 사회공학 공격을 기존 공격과 구분하기 위해 공격자의 능력, 목표, 그리고 공격 과정을 포함하는 위협 모델을 정의한다.

먼저 공격자는 공개적으로 접근 가능한 다양한 정보를 활용할 수 있는 것으로 가정한다. 여기에는 소셜 미디어,

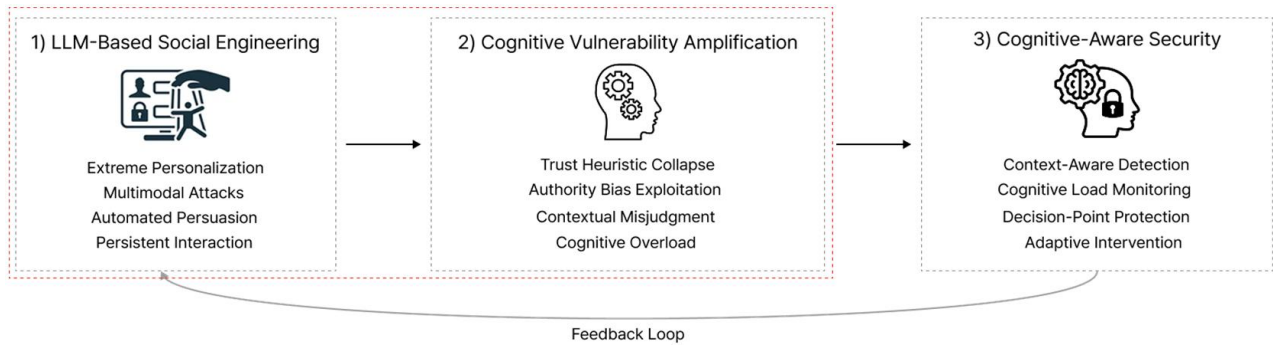


Fig. 1. Overall Architecture of LLM-Based Social Engineering and Cognitive-Aware Security Framework

조직 구조 정보, 이메일 패턴, 과거 커뮤니케이션 기록 등이 포함된다. 또한 공격자는 LLM을 활용하여 자연스러운 문맥 생성, 대상자 맞춤형 메시지 생성, 그리고 반복적 메시지 생성 자동화를 수행할 수 있다. 필요에 따라 음성 합성 및 이미지 생성과 같은 멀티모달 생성 기술을 결합하여 공격의 설득력을 강화할 수 있다.

공격자의 주요 목표는 다음 세 가지로 정의된다. 첫째, 계정 정보나 인증 코드와 같은 민감 정보 탈취, 둘째, 금전 이체와 같은 재정적 행동 유도(BEC 공격 포함), 셋째, 계정 복구 및 권한 변경을 통한 시스템 제어 권한 확보이다. 이러한 목표는 기술적 취약점이 아닌 인간의 의사결정 과정을 주요 공격 표면으로 삼는다는 특징을 가진다.

공격 과정은 다음과 같은 단계로 구성된다. (i) 대상자 정보 수집 및 프로파일링, (ii) 상황 맥락에 맞는 메시지 생성, (iii) 메시지 전달 및 사용자 반응 유도, (iv) 사용자 반응에 따른 반복적·적응적 설득.

기존 사회공학 공격이 정적 템플릿과 제한된 개인화에 의존했다면, LLM 기반 공격은 동적 적응(adaptive interaction), 대규모 개인화(extreme personalization), 그리고 상호작용 기반 설득(iterative persuasion)을 특징으로 한다. 이는 공격이 단순한 메시지 전달이 아니라, 사용자의 판단 과정을 지속적으로 조정하는 구조적 과정임을 의미한다. 이러한 위협 모델은 이후 제안되는 인지 기반 보안 프레임워크의 설계 기준으로 활용된다.

2. Cognitive-Aware Security Framework

Cognitive-Aware Security Framework는 메시지의 악성 여부만을 판정하는 모델이 아니라, 상호작용 흐름과 사용자의 판단 부담을 평가하여 고위험 의사결정 직전에 개입하는 보호 구조이다.

