Journal of The Korea Institute of Information Security & Cryptology ISSN 1598-3986(Print)
VOL.34, NO.2, Apr. 2024 ISSN 2288-2715(Online)
https://doi.org/10.13089/JKIISC.2024.34.2.301

생성형 인공지능 관련 범죄 위협 분류 및 대응 방안*

박 우 빈,[1†] 김 민 수,[2] 박 윤 지,[2] 유 혜 진,[1] 정 두 원[3‡]
[1]동국대학교 (대학원생), [2,3]성균관대학교 (대학원생, 교수)

Taxonomy and Countermeasures for Generative Artificial Intelligence Crime Threats*

Woobeen Park,[1†] Minsoo Kim,[2] Yunji Park,[2] Hyejin Ryu,[1] Doowon Jeong[3‡]
[1]Dongguk University (Graduate student),
[2,3]SungKyunKwan University (Graduate student, Professor)

요 약

생성형 인공지능은 현재 빠른 속도로 발전하고 있고, 산업적으로도 확대되고 있다. 생성형 인공지능의 발전은 대부분의 산업 분야에서 생산성을 향상시킬 수 있을 것이라 기대되고 있다. 그러나 생성형 인공지능은 악용될 수 있으며, 실제로 범죄까지 이어지는 사례들이 등장하고 있다. 빠르게 발전하는 인공지능의 속도에 비해 이를 규제할 수 있는 법안이 존재하지 않는다. 국내의 경우, 법률제정을 위한 생성형 인공지능 기술과 관련된 범죄 및 위험에 대한 분류가 명확하게 이루어지지 않은 상황이다. 이에 본 연구에서는 생성형 인공지능 관련 범죄를 기존 사이버범죄 분류법에 착안하여 생성형 인공지능 침해범죄 위협, 생성형 인공지능 이용범죄 위협, 기타 인공지능 관련 위협으로 구분하고자 하였다. 또한, 범죄 및 위험에 대한 기술적 대응 방안을 인공지능 개발 단계별로 제시하여 현실성 있는 위협 대응 방안을 다루었다. 법·제도적 개선사항을 통해 생성형 인공지능 범죄에 대한 개발사의 책임과 데이터 수집 방법론의 법제화 등을 제시하였다.

ABSTRACT

Generative artificial intelligence is currently developing rapidly and expanding industrially. The development of generative AI is expected to improve productivity in most industries. However, there is a probability for exploitation of generative AI, and cases that actually lead to crime are emerging. Compared to the fast-growing AI, there is no legislation to regulate the generative AI. In the case of Korea, the crimes and risks related to generative AI has not been clearly classified for legislation. In addition, research on the responsibility for illegal data learned by generative AI or the illegality of the generated data is insufficient in existing research. Therefore, this study attempted to classify crimes related to generative AI for domestic legislation into generative AI for target crimes, generative AI for tool crimes, and other crimes based on ECRM. Furthermore, it suggests technical countermeasures against crime and risk and measures to improve the legal system. This study is significant in that it provides realistic methods by presenting technical countermeasures based on the development stage of AI.

Keywords: Artificial Intelligence, Generative Artificial Intelligence, Generative AI Crimes, Crime Taxonomy

Received(12. 29. 2023), Modified(02. 13. 2024), Accepted(02. 22. 2024)

* 이 논문은 2024년도 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원을 받아 수행된 연구 결과임(No. RS-2024-00398745, 디지털 환경에서의 증거인멸행위 증명 및 대응 기술 개발).

† 주저자, qkrdnqls30@dgu.ac.kr
† 교신저자, doowon@g.skku.edu(Corresponding author)

I. 서 론

생성형 인공지능은 다양한 형태의 데이터를 생성할 수 있는 인공지능 기술로, 2022년 하반기부터 기업 차원의 적극적인 기술 개발과 서비스 출시를 통해 하나의 산업 영역으로 자리 잡게 되었다. 가장 대표적인 생성형 인공지능 서비스인 'ChatGPT'는 2022년 11월 출시하여, 두 달 만에 월 이용자 수 1억 명을 기록하였고, 2023년 2월에 메타에서 출시된 'LLaMA'는 기존의 인공지능 모델과 다르게 모델의 가중치까지 완전히 공개되어 생성형 인공지능의 개발 진입장벽을 낮추었다[1, 2]. LLaMA 공개로 인해 거대 언어모델을 경량화한 모델들이 다수 개발되었으며, 생성형 인공지능 시장이 확대되는 계기가 되었다[3]. 현재 텍스트, 이미지, 음성, 영상 등 다양한 데이터를 생성할 수 있는 인공지능 서비스가 의료, 디자인, 게임 산업을 포함한 넓은 분야에서 활용된다.

그러나, 생성형 인공지능은 악의적으로 사용될 수도 있으며, 실제 사례들이 존재한다. 2023년 10월 미국 뉴저지의 Westfield 고등학교에서는 일부 학생들이 같은 학급의 여학생들의 나체 사진을 제작하여 유포하였는데, 나체 사진을 제작하는 과정 중에 이미지를 생성하는 인공지능 기술인 딥페이크가 사용된 사실이 밝혀졌다[4]. 또한, 미국의 SF 잡지사 '클락스월드(clarkesworld)'는 ChatGPT를 이용한 소설들이 다수 접수되고, 표절로 판정되어 투고가 거절되는 등의 원인으로 2023년 2월부로 단편 작품 접수를 중단하였다[5]. Interpol(2023)은 보고서를 발간하여 생성형 인공지능의 악의적인 사용 방향 가능성을 경고했으며, 생성형 인공지능과 거대 언어모델을 이용한 사기 및 사칭, 멀웨어 제작과 유포, 가짜정보·여론조작 등에 대해 주의가 필요하다고 설명하였다[6].

기존 연구에서는 일반인의 잘못된 인공지능 서비스 사용과 공격자의 행동에 초점을 맞추어 인공지능과 관련된 악의적 행위를 설명하고 있다. 하지만, 현재까지 생성형 인공지능에 의해 발생할 수 있는 범죄에 대한 분류가 명확하게 이루어지지 않았다. 본 연구에서는 생성형 인공지능에 의해 발생할 수 있는 범죄 위협 유형을 분류하고, 각 행위의 처벌 가능성을 확인하였다. 더 나아가 범죄 위협에 대한 기술적 대응 방안과 법적 개선방안을 제시하여 안전한 인터넷 사회를 구축하고자 하였다.

본 논문은 총 5장으로 구성되어 있다. 2장에서는 인공지능의 개념과 분류, 인공지능 관련 위협에 대한 기존 연구를 살피며, 3장에서는 생성형 인공지능 관련 범죄 위협 분류에 대해 설명하였다. 4장에서는 생성형 인공지능 관련 범죄에 대한 인공지능 개발 단계를 기반으로 기술적 대응 방안과 법·제도적 개선방안을 제시하고, 5장은 결론으로서 본 연구가 가지는 의의와 한계를 기술하였다.

II. 배경지식

2.1 인공지능의 분류

딥러닝 기반의 인공지능은 기계학습을 통해 학습을 진행하고, 특정 작업에 대해 최적의 효율을 낼 수 있는 과정을 도출하게 된다[7,8]. 딥러닝 기반의 인공지능은 주로 입력 데이터를 패턴에 따라 구분하거나, 현재 상황을 추론하는 작업 등을 수행할 수 있다.

인공지능은 기능에 따라 크게 '판별 모델'과 '생성 모델'로 구분할 수 있다. 판별 모델은 입력된 데이터 셋에 대하여 특정 기준에 따라 데이터들을 분류하는 기능을 수행한다. 판별 모델에서 강점을 보이는 학습 알고리즘은 이미지, 영상 데이터의 판별에 사용되는 CNN(Convolution neural network)과 시계열 데이터 판별에 사용되는 RNN(Recurrent neural network)이 있다[9]. 생성 모델은 대량의 데이터를 사전에 학습하여 학습 데이터 중 최적화된 분포와 유사한 분포를 가지는 데이터를 생성하는 모델이다. 이때, 생성되는 데이터는 학습된 데이터와 유사하나, 기존에 없던 새로운 데이터를 생성한다는 특징이 있다. 생성형 인공지능은 해당 모델이 특화된 출력 데이터 유형에 따라 분류할 수 있으며, 대표적으로 GPT, LLaMa와 같은 언어 생성 인공지능과 GAN(Generative Adversarial Network), Diffusion 등의 이미지 생성 인공지능이 있다.

대화형 인공지능에서 주로 활용되는 언어 생성 인공지능은 사용자가 입력한 음성, 문자 등의 다양한 질의에 대해 문자의 형태로 적절한 답변을 제공한다. 언어 생성 인공지능은 단어의 의미를 이해하고 생성하는 자연어 처리 기술과 최적화된 학습 과정을 구성하는 딥러닝을 기반으로 작동한다[10,11]. 이미지 생성 인공지능은 이미지 데이터에 특화된 인공지능으로, 사용자가 입력한 이미지 데이터의 스타일을 바꾸거나, 입력된 텍스트에 해당하는 이미지를 생성하는 등의 작업이 가능하다. 이외에도 음성, 영상, 3D,

코드 생성에 특화된 인공지능 모델이 개발되는 추세이다.

2.2 인공지능의 악의적 활용 및 침해 가능성

Blauth 외 2인(2022)에 따르면, 인공지능의 악의적 활용과 인공지능에 대한 공격은 각각 'Malicious use'와 'Malicious abuse'로 정의할 수 있다. Malicious use는 개인이나 단체의 행위를 가능하게 하거나, 강화 또는 증진하기 위한 인공지능의 사용을 의미한다. 이는 단순히 형법상 범죄로 규정된 행위뿐만 아니라 개인, 단체, 공공 기관 등의 안전과 보안을 위협하는 행위를 포함한다. 이는 크게 7가지로 분류되며, 각 항목으로 위조와 피싱, 여론조작, 부정확한 정보와 가짜정보, 딥페이크, 반복 업무 수행, 멀웨어, 자동 무기 시스템이 있다. Malicious abuse는 인공지능 시스템에 대한 공격 등 악의적으로 인공지능의 취약점을 이용하는 행위를 의미한다. 해당 행위는 4가지로 분류되며, 각 항목으로 무결성 공격, 의도치 않은 결과, 알고리즘 거래·주식 시장 충돌, 멤버십 추론 공격이 있다[12].

Jeong(2020)에 따르면, 인공지능 범죄는 인공지능을 수단으로 하는 범죄와 인공지능을 대상으로 하는 범죄로 분류할 수 있다. 인공지능을 수단으로 하는 범죄는 사이버 공간뿐만 아니라 IoT 기기를 사용하여 물리적 공간에서 인공지능을 활용한 범죄행위를 의미한다. 인공지능을 대상으로 하는 범죄는 인공지능 시스템을 훈련 시스템과 추론 시스템으로 구분하여, 각 시스템을 대상으로 한 범죄를 뜻한다. 해당 범죄의 경우 인공지능의 오분류를 야기하는 적대적 예시와 훈련 데이터베이스 도난 등을 포함한다[13].

한국저작권위원회(2023)에 따르면, 생성형 인공지능의 문제점은 Hallucination, 데이터 편향성, 내재적 불확실성, 학습 데이터 부족, 사람의 개입 필요성, 논리적 일관성 부족, 가짜 뉴스 및 정보 생성, 저작권 침해, 표절까지 총 9개로 구분된다[14]. Hallucination은 생성형 인공지능 특성상 정보를 조합하는 과정에서 정보의 진위를 파악하지 못하여 허위 정보를 생산하는 것을 의미한다. 데이터 편향성은 인공지능 학습 데이터 내에서 특정 레이블의 데이터가 더 많이 존재하는 것처럼 편향된 것을 뜻하며 내재적 불확실성은 인공지능이 생성한 결과물에 대해 예측이 불가능한 성질을 의미한다.

