https://doi.org/10.33778/kcsa.2024.24.2.009

기업 내 생성형 AI 시스템의 보안 위협과 대응 방안

최 정 완*

요 약

본 논문은 기업 내 생성형 AI(Generative Artificial Intelligence) 시스템의 보안 위협과 대응 방안을 제시한다. AI 시스템이 방대한 데이터를 다루면서 기업의 핵심 경쟁력을 확보하는 한편, AI 시스템을 표적으로 하는 보안 위협에 대비해야 한다. AI 보안 위협은 기존 사람을 타겟으로 하는 사이버 보안 위협과 차별화된 특징을 가지므로, AI에 특화된 대응 체계 구축이 시급하다. 본 연구는 AI 시스템 보안의 중요성과 주요 위협 요인을 분석하고, 기술적/관리적 대응 방안을 제시한다. 먼저 AI 시스템이 구동되는 IT 인프라 보안을 강화하고, AI 모델 자체의 견고성을 높이기 위해 적대적 학습 (adversarial learning), 모델 경량화(model quantization) 등 방어 기술을 활용할 것을 제안한다. 아울러 내부자 위협을 감지하기 위해, AI 질의응답 과정에서 발생하는 이상 징후를 탐지할 수 있는 AI 보안 체계 설계 방안을 제시한다. 또한 사이버 킬 체인 개념을 도입하여 AI 모델 유출을 방지하기 위한 변경 통제와 감사 체계 확립을 강조한다. AI 기술이 빠르게 발전하는 만큼 AI 모델 및 데이터 보안, 내부 위협 탐지, 전문 인력 육성 등에 역량을 집중함으로써 기업은 안전하고 신뢰할 수 있는 AI 활용을 통해 디지털 경쟁력을 제고할 수 있을 것이다.

Security Threats to Enterprise Generative AI Systems and Countermeasures

Jong-woan Choi*

ABSTRACT

This paper examines the security threats to enterprise Generative Artificial Intelligence systems and proposes countermeasures. As AI systems handle vast amounts of data to gain a competitive edge, security threats targeting AI systems are rapidly increasing. Since AI security threats have distinct characteristics compared to traditional human-oriented cybersecurity threats, establishing an AI-specific response system is urgent. This study analyzes the importance of AI system security, identifies key threat factors, and suggests technical and managerial countermeasures. Firstly, it proposes strengthening the security of IT infrastructure where AI systems operate and enhancing AI model robustness by utilizing defensive techniques such as adversarial learning and model quantization. Additionally, it presents an AI security system design that detects anomalies in AI query-response processes to identify insider threats. Furthermore, it emphasizes the establishment of change control and audit frameworks to prevent AI model leakage by adopting the cyber kill chain concept. As AI technology evolves rapidly, by focusing on AI model and data security, insider threat detection, and professional workforce development, companies can improve their digital competitiveness through secure and reliable AI utilization.

Key words : Security Threats, Enterprise AI Systems, AI-specific Response System, Cyber Kill Chain

접수일(2024년 05월 16일), 수정일(1차: 2024년 06월 06일), 게재확정일(2024년 06월 30일)

* 원광대학교 기초자연과학 연구소 교수

1. 서론

1.1 연구 배경 및 목적

최근 기업들은 업무 효율성 제고와 의사결정 지원, 고객 서비스 향상 등의 목적을 달성하기 위해 AI(Artificial Intelligence), 그 중에서도 생성형(Generative) AI 시스템을 도입하고 있다. 생성형 AI는 수백억 개의 매개변수로 훈련된 초거대 언어 모델(Large Language Model)을 바탕으로 언어, 이미지, 음성 등 다양한 형태의 입력 데이터를 처리할 수 있도록 개발된 AI이다. 기업들은 이러한 초거대 모델을 자사의 데이터로 미세 조정하는 과정을 거치는데, 이 과정에서 AI는 방대한 데이터를 분석, 학습하여 기업의 핵심 경쟁력을 강화하는 한편 민감한 정보를 대량으로 보유하게 된다. 기업들은 이러한 민감한 정보를 보호하기 위해 자사 내부 서버에서 AI 시스템을 독립적으로 운영한다. 그러나 경쟁 기업의 입장에서는 여러 부서와 시스템에 분산되어 있던 기업 기밀이 AI 시스템으로 통합됨에 따라, 이를 탈취하려는 시도가 증가하고 있다. 이에 따라 AI 시스템에 대한 보안 위협이 커지고 있으며, 기업 내 AI 시스템의 보안성 확보가 중요한 과제로 주목받고 있다.

본 연구는 기업 내 AI 시스템의 보안 위험을 분석하고, 현재 적용되고 있는 접근 통제 방식의 적절성을 검토하고자 한다. 특히, 사회공학적 기법이나 피싱 공격으로 내부 네트워크가 침해되는 상황을 가정하고, 이에 대한 효과적인 대응 방안을 모색하는 것을 목적으로 한다. 이를 위해 AI 시스템 보안의 중요성과 주요 위협 요인을 살펴보고, 보안 강화를 위한 기술적, 관리적 접근 방안을 체계적으로 정리하고자 한다.

