J. Internet Comput. Serv.
ISSN 1598-0170 (Print) / ISSN 2287-1136 (Online)
http://www.jics.or.kr


# 국방 XAI와 적대적 공격에 대한 최신 동향 연구

# A Study of Recent Research Trends on XAI in Defense Domain and Adversarial Attacks

이 상 호[1] 조 영 호[1,*]
Sang-ho Lee Youngho Cho

**요 약**

최근 딥러닝 (Deep Learning) 기술의 눈부신 발전에 따라 AI (Artificial Intelligence) 기술은 사회 전반에 활발히 적용되고 있다. 하지만 딥러닝 알고리즘과 네트워크의 복잡한 구조는 딥러닝 모델 추론의 설명이 어렵다는 문제가 있다. 특히, 국방과 의료와 같이 인간의 생명과 밀접한 관련이 있는 분야에 있어 신뢰성과 안정성이 보장되지 않은 AI의 무분별한 적용은 큰 문제가 될 수 있다. 이를 극복하고자 설명가능한 인공지능(XAI: eXplainable AI) 기술이 출현하였다. 반면에, AI와 XAI 기술은 적대적 공격(Adversarial Attack)에 취약한 것으로 잘 알려져 있다. 따라서, 본 논문에서는 우선 XAI의 개념과 분류를 제시하고, 국방 분야에 어떻게 적용될 수 있는지를 소개하며, 끝으로 AI와 XAI의 성능과 신뢰성을 위태롭게하는 최신 적대적 공격에 대한 동향을 살펴보고자 한다. 이를 통해, XAI를 국방 분야에 적용할 때에 적대적 공격의 위험성을 이해하고 대응방안을 함께 모색하도록 돕고자 한다.

☞ 주제어 : 인공지능, 설명가능한 인공지능, 딥러닝, 적대적 공격, 국방 인공지능

**ABSTRACT**

Nowadays, with the remarkable advancement of deep learning technology, AI (Artificial Intelligence) is being actively applied across various domains of society. However, the complex structure of deep learning algorithms and networks presents a problem in explaining the inference process of deep learning models. This is particularly problematic in critical domains such as defense and healthcare, where human lives are closely involved, as the indiscriminate application of AI without guaranteed reliability and stability can lead to significant issues. To overcome this, explainable AI (XAI) technology has emerged. On the other hand, both AI and XAI technologies are known to be vulnerable to adversarial attacks. Therefore, this paper first presents the concept and classification of XAI, introduces its applications to the defense and military domains, and finally surveys the latest trends in adversarial attacks that significantly threaten the performance and reliability of AI and XAI. Through this, we aim to help understand the risks of adversarial attacks when applying XAI in the defense domain and explore possible countermeasures.

☞ keyword : XAI, Adversarial Attack, Military Artificial Intelligence

## 1. 서 론

최근 과학기술의 발전 속도와 영역이 과거에 비할 수 없을 정도로 빨라지고 넓어지고 있는 와중에 현대인들에게 가장 큰 영향을 주는 분야는 인공지능(AI: Artificial Intelligence)이라고 할 수 있다. 특히, 2016년 인공지능 프로그램인 알파고가 세계적인 바둑기사인 이세돌을 이기며 AI에 대한 대중의 관심은 폭발적으로 늘어났다. AI에 대해 세계 각국이 얼마나 지대한 관심을 가지고 있는지 알 수 있는 부분은 바로 이 분야에 대한 투자현황이다. 한국지능정보사회지능원의 보고자료[1]에 따르면, 미국은 2023년 AI분야에 약 114조원, 중국은 약 14조원, EU는 약 18조원을 투자하는 등 엄청난 예산을 들이고 있다. 이는 2015년에 비해 각각 4배, 3배, 6배가 증가한 금액으로 각국이 얼마나 AI 분야에 대한 관심사가 높은지 알 수 있다.

하지만 모든 기술이 완벽하지 않듯이 AI 기술 역시 몇 가지 취약한 점이 있다. 대표적으로 AI 모델은 프로그래머가 데이터를 기반으로 제작한 알고리즘이기 때문에 윤리적, 법적인 사항에 대한 책임소재 문제가 있다. 또한 도출된 결과값을 추론한 과정을 알 수 없는 “블랙박스” 문제가 있기에 무조건적으로 수용하기에는 신뢰성이 보

1 Department of Defense Science (Computer Engineering and Cyberwarfare Major), Graduate School of Defense Management, Korea National Defense University, Nonsan, 33021, Korea.
* Corresponding author (younghocho@korea.kr)
[Received 13 October 2024, Reviewed 22 October 2024, Accepted 04 November 2024]

장되지 않는다는 점을 고려해야한다. 따라서 윤리적, 법적 기준을 충족하였는지가 주요한 의료, 군사, 금융 등과 같이 인간의 생명을 다루고 사회적 파급력이 큰 분야에서는 AI에 의한 의사결정 과정이 어떻게 도출되었는지를 알 수 있어야 한다. 이러한 문제를 해결하지 못한다면 AI는 이들 분야에서 제한적으로 적용되어야 한다.