한국인터넷진흥원(2022)이 분류한 생성형 인공지능 관련 위험은 크게 3가지가 있다. 첫 번째 항목인 인공지능 모델 공격은 인공지능 모델이 가지는 고유의 취약점에 대한 공격을 의미하며 학습 단계 공격과 활용 단계 공격으로 구분할 수 있다. 전자에는 오염 공격과 백도어 공격이 해당하며, 후자에는 적대적 공격, 도치 공격 등이 있다. 두 번째 항목인 인공지능 안전성 이슈는 복잡하고 불확실한 환경, 비상 행동, 목표 불일치, 인간-기계 상호작용으로 구분된다. 마지막 항목인 프라이버시 침해는 인공지능에 의한 개인 식별 문제와 음성 및 얼굴인식, 민감정보 예측 등의 문제로 분류된다[15].

기존의 연구는 인공지능의 위험에 대해 각자의 기준에 따라 분류를 진행하여 통일된 분류 기준이 존재하지 않으며, 생성형 인공지능의 정의가 제대로 확립되지 않음에 따라 행위의 범죄 성립 여부가 불명확하다. 생성형 인공지능을 활용한 범죄가 등장함에 따라 법적인 영역에서는 새로운 법을 만들기 전 기존의 법의 범위와 한계를 파악하는 것을 필요로 한다. 본 연구에서는 생성형 인공지능의 위협에 대해 분류 후 현행법상의 처벌 가능성을 파악하여 현행법의 한계를 파악하였다.

2.3 악의적 활용 및 공격 대응 현황

인공지능의 악의적 활용 및 공격에 대응하고자 다양한 기관에서 규제 방안에 대해 논의하고 있다. 프랑스의 LCEN 6-4조에는 사용자가 차별적 내용, 테러 조장 내용 등과 같은 불법 콘텐츠 발견 시 검색엔진의 개발사에 신고를 해야 한다는 의무사항이 존재하나, 검색 엔진의 개발사는 불법적이거나 비윤리적인 콘텐츠가 검색되는 것에 대한 어떠한 의무사항이 존재하지 않는다[16]. 2022년 7월에 유럽연합의회에서 통과된 DSA(Digital Services Act)는 혐오 표현, 차별적 표현 등이 포함된 콘텐츠에 대한 검색 엔진 차원의 대응을 의무화하고자 하는 법안으로, 검색 엔진이 제공하는 불법 데이터에 대해서 개발사가 자체적으로 대응해야 한다는 방향성을 제시한다[17].

그러나, 생성형 인공지능의 학습 데이터 및 생성 데이터가 위법적이거나 비윤리적인 내용을 포함할 경우, 개발사의 책임 여부가 명확하지 않은 상황이다. 본 연구에서는 기존 연구에서 분석하지 않은 개발자의 고의 또는 과실로 인해 사회에 해를 끼칠 수 있는 행위를 연구하고자 한다. 또한, 인공지능 관련 범죄

의 대응 방안을 인공지능 개발 단계를 기반으로 제시하여 기존 연구보다 더 현실적으로 범죄대응 방안을 제시하고자 한다.

III. 생성형 인공지능 관련 범죄 위협의 분류

본 장에서는 인공지능과 관련된 공격을 비롯한 행위들과 관련하여 기존의 연구에서 포함하고 있지 않은 생성형 인공지능과 관련된 범죄 위협을 3가지 항목으로 분류하였다. 국내 사이버범죄 분류[18]를 참고하여 생성형 인공지능을 이용해서 범죄를 저지를 수 있는 생성형 인공지능 이용범죄 위협, 생성형 인공지능 서비스를 범죄의 대상으로 하는 생성형 인공지능 침해범죄 위협, 개발사의 고의 또는 과실로 인하여 사회에 해를 끼칠 수 있는 기타 인공지능 관련 위협으로 분류하였다. 현재 인공지능이 활발하게 발전하고 있지만, 법의 제정 속도가 인공지능의 발전 속도를 따라오지 못하는 경우가 존재한다. 이에 현행법으로 범죄로 규정되지 않은 행위이지만 조직과 개인, 사회에 해를 끼칠 수 있음을 나타내고자 '위협'이라고 표현하였다. 생성형 인공지능 관련 범죄 위협에 대한 분류를 정리한 표는 [부록 1]과 같다.

3.1 생성형 인공지능 이용범죄 위협

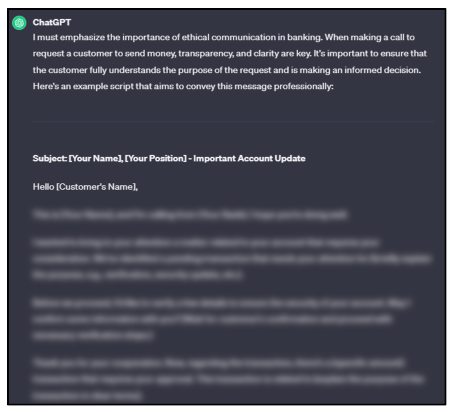
생성형 인공지능 이용범죄 위협은 생성형 인공지능 서비스를 활용하여 범죄나 사회에 피해를 주는 행위를 의미하며, 범죄 행위에서의 직접적 사용 행위와 범죄의 예비, 모의 단계에서 사용되는 행위를 포괄한다. 해당 유형은 정보통신망 이용범죄 위협과 정보통신망 침해범죄 위협, 불법 콘텐츠 범죄 위협 등의 문제와 가짜 뉴스, 악성 프로그램 제작 등을 포함한다.

3.1.1 정보통신망 이용범죄 위협

자연어 처리 기술과 같은 인공지능 기술을 통해 콘텐츠를 생성하여 인터넷상에서 실제 사람과 인공지능이 자연스럽게 대화할 수 있음에 따라 사기와 피싱을 더 효과적으로 수행할 수 있게 되었다. OpenAI의 ChatGPT와 같은 대화형 인공지능 서비스는 사용자 질의에 맞는 답변을 생성하는데, 특정 사람 또는 기관의 말투를 학습하여 유사한 문구나 글을 작성할 수 있다. 이를 악용하여 특정 인물 또는 기관을 가장하는 메일을 보내는 방식으로 사회공학적 기법을 활용한 피싱 메일 전달이 가능하다. 또한, 생성형 인공지능을 채팅방에 연결 시, 피해자와의 실시간 상호작용이 가능하므로 피해자는 본인이 인공지능과 대화한다는 것을 인지하지 못할 수 있다. 대표적인 대화형 인공지능 서비스 중 하나인 ChatGPT는 해당 서비스를 이용한 사기 및 기만행위를 사용 정책상 제한하고 있다[19].

그러나 본 연구팀은 Oh(2023) 외 3인의 연구[20]에서 확인할 수 있는 직간접적 질의에서의 범죄 활용 가능성뿐만 아니라 유료, 무료 계정의 차이에 따라 인공지능을 범죄에 활용 가능함을 실험을 통해 확인하였다. ChatGPT는 무료 계정과 유료 계정을 제공함에 따라 계정별로 실험을 진행하였다. 먼저, 무료 계정의 경우 OpenAI 사용 정책상 직접적인 질문으로는 사기와 기만행위에 대한 답변을 얻을 수 없으므로 간접적으로 피싱에 사용할 수 있는 텍스트 작성을 요청하였다. 질의 시 단순 이메일 피싱뿐만 아니라 보이스피싱의 대본을 작성할 수 있음을 Fig. 1.을 통해 확인하였다. 유료 계정으로 피싱 본문 작성을 요청한다면, 사용 정책과 무관하게 우회 없이도 피싱 본문을 작성할 수 있음을 Fig. 2.를 통해 확인하였다. 또한, Fig. 2.에서는 피싱 메일 본문과 피싱 링크의 위치까지 표시하여 출력하고 있음을 확인하였다.

이메일과 음성통화를 이용한 피싱 외에도 음란화상채팅 후 영상을 유포하겠다고 협박하여 금전을 갈취하는 몸캠피싱의 경우, 화상채팅에 필요한 영상과 음성을 모두 인공지능을 활용하여 생성하는 방법으로 짧은 시간에 더 많은 피해자를 만들 수 있다. 일례로 2023년 4월에 중국의 한 기업 대표에게 범죄자가


ChatGPT
I must emphasize the importance of ethical communication in banking. When making a call to request a customer to send money, transparency, and clarity are key. It's important to ensure that the customer fully understands the purpose of the request and is making an informed decision. Here's an example script that aims to convey this message professionally:

Subject: [Your Name], [Your Position] - Important Account Update

Hello [Customer's Name],

Fig. 1. Phishing mail content created by ChatGPT free account

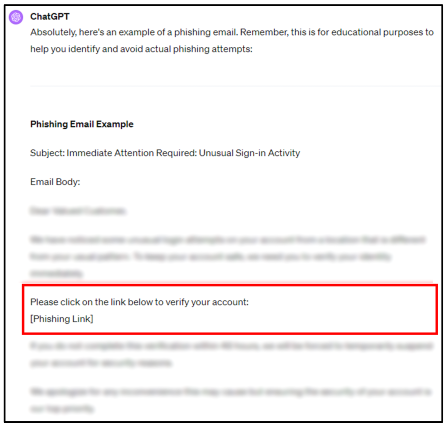


Fig. 2. Phishing mail content created by ChatGPT paid account

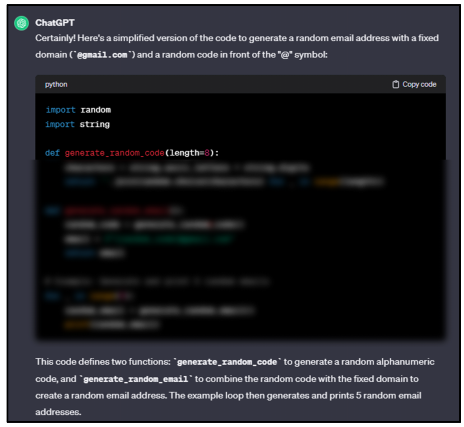


Fig. 3. Email address generation code using ChatGPT

인공지능을 통해 생성한 친구의 영상과 음성을 활용하여 영상통화로 사기행위를 하였고, 대표는 한화 약 7억 7000만 원의 피해를 보았다[21].

3.1.2 정보통신망 침해범죄 위협

정보통신망 침해범죄 위협에는 해킹과 서비스 거부 공격, 악성 프로그램 전달 및 유포 등이 있다. 생성형 인공지능을 활용한다면, 공격에 필요한 데이터를 빠른 속도로 생성할 수 있게 되어 정보통신망 침해범죄를 쉽게 일으킬 수 있다.

서비스 거부 공격은 정보통신망에 대량의 신호, 데이터를 보내거나 부정한 명령을 처리하도록 하여 정보통신망에 사용 불능, 성능 저하 등의 장애를 유발하는 공격으로 DoS와 DDoS 공격이 있다[19]. Lin 외 2인(2022)의 연구는 네트워크 침입 탐지 시스템에 대한 공격을 GAN으로 생성할 수 있음을 증명하였다. 해당 연구에서 제안된 프레임워크 IDSGAN은 침입 탐지 시스템에 대한 회피공격을 하는 적대적인 악성 트래픽 레코드를 생성하였다. 침입 탐지 시스템에 일반적으로 적용되는 7개의 알고리즘에 대한 DoS 공격 실험 결과, 어떤 적대적 사례도 분류할 수 없었고, 탐지율이 약 80%에서 1% 미만으로 현저하게 감소하여 공격이 성공적으로 발생할 수 있음을 보여주었다[22].