이를 위해 세 가지 핵심 구성 요소를 정의한다. 첫 번째는 상호작용 맥락 분석 모듈(interaction context analysis)로, 단일 메시지가 아닌 전체 상호작용 흐름을

분석한다. 메시지의 빈도, 시간 간격, 발신자 관계, 멀티모달 요소의 존재 여부, 그리고 정상 커뮤니케이션 패턴과의 차이를 종합적으로 고려하여 현재 상황의 위험도를 평가한다.

두 번째는 사용자 인지 부하 추정 모듈(cognitive load estimation)이다. 이 모듈은 메시지의 긴급성, 요구되는 작업의 복잡도, 동시에 처리해야 하는 요청의 수, 그리고 다중 채널의 중첩 여부를 기반으로 사용자의 인지 부담 수준을 추정한다. 높은 인지 부하는 사용자의 판단 정확도를 저하해 공격 성공 가능성을 증가시키는 주요 요인으로 작용한다.

세 번째는 의사결정 보호 계층(decision protection layer)으로, 금전 이체, 계정 정보 입력, 권한 승인과 같은 중요한 의사결정 시점에서 개입한다. 이 계층은 메시지를 차단하는 대신 실행 지연, 추가 인증, 상황 기반 경고, 다중 채널 검증과 같은 방법을 통해 사용자의 판단 과정을 안정화한다.

이를 기반으로 전체 시스템은 위험도를 정량적으로 평가하는 함수와 의사결정 개입 조건으로 구성된다. 시스템은 상호작용 맥락, 사용자 인지 부하, 그리고 상호작용 패턴을 입력으로 받아 전체 위험도를 계산하며, 위험도는 다음과 같이 정의된다.

$$R = w_1 C + w_2 L + w_3 I$$

여기서 R 은 전체 위험도(risk score), C 는 맥락 기반 위험도(contextual risk), L 은 인지 부하(cognitive load), I 는 상호작용 패턴 위험도(interaction pattern risk)를 의미하며, w_1, w_2, w_3 는 각 요소의 중요도를 반영하는 가중치로서 $w_1 + w_2 + w_3 = 1$ 을 만족한다.

C 는 요청이 조직 맥락에서 벗어난 정도, L 은 사용자가 요청을 검토할 때 요구되는 인지적 부담, I 는 공격자가 사용자 반응에 맞추어 설득을 반복·조정하는 정도를 의미한다.

Table 1. Observable Indicators and Example Scoring Criteria for Risk Variables.

Variable	Indicators	Scoring
C: Contextual risk	sender anomaly / role mismatch / procedure violation / high-risk request	0: normal / 0.5: partial mismatch / 1: high-risk mismatch
L: Cognitive Load	urgency / task complexity / multiple requests / channel overlap	0: low / 0.5: medium / 1: high
I: Interaction risk	repetition / pressure / strategy shift / re-persuasion	0: single / 0.5: repeated / 1: adaptive

이러한 지표는 위험 수준에 따라 0에서 1 사이의 값으로 정규화할 수 있으며, 값이 높을수록 사용자의 의사결정이 왜곡될 가능성이 높음을 의미한다. 가중치 w_1 , w_2 , w_3 는 초기 적용 단계에서는 동일 가중치로 설정할 수 있으나, 조직의 업무 특성이나 보안 정책에 따라 조정될 수 있다. 향후 실제 사용자 실험이나 조직 로그 데이터가 확보될 경우, 각 가중치와 임계값은 데이터 기반으로 보정될 수 있다.

이후 계산된 위험도가 사전에 정의된 임계값을 초과할 경우, 시스템은 사용자의 의사결정 과정에 개입한다. 이는 다음과 같이 표현될 수 있다.

$$Trigger = \begin{cases} 1 & \text{if } R > \tau \\ 0 & \text{otherwise} \end{cases}$$

이러한 구조는 특정 메시지의 악성 여부를 완벽하게 탐지하는 것을 전제로 하지 않고, 설득 기반 공격 환경에서도 사용자의 의사결정 과정을 안정적으로 보호하는 것을 목표로 한다.

