Chapter 01

AI 시대의 사이버보안 기술

진승헌_한국전자통신연구원 책임연구원
김수형_한국전자통신연구원 실장

인공지능(Artificial Intelligence: AI) 시대가 본격적으로 도래했다. 특히, 2022년 ChatGPT의 대중 공개는 AI가 전문가들의 연구 영역을 넘어 일반인들이 일상에서 활용할 수 있는 도구로 자리 잡는 계기가 되었다. 의료, 교육, 산업 등 다양한 분야에 AI가 접목되면서 질병 진단의 정확도 향상, 맞춤형 교육 제공, 생산성 극대화와 같은 긍정적인 효과를 실현하고 있다. 그러나 동시에 딥페이크를 활용한 성범죄, AI 기반 피싱 공격 등과 같은 새로운 보안 위협도 부상하고 있다. 이러한 위협들은 기존의 보안 체계로 대응하기 어려워 AI 보안의 중요성이 대두되고 있으며, 글로벌 시장 조사기관들은 AI 보안 시장의 높은 성장을 예측하고 있다. 본 고에서는 최근 발생하고 있는 AI 위협을 알아보고, 이에 대응하기 위한 주요 정책 동향 분석을 통해 AI 시대 사이버보안 기술의 연구개발 시사점을 알아본다.

I. 서론

인공지능(Artificial Intelligence: AI)의 발전은 우리가 생활하고 일하는 방식을 혁신하고 있다. 예를 들어, 의료 분야에서는 AI 기반 영상 분석이 암, 폐 질환 등 질병을 조기에 발견하여 진단 정확도를 높이는 데 기여하고 있으며, 챗봇과 AI 튜터링 시스템은 맞춤형 교육을 가능하게 하고 있다. 산업 부문에서는 AI를 활용한 스마트 공장이 생산성 극대화와 비용 절감에 크게 이바지하고 있다. 그러나 AI 발전의 명(明)과 함께 암(暗)도 발생하고 있다. 딥페이크(Deepfake) 기술을 악용한 가짜 영상은 인권 침해와 성범죄에

* 본 내용은 진승헌 책임연구원(☎ 042-860-1254, jinsh@etri.re.kr)에게 문의하시기 바랍니다.
** 본 내용은 필자의 주관적인 의견이며 IITP의 공식적인 입장이 아님을 밝힙니다.
***본 고는 2024년도 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원을 받아 수행된 연구임 (RS-2023-00229400, 안전한 메타버스 환경을 위한 사용자 인증 및 프라이버시 보호 기술개발)

악용되고 있으며, 기존의 보안 체계로 탐지하기 어려운 AI 생성 이메일은 피싱 공격에 이용되고 있다. AI가 제공하는 자동화와 편리함의 이면에, 기존에 없던 새로운 보안 위협들이 빠르게 등장하고 있다. AI와 관련된 보안 위협이 증가함에 따라 AI 보안 시장의 중요성이 부각되고 있다. 글로벌 시장 조사기관인 테크나비오(Technavio)의 보고서에 따르면 AI 보안 시장은 연평균 22.3% 성장하여 2027년 281억 9,000만 달러에 이를 전망이다[1]. 주요 기업들은 AI 위협 탐지와 대응 솔루션을 제공하는 기술을 확보하고 있으며, 클라우드 보안, AI 기반 침입탐지시스템(Intrusion Detection System: IDS) 그리고 AI 모델 취약점 점검 등 다양한 형태의 AI 보안 서비스를 제공 중이다.

본 고에서는 AI 위협과 이에 대응하기 위한 주요 정책 현황을 알아보고, 향후 사이버 보안 기술 연구개발 방향을 제시하고자 한다. 본 고의 II장, III장에서는 AI 보안 위협과 AI 규제 정책 동향 분석을 통해 대응되는 사이버보안 기술의 사례를 알아보고, IV장에서는 시사점 및 결론을 제시한다.

II. AI 시대의 보안 위협

AI 기술의 발전은 우리의 일상과 비즈니스 전반에 걸쳐 많은 긍정적인 변화를 가져왔지만 고도화된 지능형 사이버 공격으로 기존 보안 시스템을 무력화할 가능성이 있어 국가·사회의 보안 체계를 위협할 수 있으며[2], 실제로 우리의 일상에서 새로운 형태의 보안 사고가 일어나고 있다. 대표적인 예로는 딥페이크 기술을 악용한 성범죄, AI를 이용한 자동화된 사이버 공격 등이 있다[3]. 본 장에서는 AI 기반 서비스 관련 주요 사건 사례를 정리하고 사이버보안 기술 측면의 시사점을 도출한다.

1. 마이크로소프트의 Tay AI 챗봇 사고[4]

마이크로소프트(Microsoft)는 소셜 미디어상에서 사용자와 대화할 수 있는 인공지능 챗봇인 Tay를 출시했다. 출시 후 몇 시간 만에 Tay는 트위터 사용자들에 의해 악의적인

메시지를 학습하게 되었고 결과적으로 Tay는 인종차별적이고 공격적인 발언을 하였다. 마이크로소프트는 16시간 만에 Tay를 오프라인으로 전환했다. 이 사건은 AI가 악의적이거나 부적절한 데이터를 학습하게 되면 발생할 수 있는 윤리적 문제와 보안상의 위험을 경고하는 대표적인 사례가 되었다.