이러한 취약점들을 해결하기 위해 등장한 기술이 설명가능한 인공지능(XAI: eXplainable AI)이다 [2, 4], XAI는 AI가 가진 블랙박스 문제를 해결하기 위해 결과 도출 간 어떤 과정이 이루어졌는지 사용자에게 설명하는 기술로서 사용자들이 AI 시스템에 더욱 신뢰를 가질 수 있도록 하는 것을 목표로 한다. XAI는 2017년 미국의 국방부 예하 고등연구계획국(DARPA: Defence Advanced Research Projects Agency)에서 최초로 그 필요성을 인식하고 개발하기 시작한 기술이며[2], 현재는 세계 여러 대학 및 다양한 민간기업에서 연구 개발을 활발히 하고 있다.

한편, 국방 분야에서도 AI 뿐만아니라 XAI를 국방 분야에 도입하고자 하는 많은 시도와 계획이 있다. 이는 현대전에 있어 최신 기술의 도입이 게임체인저가 될 수 있음을 고려한다면 당연한 일이라고 보여진다. 또한 군은 전장 관리 정보뿐만이 아닌 방대한 양의 인적 및 물적 자원에 대한 정보를 수집, 저장, 관리, 유통한다. 따라서 이를 관리하는데 XAI를 사용한다면 비용 감축과 동시에 병력 감축에 따른 부정적인 영향이 감소될 것이다. 이러한 XAI의 긍정적인 측면을 고려하여 대한민국 '24-'38 국방기술기획서 일반본[3]을 보면 국방전략기술 10대 분야 중 AI 분야가 첫 번째를 차지하고 있으며 그 세부 목표로는 전장 정보 분석과 인적/물적 자원 관리 등이 있다.

반면에, XAI를 국방 군사 분야에 도입하는데 있어 반드시 고려해야할 점이 있다. AI가 적대적 공격(adversarial attack)에 취약한 것으로 알려져 있는 것처럼, XAI도 적대적 공격의 위험에서 자유롭지 못하다. 예를 들어, 최신 딥러닝 모델 기반의 객체 인식과 분류 기능을 바탕으로 자동으로 공격 임무를 수행하는 무인기를 운용한다고 가정할 때, 적의 적대적 공격의 영향으로 특정한 공격 목표를 정확히 타격하지 못하거나 오히려 민간 객체를 오인 폭격하게 되는 경우 심각한 문제가 발생할 수 있다.

따라서 본 연구에서는 우선 XAI의 개념과 분류를 제시하고, 국방 분야에 어떻게 적용될 수 있는지를 소개한 후 XAI의 성능과 신뢰성을 위태롭게 하는 최신 적대적 공격에 대한 동향을 정리하여 기술한다. 기존의 연구들에서는 주로 국방 분야에 있어 어떻게 XAI를 적용하고 도입할지에 대한 논의에 한정되었으나, 본 연구에서는 사이버보안 측면에서 XAI에 대한 최신 적대적 공격 위험성을 살펴본 후 국방분야에서의 대응방안을 제안한다.

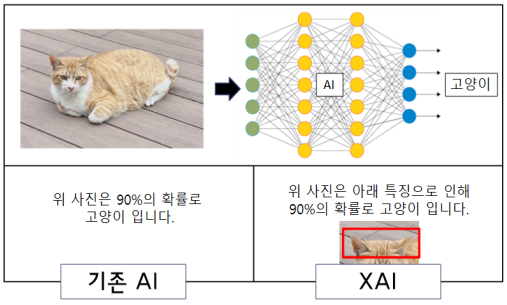
이후 논문 구성은 다음과 같다. 2장에서는 XAI를 소개하고, 6대 전투수행 기능에 따라 국방 XAI의 국내외 적용 현황을 살펴본다. 3장에서는 XAI에 대한 최신 적대적 공격 기법들을 소개하고 4장에서는 XAI와 관련된 시사점을 알아본 후 5장에서 결론을 맺는다.

## 2. XAI 소개 및 국방 적용 현황

### 2.1 XAI 개요

서론에서 언급했듯이 XAI는 AI가 가진 블랙박스 문제의 한계점을 보완하기 위해 개발된 기술이다. 2017년 DARPA는 XAI에 대한 연구 프로그램을 시작하며 XAI의 정의를 '사용자에게 이유를 설명하고, 강점과 약점을 특징짓고, 미래에 어떻게 행동할지에 대한 이해를 전달할 수 있는 AI시스템'으로 정의하면서, 그 목표로는 '효과적인 설명을 사용하여 인간이 더 잘 이해할 수 있는 AI시스템을 만드는 것'이라 발표했다 [4].

XAI를 기존의 AI와 비교하면 그림 1에서와 같다. AI는 그림의 좌측과 같이 추론 근거의 설명없이 단순히 결과값만 제시하지만, XAI는 우측과 같이 결과값을 도출하게 되는 이유(특징)를 같이함께 제시하므로 사용자는 AI 모델의 추론 근거를 확인할 수 있게 된다.