1.2 AI 도입 현황과 기업 내 활용 사례

AI 기술은 빅데이터, 클라우드 컴퓨팅, 사물인터넷(IoT) 등 디지털 기술의 발전과 함께 빠르게 진화하고 있다. 글로벌 IT 기업들은 AI 기술 개발에 막대한 투자를 하고 있으며, 다양한 산업 분야에서 AI 활용이 확산되고 있다. 이는 인공지능(AI)이 기계가 수행할 수 있는 작업의 범위를 확장하고, 인간과 기계 지능 간의 경계를 모호하게 만드는 데 기인한다[1].

2016년 Google DeepMind의 AlphaGo가 바둑 챔피언 이세돌을 물리친 것이 중요한 전환점이 되었으며, 2022년 말 ChatGPT의 초거대언어 모델을 기반으로 하는 생성형 AI 등장으로 AI 환경은 더욱 큰 도약을 경험하게 되었다. 국내외 주요 기업들은 업무 자동화, 고객 분석, 예측 분석 등에 AI를 적극적으로 도입함으로써 운영 효율성과 의사결정의 질을 높이고 있다[2].

대표적으로 Amazon은 AI 기반 추천 시스템을 통해 고객 맞춤형 상품을 제안하고 있으며, Netflix와 YouTube는 AI 알고리즘을 활용하여 사용자의 시청 패턴을 분석하고 개인화된 콘텐츠를 추천하고 있다[3]. 금융 산업에서는 AI를 활용한 고객 분석, 부정 거래 탐지, 알고리즘 트레이딩 등이 활발히 이루어지고 있다[4]. 제조업에서는 AI를 통해 제조 설비의 고장 시점을 예측하거나 품질 검사를 자동화하는 등 운영 효율성을 제고하고 있다[5].

국내에서도 대기업을 중심으로 AI 도입이 가속화되고 있다. 삼성전자는 모든 제품에 AI 기술을 적용하겠다고 발표하였으며, LG, 현대, SK등의 주요 기업들도 AI 기술 도입과 활용에 적극적인 움직임을 보이고 있다. 뿐만 아니라 중소기업들도 자사의 경쟁력 강화를 위해 AI 기술을 도입하는 사례가 늘어나고 있다[6]. 이처럼 국내 기업들은 업종과 규모를 막론하고 AI를 미래 성장의 핵심 동력으로 인식하고, AI 기술 확보와 활용에 박차를 가하고 있다.

2. 기업 내 AI 시스템 보안의 중요성

2.1 AI 시스템이 보유한 핵심 정보와 기밀성

AI 시스템은 기업의 방대한 데이터를 학습하고 분석하여 의사결정을 지원하고 업무 효율성을 높이는 역할을 수행한다[7]. 이 과정에서 AI 시스템은 기업의 민감한 정보, 영업 비밀, 고객 데이터 등 핵심 자산을 대량으로 보유하게 된다. 특히 AI를 구

현하는 알고리즘과 학습된 모델에는 기업의 노하우와 경쟁력의 원천이 내재되어 있어, 이들 정보가 외부로 유출될 경우 기업에 치명적인 피해를 초래할 수 있다.

정보 유출은 직원이 경쟁 기업으로 이직하면서 의도적으로 발생할 수도 있지만, 보안 공격에 의해 발생할 수도 있다. 더욱이 AI 시스템이 모든 데이터를 통합 관리하고 있다면, 경쟁 기업은 상대회사의 직원을 포섭하여 시간과 비용을 들여 제한된 정보를 획득하는 것보다, 상대적으로 비용이 적게 드는 사이버 보안 공격을 시도할 가능성이 높다. 특히 경쟁 기업이 해킹 능력이 뛰어난 제3국이고, 탈취하려는 기업 비밀이 반도체와 같이 국가의 명운이 걸린 사업과 관련된 경우라면 더욱 그러할 것이다.

따라서 AI 시스템이 보유한 정보의 기밀성을 유지하고 안전하게 관리하는 것이 무엇보다 중요하다. 기업은 AI 시스템을 내부의 핵심 자산으로 인식하고, 이에 상응하는 높은 수준의 보안 체계를 구축해야 한다. 아울러 AI 시스템의 취약점을 지속적으로 점검하고 보안 위협에 선제적으로 대응할 수 있는 역량을 갖추어야 한다.

2.2 AI 시스템 구성 요소 유출 사례와 위험성

AI 기술을 둘러싼 기업 간 경쟁이 치열해지면서 AI 시스템의 핵심 구성 요소인 AI 모델과 학습 데이터의 유출 사고도 증가하는 추세이다. 이러한 유출 사고는 다양한 경로로 발생할 수 있다.

2023년 Microsoft의 AI 연구팀이 GitHub에 오픈소스 AI 모델을 제공하는 과정에서 38TB의 데이터가 유출된 사건은[8] AI 개발 과정에서의 부적절한 액세스 토큰 관리가 대규모 데이터 유출로 이어질 수 있음을 보여준다. 이는 AI 시스템뿐만 아니라 전체 IT 인프라에 대한 종합적인 보안 관리가 필요함을 시사하는데, 3.1절에서는 이에 대한 대응 방안을 다룰 것이다.