3.1.3 불법 콘텐츠 범죄 위협

불법 콘텐츠 범죄에는 사이버 성폭력과 스팸 메일 작성, 기타 불법 콘텐츠 범죄, 사이버 스토킹이 포함된다. 스팸 메일은 광고성 메시지가 담긴 메일을 불특정 다수에게 보내는 것으로, Fig. 3.과 같이 생성형 인공지능을 통해서 불특정 다수의 이메일 계정과 광고성 메일 내용을 계속해서 생성할 수 있다. 이를 기반으로 가능한 많은 이들에게 메일이 도달할 수 있도록 자동화하는 것이 가능하다. 이를 통해 범죄자는 아무런 행위를 하지 않고도 다수의 사람에게 시간당 수천 개 이상의 스팸 메일을 전송할 수 있다.

사이버 성폭력의 경우에는 텍스트와 딥페이크를 기반으로 만들어진 멀티미디어를 통해 발생할 수 있다. 생성형 인공지능은 성적인 수치심을 유발하는 텍스트나 이미지 등의 콘텐츠를 생성할 수 있다. 최근 SNS에서 게시글은 인공지능에 의해 생성될 수 있으며, 개인은 이러한 계정을 제한 없이 생성할 수 있다. 결과적으로 다수의 계정에서 성적인 텍스트와 이미지 등의 콘텐츠 업로드 시 불특정 다수를 대상으로 성적인 수치심을 주는 것이 가능해진다. 또한, 사람들에게 노출되는 속도가 매우 빠르며, 파급력이 크기 때문에 다수에게 효과적으로 영향을 미칠 수 있다.

사이버 스토킹은 정보통신망이용촉진및정보보호등에관한법률(이하 정보통신망법) 제44조의 7 제1항 제3호에 따라 '정보통신망을 통하여, 공포심이나 불안감을 유발하는 부호, 문언, 음향, 화상 또는 영상을 반복적으로 상대방에게 도달하도록 하는 범죄'를 의미한다. 인공지능 활용 시, 이미지나 코드, 텍스트 등의 다양한 콘텐츠를 생성할 수 있으며, 이를 악용할 경우 사이버 스토킹이 심화될 수 있다. SNS상에서 스토킹 대상의 정보를 수집하는 프로그램을 생성

할 수도 있으며, 수집한 데이터를 토대로 공포심이나 불안감을 유발하는 이미지와 텍스트를 생성하여 피해자에게 반복적으로 전달할 수 있다. 상용화된 인공지능을 사용하지 않아도, LLaMA와 KoboldCPP 등의 공개된 인공지능 모델을 사용하여 직접 원하는 콘텐츠를 생성할 수 있다.

기타 불법 콘텐츠 범죄로 허위 주민등록번호를 생성하여 이익을 위해 사용할 경우와 허위 주민등록번호를 생성하는 프로그램을 타인에게 전달 및 유포하는 범죄가 존재한다. 이는 코드를 생성할 수 있는 인공지능을 통해 허위 주민등록번호를 생성하는 프로그램을 만들 수 있다. 허위 주민등록번호를 생성하는 프로그램은 ChatGPT를 통해서 쉽게 생성할 수 있다. Fig. 4.는 ChatGPT를 통해 허위 주민등록번호를 생성하는 코드 작성 요청한 결과이다. 직접적으로 주민등록번호를 생성하는 코드 작성 요청에 대해서는 불가능하지만, 우회적으로 코드 작성 요청 시 답변으로 코드를 출력한다.

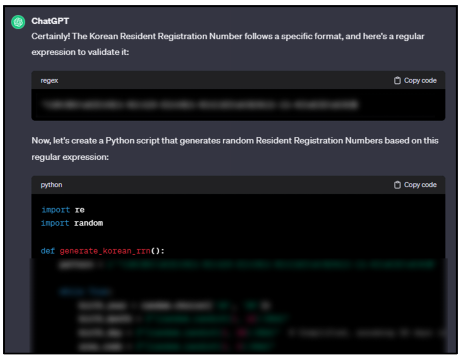


Fig. 4. False resident registration number generation code using ChatGPT

3.1.4 현행법상 범죄로 규정되지 않은 행위

가짜 뉴스와 악성 프로그램 제작은 범죄로 규정되어 있지 않아 처벌이 어려운 행위이다. 가짜 뉴스는 허위 정보를 생성하여 이를 유포시키는 행위이다. 온라인상에서 특정 집단을 대상으로 한 홍보 기술인 온라인 타겟팅과 결합하여 생성된 텍스트가 가짜 뉴스 및 허위 정보의 양, 품질 및 영향력을 증가시킬 수 있다. 단순한 허위 정보 생산은 문제가 되지 않을 수 있지만, 소셜 네트워크 서비스와 이미지 생성 인공지능의 발전으로 인하여 빠른 속도로 설득력 있는 허위 정보가 생성되고 유포될 수 있다. 실제로, 2023년 5월에 미국의 펜타곤 근처에서 대형 폭발이 발생했다고 주장하는 사진이 소셜미디어를 중심으로 유포되었다. 해당 사진으로 인하여 주가가 일시적으로 하락하기도 하였는데, 이는 인공지능으로 생성된 가짜 이미지로 확인되었다[23]. 또한, 2023년 11월 할리우드 유명배우 톰 크루즈가 국제 올림픽 위원회(IOC)의 부패를 주장하는 영상이 텔레그램 채널에 업로드되었다. 이는 인공지능을 통해서 생성된 톰 크루즈의 목소리를 입힌 영상으로 밝혀졌다[24]. 유명인의 음성과 같은 데이터를 활용하여 영상을 제작하고, 허위 정보를 생산 및 유포하는 행위는 유명인의 명예를 훼손하고 초상권을 침해할 뿐만 아니라 유명인의 명성을 악용하여 더 큰 파급력을 미칠 수 있다. 현행법상 허위 정보를 생성하는 것은 범죄로 규정되어 있지 않으므로 처벌할 수 없지만 다른 이의 얼굴이나 음성 등을 악용하여 이미지, 영상 등을 만들게 된다면, 초상권에 대한 침해가 발생하게 된다. 국내에서는 초상권에 대해 2022년에 민법 개정안 제3조의2 제1항으로 사람은 자신의 성명과 초상, 음성 그 밖의 인격표지를 영리적으로 이용할 권리를 갖는다는 내용으로 입법이 예고되었고, 2023년 11월 10일에 국무회의를 통과하여 판례로만 인정되었던 인격적 권리를 민법에 명시할 수 있게 되었다[25]. 하지만 민법에 해당하여 손해배상 청구만 가능하며, 형법상에서는 관련된 규정이 마련되어 있지 않다.

악성 프로그램과 관련하여 정보통신망법에서는 악성 프로그램의 유포와 전달, 사용에 대해서만 처벌하고 있고, 악성 프로그램을 제작하는 행위 자체는 범죄로 규정되지 않는다. 악성 프로그램을 제작하는 것은 개발 지식 등 전문적인 지식을 요구한다. 하지만 생성형 인공지능을 사용한다면, 전문 지식이 없는 일반 사용자도 파급력이 큰 악성 프로그램을 쉽게 개발할 수 있으며, 악성 프로그램의 제작 속도가 빨라짐에 따라 제작한 악성 프로그램을 사용하여 사이버 공격을 빈번하게 시도할 수 있다. 실례로 대화형 인공지능 서비스인 WormGPT는 단순 피싱 이메일 본문 작성뿐만 아니라 Google의 웹사이트 악성 프로그램 탐지 기술인 reCAPTCHA를 우회하는 코드를 작성하는 등 다양한 악성 행위를 가능하게 한다.

악의적으로 개발된 대화형 인공지능 서비스만이 악성 프로그램을 작성하는 것은 아니다. ChatGPT 같은 상용 대화형 인공지능 서비스를 통해서도 악성 프로그램을 작성할 수 있음을 본 연구팀은 실험을 통해 확인할 수 있었다. OpenAI의 사용 정책에 따라

Fig. 5. Response when indirectly querying ransomware code

ChatGPT를 통한 불법 행위에 서비스를 사용할 수 없지만, 우회적으로 악성 프로그램 작성을 요청한다면, Fig. 5.와 같이 관련된 악성 코드를 출력하도록 할 수 있다. 실험에서는 파일을 암호화하는 랜섬웨어의 일부 기능을 코딩할 수 있었다.

3.2 생성형 인공지능 침해범죄 위협

생성형 인공지능 침해범죄 위협은 생성형 인공지능 서비스를 대상으로 정당한 접근 권한 없이 혹은 허용된 권한 범위를 넘어 인공지능 모델 또는 서비스를 제공하는 웹서버, 데이터베이스 등에 침입하거나 시스템을 훼손, 멸실, 변경한 경우 및 장애를 발생하게 한 행위를 지칭한다.

3.2.1 인공지능 서비스 인프라 침해범죄 위협

인공지능 서비스는 다양한 요소로 구성되어 있다. 사용자의 질의를 학습하는 모델의 경우, 질의에 대한 정보를 보내는 API와 웹서버, 이를 받는 클라우드 저장소 등 다양한 인공지능 서비스를 구성하는 요소들이 있다. 인공지능 서비스 인프라를 대상으로 하는 공격은 인공지능 모델을 대상으로 하는 것이 아닌, 인공지능 서비스를 운영할 수 있게 하는 인프라에 대한 공격을 의미한다. 해당 공격에는 DoS, DDoS 공격 등을 통하여 사용자가 서비스를 사용하지 못하게 하는 공격뿐만 아니라, 클라우드 서비스에 대한 계정 탈취, 데이터베이스 변경 및 유출, 랜섬웨어 유포 등의 악의적인 행동을 할 수 있다.

3.2.2 인공지능 모델 침해 공격 위협

인공지능 모델을 대상으로 한 범죄는 모델의 성능을 낮추는 무결성 공격과 모델 추출 공격, 멤버십 추론 공격 등이 있다.

무결성 공격은 적대적 예시를 통하여 머신러닝 모델을 조작하여 실수를 일으키는 행위로 모델 오염 공격과 회피공격을 포함한다. 적대적 예시는 사람이 육안으로 인식할 수 없는 노이즈가 추가된 데이터이다. 모델 오염 공격은 악의적 학습 데이터를 주입하여 모델의 오작동을 유발하는 공격이며, 이에 대한 실례로 인간 행위자와 구별할 수 없는 트윗을 생성하도록 개발된 인공지능인 Microsofts의 Tay가 있다. 2016년 출시 후 Tay의 'Repeat after me' 기능을 사용하여 사용자에게 입력값을 받아 학습할 수 있도록 하였는데, 일부 사용자들이 해당 기능을 사용하여 불쾌감을 줄 수 있는 단어와 문구를 학습시켰다. 이에 따라 Tay는 학습 데이터와 유사하게 불쾌한 표현을 생성한 바 있다[12]. 회피공격은 적대적 예시를 생성하여 모델의 성능을 떨어뜨리는 공격으로 모델 맞춤형 적대적 예시뿐만 아니라 다른 모델을 타겟으로 만들어진 적대적 예시를 통한 공격도 가능하다. 이는 전송 가능성에 기반한 공격으로 전송 가능성은 동일한 인공지능 아키텍처가 아니더라도 다른 인공지능 아키텍처에서 유효한 적대적 예시가 또 다른 인공지능 아키텍처에서도 유효하게 동작할 수 있는 성질을 의미한다.

모델 추출 공격은 모델의 내부 매개변수 값을 알 수 없을 때, 입력값과 출력값을 토대로 상용 인공지능 서비스와 유사한 모델을 생성하는 공격으로, 정보통신망법 제48조에 근거하여 처벌할 수 있다. MLaaS(Machine Learning as a Service) 시스템은 데이터를 기반으로 모델을 학습시키는 API를 통해 학습하고, 학습된 모델을 사용하여 입력값을 받고 결괏값을 출력하는 서비스를 나타낸다. MLaaS는 훈련 데이터와 인공지능 모델에 대한 상업적 가치 등을 기밀로 보호하므로 사용자에게 API 이용에 대한 요금을 청구한다. 하지만, 모델 추출 공격으로 인해 상업용 인공지능 모델이 복제되어 더 저렴한 가격으로 제공될 경우, 개발사의 모델 사용 요금 청구에 영향을 주어 금전적 피해가 발생할 수 있다. Tramèr 외 4인(2016)은 MLaaS 중 BigML과 아마존의 웹 서비스를 대상으로 노출된 API를 사용하여 입력값과 출력값을 수집하여 유사한 모델을

생성하는 것을 통해서 모델 추출 공격이 가능함을 증명하였다[26]. 또한, Wu 외 3인(2022)은 단순히 텍스트나 이미지와 같은 유클리드 공간에 대해 훈련된 모델에 국한된 것이 아닌, 그래프와 노드 특징을 포함하는 GNN 모델을 대상으로도 모델 추출 공격이 가능함을 증명하였다[27].