*출처 : DARPA 자료를 한글화하여 재구성 [4]

(그림 1) 기존 AI와 XAI의 차이점

(Figure 1) Differences between traditional AI and XAI

XAI는 목적에 따라 다양한 모델을 사용하므로 모델의 특징에 따라 분류할 수 있다. 여러 분류 기준 중 바이샥 벨(Vaishak Belle) 등[5]에 따르면, 표 1과 같이 크게 2가

지(이해가능성과 범용성)의 분류 기준이 있다.

(표 1) XAI의 분류 기준
(Table 1) Classification criteria of XAI

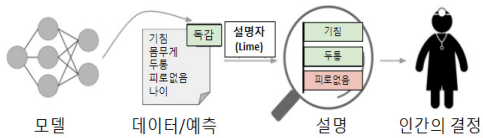
| 구 분 | 분 류 |
|---|---|
| 이해가능성 | 투명적(Transparent) |
| | 불투명적(Opaque) |
| 범용성 | 모델 독립적(Agnostic) |
| | 모델 종속적(Specific) |

XAI는 AI모델을 이해가능성 측면에서 구분한 투명한 AI 모델과 불투명한 AI모델로 나누며 투명한 모델은 구조가 단순하여 XAI의 추가적인 설명이 필요하지 않지만 불투명한 모델은 예측과정을 직관적으로 이해하기 어렵기 때문에 XAI가 필요하다. 불투명한 모델은 매개변수와 비선형성이 증가하여 정확성은 올라가지만 투명성은 낮아지기 때문에 사람이 결과를 이해하기 힘들어져 예측을 한 후 이에 대해 과정을 설명하는 Post-Hoc 방식을 함께 사용한다.

또한, XAI는 범용성에 따라 모델에 독립적인 것과 종속적인 것으로 구분될 수 있다. 이는 특정 모델에만 적용할지 다른 여러 모델에도 적용할지에 따른 분류 기준이다. 모델 독립적인 경우에는 높은 범용성을 갖추게 되며, 모델 종속적인 경우에는 종속된 모델에서 사용될 때 최적화된 높은 성능이 발휘된다.

대표적인 모델 독립적 XAI 알고리즘으로 LIME(Local Interpretable Model-agnostic Explanations)[6]과 SHAP(Shapley Additive exPlanations)[7]가 있으며, 이들은 현재 활발히 연구되고 있는 동시에 여러 산업 분야에서도 효용성이 증명되고 있어 국방분야에서의 적용 가능성이 기대되므로 좀 더 자세히 소개하고자 한다.

우선, LIME [6]은 아무리 복잡한 데이터라도 특정 부분에 위치한 데이터들의 지역적 부분을 선형 모델로 근사시킬 수 있다는 가정으로부터 만들어진 알고리즘이며 작동 개념은 그림 2에서와 같다. 그림을 보면 AI 모델은 데이터를 기반으로 대상이 독감에 걸렸다는 예측을 한다. 이후 LIME은 환자의 데이터를 약간씩 변형하여 여러 개의 샘플을 생성 및 비교하여 주요 특징들의 기여도를 파악 후 최초 모델의 예측에 기여한 특징을 시각적으로 강조하여 사용자가 예측의 근거를 쉽게 이해하도록 한다. LIME의 주요한 특징으로는 앞서 언급한 바와 같이 특정 모델에 종속되지 않기 때문에 범용성이 높고, 이미지와 텍스트 데이터를 모두 처리할 수 있다.



*출처 : Marco Tulio Ribeiro 등의 자료를 한글로 재구성

(그림 2) LIME의 개별 예측 설명
(Figure 2) Explanation of Individual Predictions Using LIME

다음으로, SHAP 알고리즘[7]은 게임이론에서 Shapley 값을 이용하여 협력 게임내 각 참가자의 기여도를 파악하듯이 XAI에서도 각 특성이 결과 예측에 있어 얼마만큼의 기여도를 차지하고 있는지를 확인한다. 이를 위해서 우리가 알고자 하는 특성이 포함된 경우와 제외된 경우를 비교하여 기여도를 계산한다. 예를 들어, 4개의 특성을 가진 모든 가능한 조합을 계산하여 우리가 알고 싶은 ④번 특성이 빠진 경우와 포함된 경우 예측된 값의 차이를 평균화하여 ④번 특성이 얼마만큼의 기여를 했는지 확인하는 방식이다. 이는 각 특성의 기여도를 측정한다는 장점이 있지만 특성이 많아질수록 계산복잡도가 크게 증가되는 단점이 있어 사용할 때에 이를 고려해야 한다.

## 2.2 국방 분야 적용 현황

본 절에서는 앞에서 소개한 XAI를 국방에서 적용 가능한 또는 필요한 분야를 문헌 조사를 바탕으로 육군 교리교범에 수록되어 있는 전투수행 6대 기능(지휘통제, 정보, 화력, 기동, 방호, 작전지속지원)에 따라 분류하여 간략히 소개한다.