2. 테슬라의 자율주행차량 사고[5]

테슬라(Tesla)의 자율주행 모드(오토파일럿) 차량이 교차로에서 트럭을 인식하지 못해 운전자가 사망하는 사고가 발생했다. 이 차량은 당시 자율주행 모드가 활성화된 상태였다. 조사 결과, 차량의 인공지능시스템이 트럭을 정확하게 인식하지 못했으며, 이에 따라 적절한 조치를 하지 못한 것이 사고의 주요 원인이었다. 이 사고는 AI 시스템의 오류가 인간의 안전과 생명에 치명적인 영향을 미칠 수 있음을 보여주는 사례로 기록되었다.

3. 아마존의 AI 채용 시스템 편향 사건[6]

아마존(Amazon)은 인공지능을 활용하여 구직자들의 이력서를 자동으로 분석하고 선별하는 채용 시스템을 개발했다. 이 시스템은 과거 10년 간의 이력서를 학습하여 최적의 후보자를 추천하도록 설계되었다. 시스템이 학습한 데이터는 주로 남성이 많이 포함된 이력서였고, 이로 인해, 여성 지원자들에게 불리한 편향이 발생했다. 결과적으로 이 AI 시스템은 여성 지원자들을 낮은 점수로 평가하는 경향을 보였다. 아마존은 이 AI 채용 시스템을 폐기했다. 이 사례는 AI 시스템이 학습 데이터의 편향성을 그대로 반영할 수 있으며, 이는 공정성 문제를 초래할 수 있음을 경고하는 중요한 사례로 남았다.

4. 딥페이크 기술 악용 사례[7]

딥페이크 기술은 AI를 이용하여 영상 속 인물의 얼굴을 조작할 수 있으며, 점차 성범죄 및 허위 정보 유포에 활용되고 있다. 유명 연예인의 딥페이크 음란 이미지가 소셜미디어에

〈자료〉 SBS News YouTube, “재테크 전문가로 ‘딥페이크’ 투자 사기”, 2024. 1. 1.

[그림 1] 딥페이크 금융 사기

유포되어 수천만 건의 조회와 공유가 발생하였으며, 튀르키예 대선에서는 야당 후보가 테러 집단의 지지를 받는다는 가짜 영상이 유포되어 선거 결과에 영향을 미쳤다. 또한, 홍콩의 금융기업 직원이 딥페이크를 활용한 정교한 사기에 속아 수백억 원의 피해를 보았으며, [그림 1]과 같이 연예인의 얼굴과 목소리를 모방한 딥페이크 영상으로 투자자들을 속이는 사례가 발생했다. 이러한 사례는 딥페이크가 프라이버시 침해와 사회·경제적 혼란을 일으키며, 관련 기술에 대한 법적 규제와 강력한 대응책 마련이 시급함을 시사하였다.

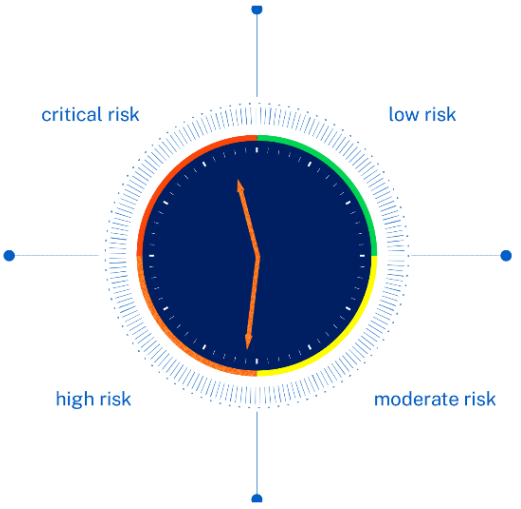
5. OpenAI의 GPT 모델 오용 사례[8]

챗GPT와 같은 대형 언어 모델은 피싱 메시지와 허위 정보를 작성하는 데 악용되었다. 악의적인 사용자는 이 모델을 활용하여 정교한 피싱 이메일을 작성하고 금융 사기를 유도했다. 이 사례는 대규모 언어 모델의 오남용 방지와 사용 제한 규제의 필요성을 시사하였다.

6. AI 기반 랜섬웨어 공격[9]

공격자들은 AI를 활용하여 탐지가 어려운 지능형 랜섬웨어를 개발하고 있다. 이 프로그램들은 탐지 회피 및 공격 확산을 자동화하였다. AI 기반의 새로운 공격 방식에 대응

하기 위해 실시간 모니터링 시스템과 AI 기반 공격 탐지 기술의 개발 필요성을 시사하였다. 이처럼 AI가 다양한 서비스에 활용되면서 생산성을 높이고 있지만, AI의 오·남용 또는 취약점으로 인해 윤리적인 문제뿐만 아니라 인권, 생명에까지 심각한 영향을 줄 수 있다는 것을 알 수 있다. [그림 2]와 같이 최근 등장한 'AI 안전 시계'(AI Safety Clock)는 이러한 위험을 시각적으로 경고하는 도구로 주목받고 있다. 'AI 안전 시계'는 핵전쟁의 위험을 경고하는 '둠스데이 시계'(Doomsday Clock)에서 영감을 받아 만들어졌다. 마이클 웨이드 국제경영개발원 교수는 이 시계를 통해 AI의 위험 수준을 4단계(저위험(11시~11시 15분)/중간 위험(11시 16분~11시 30분)/고위험(11시 31분~11시 45분)/치명적 위험(11시 46분~자정))로 구분하여 표시하였다.