2022년 구글의 전직 엔지니어 린웨이 딩(Linwei Ding)이 AI전용 프로세서인 텐서처리장치(TPU) 등 핵심 기술 정보가 포함된 500여 개의 기밀 파일을 무단으로 반출하여 중국 AI 기업에 제공한 사건은 내부자에 의한 데이터 유출 위험을 보여준다[9]. 이는 중요한 AI 기술을 다루는 기업일수록 내부 위협에 취약할 수 있음을 시사하며, 3.3절에서는 이에 대한 대응 방안으로 내부자 위협 감지 체계에 대해 논의할 것이다.

2023년 삼성전자가 ChatGPT와 같은 대화형 AI 서비스의 보안 위험을 이유로 사내 접속을 차단한 사례는 외부 AI 서비스 이용에 따른 데이터 유출 위험을 보여준다. 업무상 민감한 정보가 외부 AI 서비스에 노출되는 사고가 발생하였기 때문이다[10]. 이는 기업이 자체적인 AI 시스템을 구축하고 운영할 때에도 고려해야 할 사항으로, 3.5절에서는 이를 방지하기 위한 보안 인력 교육 및 관리 방안을 제시할 것이다.

위의 사례들은 AI 모델과 데이터 유출이 기업의 핵심 경쟁력과 직결되는 심각한 문제임을 보여준다. 3장에서는 이러한 다양한 유출 경로에 대한 구체적인 대응 방안을 논의할 것이다. 기업 내 AI 시스템의 안전성과 복원력을 확보하기 위해서는 다각적인 보안 조치가 필수적이다.

2.3 AI 보안 위협의 특징과 차별점

AI 보안 위협은 기존 사이버 보안 위협과 비교할 때 다음과 같은 고유한 특징을 갖고 있다.

<표1> 사이버 보안 위협과 AI 보안 위협의 차이점

구분	사이버 보안 위협	AI 보안 위협
공격 대상	시스템, 네트워크, 데이터 등	데이터 등 AI 모델, 학습 데이터, 추론 결과 등
공격 파급력	시스템 중단, 데이터 유출 등	AI 기반 의사결정 오류, 서비스 품질 저하 등
공격 기법	악성 코드, 피싱, DDoS 등	적대적 예제, 모델 역공학, Deepfake 기술 등
징후 포착 난이도	지능형 공격의 경우 탐지 어려움	모델 불투명성, 데이터 복잡성 등으로 인해 탐지 어려움
대응 방안	보안 솔루션, 패치 관리, 접근 통제 등	AI 보안 전략, 모델 보안 등

첫째, AI 모델 자체가 공격 대상이 될 수 있다.

세계적인 기업이 아닌 이상, 기업이 AI 모델을 처음부터 새롭게 개발하는 것은 막대한 컴퓨팅 파워와 데이터 전처리에 소요되는 시간으로 인해 쉽지 않다. 이에 많은 기업들은 허깅페이스(HuggingFace)와 같은 플랫폼에서 공개된 모델(Open source model), 예를 들어 Meta사의 Llama나 Microsoft사의 Phi 등과 같은 AI 프로그램을 기반으로 자사의 필요에 맞게 미세조정(fine-tuning)하여 고유의 AI 모델을 생성한다[11]. 그러나 이 과정에서 기반 모델이 북한에 의해 개발되어 적대적 예제(Adversarial Training)로 학습되었거나, 모델의 취약점을 역공학적으로 파고드는 공격(Model Inversion Attack)에 노출될 경우 심각한 보안 위협이 발생할 수 있다[12]. 이러한 위협에 대응하기 위해 AI 모델 자체의 견고성을 높이고, 모델 개발 과정에서 보안성 검증이 필수적이다.

둘째, AI 시스템이 의사결정에 직접 관여함에 따라 공격 파급력이 크다. 과거에는 의사결정 과정에서 주로 전문가들의 견해를 참조했으나, 최근에는 AI가 인간이 간과하거나 고려하지 못한 변수들을 제안하면서 의사결정 지원 도구로 자리매김하고 있다[7]. 경쟁 기업이 AI 시스템을 교란하여 중요한 사업 과제에 대해 잘못된 판단을 내리도록 유도할 경우, 그 영향력이 매우 클 수 있다. 이를 방지하기 위해서는 AI 시스템에 대한 접근 통제를 강화하고, 질의응답 과정에서 이상 징후를 실시간으로 탐지할 수 있는 모니터링 체계가 필요하다.

셋째, AI 기술이 빠르게 발전함에 따라 새로운 공격 기법이 지속적으로 출현하고 있다. 예를 들어, 경쟁 기업이 딥페이크(Deepfake) 기술을 이용하여 기업 내 핵심 인력을 사칭하거나[13], AI 도구를 활용하여 사회공학적 공격의 성공률을 높일 수 있다. 이러한 위협에 선제적으로 대응하기 위해서는 AI 기술 동향을 예의주시하며 전문 인력의 역량을 지속해서 강화해 나가야 한다.