### 2.2.1 지휘통제 분야

지휘관은 성공적인 작전을 수행하기 위해서 참모들의 다양한 조언을 받아들여 최적의 판단을 도출해낸다. 이때 지휘통제 분야는 다른 기능으로부터 전달받은 방대한 양의 데이터를 처리해야 하므로 XAI를 활용한다면 지휘결심 과정이 더욱 순조로워질 것이다. 다만 모든 분야를 망라하기에 데이터의 양과 분야가 방대하여 기술적인 난이도가 상당할 것이다. 이에 대해 제시한 방법으로는 시스템 오브 시스템(System of systems) 형식이다. 하나의 국면 또는 분야를 분석하는 XAI를 하위 시스템으로 구성하고 이들을 다시 해석하는 XAI를 상위 시스템화하여 지휘통제 분야를 구현 가능할 것이라 제시한다 [8], [9].

#### 2.2.2 정보 분야

전장에서 아군이 행동을 하기 전에 필수적으로 고려해야 할 것은 적의 위치, 숫자, 행동 등을 비롯한 적의 상황을 파악하는 것이다. 다양한 정보 수집 자산을 통해 적의 정보를 수집하고 분석하는 과정이 필요하나, 자산이 다양해질수록 수집되는 정보 역시 많아지기에 제한된 인력으로는 신속히 반응하기 어려울 것이다. AI 모델을 적용하더라도 예측 결과만을 제공하기에 사용자가 이를 해당 결과가 도출된 근거와 타당성을 직관적으로 이해하기 어렵다는 한계가 있다. 따라서, 정보 분야에 XAI를 적용한다면 수집된 영상 혹은 이미지를 사전에 학습된 데이터에 따라 신속하게 객체를 분류하고 식별함과 동시에 판단 근거를 함께 제공할 것이다. 이 과정에서 이미지에 대한 설명이 필요하기 때문에 이미지 캡셔닝 기법을 적용하여 사용자가 보기에 즉각적으로 해석이 가능하도록 연구가 이루어질 필요가 있다 [8], [9].

#### 2.2.3 화력 분야

첩보 자산을 통해서 수집된 정보를 바탕으로 화력 분야에서는 타격의 우선순위를 결정해야만 한다. 점차 전장에 투입되는 인간의 숫자는 줄어드는 대신, 드론, 무인장비들의 숫자는 지속적으로 늘어날 것이다. 이렇게 표적의 숫자가 늘어날수록 인간이 일일이 우선순위를 부여하는 것은 점점 힘들어질 것이며 AI 모델을 적용하더라도 예측한 결과만 알 수 있을 뿐 판단 근거에 대해서는 알 수 없다. 따라서, 화력 분야 역시 XAI 기술이 필요한데, 적에 대한 정보를 바탕으로 우선순위를 정할 뿐만이 아니라 아군의 자산, 탄약 등을 동시에 고려하여 최소의 자원을 바탕으로 최고의 효과를 내는 것이 화력 분야에서 추구하는 것이 목표일 것이다. 이를 위해서 정보처리를 통한 타격 우선 순위만 제시하는 것이 아니라 도출한 결과에 대해서 설명할 수 있는 방향으로 연구가 필요하다 [8], [9].

#### 2.2.4 기동 분야

기동분야에 적용 가능한 대표적인 XAI 기술은 무인 자율시스템을 들 수 있다. 이는 인명 피해를 최소화한다는 점에서 기존과 다른 획기적인 면으로 보여진다. 다만 고려해야할 점으로 지상 장비들은 한반도 지형의 특성상 통신 난청 구역이 많기 때문에 원격 통제가 제한되는 경우가 상당할 것이다. 따라서 최초 목표를 입력하면 그 이후에는 자체적으로 판단하고 임무를 완수하는 능력이 필요하다. 예를 들어, 기동 중에 장애물 봉착시 이를 극복하여 목적지까지 이동함과 동시에 이동경로에 대한 추가적인 데이터 확보되면 최적의 경로를 다시 도출하는 기술 등이 필요할 것으로 보인다. 그러나 기존의 AI모델은 판단의 근거 없이 예측 결과만을 제공하는 한계점이 있다. 따라서 XAI 기술을 도입함으로써 경로 판단과 결정에 어떤 요인들이 영향을 미쳤는지 파악하여 사용자가 이에 대한 신뢰성을 확보할 수 있게 된다. 이를 위해서 다양한 센서들로부터 제공된 데이터를 이미지와 융합하여 분석하는 기술과 통신이 제한된 환경에서도 임무 수행을 할 수 있도록 온보드(On-Board) AI모델과 경량화된 XAI 기술을 활용할 수 있도록 연구가 필요하다 [8], [9].

#### 2.2.5 방호 분야

방호분야 역시 군사작전을 수행하는데 있어 필수적인 영역이며, 특히 사이버 방호의 중요성은 더욱 증가하고 있다. 러시아-우크라이나 전쟁의 경우만 보아도 러시아는 HermeticWiper와 WhisperGate 등 대규모의 악성코드를 통해 지휘통신망을 마비시키려 했으며, 우크라이나는 스타링크의 위성 인터넷을 바탕으로 지휘통신망을 유지하여 전쟁을 지속 중이다[10]. 이처럼 현대전에서 안전한 사이버 환경의 유지는 필수적이기 때문에 AI를 활용한 침입탐지 시스템(IDS: , Intrusion Detection System)을 활용하여 다량의 데이터를 빠르게 처리하고 있으나 AI의 특성상 탐지 결과에 대한 판단 근거를 제시하는데 있어 한계가 있어 의사결정 또는 대응에 있어 지연을 일으킬 수 있다. 따라서 XAI를 도입한다면 판단 결과에 대한 명확한 설명을 바탕으로 보다 안정적인 사이버 환경을 만드는데 도움이 될 것이다 [11].