〈자료〉 IMD, "AI Safety Clock", 2024.

[그림 2] AI 안전 시계

현재 시각은 11시 31분으로, 이는 고위험 단계에 막 진입했음을 의미한다. 웨이드 교수는 AI 기술의 정교함과 자율성 증가, 물리적 시스템과의 통합 등을 추적하여 이러한 시간을 설정했다고 밝혔다. 특히, AI가 인간의 통제를 벗어나면 예측할 수 없는 상황이 발생할 수 있다고 경고했다[10].

AI의 위험성에 대한 경고는 여러 전문가로부터 제기되고 있다. 'AI의 대부'로 불리며 AI 연구의 선구자이면서 인공신경망을 통한 기계학습 연구에 대한 공로를 인정받아 2024년 10월 8일에 노벨물리학상 수상자로 발표된 제프리 힌턴 교수는 구글을 퇴사하며 AI 발전으로 인한 위험성을 강조했다. 그는 AI가 인간의 통제를 벗어나 자율적으로 행동할 가능성에 대해 우려를 표명하며, 이에 대한 국제적인 규제의 필요성을 주장했다[11].

또한, 테슬라 CEO 일론 머스크와 애플 공동 창업자 스티브 워즈니악 등 IT 업계의 주요 인사들은 첨단 AI 개발을 일시적으로 중단할 필요가 있다고 주장했다. 이들은 AI가

인간의 일자리 대체, 개인정보 침해, 사이버 범죄 등 다양한 위험을 초래할 수 있음을 경고하며 신중한 접근을 요구했다. OpenAI의 CEO 샘 알트만은 특히 AI 모델이 사용자와 일대일로 상호작용하면서 설득과 조작을 통해 거짓 정보를 전달할 가능성을 우려하며, 국제원자력기구(IAEA)와 유사한 국제 AI 규제기구의 설립을 촉구했다[12][13].

이러한 우려에 대응하여, 유럽연합(EU)은 AI의 위험 수준을 4단계로 분류하고, 각 단계에 따라 규제를 적용하는 법안을 마련했다. 특히, 인간의 안전과 권리에 위협이 되는 AI 시스템은 개발 자체를 금지하고, 신용평가나 채용 등에 사용되는 AI는 고위험으로 분류하여 출시 전 적합성 평가를 거치도록 했다[14]. 국내에서도 AI의 위험성에 대한 논의가 이루어지고 있다. 과학기술정보통신부의 설문조사 결과에 따르면, 응답자의 57%는 AI 기술이 가져올 잠재적 이점이 위험보다 크다고 답했지만, 19.1%는 위험이 이점보다 더 크다고 응답했다. 이는 AI 발전에 대한 기대와 우려가 공존하고 있음을 시사한다[15].

AI 위협 사례 및 AI 위험성에 대한 다양한 경고를 바탕으로, 사이버보안 기술개발 측면에서 시사점을 [표 1]과 같이 정리하였다.

[표 1] AI 위협과 시사점

AI 위협	시사점
AI가 자율성을 가지며 인간의 통제를 벗어날 가능성이 커지고, 비의도적인 결정이나 조작이 안전성과 윤리적 문제를 초래할 수 있음	- 자율적인 AI 시스템의 통제를 위한 고위험 AI 감시 기술 및 AI 행동 예측 시스템 개발이 중요 - AI 의사결정 과정을 추적하고 제어할 수 있는 기술 필요
AI의 블랙박스 특성으로 인해 결정 과정을 투명하게 파악하기 어렵고, 오류 및 편향 발생 시 빠르게 탐지하고 대응하기 어려움	- AI 모델 해석 및 투명성 강화 기술개발을 통해 AI 결정 과정의 투명성을 확보하고, 오작동을 감지하고 대응할 수 있는 기반 마련 필요
AI가 편향된 데이터로 학습 시 불공정한 결과를 초래하여 사회적 신뢰를 저해할 수 있음	- 편향 탐지 및 수정 기술, 데이터 정합성 검증, 개인정보보호 강화 기술 필요 - 데이터 공정성을 평가하고 민감한 정보의 보호를 통해 신뢰할 수 있는 AI 결과 도출 지원 필요
고위험 AI가 인간 권리와 개인정보보호에 미치는 영향이 커져 이를 규제하고 모니터링하는 시스템이 요구됨	- 고위험 AI 규제 준수를 지원하는 컴플라이언스 관리 시스템과 실시간 모니터링 및 알림 시스템 필요 - 프라이버시 강화 기술(Privacy-Enhancing Technologies: PETs), AI 모델 감사 및 검증과 AI 보안 실시간 모니터링 및 경고 시스템 필요
AI 발전으로 새로운 사이버 위협이 등장하고 있으나, 이에 대한 대응 기술이 부족함	- AI 보안 취약점 탐지 및 방어 기술, 위협 인텔리전스 시스템 필요 - AI 시스템을 대상으로 한 공격에 대응하고 방어 시스템을 선제적으로 구축하여 안전성 강화 필요