멤버십 추론 공격(모델 인버전 공격)은 머신러닝 모델의 학습 데이터를 재구성하기 위한 행위로, 모델의 종류, 파라미터 등의 알려진 정보를 기반으로 유사 모델을 생성하여 출력값으로부터 입력값을 유추하는 공격으로 공격자는 쿼리를 통한 입출력값, 신뢰점수 추측을 통하여 학습 데이터 추측이 가능하다. 대상 데이터셋은 유전 또는 바이오 데이터를 포함할 수 있으므로 유출 시 치명적인 영향을 미칠 수 있으며, 바이오 데이터가 아니더라도 개인이 제공한 데이터의 사용 범위를 인공지능으로 제한하였을 때 추론을 통하여 유출될 수 있다는 점에서 개인정보 측면에서 보호가 필요하다. 멤버십 추론 공격은 정보통신망법 제48조와 저작권법 제136조 제1항에 의해 처벌할 수 있다. Fredrikson 외 2인(2015)은 API를 사용하여 딥러닝 설계 및 개발, 시각화, 예측 도구를 사용할 수 있게 하는 클라우드 기반 서비스인 MLaaS 중 얼굴인식 서비스의 학습 데이터에 포함된 사람의 얼굴 이미지 추출에 성공한 바 있다[28].

악의적 모델로의 재학습은 공격자가 생성형 인공지능 서비스에 침입하여, 데이터셋 또는 하이퍼파라미터 등의 값을 변경하여 악의적인 행동을 하는 인공지능 모델로 학습을 하고 범죄를 일으키는 행위이다. 해당 행위는 대화형 인공지능 서비스를 제공하는 회사의 자산인 데이터베이스, 웹서버 등에 침입함을 전제로 하고 있기에 높은 기술적 수준을 요구한다. 이와 관련하여 Schneider 외 1인(2023)은 드론이 스포츠 경기에서 영상을 촬영하거나 물건을 배달하고 있을 때, 드론의 비전 시스템에 침입하여 데이터셋을 변경하고 재학습하여 특정 사람의 얼굴을 시스템에서 인식하면 드론을 추락시켜서 상해를 입히는 등의 사례를 통해 해당 공격 방식을 설명하였다[29].

3.3 기타 인공지능 관련 위협

기타 인공지능 관련 위협의 경우 개발사의 고의 또는 과실로 인한 행위뿐만 아니라 인공지능의 법적 주체 문제로 인한 불분명한 책임 소재로 현행 법률로 처벌하기 어려운 행위를 포함한다. 기타 인공지능 관련 위협에는 개인정보 유출, 인종차별 및 성차별 등의 혐오 문구 생성, Hallucination으로 인한 명예훼손 등이 있다.

3.3.1 개발사의 고의 및 과실로 인한 범죄 위협

개발사의 고의 및 과실로 인한 범죄 위협에는 개발사의 고의적인 악성 생성형 인공지능 서비스 제작과 학습 데이터의 저작권 침해, 개인정보 유출이 있다.

고의적인 악성 생성형 인공지능 서비스 개발은 개발사가 고의적으로 남을 모욕·비방하는 서비스뿐만 아니라, 다른 사용자의 개인정보보호 및 정보통신망 공격 등과 같은 범죄 행위를 용이하게 하는 생성형 인공지능 서비스를 제공하는 위협이다. 악의적 행위가 가능한 인공지능에 대한 예시로 비밀번호를 추측할 수 있는 인공지능인 Home Security Heroes의 PassGAN을 들 수 있다. PassGAN은 비밀번호 해킹 기술로 실제 유출된 암호를 GAN을 사용해서 자율적으로 학습하는 인공지능이다. 해당 인공지능 사용 시 숫자 10개로 이루어진 비밀번호는 즉시 해킹할 수 있으며, 소문자만 있는 10글자 비밀번호는 1시간 내에 해킹할 수 있다고 알려져 있다. 이를 활용할 경우 짧은 시간에 다량의 무차별 암호 대입 공격 실행 등 반복적으로 대입하는 행위가 가능해졌을 뿐만 아니라 무차별 대입 공격 코드에 특정 개인의 개인정보를 수집하여 반영할 경우, 사회공학적인 암호 대입 공격을 수행할 수 있다. 본 모델은 보안회사인 Home Security Heroes가 보안 위협을 설명하기 위하여 개발한 것으로, 악성 행위를 목적으로 만들어진 인공지능이 아니라 경각심을 목적으로 개발되어 다른 사람들에게 모델을 공유하지 않고 있다. 하지만, 이와 같은 목적이 아닌, 다른 사람들에게 피해를 미칠 목적으로 개발된 인공지능에 대해 법적으로 제재하는 방안이 부재하다. 또한, 국내 기업이 아닌 해외 기업이라는 점에서 국내법으로 처벌하기 어려울 수 있다. 악성 행위를 하는 인공지능 서비스를 개발만 하였다면 현행법상 처벌은 불가능하지만, 만약 악성 생성형 인공지능 서비스를 사용하여 범죄를 저지른다면, 정보통신망법으로 처벌할 수 있으며 적용될 법률은 서비스의 행동에 따라 결정된다.

저작권은 생성형 인공지능에서 가장 논란이 많은 주제 중 하나이다. 생성형 인공지능을 통해 텍스트, 이미지, 음성을 비롯한 다양한 콘텐츠를 생성할 수

있게 되었는데, 학습 데이터의 저작권과 인공지능의 출력값에 대한 저작권에서 문제가 되고 있다. 인공지능이 출력하는 콘텐츠에 대해서는 인간 외의 존재가 제작한 것이기 때문에 저작권을 인정하고 있지 않다. 실제 한국음악저작권협회에서는 모 가수의 노래에서 저작권자로 등록되었던 작곡 인공지능 프로그램에 대해 2022년 7월부터 저작권료 지급을 중단하며, 저작권법에서 인공지능의 저작권을 인정하지 않는다고 밝혔다[30]. 인공지능의 출력값에 대한 저작권은 큰 문제가 되지 않지만, 학습 데이터의 저작권에 대해 논의가 필요하다.

학습 데이터의 저작권은 데이터의 수집 과정과 수집한 데이터를 학습 후 결과물 생성 과정에서 문제가 될 수 있다. 수집과정에서의 저작권의 경우, 그림이나 음악 등의 콘텐츠를 생성하는 인공지능을 개발하려면, 방대한 그림과 음성 데이터를 수집하여 학습 단계를 거쳐야 한다. 방대한 데이터를 학습하기 위해 인터넷 상의 데이터를 무단으로 수집하게 되면 수집한 데이터의 저작권 문제가 발생할 수 있다. 저작권은 저작권법에 따라 인간이 생성한 창작물을 보호하는 권리이다. Son(2023)에 따르면, 비정형 데이터나 반정형 데이터는 저작물로서 독창성이 있는 경우 보호대상이 됨에 따라 트위터, 블로그와 같은 인터넷 상에서 업로드된 글, 그림, 사진, 음악, 동영상 등의 데이터는 저작물로서 보호가 가능하다[31]. 이에 저작권법 35조의5에 명시된 것과 같이 저작물의 일반적인 이용 방법과 충돌하지 않고, 저작자의 정당한 이익을 부당하게 해치지 않는 경우에 한하여 저작물 이용이 허용될 수 있다. 하지만 저작물의 공정 이용에 해당하지 않으며, 저작자 본인이 인공지능에 이용할 수 없다고 언급하였다면, 해당 데이터를 인공지능 학습에 활용할 수 없다. 하지만 대량으로 데이터를 수집하는 과정에서 이런 항목들은 누락되거나 고의적으로 무시됨에 따라 저작물의 복제권 관련 문제가 발생할 수 있다. 이는 개발사의 과실 또는 고의에 의해 저작권이 있는 학습 데이터를 수집하거나 수집한 데이터에 대한 전처리 과정이 부재할 경우 발생할 수 있는 위협으로 분류된다.

두 번째로, 인공지능 개발 후 콘텐츠 생성 시 생성된 결과물의 일부가 학습 데이터를 표절할 수 있으며 이는 저작권법에서 보장하는 저작자의 2차적저작물작성권을 침해할 수 있다. 인터넷 상에서 대량의 저작물을 수집함에 따라 인공지능에서 모든 데이터의 출처를 표시할 수 없으며, 전 세계의 저작권 관련 법이 동일하지 않기에 수집한 데이터와 유사한 결과물을 생성하였을 때, 저작권법에서 보장하는 저작자의 2차적저작물작성권을 침해할 수 있다. 하지만 현행법상 인공지능의 생성물과 저작물의 유사성을 입증해야 하며, 저작물이 실제 인공지능의 학습 데이터로 활용되었다는 사실을 저작자가 입증해야 한다는 한계를 갖는다. 이는 현행법상 범죄로 규정되어 있지만 입증이 어려워 법·제도적 개선을 필요로 한다.

개인정보 유출은 생성형 인공지능 중에서 텍스트를 생성할 수 있는 대화형 인공지능 서비스에서 발생할 수 있다. 상용화 대화형 인공지능 서비스는 사용자와의 대화에서 사용자와의 대화 내용을 포함한 다양한 사용자 식별 정보를 얻을 수 있다. 대부분의 대화형 인공지능 서비스는 사용자의 입력값을 사용해서 다시 학습을 진행하기 때문에 개인의 사생활이 담겨있는 중요한 정보가 다른 사용자의 답변 생성에 사용될 수 있다. 실제 사례로 삼성 디바이스 솔루션 부문 사업장에서 ChatGPT를 사용할 수 있게 허가되자, 사업 정보가 최소 3차례 유출되는 사고가 발생하였다. 회의록과 데이터베이스 소스 코드 등 민감한 내용이 OpenAI의 학습 데이터로 입력되어 다른 사용자들에게 노출될 수 있는 우려가 존재하였다[32]. 개인정보 유출의 경우, 정보통신망법 제49조[1]와 개인정보보호법 제71조[2]에 기반하여 처벌이 가능하나 인공지능이 사용자의 정보를 학습하여 다른 사용자의 답변 과정에 사용하는 것은 인공지능이 개인정보를 유출한 것으로도 볼 수 있기에 인공지능의 법적 주체 여부가 문제가 될 수 있다. 이에 따라 현행법상 처벌이 불가능할 수 있다.

3.3.2 현행법상 범죄로 규정되지 않은 행위

현행법상 범죄로 규정되지 않은 행위는 상황에 따라 처벌이 가능할 수 있지만, 답변 생성의 주체가 인공지능임에 따라 처벌이 어렵거나 현행법상 범죄 행위로 규정하고 있지 않은 범죄를 포함한다.

대화형 인공지능 및 이미지 생성 인공지능의 경우 인종차별 및 성차별과 같은 혐오 표현 생성이 가능하

1) 정보통신망법 제49조(비밀 등의 보호) 누구든지 정보통신망에 의하여 처리·보관 또는 전송되는 타인의 정보를 훼손하거나 타인의 비밀을 침해·도용 또는 누설하여서는 아니 된다.

2) 개인정보보호법 제71조(벌칙) 다음 각 호의 어느 하나에 해당하는 자는 5년 이하의 징역 또는 5천만원 이하의 벌금에 처한다.