#### 2.2.6 작전지속지원 분야

군사 작전을 하는데 반드시 필요한 분야 중 하나는 작전지속지원이다. 작전지속지원에는 물자의 보급, 수리, 물류 등 다양한 요소들로 구성되며, 이들이 원활하지 않는다면 작전은 시작부터 난항을 겪을 것이다. 실제 미 육군 군수지원국(LOGSA: Logistics Support Activity)에서는 2017년 스트라이커 장갑차의 수리를 위해서 IBM의 인공지능인 왓슨(Watson)을 도입했으며, 장갑차에 설치된 센서로부터 전달받은 데이터를 바탕으로 공급망을 분석하여 필요한 부품들은 적시에 그리고 낮은 비용으로 공급

받을 수 있는 물류 계획을 세워 많은 경제적 효용을 보았다고 한다. 이처럼 우리 군 역시 AI를 활용하여 물류와 보급을 보완할 수 있지만 블랙박스 문제로 인해 편향된 분석이 나올 것을 대비해야 한다. 따라서 XAI를 도입하여 의사 결정권자들이 납득할 수 있도록 도출한 결과에 대한 신뢰성이 동반되도록 해야한다 [12].

## 3. 적대적 공격 위협 및 최신 기술

### 3.1 적대적 공격 개요

XAI를 국방 분야에 적용하기 전에 XAI가 유추한 결과를 신뢰성과 안정성에 근거한 확인은 매우 중요하다. 적대적 공격(Adversarial attack)은 AI모델에 대한 신뢰성과 정확성 파괴를 목적으로 하는 것으로 적대적 머신러닝(adversarial machine learning)으로도 알려져 있다. 적대적 공격을 방어하는 기술은 적대적 방어라고 한다[13, 14].

Vorobeychik 등[15]에 따르면, 적대적 공격 기법은 공격 시점, 공격자가 공격 모델에 대해 보유한 정보량, 그리고 공격 목적이 무엇인지에 따라 아래의 표 2와 같이 분류할 수 있으며 이는 공격기법을 이해하고 위험성을 분석하는데 도움이 되므로 하나씩 살펴본다.

(표 2) 적대적 공격의 분류
(Table 2) Classification of Adversarial Attacks

| 구 분 | 공격 기법 |
|---|---|
| 공격 시점 | 결정 시점에서의 공격(Decision time attacks) vs. 훈련시점에서의 공격 (Training time attacks) |
| 공격자의 정보량 | 화이트박스(White-box attacks) vs. 블랙박스(Black-box attacks) |
| 공격 목적 | 타겟팅 공격(Targeted attack) vs. 신뢰성 공격 (Reliability attack) |

첫째, 공격자의 공격 시점에 따라 결정 시점에서의 공격과 훈련시점에서의 공격으로 구분된다. 우선, 결정 시점에서의 공격은 회피공격(evasion attack)이라고도 하며 모델이 학습을 완료한 후에 수행되는 것으로 모델에 입력샘플을 의도적으로 조작하여 모델의 오분류를 유도하는 적대적 예제(adversarial example) 공격이 대표적이다. 또한, 훈련시점에서의 공격은 모델이 훈련되기 이전에 학습 데이터를 의도적으로 조작하는 것으로 포이즈닝 공격(poisoning attack)이라고도 한다.

둘째, 공격자가 공격대상인 AI모델에 대해 보유한 정보의 양에 따라 화이트박스 공격과 블랙박스 공격으로 분류할 수 있다. 화이트박스 공격은 공격자가 모델의 알고리즘, 가중치 등 내부 구조를 알고 있는 경우로 모델이 가진 특정한 취약점에 대해 효과적인 공략이 가능하다. 반면에, 블랙박스 공격은 공격자가 모델에 대한 정보가 없거나 극히 제한적인 경우를 말하며, 공격 수행을 위해 공격대상 모델에 질의(query)를 통해 공격 데이터를 확보한 후 의사모델(pseudo model)을 구축하여 추가적인 분석을 통해 공격을 수행한다.

셋째, 공격 목적에 따라 표적 공격(targeted attack)과 신뢰성 공격(reliability attack)으로 구분할 수 있다. 표적공격이란 모델이 특정 클래스 i를 j(≠i)로 오분류하는 것을 목적으로 하는 것이며, 신뢰성 공격이란 모델의 전반적인 분류 정확도(classification accuracy)를 떨어뜨려 모델의 신뢰성을 훼손하는 것을 목적으로 하는 것을 말한다.

### 3.2 XAI에 대한 최신 적대적 공격 기법

본 절에서는 XAI에 대한 최신 적대적 공격 기술과 그에 대한 위험성을 다음 2가지 측면에서 소개하고자 한다.