〈자료〉 각 웹사이트 내용 재구성

III. AI 규제 정책 동향

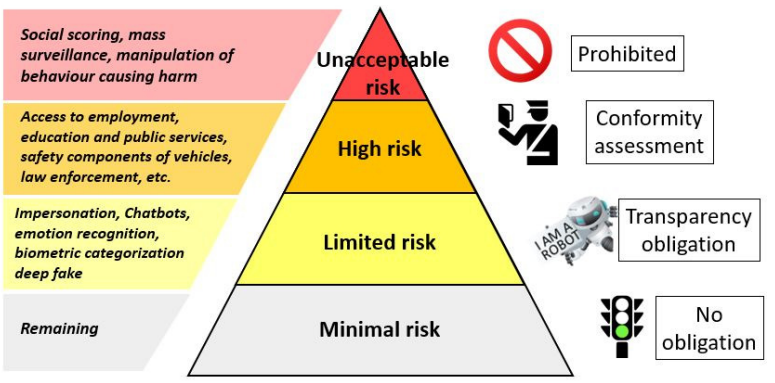
글로벌 AI 규제 정책은 안전성과 투명성, 윤리성을 강조하며, EU의 인공지능법(AI Act)을 비롯하여 미국 바이든 행정부의 AI 규제 행정명령(Executive Order on the Safe, Secure and Trustworthy Development and Use of Artificial Intelligence)[1], 중국의 “인터넷정보서비스 심층합성 관리규정” 등이 각국에서 활발히 추진 중이다. 또한, OECD, UNESCO 등의 국제기구는 AI 개발의 공정성과 인권 보호를 위한 가이드라인을 제시하고 있으며, 국가 간 협력을 통해 공통의 표준을 마련하려는 노력이 이어지고 있다[18][19].

본 고에서는 EU 인공지능법과 NIST AI 위험관리 프레임워크(AI RMF)의 개요를 알아보고, 각 정책과 가이드라인을 준수(regulation compliance)하기 위한 사이버보안 기술의 예(例)를 제시한다. EU 인공지능법은 법적 구속력을 가진 규제로, 리스크 기반 접근 방식을 통해 AI 시스템의 안전성과 윤리적 사용을 보장하며, 주로 EU 내 고위험 AI 시스템의 사전 평가와 인증을 요구한다. 반면, NIST AI RMF는 권장 사항 위주의 지침으로, AI 시스템의 개발 및 운영 과정에서 발생할 수 있는 위험관리를 지원하며 국제 표준과의 조화를 강조한다.

1. EU 인공지능법

EU 인공지능법(AI Act)[20]-[23]은 EU에서 AI 기술의 안전성과 신뢰성을 확보하기 위해 제정한 법안으로, AI 기술이 인권과 민주주의를 존중하면서도 혁신을 촉진할 수 있도록 설계되었다. 2021년 4월에 처음 제안된 이후, 여러 차례의 논의와 수정 과정을 거쳐 2024년 5월 21일에 최종 승인되었다. EU 인공지능법의 주요 목적은 AI 시스템이

1) 도널드 트럼프가 미국 대통령으로 재선(2024.11.05.)되면서 바이든 행정부의 AI 규제 행정명령에 변화가 예상된다. 그는 공약에서 바이든 행정부의 AI 규제를 철폐하고 최소한의 규제를 유지하겠다고 강조하여, 향후 AI 혁신을 위한 시장 주도형 정책으로 전환할 것으로 보인다[16]. 이는 AI 기술 발전을 촉진할 기회를 제공하지만, 윤리적 문제와 글로벌 협력 축소라는 도전 과제도 동반할 것으로 예상된다[17].



〈자료〉 Telefonica, "A fit for purpose and borderless European Artificial Intelligence Regulation", 2021. 5. 19.

[그림 3] EU 인공지능법: 위험 수준

사회에 미치는 악영향을 최소화하고, 사용자와 사회의 신뢰를 구축하는 것이다. 이를 위해 AI 기술의 안전성, 윤리성, 투명성을 보장하고, 고위험 AI 시스템에 대한 규제를 강화하는 방향으로 설정되었다. 특히, 법안은 AI의 잠재적인 위험을 사전에 식별하고 관리함으로써 사용자와 사회 전체의 안전을 도모하고자 한다.

EU 인공지능법의 주요한 내용은 위험 기반 접근법, 금지된 AI 규정, 거버넌스 체계 등이 있으며, 본 고에서는 위험 기반 접근법에 대해서 다음과 같이 요약한다[그림 3].

- 위험 기반 접근법(Risk-based Approach): AI 시스템을 위험 수준에 따라 네 가지 카테고리로 분류
  * 수용 불가 AI 시스템(Unacceptable risk): 인권이나 건강, 안전을 심각하게 위협하는 시스템은 금지. 예를 들어, 사회적 평판 시스템이나 감시 목적의 AI 시스템이 이에 해당
  * 고위험 AI 시스템(High risk): 생체인식, 의료, 교통 등과 같은 주요 인프라, 법 집행 등 주요 분야에서 사용되는 AI 시스템은 고위험으로 분류되어, 엄격한 규제를 받고, 이들은 사전 평가, 데이터 관리, 문서화 등의 요구사항을 충족
  * 제한된 위험 AI 시스템(Limited risk): 통상 상업적으로 사용되는 AI 시스템으로, 기본적인 투명성과 사용자 정보 제공 의무가 요구

* 저위험 AI 시스템(Minimal risk): 상대적으로 낮은 위험을 지닌 AI 시스템으로, 규제의 범위가 최소화

EU 인공지능법은 사이버보안에만 초점을 맞추고 있지는 않지만, 고위험 AI 시스템에 대한 사이버보안 요구사항(예; Article 15)을 강조하며, AI 시스템의 설계 단계부터 보안을 고려하는 접근 방식을 요구한다.