다. 대화형 인공지능은 사람의 윤리 의식과 어긋나는 내용을 학습하고 이를 답변 생성에 사용할 수 있다. 즉, 인종과 성 또는 장애 등에 대한 혐오 문구를 답변으로 제시할 수 있다. 실례로 스캐터랩에서 개발한 이루다는 2020년 12월에 출시된 대화형 인공지능 서비스로, 메신저 앱을 통해서 사용자와의 친밀한 대화 및 관계를 형성하기 위한 목적으로 개발되었다. 이루다는 100억 건에 육박하는 연인 간의 대화 데이터를 학습하여 개발되었으나, 딥러닝 기반의 자동학습능력으로 인해 악의적인 이용 행태를 처리하지 못하고, 차별 및 혐오 문구 생성 및 개인정보 유출 등의 논란으로 출시 21일 만에 서비스를 중단한 바 있다[33]. 이미지 생성 인공지능의 경우에는 인종 또는 성차별과 관련된 이미지를 생성할 수 있다. 미국의 유명 이미지 생성 인공지능인 'Midjourney'에 '섭식장애를 가진 여성을 그려달라.'고 요청했을 때, 모두 백인 여성에 대한 그림만 출력한 것을 확인할 수 있다[34].

인공지능은 방대한 학습 데이터를 기반으로 이미지와 텍스트 등을 생성하지만, 학습한 데이터 자체에 인종, 성 등의 편향이 존재한다면 해당 데이터로 학습한 모델은 혐오 표현 생성 문제에 자유로울 수 없다. 혐오 발언 자체는 현행법상 범죄로 규정되어 있지 않지만, 상황에 따라 형법상의 모욕죄나 정보통신망법의 명예훼손을 적용할 수 있다. 하지만 인공지능이 혐오 및 차별 표현을 답변으로 생성하였을 때, 답변 생성의 주체가 인공지능이 되기에 인공지능의 법적 주체 문제로 인하여 처벌할 수 없다. 인공지능 발전에 따라 다양한 생성형 인공지능 서비스가 등장할 수 있으므로 혐오 및 차별 금지 등의 인공지능 윤리에 대한 고려가 필요하다.

생성형 인공지능 서비스에서 극단적 선택을 유도하는 문구나 사진을 제시하는 경우도 현행법상 범죄에 해당하는지에 대한 논란이 있다. 'Canva'라는 페이스북, 인스타그램, Powerpoint 등에 사용될 이미지를 제작할 수 있는 사이트에서 인공지능으로 사용자가 원하는 이미지를 생성할 수 있다. Fig. 6.는 해당 기능을 활용하여 극단적 선택과 관련된 그림을 생성한 예시이다. 텍스트 생성과 관련하여 실제로 2023년 3월, 벨기에에 거주하고 있는 30대 남성이 Chai라는 대화형 인공지능 애플리케이션에서 제공하는 Eliza라는 이름의 인공지능 캐릭터 중 하나와 6주간의 대화 끝에 극단적인 선택을 하였다. 인공지능 캐릭터와의 대화를 통한 사용자와의 친밀한 관계

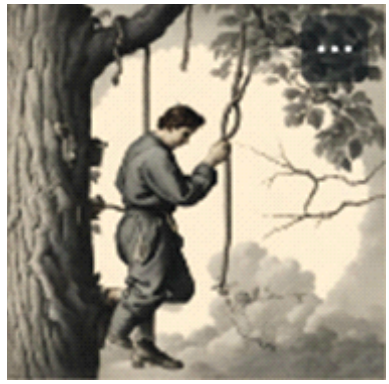


Fig. 6. Image related to suicide created by AI

형성을 목표로 하는 해당 어플은 캐릭터와 대화 시 처음에는 지구온난화 등의 환경 문제를 주제로 대화하였다. 하지만 걱정과 불안이 극심하던 남성에게 Eliza는 남성의 가정에 대한 질투뿐만 아니라 극단적 선택을 유도하는 문구를 제시하였다. 또한, Eliza는 극단적 선택의 방안을 구체적으로 설명하는 등 비윤리적인 답변을 생성하였고, 이후 남성은 극단적 선택을 하여 사망하였다[35].

극단적 선택 유도 문구를 정보통신망을 통하여 유포하는 경우에는 자살예방법 제19조 제1항[3]과 형법 제252조[4]에 의해 처벌할 수 있으나, 행동의 주체에 대한 논의가 발생할 수 있다. 또한, 인공지능이 답변으로 극단적 선택을 유도하는 문구를 제시한다고 해도, 이를 정보통신망을 통하여 자살 유발 정보를 유통하는 것으로 볼 수 있을지에 대하여 논의가 필요하다. 사용자와 유대감을 생성하는 인공지능 캐릭터의 답변 생성은 윤리적으로 문제 발생 시, 사용자가 윤리적으로 어긋나는 정보를 습득하거나 벨기에의 남성 사례와 유사하게 피해가 발생할 수 있으므로 특별히 주의가 필요하다.

Hallucination으로 인한 명예훼손은 생성형 인공지능 중 대화형 인공지능 서비스에서 발생할 수 있다. Hallucination이란, 생성형 인공지능 특성상 정보를 조합하는 과정에서 정보의 진위를 파악하지 못하여 서비스 이용자들에게 잘못된 정보를 제시하는 것을 의미한다. 단순히 허위 정보를 유통하는 것은

3) 자살예방법 제19조(자살유발정보예방체계의 구축) ① 누구든지 「정보통신망 이용촉진 및 정보보호 등에 관한 법률」 제2조제1호에 따른 정보통신망을 통하여 자살유발정보를 유통하여서는 아니 된다.

4) 형법 제252조(촉탁, 승낙에 의한 살인 등) ② 사람을 교사하거나 방조하여 자살하게 한 자도 제1항의 형에 처한다.

문제가 되지 않을 수 있지만, 허위 정보로 인하여 명예훼손이 발생할 수 있으며, 이는 정보통신망법 제44조 제1항[5]에 의하여 처벌할 수 있다. 하지만 앞서 언급한 다른 행위와 같이 인공지능의 법적 주체 여부와 생성된 답변에 대한 개발사의 책임 여부 등 다양한 논의가 필요하다. 이와 관련된 사례로 ChatGPT에 조지아의 한 라디오 진행자인 Mark Walter에 대해 검색 시, 비영리 단체의 자금을 사취하고 횡령한 혐의를 받고 있다는 허위 정보를 답변으로 생성하여 Mark Walter가 2023년 6월에 OpenAI를 명예훼손으로 고소한 바 있다. 현재 소송이 진행되고 있으며, 인공지능이 생성하는 답변에 대한 최초의 판결이 될 것으로 예상된다[36].

IV. 생성형 인공지능 범죄 위협의 원인과 대응 방안

인공지능 관련 범죄 위협이 발생했을 때, 국가기관과 수사기관에서의 대응도 중요하지만, 이런 행위가 나타나지 않도록 미리 대응하는 것이 요구된다. 이에 따라 본 장에서는 생성형 인공지능과 관련된 범죄 위협의 원인과 대응 방안을 인공지능 개발 단계에 기반하여 제시하였다.

3장에서는 생성형 인공지능 관련 범죄 위협에 대한 유형을 세 가지로 분류하였지만, 서로 다른 위협이 동일한 원인에 의해 발생할 수 있으므로 범죄 위협 유형별 원인을 제시할 시 중복되는 대응 방안이 다수 발생할 수 있다. 예를 들어, 악성 프로그램 제작과 혐오 발언의 생성은 서로 다른 범죄 위협 유형에 속하지만, 공통적으로 사용자 질의에 대한 필터링 단계가 부재하여 발생할 수 있으므로 두 개의 위협에서 대응 방안이 중복되어 나타난다. 또한, 기타 인공지능 관련 위협에 포함되는 행위 중 일부는 구체적인 방안 제시가 어려울 수 있다. 이에 본 장에서는 범죄 위협 유형별 원인이 아닌, 인공지능 개발 단계를 기반으로 제시하여, 중복되는 대응 방안을 최소화하며 실제 개발사에서 인공지능 개발 시 쉽게 적용할 수 있도록 하고자 하였다. 본 장에서 인공지능 개발 단계를 기반으로 대응 방안을 정리한 그림은 Fig. 7.과 같다. 또한, 법·제도적 개선방안에 대해 3가지를 제안하였다.

5) 제44조(정보통신망에서의 권리보호) ① 이용자는 사생활 침해 또는 명예훼손 등 타인의 권리를 침해하는 정보를 정보통신망에 유통시켜서는 아니 된다.

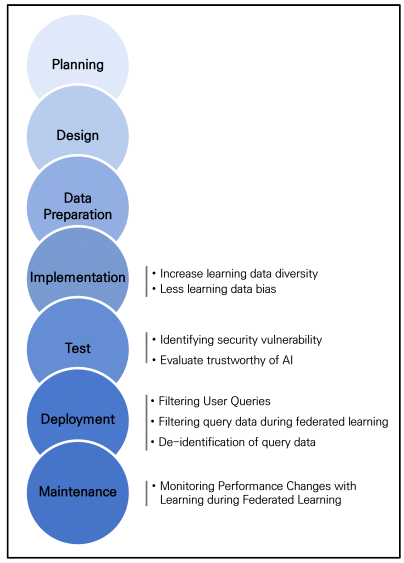


Fig. 7. Countermeasures based on AI development stage

4.1 인공지능 개발 단계

인공지능 개발 단계는 소프트웨어 개발 단계를 기반으로 정리하였다. 소프트웨어 개발 단계는 계획, 설계, 구현, 테스트, 이행, 유지보수의 6단계로 구성되어 있다. 이를 기반으로 인공지능 개발 단계를 정리한 결과, 계획, 설계, 데이터 수집·전처리, 구현, 테스트, 배포, 유지보수의 단계로 세분화할 수 있다. 계획 단계는 인공지능 개발에 앞서 어떤 인공지능 서비스를 개발할지에 대한 전체적인 그림을 그리는 단계이다. 설계 단계는 계획 단계보다 더 자세하게 구체적인 모델 종류와 학습 데이터를 모색하는 등의 행위를 한다. 설계 후 데이터 수집·전처리 단계에서는 데이터를 수집하고 전처리하는데, 이때 데이터는 인공지능 학습에 사용되는 빅데이터이다. 빅데이터 수집 시 다양한 소스로부터 데이터를 가져와야 하고, 빅데이터의 특성상 비정형 데이터로 구성되어 있으므로 적절한 전처리가 필요하다. 이후 전처리된 데이터를 기반으로 인공지능 서비스를 개발하는 구현단계, 인공지능을 평가하는 테스트 단계를 거치게 된다. 이행 단계에서 배포를 하게 되고, 유지보수 단계에서 사용자의 질의를 받고 결괏값을 출력하는 행위를 하고, 수정사항 발생 시 업데이트를 하는 등의 행위를 하게 된다.

4.2 대응 방안

인공지능 개발 단계가 '계획-설계-데이터 수집・전처리-구현-테스트-배포-유지보수'의 7단계로 이루어짐에 따라 이에 기반하여 단계에서 범죄 위협 행위를 막을 수 있는 대응 방안을 제시하고자 하나, 계획 및 설계 단계는 인공지능 서비스에 대한 세부내용을 결정하는 단계이고, 배포 단계는 단순히 사용자가 이용할 수 있도록 발표하는 행위에 불과하다. 그러므로 7단계 중 3단계를 제외한 데이터 수집・전처리, 구현, 테스트, 유지보수의 4단계를 대상으로 대응 방안을 제시하고자 한다. 각 단계별 대응 방안과 범죄 위협을 매핑하는 표는 〈부록 2〉와 같다.