제안된 프레임워크의 전체 구조는 Fig. 1에 나타나 있다. Fig. 1은 LLM 기반 사회공학 공격의 주요 특성과 인간 인지 취약성, 그리고 이에 대응하기 위한 Cognitive-Aware Security Framework의 구성 요소를 함께 보여준다. 먼저 공격 측면에서 LLM은 초개인화된 메시지 생성, 멀티모달 공격, 자동화된 설득, 지속적 상호작용을 가능하게 한다. 이러한 특성은 공격자가 단일 메시지를 전달하는 데 그치지 않고, 사용자의 상황과 반응에 따라 설득 전략을 조정할 수 있음을 의미한다.

이러한 공격 특성은 인간의 인지 취약성과 연결된다. 초개인화된 메시지는 사용자의 맥락 의존성을 자극하고, 자연스러운 언어 표현은 언어 기반 신뢰를 형성하며, 권위 있는 발신자 표현은 권위 편향을 강화한다. 또한 반복적 요청과 긴급성 표현은 사용자의 인지 부하를 증가시켜 비판적 검토보다 직관적 판단에 의존하도록 만들 수 있다.

이에 대응하여 제안 프레임워크는 상호작용 맥락 분석, 인지 부하 추정, 의사결정 보호 계층으로 구성된다. 상호작용 맥락 분석은 메시지의 빈도, 시간 간격, 발신자 관계, 정상 커뮤니케이션 패턴과의 차이를 평가하고, 인지 부하 추정은 긴급성, 작업 복잡도, 동시 요청 수, 다중 채널 중첩 여부를 고려한다. 이후 위험도가 높다고 판단되는 경우, 의사결정 보호 계층은 금전 이체, 계정 정보 입력, 권한 승인과 같은 고위험 행동이 실제로 수행되기 전에 실행 지연, 추가 인증, 상황 기반 경고, 다중 채널 검증과 같은 방식으로 개입한다.

또한 Fig. 1의 피드백 루프는 LLM 기반 사회공학 공격과 방어가 정적인 관계가 아니라, 사용자 반응과 상호작용 흐름에 따라 동적으로 변화하는 구조임을 보여준다. 따라서 효과적인 대응을 위해서는 개별 메시지의 악성 여부만을 판단하는 방식에서 벗어나, 사용자 상태와 상호작용 맥락을 지속적으로 반영하는 적응적 개입이 필요하다.

3. Case Study

본 절에서는 제안된 인지 기반 보안 프레임워크의 필요성을 설명하기 위해, LLM 기반 사회공학 공격 시나리오를 구조적으로 분석한다. 이를 위해 가상의 기업 환경을 가정하고, 공격자가 특정 직원을 대상으로 수행하는 기업 이메일 침해(BEC) 공격을 사례로 제시한다.

공격자는 먼저 조직의 공개 정보를 수집하여 재무 담당자와 같은 특정 직원을 타깃으로 설정한다. 이 과정에서 소셜 미디어, 조직 구조 정보, 그리고 기존 커뮤니케이션 패턴을 분석하여 대상자의 역할과 관계를 파악한다. 이러한 정보 수집은 이후 공격 메시지의 맥락적 정합성과 신뢰도를 강화하는 기반으로 작용한다.

이후 공격자는 LLM을 활용하여 CEO의 언어 스타일과 조직 내부 맥락을 반영한 메시지를 생성한다. 생성된 메시지는 자연스럽고 실제 업무 커뮤니케이션과 유사한 형태를 가지며, 긴급한 자금 이체 요청과 같은 상황을 포함하여 사용자의 즉각적인 대응을 유도한다.