넷째, 데이터 투명성과 모델 설명 가능성이 부족하여 공격 징후를 포착하기 쉽지 않다. 데이터 투명성 문제는 AI 모델이 방대한 데이터로 학습하는 과정에서 어떤 근거로 특정 결론에 도달했는지 설명하기 어려운 블랙박스와 같은 상태가 되기 때문에 발생한다[14]. AI 모델의 투명성이 낮아지면 이상 징후를 감지하기가 어려워지고, 사후 분석도 쉽지 않은 경우가 많다.

이처럼 AI 보안 위협은 여러 고유한 특징을 가지고 있어, 기존의 사람 중심 보안 체계와는 차별화된 대응 전략이 필요하다. 전통적인 보안 체계는 직원을 신뢰하고 검증된 인력에 대해 일정 수준의 권한을 부여하는 방식으로 운영되었다. 그러나 기업의 핵심 기밀이 AI 시스템으로 이관되고, AI가 다른 프로그램과 플러그인(Plug-in)을 통해 상호작용하는 환경이 되면서 AI에 특화된 보안 체계를 도입해야 한다.

AI 보안 체계 구축을 위해서는 AI에 접근하는 주체의 의도와 질의 목적을 파악할 수 있어야 한다. 이는 기존의 보안 체계에서 충분히 다루지 않았던 영역으로, 접근 권한을 부여받은 사용자의 의도를 지속적으로 모니터링하고, 비정상적인 활동을 실시간으로 탐지할 수 있는 시스템이 필요하다. 과거에는 접근 권한만 있다면 데이터 활용 목적에 대해 묻지 않았는데, 그것이 가능했던 이유는 엄격한 검증을 거친 인력에 대해서만 접근을 허용했기 때문이다. 그러나 이제는 AI가 데이터에 직접 접근하고 사람은 컴퓨터를 매개로 질의하는 구조가 되었다. 만약 질의에 사용된 컴퓨터가 악성코드에 감염되었는데 이를 탐지하지 못하고, 동시에 해당 컴퓨터 사용자가 기밀에 접근할 수 있는 고위 권한을 가진 인력이라면 내부자에 의한 위협이 발생할 수 있다. 실제로 아무리 강력한 보안 시스템이라도 내부자의 침투를 수개월 동안 탐지하지 못하는 사례가 존재한다[15]. 이는 AI 보안 체계를 구축할 때 IT 중심의 보안 체계와 더불어 OT(Operational Technology) 보안 체계도 함께 고려해야 함을 의미한다.

따라서 기업은 AI 모델의 견고성을 높이고, 데이터 투명성을 확보하며, 설명 가능한 AI를 구현하는 동시에 내부 위협 탐지, AI 기술 발전 동향 분석 등에 역량을 집중함으로써 AI 고유의 보안 위험에 효과적으로 대응할 수 있어야 한다.

3. AI 시스템 보안 강화 방안

3.1 AI 시스템 및 IT 인프라 보안 강화

외부의 생성형AI 시스템 대신 자사 내부 서버의 생성형 AI 시스템을 운용하는 기업 입장에서는 AI 시스템의 안전성 확보를 위해서 우선 IT 인프라 보안 수준을 제고하는 것이 중요하다. 이를 위해 정기적인 보안 점검과 패치 관리를 수행하고, 접근 통제 및 권한 관리를 강화해야 한다. 특히 중요 시스템과 데이터에 대해서는 다중 인증(Multi-Factor Authentication, MFA)을 적용하여 인증 강도를 높이는 것이 바람직하다.

나아가 Zero Trust 아키텍처를 도입하여 내외부 사용자와 디바이스를 상시 검증함으로써 내부 위협에 대한 방어 능력을 한층 높일 수 있다[16]. Zero Trust는 '네트워크 내부에서도 신뢰하지 않는다'는 원칙 하에 모든 접근 시도를 감시하고 통제하는 개념으로, 최근 클라우드, 모바일 환경에서 각광받고 있다. 2.2절 Microsoft 사례에서 볼 수 있듯이, 부적절한 액세스 토큰 관리는 대규모 데이터 유출로 이어질 수 있는 만큼, Zero Trust를 통해 엄격한 접근 통제를 적용하는 것이 중요하다. 다만 Zero Trust는 데이터 접근 권한 관리에는 효과적이지만, AI 모델 자체의 취약점을 완벽히 방어하기에는 한계가 있다. Zero Trust는 어떤 데이터에 접근 가능한지 여부는 파악할 수 있지만, 내부자가 정상적인 권한으로 AI 모델에 접근한 뒤 모델을 무단 반출하는 경우까지 통제하기는 어렵기 때문이다. 이를 보완하기 위해서는 AI 모델의 견고성을 높이기 위한 별도의 기술적 조치가 필요하다.