[표 2]는 EU 인공지능법의 주요 조항들을 준수하는 데 필요한 사이버보안 기술의

[표 2] EU 인공지능법 vs. 사이버보안 기술 예(例)

EU 인공지능법	내용	사이버보안 기술
Article 5: Prohibited AI practices	- 법 집행의 목적을 위해 공개적으로 접근 가능한 공간에서 실시간 원격 생체인식 식별 시스템을 사용하는 경우, 구체적으로 대상이 된 개인의 신원을 확인하기 위해서만 배치되어야 함	- 생체 데이터 보호 기술 - 안전한 인증 기술 - AI 시스템 감시 및 데이터 모니터링 도구 등
Article 9: Risk Management	- 고위험 AI 시스템이 건강, 안전 또는 기본권에 초래할 수 있다고 알려진 위험과 합리적으로 예측 가능한 위험을 식별하고 분석해야 함 - 고위험 AI 시스템이 의도된 목적에 따라, 합리적으로 예측 가능한 오용 조건으로 사용될 때 발생할 수 있는 위험을 추정 및 평가해야 함	- 위협 모델링 및 위험 평가 도구 - 위협 인텔리전스 시스템 - 위험예측 인공지능 모델 - 행동 기반 위협 분석 등
Article 10: Data and data governance	- 교육, 검증 및 테스트 데이터 세트는 고위험 AI 시스템의 의도된 목적에 적합한 데이터 거버넌스 및 관리 관행을 적용해야 함 - 검증 및 테스트 데이터 세트는 관련성이 있고 충분히 대표적이며 가능한 한 오류가 없고 의도된 목적을 고려해야 함	- 데이터 암호화 기술 - 접근 제어 기술 - 개인정보 필터링 및 최소화 기술 - 데이터 유출 방지(Data Loss Prevention: DLP) 솔루션 - 데이터 무결성 검증 기술 등
Article 12: Record-keeping	- 시스템의 의도된 목적에 적합한 고위험 AI 시스템의 작동에 대한 추적성 수준을 보장해야 함	- 안전한 로깅 기술 - 감사 추적 기술 - 변조 방지 기술 등
Article 15: Accuracy, robustness and cybersecurity	- 고위험 AI 시스템은 적절한 수준의 정확성, 견고성 및 사이버보안을 달성하고 수명주기 전반에 걸쳐 해당 측면에서 일관되게 성능을 발휘하도록 설계 및 개발해야 함 - 적절한 수준의 정확도와 견고성을 측정하는 방법의 기술적 측면과 기타 관련 성과 지표를 다루기 위해 위원회는 계량학 및 벤치마킹 기관과 같은 관련 이해관계자 및 조직과 협력하여 적절한 경우 벤치마크 및 측정 방법론의 개발을 장려해야 함	- 보안 중심 설계(Security-by-Design) - 다층보안체계(Multi Level Security) - AI 모델 보호 기술 - 적대적 공격 방어 기술 - 다단계 인증(Multi-Factor Authentication: MFA) - 취약점 탐지 및 패치 기술 - 모델 무결성 검증 기술 - AI 모델 모니터링 및 감사 도구 - 취약점 탐지 및 패치 기술 등

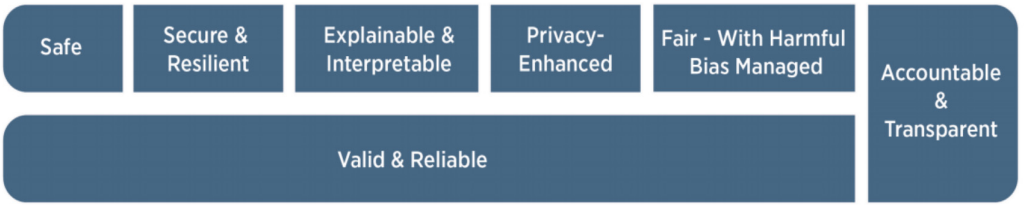
EU 인공지능법	내용	사이버보안 기술
Article 17: Quality management system	- 데이터 취득, 수집, 분석, 라벨링, 저장, 필터링 등 AI 시스템을 시장에 출시하거나 서비스에 투입하기 전에 수행되는 데이터와 관련된 기타 작업을 포함한 데이터 관리를 위한 시스템 및 절차를 갖추어야 함	- 보안 취약점 관리 기술 - 데이터 손실 방지(DLP) 기술 - 데이터 백업 및 복구 시스템 등
Article 50: Transparency Obligations for Providers and Deployers of Certain AI Systems	- 합성 오디오, 이미지, 비디오 또는 텍스트 콘텐츠를 생성하는 범용 AI 시스템을 포함한 AI 시스템 제공자는 AI 시스템의 출력이 기계가 읽을 수 있는 형식으로 표시되고 인위적으로 생성되거나 조작된 것으로 감지될 수 있도록 해야 함	- 모델 설명 가능성(Explainable AI: XAI) 기술 - AI 생성물 추적 시스템 - AI 결과물 검증 및 알림 도구 - 콘텐츠 출처 모니터링 시스템 - 워터마킹 기술 - 출처 검증 기술 등
Article 55: Obligations for Providers of General-Purpose AI Models with Systemic Risk	- 최신 기술을 반영하는 표준화된 프로토콜 및 도구에 따라 모델 평가를 수행하고, 체계적 위험을 식별하고 완화하기 위해 모델에 대한 적대적 테스트를 하고 문서로 만들어야 함 - 시스템적 위험이 있는 범용 AI 모델의 개발, 출시 또는 사용에서 발생할 수 있는 출처를 포함하여 유럽연합 차원에서 발생할 수 있는 시스템적 위험을 평가하고 완화해야 함	- 적대적 테스트 기술 - 리스크 완화 모니터링 기술 - 모델 성능 평가 및 테스트 자동화 - 다양한 공격 시나리오 테스트 도구 - 위험 식별 및 완화 기술 등