생성된 메시지는 이메일 또는 메신저를 통해 전달되며, 긴급성과 중요성을 강조하는 방식으로 구성된다. 이 단계에서는 시간 압박과 업무 맥락이 결합해 사용자의 인지 자원을 제한하고, 비판적 검토보다는 직관적 판단을 유도한다. 이후 공격자는 사용자 반응에 따라 메시지를 지속적으로 수정하며 추가적인 설득을 수행한다. 이러한 적응적 설득 과정은 기존의 단일 메시지 기반 공격과 달리, 상호작용을 통해 신뢰를 점진적으로 강화하는 특징을 가진다.

이러한 공격 과정은 단계별로 특정한 인지 취약성을 체

계적으로 공략한다. 이를 정리하면 다음과 같다.

Table 2. Mapping between Attack Stages and Cognitive Vulnerabilities.

Attack Stage	Attack Strategy	Cognitive Vulnerability
Step 1	contextual information gathering	contextual dependence
Step 2	natural language generation	language-based trust
Step 3	message delivery	authority bias
Step 4	iterative persuasion	cognitive overload

위 표에서 볼 수 있듯이, 공격자는 단일 취약점을 공략하는 것이 아니라 각 단계마다 서로 다른 인지 요소를 결합하여 사용자의 판단을 점진적으로 왜곡한다. 특히 초기 단계에서는 맥락 의존성을 통해 신뢰 형성을 유도하고, 이후 권위와 긴급성을 결합하여 의사결정을 가속하며, 마지막 단계에서는 반복적 상호작용을 통해 인지 자원을 소모하는 전략을 사용한다. 이러한 구조는 공격이 단순한 메시지 전달이 아니라 인간의 인지 과정을 단계적으로 조작하는 과정임을 보여준다.

해당 시나리오에는 기존 콘텐츠 기반 보안 모델이 가지는 한계를 보여준다. 공격 메시지는 문법적으로 자연스럽고 정상적인 커뮤니케이션과 유사하므로 단순한 텍스트 분석이나 패턴 기반 탐지만으로는 식별이 어렵다. 또한 시간 압박과 권위 기반 설득은 사용자의 판단 과정을 왜곡하여 보안 경고의 효과를 약화할 수 있다.

기존 탐지 방식은 주로 개별 메시지의 텍스트 패턴, URL, 발신자 정보와 같은 콘텐츠 수준의 신호를 분석한다. 반면 제안 프레임워크는 동일한 시나리오를 상호작용 흐름, 사용자 인지 부하, 의사결정 시점이라는 관점에서 분석한다. 예를 들어 공격자가 사용자의 반응에 따라 메시지를 반복적으로 수정하거나 긴급성을 강화하는 경우, 이는 단일 메시지의 악성 여부보다 상호작용 패턴과 인지 부하 측면에서 중요한 위험 신호로 해석될 수 있다.

따라서 사회공학 방어는 메시지를 차단하는 기존 접근에서 벗어나, 사용자의 의사결정 과정이 공격에 의해 왜곡되는 지점을 식별하고 보호하는 방향으로 확장될 필요가 있다. 본 연구에서 제안한 인지 기반 보안 프레임워크는 상호작용 맥락과 사용자 상태를 통합적으로 고려함으로써, 기존 탐지 방식이 포착하기 어려운 인지적·상호작용적 위험 요소를 설명하고 대응 설계에 반영할 수 있는 구조를 제시한다.

IV. Structural Analysis and Discussion

본 장에서는 제안된 Cognitive-Aware Security Framework를 바탕으로, LLM 기반 사회공학 공격이 기존 보안 모델에 제기하는 한계와 대응 방향을 논의한다. 본 연구는 실험적 성능 검증을 수행하기보다, 공격 메커니즘과 방어 모델의 관계를 분석하고 사례 기반 논의를 통해 제안 프레임워크의 적용 가능성을 검토한다.