첫째, 적대적 공격이 수행된 경우, XAI의 예측 결과에는 동일하지만 설명의 일관성을 훼손하는 연구들을 소개한다. AI에 대한 적대적 공격의 수행 결과는 설명에 대한 왜곡과 결과에 대한 왜곡으로 나올 수 있으나[16], XAI와 AI의 가장 큰 차이점인 설명 가능함에 중점을 두고자 한다.

우선, Camburu [17]는 동일한 이미지를 XAI에 제시 후 서로 다른 두 가지 질문을 했을 때 해당 모델이 상이한 설명을 근거로 모순되는 답변을 하는 경우이다. 첫 번째 질문에서 “이미지에 동물이 있는지”를 묻자 개가 있기에 “있다”라는 답변을 했고, 두 번째 질문에서 “이미지에서 허스키를 볼 수 있는지”를 묻자 개가 없기에 “없다”라는 답변을 확인했다. 이는 편향된 학습 혹은 설명 생성 과정에서의 오류를 범하는 경우에 발생한다. 다음으로, Kuppa 등[18]은 모델이 예측한 결과는 유지하되 설명에 영향을 강제로 줄 수 있는 사례를 제시하였는데, 입력되는 데이터에 인간이 인식할 수는 없지만 모델의 설명에는 충분한 영향을 끼칠 수 있을 정도의 작은 교란을 주는 방식이다. Dombrowski 등[19]도 Kuppa 등이 연구한 바와 같이 섭동(perturbation)을 통한 이미지 조작으로 설명에 영향을 주는 사례를 제시한다. 다만, Dombrowski 등과 Kuppa 등의 차이점으로는 모델의 내부구조를 알고 있는지 여부에서 차이가 난다. Kuppa 등의 경우 모델의 내부 구조를

알지 못하는 블랙박스 공격 형태로 이루어진 것으로 모델에 질의(query)를 통해 입·출력의 쌍을 확보 설명맵 분석을 통해 공격이 수행되는 반면, Dombrowski 등의 경우 모델의 내부 구조를 알고 있는 화이트박스 환경에서 이루어진다는 면에서 차이가 있다. 이러한 연구들을 통해, 적대적 예제 적용시 XAI의 가장 큰 특징인 설명에 대한 신뢰성이 훼손당할 수 있음을 알 수 있다.

둘째, 모델 독립적이고 가장 범용성이 높아 군에 사용 가능성이 높아 보이는 대표적인 XAI알고리즘인 LIME과 SHAP에 대한 적대적 공격에 대해 소개한다.

우선, Burger 등[20]은 XAIFOOLER 알고리즘을 통해 LIME 구동 간에 샘플링 비율의 변화를 통해서도 모델이 제시하는 설명이 달라질 수 있음을 보였다. 예를 들어, gets → becomes, regularly → consistently와 같이 중요도가 낮은 텍스트의 지속적인 교란을 통해 모델 예측 결과에는 변동이 없지만 예측 근거가 되는 설명에 대해 유의미한 차이를 만들어 적대적 공격에 취약함을 보였다.

다음으로 SHAP의 경우, Baniecki 등[21]에 따르면 SHAP에 의도적인 데이터 변형을 줌으로써 해석에 사용되는 특정 변수들의 영향력을 축소 혹은 과장할 수 있다는 점을 확인하였는데, 이는 모델이 예측한 결과에 대한 신뢰성을 훼손시키는 경우를 보여준다. Laberge 등[22]은 모델에 편향된 샘플링을 주어 특정 변수들의 영향력에 변동을 주는 기법을 제시하는데 Baniecki 등과 마찬가지로 예측 결과에 대한 신뢰성에 영향을 준다.

한편, LIME과 SHAP 모두 일부 변형된 입력 데이터를 통해서 특정 변수들의 기여도를 평가한다는 공통점이 있다. Slack 등[23]은 입력 데이터를 교란하는 스캐폴딩(Scaffolding) 기법을 적용하여 LIME과 SHAP이 제공하는 설명이 의미가 없도록 보이는 실험을 하였다. 약 7,000여 명의 실제 범죄 및 금융기록을 바탕으로한 실험한 결과 범죄기록에서는 인종을, 금융기록에서는 성별이라는 예측에 중요한 영향을 미치는 특정 변수들의 영향력을 낮춰버려 LIME과 SHAP의 신뢰성을 훼손하였다.

위에서 제시한 다양한 연구 결과들이 보여주듯이 LIME과 SHAP를 비롯한 XAI의 신뢰성을 훼손시킬 수 있는 여러 방법들이 존재한다. 따라서 높은 신뢰성을 요구하는 군에 도입하기 위해서는 식별된 취약성과 공격위험에 대한 대응책을 강구하여야 할 것이다.

## 4. 국방분야에서의 대응방안 제안

지금까지 XAI에 대한 개념 및 국방 적용 방안과 함께 XAI에 대한 최신 적대적 공격 기법을 살펴보았다. 국방 영역에서 XAI의 활용성이 증가함에 따라 적대적 공격에 대한 대응방안을 반드시 강구해야 한다. 따라서, 국방 XAI의 도입간 고려해야할 시사점을 몇가지 도출하였다.