3.2 AI 모델 보안성 검증 및 적대적 공격 대응

AI 모델은 설계 및 학습 과정에서 다양한 취약점이 내재될 수 있으므로, 이를 사전에 식별하고 평가하는 포괄적인 보안성 검증 프로세스가 필수적이다. 먼저 데이터 중독 공격(Data Poisoning)과 같이 악의적인 데이터를 학습 데이터에 삽입하여 모델의 성능을 저하시키는 공격에 대응하기 위해, 학습 데이터의 오염이나 조작 여부를 확인하는 데이터 무결성 검증이 필요하다. 다음으로 특정 트리거를 삽입해 모델이 의도와 다르게 동작하도록 유도하는 백도어 공격(Backdoor Attack)에 대비하기 위해, 모델 구조와 학습 과정에서의 오류나 결함을 점검하는 모델 완전성 검사가 수행되어야 한다. 또한 입력 데이터에 약간의 변화를 주어 모델이 잘못된 예측을 하도록 유도하는 적대적 예제(Adversarial Example) 공격에 대응하는 한편, 모델의 과도한 복잡성으로 인해 발생하는 과적합(Overfitting) 문제 역시 해결해야 한다.

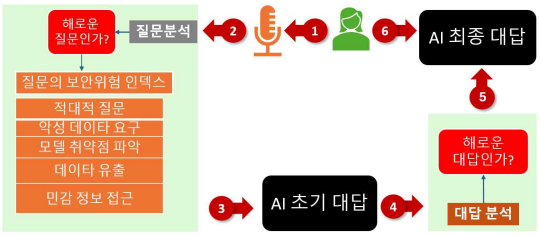
이러한 다양한 위협에 대응하기 위해서는 각 공격 유형에 특화된 방어 기술들이 활용되어야 한다. 데이터 중독 공격을 방어하기 위해서는 입력 데이터에 대한 검증과 필터링을 강화하고, 이상치 탐지 기법 등을 통해 악의적인 데이터를 사전에 차단하는 것이 중요하다. 백도어 공격에 대해서는 모델 구조와 학습 과정에 대한 투명성을 제고하고, 지속적인 모니터링과 검증을 통해 의심스러운 트리거나 악성 코드를 탐지해 내는 것이 핵심이다. 적대적 예제 공격에 대응하기 위해서는 적대적 학습(Adversarial Learning)을 통해 모델을 보다 강건하게 만드는 것이 도움이 된다. 과적합 문제를 완화하기 위해서는 모델 경량화(Model Quantization) 등을 통해 불필요한 복잡성을 제거하고 일반화 성능을 높이는 방안을 고려할 수 있다.

따라서 기업은 자사의 데이터 특성, 모델 아키텍처, 가용 자원 등을 종합적으로 고려하여 최적의 AI 모델 보안 전략을 수립해야 한다.

3.3 내부자에 의한 위협 감지

내부자에 의한 위협을 감지하기 위해서는 컴퓨터 접근 행위의 정상 여부를 판단하는 것이 출발점이 된다. 행위의 정상성은 크게 두 가지 측면에서 평가할 수 있는데, 하나는 컴퓨터 사용량과 시간을 살피는 정보기술(Information Technology, IT) 측면이고, 다른 하나는 질의 내용을 분석하는 운영기술(Operational Technology, OT) 측면이다. IT 관점에서는 주로 CPU 사용량, 접근 횟수, 데이터 사용량 등을 모니터링하면서 이전 접근 패턴과 비교하

여 정상 범위 내에 있는지 확인한다. 그러나 IT 영역에서는 통신 암호화로 인해 데이터 내용 자체를 살펴보기 어려운 경우가 많으므로, AI 보안 체계 내에서 내부자 위협을 방지하기 위해서는 OT 측면에 더욱 주목할 필요가 있다. OT는 데이터의 실질적인 내용을 분석한다. 이를 통해 사용자가 어떤 질문을 던지고, 그 결과로 무엇을 얻으려 하는지 행동 패턴을 면밀히 파악하는 것이 핵심이다. 따라서 AI 보안 체계는 질의 입력 단계와 응답 출력 단계를 구분하여, 질문 내용을 평가하고 AI의 답변도 분석할 수 있도록 설계되어야 한다.



(그림 1) 내부자에 의한 위협 감지를 위한 AI 보안체계 설계도

그림 1은 내부자에 의한 위협 감지를 위해 제안하는 AI 보안 체계의 개념도이다. 질의 입력부와 응답 출력부를 명확히 분리하고, 각 단계에서 질문과 답변을 분석하여 내부자 위협을 탐지하고 차단하는 구조를 취하고 있다. 동시에 AI 모델에 대한 접근과 반출 행위를 면밀히 모니터링하고 통제함으로써 모델 유출 자체를 원천 방지하고자 한다. 이를 통해 악의적인 내부자가 AI 시스템을 악용하려는 시도를 조기에 인지하고 선제적으로 대응할 수 있을 것으로 기대된다.

3.4 Cyber Kill Chain을 통한 AI 보호

기업 맞춤형 AI 모델의 유출을 방지하기 위해 사이버 킬 체인(Cyber Kill Chain) 개념을 적극 활용할 필요가 있다. 사이버 킬 체인은 공격자의 행위를 7단계로 구분하여, 각 단계별 대응 전략을 수립하는 프레임워크이다[17].