이를 위해 본 장에서는 세 가지 기준을 중심으로 논의를 전개한다. 첫째, LLM 기반 사회공학 공격을 개별 메시지가 아니라 상호작용 흐름으로 파악할 수 있는지 살펴본다. 둘째, 사용자 인지 부하와 판단 왜곡 가능성을 위험 평가 요소로 반영할 수 있는지 검토한다. 셋째, 금전 이체, 계정 정보 입력, 권한 승인과 같은 고위험 의사결정 시점에서 예방적 개입이 가능한지 논의한다. 이러한 기준은 제안 프레임워크의 의의를 정량적 성능 우위가 아니라, 기존 보안 모델이 충분히 포착하지 못한 인지적·상호작용적 위험 요소를 설명하고 대응 설계에 반영할 수 있는지의 관점에서 평가하기 위한 것이다.

기존 콘텐츠 기반 보안 모델은 LLM 기반 사회공학 공격 환경에서 여러 한계를 드러낸다. 전통적인 탐지 방식은 메시지의 텍스트 패턴, URL 구조, 발신자 정보 등 비정상적 특징에 의존한다. 그러나 LLM이 생성한 공격 메시지는 문법적 오류나 비정상적 표현이 거의 없고, 정상적인 업무 커뮤니케이션과 높은 유사성을 가진다. 이로 인해 기존 모델이 전제하는 “이상 패턴 기반 탐지” 가정은 약화되며, 탐지 성능 역시 저하될 수 있다. 또한 공격이 단일 메시지가 아니라 반복적 상호작용을 통해 진행되는 경우, 개별 메시지 분석만으로는 전체 공격 의도를 파악하기 어렵다.

사례 기반 논의에서도 LLM 기반 사회공학 공격은 단순히 메시지 품질을 높이는 수준을 넘어, 사용자의 인지 취약성을 단계적으로 악용할 수 있음을 보여준다. 공격자는 자연스러운 언어 표현을 통해 신뢰를 형성하고, 권위 있는 발신자로 가장하여 사용자의 검증 의지를 약화할 수 있다. 여기에 긴급성, 반복적 요청, 상황 맥락과의 일치성이 결합되면 사용자는 비판적 검토보다 직관적 판단에 의존할 가능성이 높아진다. 따라서 LLM 기반 사회공학 공격은 콘텐츠 탐지만으로 설명하기 어려우며, 사용자의 판단 환경과 의사결정 흐름을 함께 고려해야 한다.

Table 3. Comparison of Existing Security Models and the Proposed Cognitive-Aware Security Framework.

Aspect	Conventional Security	Human-Centered Security	Proposed Framework
Analysis Unit	message level	user-interaction level	decision-flow level
Focus	content detection	usability & behavior	cognition & context
Threat View	malicious content	user error	cognitive manipulation
Assumption	detectable anomalies	supportable users	vulnerable decisions
Defense Strategy	blocking / filtering	education / warning	decision protection
Intervention Timing	message delivery	user interaction	high-risk decision
Key Risk Factor	text / URL / sender	usability / awareness	load / context / persuasion
Main Limitation	evasion by natural text	limited persuasion modeling	need for empirical validation

위 비교에서 볼 수 있듯이, 기존 모델은 콘텐츠 기반 탐지에 의존하는 반면, 제안된 프레임워크는 사용자 중심의 의사결정 보호를 핵심으로 한다. 기존 사회공학 대응 연구는 주로 메시지, URL, 발신자 정보와 같은 콘텐츠 수준의 신호를 분석하거나, 사용자 교육과 보안 경고를 통해 공격 인식을 높이는 데 초점을 두어 왔다. 또한 Human-Centered Security 접근은 사용자 행동, 보안 경고, 사용성, 정책 준수와 같은 인간 요소를 보안 설계에 반영한다는 점에서 중요한 의미가 있다.