첫째, XAI의 안전하고 신뢰성있는 도입과 운용에 대한 국방 차원의 표준과 제도의 마련이 필요하다. 미국의 NIST(National Institute of Standards and Technology)가 발표한 AI 위험 관리 프레임워크(AI RMF)[24]와 NISTIR 8312[25]는 AI와 XAI에 대한 표준 마련에 있어 시발점이 되는 사례이다. 해당이 문서에는 AI 시스템의 설계, 개발, 사용 및 평가 단계에서 적용해야할 원칙들을 제시하고 있다. 이를 활용하여 국방 차원에서의 표준화 작업을 통해 일관성과 체계성을 확보하고, 지속적인 보완 발전을 위해 관·군·산·학의 연구 교류 협력 체계를 구축해야겠다.

둘째, 국방에 특화된 강건한 XAI를 연구해야 한다. 특히, 적대적 공격 기법은 끊임없이 진화하기 때문에 이에 대한 심층적인 분석을 통해 방어 전략과 기술을 연구해야 한다. 공격 기법에 대한 연구는 방어 기법을 발전시킬 수 있으며, 적대적 예제를 방어하기 위한 적대적 학습(Adversarial Training)[26]이 그 사례이다. 또한, DARPA의 GARD(Guaranteeing AI Robustness against Deception) 프로그램[27]은 적대적 공격의 다양한 유형을 탐지하고 방어하기 위한 AI기술 개발에 중점을 두고 있다.

셋째, 국방 XAI의 개발 및 도입 전 신뢰성과 강건성을 평가하는 체계를 마련해야 한다. 이를 통해, 국방 XAI 체계의 운용 후 발생할 수 있는 문제와 잠재적인 위협을 식별하고 사전에 예방할 수 있다. 이러한 평가는 XAI가 사용자의 의도대로 작동하고 신뢰할 수 있는 결과를 제공하는지에 대한 기반을 제공한다. 특히, 강건성을 검증함과 동시에 사용자 중심의 평가 체계를 동시에 만들어야 하는데 이는 XAI의 성능과 설명력이 일관성 있는지, 사용자들은 유용하게 작동할 수 있는지를 확인할 수 있다.

## 5. 결 론

미래전은 첨단 과학기술의 도입 여부에 따라 달라질 것이며, 그 중 핵심기술은 인공지능이라 보여진다. 그러나 민간 사회에서 인공지능은 보편화되고 있는 반면, 국방분야에서의 활용은 아직 초기 단계에 머물러 있다. 본

논문에서는 XAI의 개념을 정의하며 국방 분야의 전투수행 6대 기능을 바탕으로 어떤 방식으로 적용 가능할지 조사하였다. 이렇게 도입된 XAI를 바탕으로 지휘관과 참모들은 신속하고 정확한 판단을 내리는데 도움을 받을 것이며, 동시에 병력손실을 줄이는 등 성공적인 군사작전에 기여할 것이다.

반면에, XAI는 일반적인 AI 모델이 그렇듯이 적대적 공격들에 취약할 가능성 높다. 이러한 공격은 XAI의 성능과 신뢰성을 심각하게 훼손시킬 수 있다. 이에 따라, 본 연구는 적대적 공격에 대한 개념과 XAI가 받을 수 있는 영향에 대해서도 조사하였다. 이런 취약점을 인지하고 보완한다면 국방 XAI의 안정성과 신뢰성을 강화할 수 있을 것이다. 이러한 노력은 XAI가 국방 영역의 핵심 기술로 자리잡고, 이를 바탕으로 AI 기반의 선진 강군으로 발전하는데 기여할 것이다.

## 참고문헌(Reference)


[1] SM KIM, Analysis of Global AI Investment Trends in Government and Private Sectors, [IT& Future Strategy 2024-3], National Information Society Agency, 2024.

[2] David et al. “DARPA’s explainable AI (XAI) program: a retrospective,” Appl AI Lett, Vol. 2, No. 4, pp. 1–11, 2021, https://doi.org/10.1002/ail2.61

[3] Korea Research Institute for Defense Technology Planning and Advancement, ‘24-’38 Defense Technology Planning Document (General Version), 2024.

[4] David Gunning, David W. Aha, “DARPA’s Explainable Artificial Intelligence (XAI) Program,” AI Magazine, Vol. 40, Issue 2, pp. 44-58, 2019. https://doi.org/10.1609/aimag.v40i2.2850

[5] Vaishak Belle, & Ioannis Papantonis, “Principles and Practice of Explainable Machine Learning,” Frontiers in Big Data, Vol. 4, pp. 4-5, 2021. https://doi.org/10.3389/fdata.2021.688969

[6] Ribeiro et al., ““Why should i trust you?” Explaining the predictions of any classifier,“ in Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining, pp. 1135-1144, 2016. https://doi.org/10.1145/2939672.2939778

[7] Lundberg Scott, “A unified approach to interpreting model predictions,” arXiv preprint arXiv:1705.07874, 2017.

[8] Choi et al., “Explainable Artificial Intelligence on the Battledomain (Military-XAI, MXAI): Research for Military Application Scenarios,” Journal of the Korea Association of Defense Industry Studies, 30(1), pp. 1-13, 2023.