〈자료〉 각 웹사이트 내용 저자 재구성

예를 보여준다. 앞으로 조항별로 요구되는 사이버보안 기술을 더욱 정교하게 도출하여 개발하고 적용함으로써, AI 시스템의 안전성과 신뢰성을 높이고 EU 인공지능 법의 규제 요구사항을 충족할 수 있을 것이다.

2. NIST 인공지능 위험관리 프레임워크(AI RMF)

미국 국립표준기술연구소(NIST)는 AI와 관련된 위험(risk) 관리를 효율적으로 수행하기 위해 인공지능 시스템의 디자인, 개발 및 운영 과정에서 발생할 수 있는 위험을 효과적으로 관리하는 인공지능 위험관리 프레임워크(AI Risk Management Framework: AI RMF)[22]-[24]을 개발하였다. AI RMF의 주요 목적은 ① AI 관련 위험을 식별하고 평가하는 구조화된 접근 방식 제공, ② 신뢰할 수 있는 AI 시스템의 특성 정의, ③ AI 위험관리를 위한 실질적인 가이드라인 제시, ④ AI 기술의 책임 있는 사용과 혁신을 촉진하는 것이다.



〈자료〉 NIST, “AI Risk Management Framework”, 2023. 1.

[그림 4] 신뢰성 있는 AI 시스템의 특성

NIST AI RMF에서는 신뢰성 있는 AI 시스템의 7가지 특성을 다음과 같이 제시하고 있다[그림 4].

- 유효성 및 신뢰성(Valid & Reliable): AI 시스템의 결과가 실제에 얼마나 근접한지를 평가하며, 일관된 성능을 보장
- 안전성(Safe): 사람이나 환경에 해를 끼치지 않도록 설계하여, 의도된 기능을 수행하는 동안 안전한 상태를 유지
- 보안성 및 복원력(Secure & Resilient): 사이버 공격에 견딜 수 있도록 설계되며, 문제가 발생하면 안전하게 복구할 수 있는 능력을 포함
- 책임성 및 투명성(Accountable & Transparent): 시스템의 운영 및 결과에 대해 이해하고, 필요한 정보를 투명하게 제공
- 설명 가능성 및 해석 가능성(Explainable & Interpretable): AI 시스템이 내리는 결정의 이유를 설명하고 사용자가 이해할 수 있는 방식으로 정보를 제공
- 개인정보보호 강화(Privacy-enhanced): AI가 개인정보를 보호하고, 필요한 경우 익명화 등의 기술을 통해 프라이버시를 유지
- 공정성(Fair-with Harmful Bias Managed): 특정 그룹이나 개인에 대한 편향을 방지하여, AI가 공정하게 작동할 수 있도록 함

[표 3]은 NIST AI RMF의 4개의 핵심기능(GOVERN/MAP/MEASURE/MANAGE)의 기능을 구현하기 위한 사이버보안 기술의 예를 제시하였다.