그러나 LLM 기반 사회공학 공격은 단순히 사용자가 악성 메시지를 인식하는지의 문제가 아니라, 공격자가 자연스러운 언어 표현, 권위 있는 발신자 표현, 긴급성, 반복적 설득을 결합하여 사용자의 판단 환경을 변화시키는 문제이다. 따라서 본 연구는 공격 메시지 자체보다 해당 메시지가 사용자의 인지 부하와 의사결정 과정에 어떤 영향을 미치는지에 주목한다는 점에서 기존 접근과 차별화된다.

이에 따라 제안된 Cognitive-Aware Security Framework는 상호작용 맥락, 사용자 인지 상태, 의사결정 시점이라는 요소를 통합하여 LLM 기반 사회공학 공격에 대응하고자 한다. 이는 기존 콘텐츠 기반 탐지 연구, 사용자 교육 중심 연구, Human-Centered Security 접근을 대체하는 것이 아니라, 생성형 AI 환경에서 기존 연구가 충분히 다루지 못했던 인지적·상호작용적 위험 요소를 보완하는 확장적 접근이라 할 수 있다.

이러한 관점에서 본 연구는 사회공학 공격을 단순한 사용자 부주의의 문제가 아니라 인간 인지 구조의 취약성을 활용하는 구조적 문제로 해석한다. 이에 따라 향후 보안

시스템은 콘텐츠 분석 중심 접근을 넘어, 사용자 상태, 상호작용 맥락, 그리고 의사결정 과정을 통합적으로 고려하는 방향으로 설계될 필요가 있다.

V. Conclusion

본 논문은 대규모 언어 모델(LLM)을 기반으로 한 생성형 AI가 사회공학 공격의 구조를 근본적으로 변화시키고 있음을 분석하였다. 기존 사회공학 공격이 공격자의 언어 능력과 시간적 제약으로 제한되었던 반면, LLM 기반 공격은 초개인화된 메시지 생성, 멀티모달 정보 결합, 그리고 반복적 상호작용을 통한 설득을 가능하게 함으로써 인간의 의사결정 과정을 직접적으로 공략하는 형태로 진화하고 있다.

본 연구는 이러한 변화를 단순한 공격 자동화의 문제가 아닌, 인간 인지 취약성의 구조적 증폭 문제로 재정의하였다. 이를 위해 언어 기반 신뢰, 권위 편향, 맥락 의존성, 인지 과부하와 같은 핵심 인지 요소를 중심으로 공격 메커니즘을 분석하였으며, 사례 기반 구조적 분석을 통해 이러한 요소들이 결합할 때 사용자 판단이 단계적으로 왜곡될 수 있음을 논의하였다.

본 연구의 주요 기여를 종합하면 다음 세 가지 범주로 정리할 수 있다. 첫째, LLM 기반 사회공학 공격을 콘텐츠 탐지의 문제가 아니라 인간 의사결정 보호의 문제로 재정의하였다. 둘째, 기존 콘텐츠 기반 보안 모델이 충분히 다루지 못했던 상호작용 맥락, 사용자 인지 상태, 의사결정 시점을 통합적으로 고려하는 Cognitive-Aware Security Framework를 제안하였다. 셋째, 사례 기반 구조적 분석을 통해 LLM 기반 공격이 단일 메시지의 악성 여부보다 반복적 설득과 인지 부하의 누적을 통해 사용자 판단을 왜곡할 수 있음을 설명하였다.

실무적 관점에서 제안된 Cognitive-Aware Security Framework는 커뮤니케이션을 통해 형성된 판단이 금전이체, 중요 정보 제공, 계정 변경 또는 권한 승인과 같은 고위험 행동으로 이어지는 업무 환경에 활용될 수 있다. 대표적으로 기업 이메일·업무용 메신저·전자결재 환경에서는 발신자와 수신자의 관계, 요청 내용과 담당 업무의 일치 여부, 정상 결재 절차와의 차이, 반복적인 긴급 요청 등을 상호작용 맥락의 관점에서 분석할 수 있다. 금융 거래 및 고객 상담 환경에서는 즉각적인 송금 요구, 여러 채널을 통한 반복적 연락, 복잡한 처리 요구가 사용자나 상담원의 판단 부담을 높이는지를 검토할 수 있으며, 계정 복