<표 2> 록히드 마틴사의사이버킬체인모델과 AI 대응용 사이버 킬체인 비교

단계	기존 사이버 킬 체인	AI 대응용 사이버 킬 체인
1단계 정찰	공격 목표와 표적 조사, 식별 및 선정	AI 모델과 시스템에 대한 정보 수집 시도 탐지
2단계 무기화	자동화 도구 등을 이용하여 공격 무기 준비	AI 모델 취약점 탐색 시도 탐지, 공격 도구 제작 시도 차단
3단계 전달	시스템에 사이버 무기를 전달	AI 시스템 접근 및 공격 도구 전달 시도 차단
4단계 악용	무기의 작동 촉발	AI 모델 유출 시도 무력화, 모델 무결성 검증, 실행 프로세스 통제 강화
5단계 설치	시스템에 악성 프로그램 설치	비정상적인 설치 행위 탐지 및 악성코드 제거
6단계 명령 및 제어	표적 시스템 원격 조작 채널 구축	AI 모델 변경에 대한 승인 절차 확립, 변경 시 감지를 통해 차단
7단계 목표 달성	정보 수집, 시스템 파괴 등 목적 달성	AI 모델 유출 피해 최소화, 재발 방지, 사고 원인 분석, 이해관계자 소통

본 연구에서는 이러한 사이버 킬 체인의 기본 개념을 유지하면서도, AI 모델 탈취 과정에 맞게 각 단계를 재정의하고 AI 고유의 보안 이슈를 반영한 대응 전략을 제시하고자 한다. 특히, 기존의 침입 탐지 시스템(Intrusion Detection System, IDS)에서 한 단계 더 나아가, AI 모델 변경에 대한 승인 절차를 도입한 점이 본 연구의 차별점이다. 이는 AI 모델의 무결성을 보호하고 무단 변경을 방지하는 데 효과적인 방안이 될 것으로 기대된다. 이러한 관점에서 AI 모델을 표적으로 하는 공격 시나리오를 분석하고 선제적 대응 방안을 마련하고자 한다. 구체적으로, AI 모델 탈취 과정을 정찰, 무기화, 전달, 악용, 설치, 명령 및 제어, 목표 달성 등의 단계로 구분하고, 각 단계에서의 공격 징후를 포착하여 맞춤형 방어 전략을 수립함으로써 AI 모델 유출 위험을 체계적으로 관리할 수 있을 것이다.

정찰(Reconnaissance) 단계에서는 AI 모델과 시스템에 대한 정보 수집 시도를 탐지하고 차단해야 한다. 이를 위해 AI 자산 목록을 상시 관리하고, 비인가 접근에 대한 모니터링을 강화해야 한다. 무기화(Weaponization) 단계에서는 AI 모델의 취약점을

찾아 공격 도구를 제작하려는 시도를 인지하고 무력화해야 한다. 이를 위해 AI 모델에 대한 보안 검토를 주기적으로 수행하고, 취약점이 발견되면 신속히 보완해야 한다. 전달(Delivery) 단계에서는 공격자가 AI 시스템에 접근하여 공격 도구를 전달하려는 행위를 차단해야 한다. 네트워크 접근 통제를 강화하고, 이상 행위 탐지 시스템을 고도화하여 의심스러운 접근을 실시간으로 파악해야 한다. 악용(Exploitation) 단계에서는 공격자가 전달한 공격 도구를 실행하여 AI 모델을 유출하려는 시도를 무력화해야 한다. AI 모델의 무결성을 검증하는 기술을 도입하고, 실행 프로세스에 대한 통제를 강화할 필요가 있다. 설치(Installation) 단계는 공격자가 AI 시스템 내부에 침투하여 지속적인 통제권을 확보하려는 단계이다. 비정상적인 설치 행위를 탐지하고 악성 코드를 제거하는 등 적극적인 대응이 필요하다.

명령 및 제어(Command & Control) 단계에서는 AI 모델 변조, 무단 반출 등 목적 달성을 위한 공격자의 시도를 차단해야 한다. 본 연구에서 제안하는 바와 같이, AI 모델 변경에 대한 엄격한 승인 절차를 마련하고 무단 변경 시도를 실시간으로 감지 및 차단하는 것이 핵심이다. 구체적으로, AI 모델 접근이나 변경 요청 시 사전 정의된 전자 결재 시스템을 통해 승인 절차를 진행하고, 승인 없는 접근이나 변경 시도 시 경고 발생과 함께 해당 행위를 차단해야 한다. 또한 접근 가능한 단말기를 사전 등록하고 인증하는 장치 제어 정책을 적용하며, 승인된 AI 모델은 지정된 스토리지에만 저장하도록 해야 한다. 아울러 AI 모델 파일에 버전 정보와 메타데이터를 포함하여 무단 변경을 탐지하고, 비정상적인 AI 모델 이동 시 데이터 손실 방지(DLP) 시스템을 통해 차단하는 방안도 고려할 수 있다. 이러한 일련의 조건들이 충족되지 않을 경우, 제안된 사이버 킬 체인 매커니즘이 작동하여 관련 행위를 차단하고 관리자에게 경고를 발송하게 된다. 이는 기존 침입 탐지 시스템(Intrusion Detection System, IDS) 대비 변경 승인 절차를 포함하여 한층 강화된 AI 모델 보안 체계라 할 수 있다.