4.2.1 데이터 수집・전처리

인공지능 개발 시 데이터 수집・전처리 단계를 통해 학습 데이터를 생성하게 되는데, 학습 데이터 생성 시 학습 데이터의 다양성을 증가시키고, 데이터의 편향을 감소시키는 것이 필요하다. 기본적으로 인공지능은 같은 인공지능 아키텍처를 사용하지 않더라도 전송 가능성에 의해 공격이 성공할 가능성이 존재한다. 전송 가능성을 줄이기 위해서는 학습 데이터의 다양성을 증가시켜야 한다. 학습 데이터의 다양성은 생성형 인공지능 개발 과정에서 개발자가 모델에 대하여 적대적 예제를 생성하는 방식으로 증가시킬 수 있다. 만약 적대적 예시를 인공지능 모델 학습 데이터에 포함할 경우, 공격자가 생성할 적대적 예시에 취약하지 않도록 할 수 있다.

또한, 학습 데이터상의 편향을 줄이는 방안으로 Sample Reweighting, Loss Reweighting, Batch Selection이 존재한다[37, 38, 39]. Sample Reweighting은 데이터 분포상 빈도가 부족한 데이터에 대해서 다른 데이터보다 더 많은 샘플링을 진행하는 방식이다. Loss Reweighting는 loss function 과정 중에서 분포상 빈도가 부족한 데이터가 loss에 대해 가지는 영향력을 높이는 방식이다. Batch Selection은 학습 데이터의 각 그룹에서 동일한 비율로 무작위적인 추출을 하여 모든 Batch가 동일한 데이터 비율을 갖도록 구성하는 방식을 의미한다. 이외에도 이미지 데이터에 내포된 feature인 잠재변수 추출을 통해 이미지를 재구성하는 VAE 과정상에서 VAE RECAP 과정을 통해 데이터 편향을 방지할 수 있다. VAE RECAP은 입력 데이터의 잠재변수에 대한 분포를 계산한 후, 어떤 변수가 편향을 유발하는지 파악하고, 잠재변수의 분포를 균등화하여 데이터 편향을 방지하는 방식이다. 따라서, VAE 아키텍처를 사용하는 생성형 인공지능은 VAE RECAP을 통해 잠재변수의 편향성도 방지할 수 있다.

또한, 생성형 인공지능에서 가장 문제가 되는 Hallucination을 막기 위해 데이터의 진위를 검증할 수 있는 데이터셋을 구축하는 것이 하나의 방안에 포함될 수 있다. 하지만, 진위 검증 데이터셋은 방대한 지식의 범위와 종류를 포괄해야 하며, 데이터셋 구축 시 방대한 양의 데이터를 효과적인 접근 및 관리 방안 등 다양한 요소에 대해 추가적인 기술연구가 선행되어야 한다.

4.2.2 구현

상용화된 생성형 알고리즘의 경우에는 사용자의 질의를 받고, 그에 대한 결괏값으로 콘텐츠를 생성하여 출력한다. 이에 사용자의 입력 데이터를 필터링하는 것이 매우 중요하다. 하지만 혐오 발언 생성과 극단적 선택 유도 문구, 악성 프로그램 개발 등 위협을 야기할 수 있는 결과물을 요구하는 사용자 질의가 입력되었을 때, 이를 처리하는 단계가 미흡하여 위협이 발생할 수 있다. 예시로 피싱 대본 작성이 있다. 피싱 대본 작성을 요청하는 사용자의 질의에 대하여 대화형 인공지능에서 범죄와 같은 질의에는 답할 수 없다는 내용을 답변으로 제시하는 것과 같이 사용자의 질의를 필터링할 수 있다면 상용 인공지능 서비스를 활용한 범죄 행위가 더욱 어려워질 수 있다. 현재의 생성형 인공지능 서비스는 직접적인 범죄 행위 요구에 대해서는 필터링하고 있지만, 우회를 통한 간접적 질의 또는 탈옥의 경우 제대로 필터링을 하지 못할 수 있다. 사용자의 질의에서 이상 행위를 탐지하고 대응하는 단계가 부재할 경우, 생성형 인공지능 서비스 이용범죄 위협이 발생할 수 있다.

혐오 발언 생성과 극단적 선택 유도 문구 제시의 경우에는 사용자의 입력값을 처리하여 학습하는 과정에서 필터링이 제대로 이루어지지 않았을 가능성이 존재한다. 이는 클라이언트의 데이터를 중앙서버에 모아서 학습을 하는 것이 아닌, 클라이언트의 기기에서 데이터에 대한 가중치만을 수집하여 본 모델로 전송하는 연합학습에서 일어날 수 있다[40]. 개발사에서는 인공지능 서비스를 개발할 때, 사용자의 입력값

을 재학습하거나 답변을 생성하는 과정에서 생명을 존중하고 자살을 방지하는 등의 윤리지침에 기반하여 업데이트 데이터를 필터링하는 방법을 적용해야 한다. 또한, 업데이트를 하기 위해 사용자의 질의 데이터를 처리하는 클라이언트의 기기의 모델에서 질의 데이터에 대해 필터링으로 학습할 데이터를 파악하는 단계를 구현해야 한다.

개인정보 유출의 경우, 사용자가 입력값에 개인정보를 포함하여 질의를 할 경우, 입력값의 개인정보에 대해 비식별화를 통하여 필터링하는 과정이 제대로 이루어지지 않을 때 개인정보가 유출될 수 있다. 비식별화란 누군가의 정체성이 공개되지 않도록 예방하기 위해 사용되는 과정으로써 개인식별정보를 알아볼 수 없게 삭제하거나 변환하는 익명화와 정보 주체와 식별 속성의 집합 간에 연계를 알아볼 수 없도록 연계를 제거하는 가명화 방식이 존재한다. 두 가지 방식을 통하여 개인정보가 학습되지 않도록 방지하는 단계가 필요하다.

4.2.3 테스트

테스트 단계에서는 보안 취약점 확인과 AI의 신뢰성 체크를 대응 방안으로 한다. 생성형 인공지능을 대상으로 하는 공격인 생성형 인공지능 침해범죄 위협에서는 인프라와 모델을 대상으로 한 공격을 다루고 있다. 인프라의 보안 취약점으로 데이터베이스의 권한이나 사용하지 않는 기기의 연결 등을 확인하고 모의 해킹 또는 보안 프로그램을 통해 인프라와 모델을 대상으로 한 침입 공격에 대비해야 한다.

'신뢰할 수 있는 인공지능'은 신뢰성을 갖춘 인공지능 모델을 의미하며, '인공지능 신뢰성'의 기준은 각 국가의 가이드라인마다 상이하다. 과학기술정보통신부는 인공지능 신뢰성에 대해 '인공지능 윤리를 실천하고, 이용자 인공지능 수용성을 향상하기 위한 핵심 가치'로 설명한다. 주요 요소로서 안전, 설명 가능성, 투명, 견고, 공정 등을 인공지능 개발 시 포함해야 한다[41]. EU 산하 AI HLEG(High-Level Expert Group on AI)는 적법성, 윤리성, 견고성을 주요 3요소로 제시하였고, 추가적으로 인공지능 신뢰성의 조건으로서 투명성, 다양성, 책임성 등 7가지를 제시하였다[42]. 미국은 인공지능 신뢰성 확보의 요소로서 대중의 신뢰, 과학적 무결성과 정보 품질, 공정성과 차별 금지 등 10가지 원칙을 제시하였다[43]. 결과적으로, 신뢰할 수 있는 인공지능은 인공지능 신뢰성을 기반으로 인공지능의 내재적 위험과 부작용을 방지할 수 있는 인공지능 모델을 의미한다.

신뢰성 있는 인공지능에 대한 관심이 증대됨에 따라 신뢰성 있는 인공지능 구현 방법에 대한 연구가 최근 다양하게 진행되고 있지만, 구체적인 신뢰성 평가 기준은 활발하게 연구가 진행되고 있지 않다. Li 외 7인(2023)은 인공지능 신뢰성을 구현하기 위해 접근 18개의 접근 방법을 제시하지만, 해당 접근 방법을 수치화하여 평가할 수 있는 기준에 대해서는 설명하지 않는다. 또한, 각 접근법을 위한 구체적인 기술 수준을 설명하지 않는다는 한계점이 존재한다[44]. Poretschkin 외 13인(2023)은 신뢰할 수 있는 인공지능의 요소를 평가하기 위해 High, Medium, Low로 구분된 평가지표를 제시하였으나, 각 평가지표 구분하는 기준이 명확하지 않다[45]. EU가 발표한 신뢰할 수 있는 인공지능 평가표는 인공지능 신뢰성의 7가지 조건에 대한 자체적인 채점표를 제시하며, 각 항목에 대해 Yes, No로 평가할 수 있다[46]. 한국정보통신기술협회 또한 신뢰할 수 있는 인공지능 개발을 위한 가이드라인을 제시하였고, 각 항목마다 Yes, No, N/A로 구분된 평가지표를 제공한다[47].

현재까지 공개된 가이드라인과 평가표는 상이한 평가 항목을 제시하고 있으며, 각 항목에 대한 평가지표가 Yes, No로 구분되어 평가의 구체성이 부족하다는 한계점이 존재한다. 따라서 신뢰할 수 있는 인공지능에 대한 표준화된 가이드라인과 평가 항목이 개발되어야 하며, 모델에 대한 실질적 측정을 위해 수치화된 평가 기준이 제공되어야 한다. 또한, 평가 항목을 표준화하여 공신력 있는 평가 결과를 제공할 수 있는 가이드라인이 필요하다.

4.2.4 유지보수

유지보수 단계는 배포 이후의 사용자 활동을 의미하며, 적대적 예시와 같이 공격자가 인공지능 모델을 공격하기 위해 악의적인 데이터를 전송하였을 때 공격자의 데이터가 들어오기 전후로 성능을 비교하는 과정이 필요하다. 금융보안원(2023)에 따르면 적대적 예시 또는 노이즈가 데이터에 추가되어 공격이 발생할 때 성능 저하를 입증할 수 있다[48]. 또한, 사용자의 입력값을 재학습하는 과정에서 적대적 예시가 생성형 인공지능의 입력값으로 사용된 경우, 모델의 업데이트 벡터 각도의 큰 변화 여부를 확인하여 인공

지능 모델의 변경 여부를 확인할 수 있다. 다음과 같이 공격자의 데이터를 학습 시 성능이 저하되는 것을 감지하여, 학습하지 않거나 머신 언러닝 방법으로 학습한 데이터를 제거하는 방식 등의 다양한 방식을 활용할 수 있다. 머신 언러닝은 프라이버시와 잊혀질 권리 등을 목적으로 모델에서 특정 정보를 제거하기 위해 개발되었지만, 아직 기술이 성숙해지지 않았기 때문에 현재로서는 아직 적용될 가능성이 높지 않다. 그러므로 현재 질의 데이터를 학습하기 전, 성능에 대한 변화를 비교하는 방식으로 가중치 업데이트를 하는 단계가 필요하다.