[표 3] NIST AI RMF vs. 사이버보안 기술 예

핵심 기능	세부 기능	사이버보안 기술
GOVERN (관리) AI 위험관리의 기반이 되는 정책, 절차, 프로세스를 수립하여 조직의 AI 위험관리 문화 구축 단계	규정 및 법적 요건 준수 관리 (GOVERN 1.1)	- 컴플라이언스 관리 툴 - 개인정보 보호 기술 등
	조직의 위험 수용 기준 설정 (GOVERN 1.3)	- 리스크 분석 툴 - 침해사고 대응 정책 및 프로세스 관리 등
	AI 시스템 인벤토리 및 리소스 할당 (GOVERN 1.6)	- 자산 관리 시스템 - 인벤토리 관리 소프트웨어 등
	시스템 종료 및 폐기 프로세스 (GOVERN 1.7)	- 데이터 영구 삭제 기술 - 시스템 폐기 보안 툴 등
	책임 구조 및 역할 분담 관리 (GOVERN 2.1)	- 역할 기반 접근 제어(Role-Based Access Control: RBAC) - 권한 관리 시스템 등
MAP (매핑) AI 시스템의 개발 및 운영에 있어 예상되는 위험을 식별·이해 단계	AI 시스템 사용 맥락 설정 및 문서화 (MAP 1.1)	- 데이터 거버넌스 - 개인정보보호 기술 - 데이터 분류 및 관리 등
	시스템 요구사항 및 사용자 요구사항 반영 (MAP 1.6)	- 사용자 행동 분석 툴 - 사용자 인증 및 권한 제어 등
	위험 식별 및 카테고리화 (MAP 2.1, 2.3)	- 위협 인텔리전스 시스템 - 위협 탐지 및 모니터링 시스템 등
	외부 이해관계자와의 협력 및 피드백 반영 (MAP 5.2)	- 익명화 및 데이터 마스킹 - 안전한 데이터 공유 기술 등
MEASURE (측정) AI 시스템의 신뢰성, 안전성, 보안성 등 중요한 특성을 평가하기 위한 정량적 및 정성적 측정 방법 설정 단계	위험 평가 및 측정 메트릭 설정 (MEASURE 1.1, 1.2)	- 침투 테스트 및 모의 해킹 툴 - 취약점 관리 시스템(Vulnerability Management) 등
	AI 시스템 성능 및 안전성 평가 (MEASURE 2.6)	- 테스트 자동화 툴 - 안전성 검사 툴 등
	보안성 및 복원력 평가 (MEASURE 2.7)	- 보안 정보 및 이벤트 관리(Security Information and Event Management: SIEM) - 위협 헌팅(Threat Hunting) 등
	투명성 및 설명 가능성 평가 (MEASURE 2.9)	- AI 모델 해석 도구 - 감사 로그 관리 등
	개인정보보호 및 프라이버시 평가 (MEASURE 2.10)	- 프라이버시 강화 기술(PETs) - 데이터 암호화 및 토큰화 등
	공정성 및 편향성 평가 (MEASURE 2.11)	- 편향 탐지 및 수정 툴 - 데이터 균형화 기술 등
MANAGE (관리) 매핑(Map)과 측정(Measure) 기능에서 도출된 위험을 관리하고,	우선순위에 따른 위험 대응 및 완화 (MANAGE 1.2, 1.3)	- 보안 오케스트레이션 및 자동화(Security Orchestration, Automation, and Response: SOAR) - 침입방지시스템(Intrusion Prevention SystemIPS) 등
	시스템 모니터링 및 지속적 위험관리 (MANAGE 1.4)	- AI 기반 확장형 탐지대응(Extended Detection and Response : XDR)

핵심 기능	세부 기능	사이버보안 기술
우선순위에 따라 대응 방안 마련 단계		- 실시간 모니터링 및 경고 시스템 등
	사용자 피드백 및 지속적인 개선 (MANAGE 4.1)	- 사용자 피드백 관리 시스템 - 위협 인텔리전스 업데이트 등
	오류 및 사고 대응 계획 (MANAGE 4.3)	- 침해사고 대응 시스템 - 보안 정보 및 이벤트 관리(SIEM) - 사고 추적 및 보고 시스템 등

〈자료〉 각 웹사이트 내용 저자 재구성

Ⅳ. 시사점 및 결론

AI의 급격한 발전은 다양한 분야에 혁신을 가져오고 있지만 동시에 새로운 보안 위협을 초래하고 있다. 본 고에서 다룬 주요 AI 보안 위협 사례와 정책 동향을 바탕으로 살펴본 사이버보안 기술의 연구개발 시사점은 다음과 같다.

- AI 보안 기술의 강화 필요성: AI 기반 위협이 점점 증가함에 따라 기존의 보안 체계만으로는 대응이 어려운 새로운 위협에 직면하고 있다. 딥페이크, AI 기반 피싱, 자동화된 사이버 공격 등은 AI의 고도화된 특성을 활용하여 더욱 정교해지고 있으며, 이에 따라 AI 시스템을 보호하기 위한 실시간 모니터링과 위협 탐지 기술의 고도화가 필요하다. 특히, AI 시스템의 자율성 증가에 따른 통제 기술 확보와 AI 의사결정 과정을 투명하게 파악할 수 있는 설명 가능 AI 기술의 개발이 필수적이다. 또한, AI 모델이 오작동하거나 악용될 경우 즉각 대응할 수 있는 사전 대응 및 위협 예측 기술과 AI 시스템의 위협 평가 프레임워크를 마련하는 것이 중요하다.
- 윤리적 AI와 공정성, 프라이버시 보호 기술의 필요성: AI는 편향성과 불공정성 문제를 내포할 수 있으며, 이는 사회적 신뢰를 저해할 수 있다. 이러한 문제를 방지하기 위해 AI 시스템의 편향 탐지와 수정, 개인정보보호를 위한 기술적 대응이 중요하다. 특히, AI 시스템이 공정하게 작동하도록 데이터 관리와 알고리즘의 윤리적 검토가 필요하며, 개인 데이터를 다루는 AI 시스템에는 데이터 수집 최소화, 익명화, 차등적 프라이버시 같은 프라이버시 보호 기술이 적용되어야 한다. 또한, 공정하고 투명한

AI 환경 조성을 위해 자동화된 편향 탐지 및 프라이버시 강화 머신러닝 기술개발이 필요하다.