구 및 접근 권한 관리 환경에서는 비밀번호 초기화, 다중 요소 인증 수단 변경, 관리자 권한 부여와 같은 요청이 실제로 처리되기 전에 추가적인 확인이 필요한 상황을 식별하는 기준으로 활용할 수 있다. 즉, 제안 프레임워크는 개별 메시지가 정상적인 형태를 보이더라도 전체 상호작용 과정에서 사용자의 판단이 왜곡될 가능성을 분석하고, 중요한 행동이 실행되기 전에 의사결정 보호 계층을 작동시키는 데 적용될 수 있다.

또한 본 프레임워크는 이메일 요약, 업무 요청 분류, 결재 지원 및 후속 행동 추천 기능을 제공하는 생성형 AI 기반 업무 지원 시스템이나 LLM 에이전트에도 적용될 수 있다. 외부에서 전달된 메시지가 AI 시스템을 통해 요약되거나 중요 요청으로 추천되고, 나아가 실제 업무 수행으로 연결되는 경우, 해당 메시지의 내용뿐 아니라 발신자 관계, 업무 맥락, 긴급성, 반복성 및 요구 행동의 위험성을 함께 검토하는 기준으로 활용할 수 있다. 본 연구의 적용상 차별점은 새로운 인증이나 차단 수단을 제안하는 데 있는 것이 아니라, 기존 콘텐츠 탐지 시스템이 포착하기 어려운 인지적·상호작용적 위험을 평가하고, 그 결과를 기존의 추가 인증, 독립적 확인 또는 승인 절차와 연결하는 데 있다. 이를 통해 제안 프레임워크는 사회공학 대응의 범위를 메시지 탐지에서 사용자의 의사결정 과정 보호로 확장하는 실무적 분석 기준으로 활용될 수 있다.

다만 본 연구는 개념적 모델링과 구조적 분석에 초점을 맞추고 있으며, 실제 사용자 실험이나 시스템 구현을 포함하지 않는다는 한계를 가진다. 따라서 본 연구의 결과는 제안 프레임워크가 기존 보안 모델보다 정량적으로 우수함을 입증하는 것이 아니라, LLM 기반 사회공학 공격 환경에서 기존 콘텐츠 기반 보안 모델이 충분히 다루지 못하는 인지적·상호작용적 위험 요소를 보완하는 분석 틀로 해석되어야 한다. 향후 연구에서는 사용자 실험, 전문가 평가, 시뮬레이션 기반 평가, 그리고 실제 조직 환경에서의 적용을 통해 제안 프레임워크의 효과를 정량적·정성적으로 검증할 필요가 있다. 또한 멀티모달 공격 환경과 다양한 사용자 특성을 반영한 모델로의 확장이 요구된다.

결론적으로, 생성형 AI는 사회공학 공격을 정교화하는 도구를 넘어 인간 인지 취약성을 구조적으로 증폭시키는 핵심 요인으로 작용하고 있다. 이에 따라 향후 보안 연구는 기술적 탐지를 넘어 인간의 의사결정 과정을 보호하는 방향으로 확장되어야 하며, 본 연구는 이러한 전환을 위한 기초적 틀을 제공한다.

ACKNOWLEDGEMENT

This work was supported by the Institute of Information & Communications Technology Planning & Evaluation(IITP)-Innovative Human Resource Development for Local Intellectualization program grant funded by the Korea government (MSIT)(IITP-2026-RS-2022-00156334).