사이버 킬 체인의 마지막 단계인 목표 달성(Actions on Objectives) 단계에서는 AI 모델 유출로 인한 피해를 최소화하는 방안을 마련한다. 예를 들어, AI 모델에 기기 단위 인증서(Machine Level Certificate)를 부여해서, 인가된 서버에서만 운용되도록 하는 것이다. 또한 재발 방지를 위한 사후 관리 활동이 중요하다. 사고 원인을 철저히 분석하여 대응 체계를 보완하고, 피해 확산 방지를 위해 이해관계자와의 소통을 강화해야 한다.

이처럼 사이버 킬 체인의 모든 단계에서 공격 시도를 감지하고 선제적으로 대응함으로써, 기업은 내외부의 AI 모델 유출 위협으로부터 핵심 자산을 보호할 수 있다. 나아가 AI 모델 보안 프로세스와 거버넌스를 확립하고, AI 윤리 준수 여부를 상시 점검하는 등 AI 모델 관리 체계를 고도화함으로써 기업은 안전하고 책임감 있는 AI 활용 환경을 조성할 수 있을 것이다. 따라서 기업은 본 연구에서 제안한 사이버 킬 체인 기반의 AI 모델 보호 전략을 참고하여, 자사에 최적화된 대응 방안을 수립하는 데 역량을 집중해야 한다. 특히 명령 및 제어 단계에서 강조된 AI 모델 변경에 대한 철저한 승인 절차와 감사 체계 확립이 무단 변경 차단의 핵심임을 인지하고, 각 단계별 위험 요인과 징후를 식별하여 맞춤형 대응 전략을 마련해야 할 것이다. 아울러 새로운 공격 기법에 선제적으로 대비하기 위해, 위험 관리 체계를 지속해서 점검하고 개선해 나가는 자세 또한 필요하다. AI 기술이 빠르게 발전하는 만큼, AI 보안 역량을 지속적으로 고도화하려는 기업의 노력이 그 어느 때보다 중요하다.

3.5 보안 인력 교육 및 관리

앞서 살펴본 AI 시스템 보안 강화 방안들을 실효성 있게 이행하기 위해서는 전문 인력 확보와 조직 체계 정비가 필수적이다. 기업은 AI보안 전문가를 영입하고 지속적인 교육을 통해 실무 능력을 제고해야 한다. 특히 AI 모델 개발자, 데이터 엔지니어, 보안 담당자 간 긴밀한 협업이 이루어지도록 조직 체계를 정비하고 소통 채널을 마련해야 한다. 또한 2.2절에서 언급한 삼성전자의 사례처럼, 외부 AI 서비스 이용에 따른 데이터 유출 위험을 방지하기 위해 전 직원을 대상으로 한 AI 보안 인식 제고 교육

을 정기적으로 실시해야 한다. 이 교육에서는 민감한 기업 정보를 외부 AI 서비스에 노출시키지 않도록 주의할 것과, 만약 자체적인 AI 시스템을 구축하여 운영할 경우 어떤 보안 지침을 준수해야 하는지 등 실무에 필요한 내용을 다뤄야 한다. 이를 통해 전사적인 AI 보안 문화를 정착시킬 수 있다. 나아가 우수 인력 유치와 유지를 위한 동기 부여 체계를 강화하고, 직무 순환과 경력 개발 기회를 제공함으로써 AI 보안 조직의 지속 가능성을 확보할 필요가 있다.

3.6 제안 방안의 한계 및 향후 과제

본 연구에서 제안한 AI 시스템 보안 강화 방안들은 나름의 타당성과 실효성을 갖추고 있으나, 실제 적용 과정에서는 한계와 어려움이 예상된다. 적대적 학습이나 모델 경량화 등의 기법은 개발 비용과 성능 저하의 위험성을 내포하고 있으며, 기업별 특성에 맞는 선별적 접근이 필요하다. 제안된 AI 보안 체계가 내부자 위협에 특화되어 있다는 점도 한계로 지적될 수 있어, 내외부 위협을 아우르는 통합적 보안 체계로의 발전이 요구된다. 무엇보다 AI 기술 자체가 빠르게 진화하고 있는 만큼, 제안 방안들도 지속적인 업데이트와 보완이 필요하다. 이를 위해서는 AI 보안 연구개발 투자 확대, 전문 인력 양성, 기업 차원의 대응 체계 구축 등 장기적 노력이 뒷받침되어야 한다. 본 연구는 이러한 한계점들을 인식하고 향후 보완해야 할 방향을 제시하는 데 그 의의가 있다. AI 기술의 중요한 전환기를 맞아, AI 보안 분야의 심화 연구가 더욱 활발히 이루어져야 한다.

4. 결론

AI 시스템이 기업의 핵심 자산으로 인식되면서, AI 모델과 데이터에 대한 보안 강화가 매우 중요해졌다. 기존의 보안 체계로는 AI 고유의 취약점과 새로운 공격 기법에 효과적으로 대응하기 어려우므로, AI에 특화된 보안 기술과 체계 구축이 시급하다.