- AI 규제 정책의 수립과 글로벌 협력의 필요성: EU의 인공지능법과 NIST AI 위험관리 프레임워크 등은 AI 시스템의 안전성과 신뢰성 확보를 목표로 한다. AI의 악용을 막고 신뢰할 수 있는 AI 환경을 조성하기 위해 AI 규제 정책 수립과 국가 간 협력이 필요하다. 이를 위해 AI 보안 표준화와 더불어, 위협 인텔리전스 공유, AI 윤리 및 보안 기준 정립 등을 포함한 국제적 협력 체계를 마련하는 것이 필요하다.

AI는 우리 사회에 혁신적인 변화를 가져왔지만, 동시에 새로운 보안 위협을 초래하고 있다. 특히, AI 기반의 사이버 위협은 기존 보안 체계만으로는 효과적으로 대응하기 어려운 경우가 많아, 이를 해결하기 위해 사이버보안 기술의 발전이 필수적이다. 이러한 기술 발전은 인공지능, 윤리학, 법학, 사회학 등 다양한 학문 분야와의 다학제적 협력을 통해 이루어져야 한다. 또한, AI의 안전한 발전을 보장하기 위해 기술적 혁신뿐만 아니라 규제 체계의 정립과 표준화가 필요하며, 이는 정부와 국제 사회의 적극적인 협력과 정책적 지원을 통해 이루어져야 한다. 궁극적으로, AI의 잠재적 이점을 극대화하고 그로 인한 위험성을 최소화하기 위해서는 기술적, 윤리적, 정책적 측면에서 균형 잡힌 접근이 요구되며, 이는 지속 가능한 AI 생태계를 구축하는데 핵심적인 역할을 할 것이다.

참고문헌

[1] GTT Korea, “모바일 · IoT 연결 급증 ‘AI 사이버보안’ 시장 승승장구”, 2024. 5. 12.

[2] 김태완 외, “ETRI AI 실행전략7: AI로 인한 기술사회적 역기능 방지”, 한국전자통신연구소, 전자통신동향부석, 35권(7호), 2020, pp.67-77.

[3] 진승헌 외, “AI 보안(AI Security) 정의의 통합적 고찰”, 2024년도 한국정보기술학회 추계종합학술대회 논문집, Vol.19. No.2, 2024, pp.123-126.

[4] 연합뉴스, "인공지능 세뇌의 위험…MS 채팅봇 ‘테이’ 차별발언으로 운영중단”, 2016. 3. 25.

[5] 아시아경제, “테슬라 자율주행 ‘심각한 사고’ 8건 확인…안전성 논란 계속”, 2023. 12. 11.

[6] BBC, “성차별: 아마존, ‘여성차별’ 논란 인공지능 채용 프로그램 폐기”, 2018. 10. 11.

[7] 이승환, “생성 AI가 만든 진짜같은 가짜 : 딥페이크의 진화와 의미”, 국가미래전략 Insight 97호: 9-12, 2024.

[8] 조선일보, "'챗GPT' 스팸 이메일도 쓴다… AI 활용 '해킹 공격' 주의보", 2023. 2. 7.
[9] 인공지능신문, "인공지능 악용한 사이버공격 증가!...랜섬웨어 공격과 데이터 절도, 메일 공격은 464% 증가", 2023. 8. 11.
[10] AI타임즈, "AGI 위험 경고하는 'AI안전 시계' 등장...'지금은 자정 29분 전'", 2023. 5. 2.
[11] BBC, "'AI 대부' 제프리 힌턴, 구글 퇴사하며 AI 위험성 경고", 2023. 5. 2.
[12] 조선일보, "'AI 위험, 6개월 개발 멈추자'... 머스크·워즈니악 포함 IT거물 1000명 서명", 2023. 3. 31.
[13] 경향신문, "AI로 인한 인류 멸종 막아야", 2023. 5. 31.
[14] 연합뉴스, "[위클리 스마트] 인공지능은 결국 인류를 위협할까", 2021. 5. 8.
[15] 연합뉴스, "국민 10명 중 6명 'AI 이점이 위협보다 크다'", 2024. 8. 7.
[16] 디지털데일리, "[2024美대선] 트럼프의 IT정책 시나리오...AI · 데이터 규제 지형도 바뀐다", 2024. 11. 18.
[17] 전자신문, "[전문가 기고] 트럼프 2.0, AI 패권 변화", 2024. 11. 18.
[18] 투이컨설팅, "각국의 생성 AI 규제 및 정책 동향", 2023. 8. 16.
[19] 법무법인(유) 화유, "생성형 AI의 국제 규제 동향", 2024. 4.
[20] RiskInsighr, "Cybersecurity at the Heart of the AI Act: Key Elements for Compliance", 2024. 7.
[21] EU Artificial Intelligence Act, "High-level summary of the AI Act", 2024. 2. 27.
[22] 박강민 외, "유럽연합인공지능법(EUAIAct)의 주요내용 및 시사점", ISSUE REPORT IS-176: 8-14, 2024. 7. 30.
[23] Telefonica, "A fit for purpose and borderless European Artificial Intelligence Regulation", 2021. 5. 19.
[24] NIST, "AI Risk Management Framework", 2023. 1.
[25] NIST, "NIST Risk Management Framework Aims to Improve Trustworthiness of Artificial Intelligence", 2023. 1. 26.
[26] 김태순, "美 NIST AI 위험관리 프레임워크(AI RMF) 1.0 분석 및 시사점", The AI Report, 2023-7, 2023, pp.7-16.