REFERENCES

- [1] D. Goel, "Detection and Prevention of Smishing Attacks," arXiv, Dec. 31, 2024. <https://arxiv.org/abs/2501.00260>
- [2] D. Timko, D. H. Castillo and M. L. Rahman, "A Quantitative Study of SMS Phishing Detection," arXiv, Nov. 12, 2023. <https://arxiv.org/abs/2311.06911>
- [3] T. Mahmud, M. A. H. Prince, M. H. Ali, M. S. Hossain, and K. Andersson, "Enhancing Cybersecurity: Hybrid Deep Learning Approaches to Smishing Attack Detection," Systems, vol. 12, no. 11, Art. no. 490, Nov. 2024. DOI: 10.3390/systems12110490
- [4] D. Timko, D. H. Castillo, and M. L. Rahman, "Understanding Influences on SMS Phishing Detection: User Behavior, Demographics, and Message Attributes," in Proc. Symposium on Usable Security and Privacy (USEC 2025), San Diego, CA, USA, Feb. 2025. DOI: 10.14722/usec.2025.23027
- [5] A. K. Jain and B. B. Gupta, "Rule-Based Framework for Detection of Smishing Messages in Mobile Environment," Procedia Computer Science, vol. 125, pp. 617-623, 2018. DOI: 10.1016/j.procs.2017.12.079
- [6] S. M. Sanjari, J. Roberts, and M. M. A. Pritom, "Poster: A Few-Shot Learning Method for SMS Phishing Detection and Explanation Using SLMs," in Proc. 46th IEEE Symposium on Security and Privacy, San Francisco, CA, USA, May. 2025. <https://sp2025.ieee-security.org/downloads/posters/sp25posters-final23.pdf>
- [7] R. Dhamija, J. D. Tygar, and M. A. Hearst, "Why Phishing Works," Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pp. 581-590, 2006. DOI: 10.1145/1124772.1124861
- [8] A. Vishwanath, "Examining the distinct antecedents of e-mail habits and its influence on the outcomes of a phishing attack," Journal of Computer-Mediated Communication, vol. 20, no. 5, pp. 570-584, 2015. DOI: 10.1111/jcc4.12126
- [9] P. Lawson, C. J. Pearson, A. Crowson, and C. B. Mayhom, "Email phishing and signal detection: How persuasion principles and personality influence response patterns and accuracy," Applied

- Ergonomics, vol. 86, Art. no. 103084, Jul. 2020. DOI: 10.1016/j.apergo.2020.103084
- [10] Y. Yao, J. Duan, K. Xu, Y. Cai, Z. Sun, and Y. Zhang, "A survey on large language model (LLM) security and privacy: The Good, The Bad, and The Ugly," *High-Confidence Computing*, vol. 4, no. 2, Art. no. 100211, Jun. 2024. DOI: 10.1016/j.hcc.2024.100211
- [11] A. Almutairi, B. Kang, and N. Alhashimy, "Business email compromise: A systematic review of understanding, detection, and challenges," *Computers & Security*, vol. 158, Art. no. 104630, Nov. 2025. DOI: 10.1016/j.cose.2025.104630
- [12] L. Schöni, M. Strohmeier, I. Sluganovic, and V. Zimmermann, "Stop the clock—Counteracting bias exploited by attackers through an interactive augmented reality phishing training," in *Proc. 2025 CHI Conference on Human Factors in Computing Systems (CHI '25)*, Yokohama, Japan, Apr. 26–May 1, 2025, Art. no. 594, pp. 1–23, DOI: 10.1145/3706598.3714023
- [13] J. Black and D. M. Sarno, "The influence of time pressure and persuasion principles on phishing detection," *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, vol. 67, no. 1, pp. 1977–1982, Sep. 2023, DOI: 10.1177/21695067231192442

Authors



Woo Jin Jung received the B.S. degree in Cyber Security from Pai Chai University, South Korea in 2026. He is currently pursuing the M.S. degree in Information Security at Korea University, South Korea.

His research interests include AI-driven security analytics, cyber threat intelligence, malware detection, data-efficient learning and medical data analytics.



Ah Reum Kang received the M.S. and Ph.D. degrees in information security from Korea University, South Korea, in 2012 and 2016. She is a professor in the Department of Information Security at Pai Chai University

in Daejeon, South Korea. Her current research interests include security, artificial intelligence, malware, medical data analysis, and online game security.