기술적으로는 적대적 훈련, 모델 경량화 등을 통해 AI 모델의 견고성을 높이고, 데이터 프라이버시를 강화해야 한다. 또한 설명 가능한 AI로 모델의 투명성을 제고하고, 지속적인 모니터링으로 이상 징후를 조기에 포착할 수 있어야 한다. 제안된 AI 보안 체계 설계도는 내부자 위협 감지에 효과적이며, 사이버 킬 체인 개념을 활용한 AI 모델 유출 방지 방안도 주목할 만하다.

AI 보안 고도화를 위해서는 전문 인력 확보와 역량 강화, 구성원 간 협업, 보안 인식 제고 활동 등이 종합적으로 이루어져야 한다. 급변하는 AI 기술 발전 속도에 맞춰 보안 체계도 지속적으로 진화해야 한다. 경영진의 강력한 의지를 바탕으로 선제적인 AI 보안 활동을 해 나간다면, 기업은 디지털 시대의 경쟁력을 한층 강화할 수 있을 것이다.

참고문헌

[1] 백승익, 임규건, 여등승, “인공지능과 사회의 변화”, 한국지능정보사회진흥원, 제23권, 제4호, pp. 3-23, 2016.

[2] S. Wamba-Taguimdje, S. F. Wamba, J. R. K. Kamdjoug, C. E. T. Wanko. “Influence of artificial intelligence (AI) on firm performance: the business value of AI-based transformation projects”, Business process management journal, Vol 26, Issue 7, 1893-1924, 2020.

[3] S. Pattanayak, V. K. Shukla. “Review of recommender system for OTT platform through artificial intelligence”, International Conference on Artificial Intelligence and Smart Systems (ICAIS), IEEE, 1-5, 2021.

[4] Y. Han, J. Chen, M. Dou, J. Wang, K. Feng. “The Impact of Artificial Intelligence on the Financial Services Industry”, Academic Journal of Management and Social Sciences, Vol. 2, No. 3, pp. 83-85, 2023.

[5] N. Nebelung, M. D. de Oliveira Santos, S. T. Helena, A. F. de Moura Leite, M. B. Canciglieri, A. L. Szejka. “Towards Real-Time Machining Tool Failure Forecast Approach for Smart

Manufacturing Systems", IFAC-PapersOnLine, Vol. 55, No. 2, pp. 548-553, 2022.
[6] 이유환. "디지털 전환 핵심기술 (ICBM+AI) 도입이 제조기업의 혁신성과에 미치는 영향 연구: 혁신전략의 조절효과를 중심으로" 기업경영연구, 제30권, 제5호, pp. 53-74, 2023.
[7] N. Gudigantala, S. Madhavaram, P. Bicen. "An AI decision making framework for business value maximization", Wiley Online Library, Vol. 44, No. 1, pp. 67-84, 2023.
[8] 보안뉴스, "MS의 인공지능 개발자들, 실수로 38TB의 데이터 노출시켜", https://boannews.com/media/view.asp?idx=122106&page=59&mkind=1&kind=1, 2023.
[9] CNN, "Google employee charged with stealing AI trade secrets", https://edition.cnn.com/2024/03/06/tech/google-employee-charged-stealing-ai-trade-secrets/index.html, 2024.
[10] AI타임스, "삼성, chatGPT 데이터 유출 후 임직원 AI 사용금지", https://www.aitimes.com/news/articleView.html?idxno=150837, 2023.
[11] Y. Kishore. "Optimizing Enterprise Conversational AI: Accelerating Response Accuracy with Custom Dataset Fine-Tuning", Intell. Inf. Manag., Vol. 16, No. 2, pp. 65-76, 2024.
[12] 박재경, 장준서. "AI 모델의 적대적 공격 대응 방안에 대한 연구", 한국컴퓨터정보학회 학술발표논문집, 제31권, 제2호, 619-620, 2023.
[13] CNN, "Finance worker pays out $25 million after video call with deepfake 'chief financial officer'", https://edition.cnn.com/2024/02/04/asia/deepfake-cfo-scam-hong-kong-intl-hnk/index.html, 2024.
[14] N. Thalpage. "Unlocking the Black Box: Explainable Artificial Intelligence (XAI) for Trust and Transparency in AI Systems", J. Digit. Art Humanit., Vol. 4, No. 1, pp. 31-36, 2023.
[15] J. D. Connor. "The Sony hack: Data and decision in the contemporary studio, Michigan Publishing", University of Michigan Library, Vol. 2, No. 2, 2015.
[16] V. A. Stafford. "Zero trust architecture", NIST special publication, Vol. 800, pp. 207, 2020.
[17] L. Martin, "Cyber kill chain", Lockheed Martin, URL: http://cyber.lockheedmartin.com/hubfs/Gaining the Advantage Cyber Kill Chain.pdf, 2014.

〔저자소개〕

최 정 완 (Jong-woan Choi)
2006년 원광대학교 전산물리학 학사
2011년 원광대학교 방사광응용 석사
2016년 원광대학교 방사광응용 박사
email : jangja21@wku.ac.kr