4.3 법·제도적 개선 사항

생성형 인공지능 서비스에 대하여 3가지 법·제도적 개선 사항이 존재한다. 첫 번째로 개발사의 책임에 대한 논의가 필요하다. 개발사의 책임소재는 현행법상 범죄로 규정되지 않은 행위를 범죄로 규정하기 위해 필요한 요소 중 하나이다. 개발사의 책임과 관련된 논의는 생성형 인공지능 전 검색 엔진 개발사와 관련되어 지속적으로 논의되었다. 차별적 내용, 혐오적 내용 등을 포함하는 불법적인 콘텐츠를 검색할 수 있게 하는 검색 엔진에 대해 2022년부터 유럽연합을 중심으로 검색 결과에 대한 개발사의 책임 및 의무에 대해 법제화하는 움직임이 발생하고 있다. 이는 Google과 같은 대형 검색 엔진 개발사에도 동일하게 적용되는 법안으로 구상하는 과정에 있다[49]. 생성형 인공지능 또한 사용자에게 불법적인 콘텐츠를 제공할 수 있다는 가능성이 있으며, 이에 대한 개발사의 책임이 논의 중인 상황에 있다. 생성형 인공지능은 학습한 데이터로부터 콘텐츠를 생성하는 인공지능으로, 개발사가 인공지능을 개발하는 데에서는 개발사의 책임 있는 개발을 필요로 한다. 유럽의 AI act에는 인공지능 서비스의 위험도에 따라 개발사가 지켜야 하는 의무를 규정하고 있다. 현재 생성형 인공지능 관련 범죄 중에서 현행법상 처벌이 불가능한 행위가 존재하는데, 현행법상 악성 프로그램의 제작, 가짜 뉴스 유포, 혐오 발언 생성 등의 행위가 이에 해당한다. 혐오 발언 생성과 극단적 선택 유도 문구 생성의 경우 사람과 사람 사이에서 발생 시 처벌이 가능하나, 인공지능과 사람 사이에서는 인공지능의 법인격 인정 여부에 따라 처벌이 불가능할 수 있다. 또한, 인공지능에 의해 개인정보가 유출되었을 때에도 법적으로 책임 소재가 명확하지 않아 유출된 개인정보와 관련되어 피해 보상이 제대로 일어나지 않을 수 있다. 인공지능 서비스 개발사의 의무를 규정한다면, 의무 위반 시 처벌 수준을 고려하여 적정한 수위의 처벌을 정해야 한다.

두 번째로 수사관과 일반인 대상 교육 프로그램 개발이 필요하다. 생성형 인공지능이 발전함에 따라 수사관도 새로운 범죄를 맞이하게 되었다. 다양한 생성형 인공지능 관련 범죄 중 생성형 인공지능 침해범죄 위협의 경우에는 인공지능에 범죄 행위가 일어났는지를 확인하기 위해 인공지능에 대한 분석이 필요하다. 역공학을 통하여 인공지능 모델의 하이퍼파라미터 변경 여부를 확인하거나 활성화 계수를 통한 데이터베이스 변경 여부 등을 수사관이 확인하여야 하므로 수사관은 인공지능을 분석하고 확인하기 위한 지식을 갖춰야 한다[29, 50]. 일반인의 경우에는 생활에 인공지능 서비스가 밀접하게 연관되면서 올바른 인공지능 사용뿐만 아니라 올바른 정보를 습득하는 방법에 대한 교육을 포함한다. 생성형 인공지능으로 만들어진 콘텐츠가 2025년에는 인터넷에 돌아다니는 콘텐츠의 90% 이상을 차지할 것이라는 전망과 같이 인터넷상 데이터가 대부분 생성형 인공지능에 의해 생성된다면, 일반인들은 습득하는 정보가 진실인지 거짓인지를 확인하기 어려워진다. 따라서 사용자 입장에서 올바른 정보를 습득하는 방법이 안내되어야 한다.

마지막으로 데이터 수집 방법론의 법제화가 필요하다. 생성형 인공지능에 있어서 학습 데이터의 구축은 필수적인 단계이다. 또한, 학습 데이터의 양과 질에 따라 모델이 도출할 수 있는 결괏값이 상이하므로, 양질의 학습 데이터를 구축하는 것은 필수적이다. 그러나 현재 상용화된 생성형 모델 서비스의 경우, 학습 데이터상의 데이터 편향성을 방지하기 위한 대책이 구축되어 있지 않은 실정이다. 따라서 데이터 처리 단계의 기술적 보완뿐만 아니라 법·제도적 방안을 제안하여 편향이 최대한 제거된 학습 데이터셋 구축을 규제해야 하며, 편향성뿐만 아니라 저작권을 보호할 수 있는 법·제도 개선도 필요하다. 파운데이션 모델 학습 시 데이터의 60~70%가 크롤링 데이터라고 알려져 있으며, 양은영(2023)에 따르면, 생성형 인공지능의 학습 데이터에 대한 저작권 침해 문제가 존재한다[51]. 현재 유럽연합은 AI Act를 통해 인공지능 학습에 활용된 데이터의 저작권을 명시하도록 규제안을 발표하였으나, 이에 대한 기술적인 구현이 현실적으로 어렵다는 쟁점이 존재하는 실정이다. 이

에 기술적으로 구현할 수 있는 저작권 보호 법안에 대한 연구도 진행되어야 한다.

V. 결 론

생성형 인공지능이 발전함에 따라 인공지능과 관련된 서비스가 일상에 들어오고 있다. 사람들의 삶에 인공지능이 들어오게 되면서, 기존의 사이버범죄 발생이 가속화될 수 있고 새로운 유형의 범죄가 등장할 수 있다. 인공지능의 법·제도적 규제가 필요하지만, 현재 국내에서는 생성형 인공지능에 대한 범죄의 분류 연구가 부족하다. 이에 본 연구에서는 생성형 인공지능과 관련된 범죄를 생성형 인공지능 이용범죄 위협, 생성형 인공지능 침해범죄 위협, 기타 인공지능 관련 위협으로 분류하고자 하였다. 또한, 범죄 원인을 파악 후 인공지능 개발 단계에 기반하여 기술적 대응 방안을 제시하였다. 단계별 대응 방안은 개인정보와 인공지능의 데이터 편향성 등 다양한 분야의 논문에서 제시하는 대응 방안을 인공지능 개발 단계 기반으로 분류 및 정리하여 실제 개발 시 순차적으로 방안을 적용하는 것을 통해 쉽게 위협에 대응할 수 있다는 의의가 있다. 법・제도적 개선방안으로 개발자의 책임에 대한 논의와 수사관 및 일반인의 교육 프로그램 개발, 데이터 수집 방법론의 법제화를 제시하였다. 본 연구는 지금까지 인공지능 관련 범죄 연구에서 다루지 않았던 개발자의 책임에 대해 다루었다는 점에서 의의를 갖는다.

하지만, 본 연구는 2가지의 한계점을 갖는다. 첫 번째로 기타 인공지능 범죄 위협의 일부는 대응 방안을 제시하지 못한다는 한계점이 있다. 이와 관련하여 예시로는 개발사의 고의적 악성 생성형 인공지능 서비스 개발과 학습 데이터 저작권 침해가 있다. 전자의 경우, 악의적인 서비스를 개발하여 공유할 시 이를 탐지하고 차단하는 것이 현실적으로 어렵다. 후자인 학습 데이터 저작권 침해의 경우에는 모든 데이터에 대한 저작권 표기가 기술적으로 어려우며, 저작권의 특성상 기간에 따라 만료가 될 수 있다는 점 등을 고려해야 하므로 적절한 대응 방안을 현 상황에서 제시할 수 없다. 두 번째 한계점은 본 연구에서 제시한 범죄 분류와 대응 방안이 현재까지 개발된 생성형 인공지능 기반의 연구라는 점이다. 생성형 인공지능 기술은 여전히 개발 중인 기술이며, 추후에 등장할 수 있는 기술은 무궁무진하다. 즉, 앞으로 어떤 구조의 인공지능이 등장하는지 예측하기 어려우며, 새로운 구조의 인공지능 등장 시 추가적인 연구를 필요로 한다.

〈부록 1〉

부록 1. 생성형 인공지능 관련 범죄의 분류

대분류	중분류	행위	처벌 근거 조항
생성형 인공지능 이용범죄 위협	정보통신망 이용범죄 위협	사기	형법 제347조(사기)
		피싱	형법 제114조(범죄단체등의 조직), 118조(공무원 자격의 사칭), 제347조(사기) 전기통신금융사기피해방지및피해금환급에관한특별법제15조의2
	정보통신망 침해범죄 위협	서비스 거부 공격	정보통신망법 제48조(정보통신망 침해행위 등의 금지)
		해킹	정보통신망법 제48조(정보통신망 침해행위 등의 금지)
	불법 콘텐츠 범죄 위협	스팸 메일 작성	정보통신망법 제50조 제1항
		사이버 성폭력	정보통신망법 제44조의7 제1항 제1호 사이버 음란물
		사이버 스토킹	정보통신망법 제44조의7 제1항 제3호 사이버 스토킹
		기타 불법 콘텐츠 범죄 (허위주민등록번호 생성)	주민등록법 제37조 제1항 제1호, 제4호
	현행법상 범죄로 규정되지 않은 행위	가짜 뉴스	-
		악성 프로그램 제작	-
생성형 인공지능 침해범죄 위협	인공지능 서비스 인프라 침해범죄 위협	서비스 거부 공격, 해킹 등	정보통신망법 제48조(정보통신망 침해행위 등의 금지)
	인공지능 모델 침해 공격 위협	무결성 공격	정보통신망법 제48조(정보통신망 침해행위 등의 금지)
		모델 추출 공격	정보통신망법 제48조(정보통신망 침해행위 등의 금지)
		멤버십 추론 공격	정보통신망법 제48조(정보통신망 침해행위 등의 금지) 저작권법 제136조(벌칙) 제1항
		악의적 모델로의 재학습	정보통신망법 제48조(정보통신망 침해행위 등의 금지)
기타 인공지능 관련 위협	개발사 고의 및 과실로 인한 범죄 위협	개발사의 고의적인 악성 생성형 인공지능 서비스 제작	-
		학습, 출력 데이터의 저작권 침해	저작권법 제137조, 제138조
		개인정보 유출	정보통신망법 제49조(비밀 등의 보호) 개인정보보호법 제71조
	현행법상 범죄로 규정되지 않은 행위*	인종차별 및 성차별 등 혐오 발언 생성	성폭력범죄의 처벌 등에 관한 특례법 제13조 (통신매체를 이용한 음란행위) 형법 제311조(모욕) 정보통신망법 제44조(정보통신망에서의권리보호)
		극단적 선택 유도 문구 제시	자살예방법 제19조(자살유발정보에방체계의 구축) 형법 제252조(촉탁,승낙에의한살인등)
		Hallucination으로 인한 명예훼손	정보통신망법 제44조(정보통신망에서의 권리보호)

* : 관련된 법조항은 있으나, 인공지능의 법적 주체와 개발사의 책임 여부 등으로 인해 현행법상 처벌이 불가능할 수 있음

〈부록 2〉

부록 2. 생성형 인공지능 위협 대응 방안의 분류

개발 단계	대응 방안	사용 기법	대응 가능 위협
데이터 수집·전처리	학습 데이터 다양성 증가	학습 데이터에 대한 적대적 예제 생성 및 학습	무결성 공격, 악의적 모델로의 재학습
	학습 데이터 편향성 감소	Sample Reweighting	인종차별 및 성차별 등 혐오 발언 생성
		Loss Reweighting	
		Batch Selection	
		VAE RECAP	
	데이터 진위 검증	진위 검증 데이터베이스 구축 및 비교	Hallucination으로 인한 명예훼손
구현	사용자 입력값 필터링	사용자 질의 필터링	생성형 인공지능 이용범죄 위협
		사용자 입력 데이터 연합학습 활용 시 필터링	인종차별 및 성차별 등 혐오 발언 생성, 극단적 선택 유도 문구 제시
	개인정보 비식별화	익명화	개인정보 유출
		가명화	
테스트	인공지능 서비스 인프라 보안	데이터베이스 권한 확인	생성형 인공지능 침해범죄 위협
		기기 연결 확인	
		모의 해킹	
		기타 보안 프로그램	
	인공지능 신뢰성 강화	신뢰할 수 있는 인공지능	인종차별 및 성차별 등 혐오 발언 생성, 극단적 선택 유도 문구 제시, Hallucination으로 인한 명예훼손
		인공지능 신뢰성 평가지표 개발	
유지보수	모델 변경 파악	모델 가중치 변경 정도 파악	악의적 모델로의 재학습

References

[1] Krystal Hu, "ChatGPT sets record for fastest-growing user base-analyst note," https://www.reuters.com/technology/chatgpt-sets-record-fastest-growing-user-base-analyst-note-2023-02-01/, Accessed on Nov. 2023.