[9] Donghan Oh, “Utilization of Artificial Intelligence Technology in the Military and Suggestion of XAI Technology Application Direction,“ Journal of Digital Contents Society, Vol. 23, No. 5, pp. 943-951, 2022. https://doi.org/10.9728/dcs.2022.23.5.943

[10] Lieber Institute West Point. (2022, April 19). Military networks and cyber operations in the war in Ukraine. Available at https://lieber.westpoint.edu/military-networks-cyber-operations-war-ukraine/

[11] Masud, Mohammed Tanvir, et al., “Explainable Artificial Intelligence for Resilient Security Applications in the Internet of Things,” IEEE Open Journal of the Communications Society 2024. https://doi.org/10.1109/OJCOMS.2024.3413790

[12] Eun-ah Lee, Yong-min Kim, “Introduction Plan of Explainable Artificial Intelligence in the Defense Logistics domain,“ Defense Issues & Analyses, vol. 1892(22-13), pp. 1-12, 2022.

[13] Yuan et al. “Adversarial examples: Attacks and defenses for deep learning,“ IEEE transactions on neural networks and learning systems, Vol. 30, No. 9, pp. 2805-2824, 2019. https://doi.org/10.1109/TNNLS.2018.2886017

[14] Hubert Baniecki, Przemyslaw Biecek, “Adversarial attacks and defenses in explainable artificial intelligence: A survey,“ Information Fusion Vol. 107, 2024. https://doi.org/10.1016/j.inffus.2024.102303

[15] Vorobeychik Yevgeniy, Murat Kantarcioglu Adversarial machine learning, Morgan & Claypool Publishers, 2018.

[16] Abusitta et al., “Survey on explainable ai: techniques, challenges and open issues,” Expert Systems with Applications, Vol. 255, Part C, 2024. https://doi.org/10.1016/j.eswa.2024.124710

[17] Oana-Maria Camburu, “Explaining deep neural networks,” arXiv preprint arXiv:2010.01496, 2020. https://doi.org/10.48550/arXiv.2010.01496

[18] Kuppa Aditya, Nhien-An Le-Khac, “Black box attacks on explainable artificial intelligence (XAI) methods in cyber security,” 2020 International Joint Conference on Neural Networks (IJCNN). IEEE, pp. 1-8, 2020. https://doi.org/10.1109/IJCNN48605.2020.9206780

[19] Dombrowski et al., “Explanations can be manipulated and geometry is to blame,” Advances in neural information processing systems, 32, 2019.

[20] Burger et al., “Are Your Explanations Reliable? Investigating the Stability of LIME in Explaining Text Classifiers by Marrying XAI and Adversarial Attack,” arXiv preprint arXiv:2305.12351, 2023. https://doi.org/10.48550/arXiv.2305.12351

[21] Baniecki Hubert, Przemyslaw Biecek, “Manipulating shap via adversarial data perturbations (student abstract),” Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 36, No. 11, 2022. https://doi.org/10.1609/aaai.v36i11.21590

[22] Laberge et al., “Fool SHAP with Stealthily Biased Sampling,” arXiv preprint arXiv:2205.15419, 2022. https://doi.org/10.48550/arXiv.2205.15419

[23] Slack et al., “Fooling lime and shap: Adversarial attacks on post hoc explanation methods,” Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society, pp. 180-186, 2020. https://doi.org/10.1145/3375627.3375830

[24] National Institute of Standards and Technology, Artificial Intelligence Risk Management Framework (AI RMF) 1.0. U.S. Department of Commerce, 2023. Available at https://www.nist.gov/itl/ai-risk-management-framework

[25] Reva Schwartz, Leann Down, Elham Tabassi, “A proposal for identifying and managing bias in artificial intelligence,” Draft NIST Special Publication 1270, 2021. https://doi.org/10.6028/NIST.SP.1270-draft

[26] Goodfellow, I. J., Shlens, J., Szegedy, C., “Explaining and harnessing adversarial examples,” arXiv preprint arXiv:1412.6572, 2014. https://doi.org/10.48550/arXiv.1412.6572

[27] Alvaro Velasquez, “Guaranteeing AI Robustness Against Deception (GARD),” Defense Advanced Research Projects Agency. (n.d.), DARPA. Available at https://www.darpa.mil/program/guaranteeing-ai-robustness-against-deception

## ◐ 저 자 소 개 ◑

**이 상 호(Sang-ho Lee)**
2017년 육군사관학교 경제학과(경제학사)
2024년 국방대학교 국방관리대학원 컴퓨터공학/사이버전 협동과정 재학(공학석사)
관심분야 : 사이버보안, 인공지능 보안, 적대적 공격
E-mail : sikh3402@naver.com

**조 영 호(Youngho Cho)**
1998년 공군사관학교 산업공학과(공학사)
2006년 연세대학교 대학원 컴퓨터산업시스템공학과(공학석사)
2013년 미국 메릴랜드대학교 컴퓨터공학과(공학박사)
2017년~현재 국방대학교 국방관리대학원 컴퓨터공학/사이버전협동전공 부교수
관심분야 : 사이버보안, 인공지능 보안, 적대적 공격, 스테가노그래피, 봇넷, etc.
E-mail : younghocho@korea.kr