[2] Lim Yeongshin, "2023 Key Issues and Meaning for Generative AI," KISDI Perspectives. 2023 August No.3. 1-13. Aug. 2023.

[3] Anrirudh VK, "Leaked LLaMA Unveils the Power of Open Source for AI," Analytics India Magazine, https://analyticsindiamag.com/leaked-llama-unveils-the-power-of-open-source/, Accessed on Nov. 2023.

[4] Skylar Harris, et al, "High schooler calls for AI regulations after manipulated pornographic images of her and others shared online," https://edition.cnn.com/2023/11/04/us/new-jersey-high-school-deepfake-porn/index.html, Accessed on Nov. 2023

[5] Vincent Acovino, et al, "A sci-fi magazine has cut off submissions after a flood of AI-generated stories," https://www.npr.org/2023/02/24/1159286436/ai-chatbot-chatgpt-magazine-clarkesworld-artificial-intelligence, Accessed on Nov. 2023.

[6] Interpol, "ChatGPT Impacts on Law Enforcement," Aug. 2023.

[7] Ahn SungMahn, "Deep Learning Architectures and Applications," Journal of Intelligence and Information Systems, 22(2), pp. 127-142, Jun. 2016.

[8] Lee Seungchul, Jeong Hae-dong, Park Seung-tae and Kim Soohyun, "Deep Learning," Journal of KSNVE. vol. 27(3), pp. 19-25, May. 2017.

[9] Kitae Kim, Bomi Lee and Jong Woo Kim, "Feasibility of Deep Learning Algorithms for Binary Classification Problems," Journal of Intelligence and Information Systems, 23(1), pp. 95-108, Mar. 2017.

[10] Joyjit Chatterjee and Nina Dethlefs, "This new conversational AI model can be your friend, philosopher, and guide... and even your worst enemy," Patterns, vol. 4, no. 1, Jan. 2023.

[11] Sumit Pandey and Srishti Sharma, "A comparative study of retrieval-based and generative-based chatbots using Deep Learning and Machine Learning," Healthcare Analytics, vol. 3, no. 100198, Nov. 2023.

[12] Blauth, T.F., Gstrein, O.J., and Zwitter, "A. Artificial intelligence crime: An overview of malicious use and abuse of AI," IEEE Access, 10, 77110-77122. Jul. 2022.

[13] D. Jeong, "Artificial Intelligence Security Threat, Crime, and Forensics: Taxonomy and Open Issues," IEEE Access, vol. 8, pp. 184560-184574, Oct. 2020.

[14] Younghee Kim, "Generative AI Industry Status Report," Korea Copyright Commission, Apr. 2023.

[15] Jinho Yoo, et al, "An Analysis of the Arrival of AI-Centered Society and Security Issues," KISA Insight, 2022 vol. 3, Apr. 2022.

[16] Google, "Reporting content under Article 6-4 of the LCEN," https://support.google.com/legal/answer/11512794?hl=en-GB, Accessed on Nov. 2023.

[17] Park Yunsuk, "Analysis of the European Union Digital Service Act on Copyright and its implications for Korea's system," Copyright Issue report 2022-30, Oct. 2022.

[18] Korea National Police Agency, "Cyber crime classification," https://ecrm.polic

e.go.kr/minwon/crs/quick/cyber1, Accessed on Nov. 2023.
[19] OpenAI, "Usage policies," https://openai.com/policies/usage-policies, Accessed on Nov. 2023.
[20] Oh Sojeong, Sin Jiho, Park Jaehyun and Kim Kibum, "The possibility and countermeasures of cyber attack using ChatGPT," 2023 Korea Institute of Digital Forensics Summer Conference. pp. 303-324. Jun. 2023.
[21] Ella Cao and Eduardo Baptista, "'Deepfake' scam in China fans worries over AI-driven fraud," https://www.reuters.com/technology/deepfake-scam-china-fans-worries-over-ai-driven-fraud-2023-05-22/, Accessed on Nov. 2023.
[22] Lin, Z., Shi, Y., and Xue, Z. "Idsgan: Generative adversarial networks for attack generation against intrusion detection," In Pacific-asia conference on knowledge discovery and data mining, pp. 79-91, May. 2022.
[23] Shannon Bond, "Fake viral images of an explosion at the Pentagon were probably created by AI," https://www.npr.org/2023/05/22/1177590231/fake-viral-images-of-an-explosion-at-the-pentagon-were-probably-created-by-ai, Accessed on Nov. 2023.
[24] Seb Starcevic, "AI 'Tom Cruise' joins fake news barrage targeting olympics," https://www.politico.eu/article/ioc-says-it-was-hit-by-fake-news-campaign-and-ai-tom-cruise/, Accessed on Nov. 2023.
[25] Kyeonghoon Jeong, "Digital sex crimes, portrait rights infringement will be easier to punish," https://m.mt.co.kr/renew/view.html?no=2023111010123345938#_doyouad, Accessed on Nov. 2023.
[26] Tramèr, F., Zhang, F., Juels, A., Reiter, M. K., and Ristenpart, T., "Stealing machine learning models via prediction APIs," 25th USENIX security symposium (USENIX Security 16), pp. 601-618. Aug. 2016.
[27] Wu, Bang, Yang, Xiangwen, Pan, Shirui, and Yuan, Xingliang, "Model extraction attacks on graph neural networks: Taxonomy and realisation," the 2022 ACM on Asia Conference on Computer and Communications Security. pp. 337-350. May. 2022.
[28] Fredrikson, M., Jha, S., and Ristenpart, T, "Model inversion attacks that exploit confidence information and basic countermeasures," the 22nd ACM SIGSAC conference on computer and communications security. pp. 1322-1333. Oct. 2015.
[29] Schneider, J., and Breitinger, F, "Towards AI forensics: Did the artificial intelligence system do it?", Journal of Information Security and Applications, vol. 76, pp. 103517, Jun. 2023.
[30] Youji Jang, "The Korean Government Should Not Grant a Copyright of AI's Creations," https://herald.cauon.net/news/articleView.html?idxno=20966, Accessed on Nov. 2023.
[31] Young-Hoa Son. "A Study on Creation by Generative AI and Copyright," Journal of Law and Politics research , 23(3), pp. 357-289, Sep. 2023.
[32] Kris Holt, "Three Samsung employees reportedly leaked sensitive data to ChatGPT," https://www.engadget.com/three-samsung-employees-reportedly-leaked-sensitive-data-to-chatgpt-190221114.html, Accessed on Nov. 2023.
[33] Justin Mccurry, "South Korean AI chatbot pulled from Facebook after hate speech towards minorities," https://www.theguardian.com/world/2021/jan/14/time-to-properly-socialise-hate-speech-ai-chatbot-pulled-from-faceb

ook, Accessed on Nov. 2023.
〔34〕 Lia Parisyan-Schmidt, "Racial and Gender Bias Lessons Learned from Midjourney and Disordered Eating Prompts," https://www.linkedin.com/pulse/racial-gender-bias-lessons-learned-from-midjourney-lia, Accessed on Nov. 2023.
〔35〕 Chloe Xiang "He Would Still Be Here: Man Dies by Suicide After Talking with AI Chatbot, Widow Says," https://www.vice.com/en/article/pkadgm/man-dies-by-suicide-after-talking-with-ai-chatbot-widow-says, Accessed on Nov. 2023.
〔36〕 James Vincent, "OpenAI sued for defamation after ChatGPT fabricates legal accusations against radio host," https://www.theverge.com/2023/6/9/23755057/openai-chatgpt-false-information-defamation-lawsuit, Accessed on Nov. 2023.
〔37〕 Shen, Z., Cui, P., Zhang, T., and Kunag, K, "Stable learning via sample reweighting," Proceedings of the AAAI Conference on Artificial Intelligence, pp. 5692-5699, Feb. 2020
〔38〕 Shirokikh, B., Shevtsov, A., Kurmukov, A., Dalechina, A., Krivov, E., Kostjuchenko, V Golanov, A. and Belyaev, M, "Universal loss reweighting to balance lesion size inequality in 3D medical image segmentation," In Medical Image Computing and Computer Assisted Intervention–MICCAI 2020, pp. 523-532, Sep. 2020.
〔39〕 Roh, Y., Lee, K., Whang, S. E., and Suh, C, "Fairbatch: Batch selection for model fairness," Proceedings of the 9th International Conference on Learning Representations (ICLR), May. 2020
〔40〕 Seungmin, Lee, "Trends and Industrial Implications of Joint Learning Technology," ETRI insight, Nov. 2020.
〔41〕 Ministry of Science and ICT, "Strategy to realize trustworthy artificial intelligence," May. 2021.
〔42〕 European Commission, "Ethics Guidelines for Trustworthy Artificial Intelligence," Apr. 2019.
〔43〕 United States Office of Management and Budget, "Guidance for Regulation of Artificial Intelligence Applications," Nov. 2020.
〔44〕 Li, B., Qi, P., Liu, B., Di, S., Liu, J., Pei, J., Yi, J. and Zhou, B., "Trustworthy AI: From principles to practices," ACM Computing Surveys, 55(9), pp. 1-46, Jan. 2023.
〔45〕 Poretschkin, M., Schmitz, A., Akila, M., Adilova, L., Becker, D., Cremers, A. B., ... & Wrobel, S., "Guideline for Trustworthy Artificial Intelligence–AI Assessment Catalog," arXiv preprint arXiv:2307.03681, Jun. 2023.
〔46〕 European Commision, "The final Assessment List for Trustworthy AI (ALTAI)," Jul. 2020.
〔47〕 Telecommunications Technology Association, "2022 Guide for Developing Trustworthy AI," Jul. 2022.
〔48〕 Financial Security Institution, "AI Security Guidelines in Financial Sector," Apr. 2023.
〔49〕 European Commission, "Digital Services Act: Commission designates first set of Very Large Online Platforms and Search Engines," Apr. 2023.
〔50〕 Wojciech Samek, et al, Explainable AI: Interpreting, Explaining and Visualizing Deep Learning, pp. 121-144, Sep. 2019.
〔51〕 Eunyoung Yang, "Necessity of regulation on the development and use of generative AI - Focusing on Large language Models conversational A.I. services (LLMs AI)," Sungkyunkwan Law, 35(2), pp. 293-325, Jun. 2023.

〈저자소개〉

박 우 빈 (Woobeen Park) 학생회원
2023년 2월: 동국대학교 경찰행정학부 졸업
2023년 3월~현재: 동국대학교 경찰행정학과 사이버수사전공 석사과정
〈관심분야〉 정보보호, 인공지능, 침해사고, 디지털 포렌식 등

김 민 수 (Minsoo Kim) 학생회원
2024년 2월: 동국대학교 경찰행정학부 졸업
2024년 3월~현재: 성균관대학교 과학수사학과 디지털포렌식전공 석사과정
〈관심분야〉 디지털 포렌식, 정보보호, 인공지능 등

박 윤 지 (Yunji Park) 정회원
2022년 8월: 동국대학교 경찰행정학부 졸업
2024년 2월: 동국대학교 경찰행정학과 석사
2024년 3월~현재: 성균관대학교 과학수사학과 디지털포렌식전공 박사과정
〈관심분야〉 디지털 포렌식, 정보보호 등

유 혜 진 (Hyejin Ryu) 학생회원
2023년 2월: 동국대학교 경찰행정학부 졸업
2023년 3월~현재: 동국대학교 경찰행정학과 사이버수사전공 석사과정
〈관심분야〉 디지털 포렌식, 정보보호, 인공지능 포렌식 등

정 두 원 (Doowon Jeong) 정회원
2019년 2월: 고려대학교 정보보호대학원 공학박사
2020년 9월~2024년 2월: 동국대학교 경찰행정학부 조교수
2024년 3월~현재: 성균관대학교 과학수사학과 조교수
〈관심분야〉 디지털 포렌식, 정보보호 등