



Received: 2026/03/26
Revised: 2026/04/06
Accepted: 2026/06/01
Published: 2026/06/30

***Corresponding Author:**

Chul Kyu Kim

Chungcheongnam-do, Gyeryong-si, 32800,
Republic of Korea

Tel: +82-42-553-2191

E-mail: chul3812@naver.com

AI 기반 무기체계의 특성을 고려한 시험평가 방법론 발전 방향 연구

Development of Test and Evaluation Methodologies for AI-Enabled Weapon Systems

김철규^{1*}, 심민건², 김도현³

¹해군 중령/해군본부 기획관리참모부 첨단전력과 첨단전력소요담당

²해군 소령/해군본부 기획관리참모부 첨단전력과 첨단전력사업담당

³해군 소령/해군본부 기획관리참모부 첨단전력과 감시정찰전력담당

Chul Kyu Kim^{1*}, Min Gun Sim², Do Hyun Kim³

¹CDR, ROK Navy/Advanced Force Requirements Officer, Advanced Force Branch, Planning & Management Staff, ROK Navy Headquarters

²LCDR, ROK Navy/Advanced Force Programs Officer, Advanced Force Branch, Planning & Management Staff, ROK Navy Headquarters

³LCDR, ROK Navy/ISR Capability Officer, Advanced Force Branch, Planning & Management Staff, ROK Navy Headquarters

Abstract

본 연구는 인공지능이 적용된 무기체계의 확률적·비결정론적 특성을 고려한 시험평가(test and evaluation) 방법론의 발전 방향을 제시한다. 기존 무기체계에 적용되던 결정론적 성능 중심 평가 방식은 지속적 학습과 환경 의존성이 강한 AI 무기체계의 특성을 충분히 반영하는데 한계가 있다. 이에 본 연구는 미국, 유럽, 중국 등 주요 국가의 AI 무기체계 평가 사례와 우리 군의 실제 적용 사례를 분석하여, 데이터 품질 평가, 알고리즘 강건성 검증, 인간 개입 기반 통제 체계(human-in-the-loop), 전 주기적 평가 개념의 필요성을 도출하였다. 이를 바탕으로 우리 군에 적합한 AI 무기체계 시험평가 방향과 제도적·기술적 개선 방안을 제언한다.

This study presents a framework for developing test and evaluation (T&E) methodologies tailored to the probabilistic and non-deterministic characteristics of AI-enabled weapon systems. Conventional deterministic evaluation approaches have limitations in assessing AI systems due to their data dependency, continuous learning, and environmental sensitivity. By analyzing T&E frameworks of major countries and case studies from the Republic of Korea's military, key limitations of current practices are identified. The study emphasizes the need for data-centric evaluation, algorithm robustness assessment, human-in-the-loop control, and lifecycle-based continuous evaluation. Based on these findings, policy and technical recommendations are proposed for improving T&E of AI-enabled weapon systems.

Keywords

인공지능(Artificial Intelligence),
인공지능 적용 무기체계
(AI-Enabled Weapon Systems),
시험평가(Test and Evaluation),
휴먼인더루프(Human-in-the-Loop),
신속시범사업(Rapid Prototyping)

1. 서론

Opne AI, Gemini, 팔란티어 등 인공지능 개발 회사의 유명세가 예사롭지 않다. 그만큼 인공지능은 우리의 일상에 점점 더 깊이 스며들고 있다. 군에서도 사회의 변화 흐름에 발맞추어 인공지능을 적용한 무기체계 확보를 위해 노력을 기울이고 있다. 국방부는 2021년 국방 인공지능 추진전략을 수립하여 발전시키고 있으며, 국방기술진흥연구소(이하 국기연)는 ‘미래국방 2030 기술전략(AI)’을 작성하여 국내기술수준을 분석하였고, 우리나라의 기술수준은 선진국대비 69% 수준으로 명시하였다. 또한, 국방 분야에서의 인공지능은 전장환경을 인식하고(전장인식), 상황을 분석하여 판단하고(자율판단), 이를 토대로 한 의사결정(지휘결심) 능력, 환경과 상호작용을 하는 물리적 지능(임무수행)을 의미한다고 정의[1]하는 등 국방 분야에서도 인공지능에 대한 관심이 구체화되고 있다.

각 군도 이러한 사회적 변화 시류에 맞추어 각 군 본부에 인공지능을 담당하는 조직의 강화와 더불어 인공지능이 적용된 전력을 확보하는데 최선을 다하고 있다. 해군의 경우 해상인공지능와치

봇, AI기반 기뢰탐지체계 등 전장인식과 연관된 전력 확보를 추진 중이다.

AI 적용 무기체계 확보가 가시화되어가는 시점에 현장의 가장 어려운 부분 중 하나는 AI가 적용된 무기체계를 어떻게 평가하여 신뢰성을 확보할 수 있는가에 대한 것이다.

전력건설 현장은 아직까지 인공지능에 대한 이해도가 부족한 상황으로 우선 인공지능 평가수행의 기본개념과 방법을 제시하는 것이 본 연구의 목표이다.

본 연구에서는 AI의 평가수행을 위한 기본개념과 방법을 제시하기 위해 국외에서의 AI 적용 무기체계 평가와 관련된 기본개념과 방법을 조사하여 제시하였고, 이를 우리 군에 적합하게 적용하기 위한 방안을 제시하였다.

2. 본론

미국 등 선진국에서는 기존의 확정적(deterministic) 성능을 보유한 무기체계와 달리 성능 발휘의 결과가 확률적(probabilistic)이고, 설치가 완료된 이후에도 지속적으로 학습하는 인공지능이 적용된 무기체계의 특성에 맞춰 새로운 평가 방법론을 정립하고 있다.

인공지능 적용 무기체계의 개발이 가장 활발한 미국의 경우 성능, 제원 등 확정적으로 수치화된 능력을 평가하는 것이 아니라 Table 1과 같이 인공지능 모델, 시스템, 인간-시스템통합 등 4개 분야를 중점적으로 평가하고 있다.

Table 1. Key focus areas of AI test and evaluation in the United States

Category	Description
Model Evaluation	Verification of data bias and logical robustness of the algorithm itself
System Integration Evaluation	Assessment of whether the AI operates seamlessly with the overall weapon system (hardware and software)
Human-System Integration	Evaluation of the level of trust operators have in AI-driven decisions
Operational Evaluation	Final performance validation in environments similar to real battlefield conditions

하지만 우리 군은 인공지능 평가에 있어 아직까지 과거에 수행했던 확정적 성능을 평가하는 방식의 관성에서 벗어나지 못하고 있다. 이로 인해 인공지능이 적

용된 무기체계를 확보하고, 평가하는 현장에서는 큰 어려움을 겪고 있으며, 하루빨리 인공지능의 특성을 고려한 평가방식의 정립이 이뤄지길 기대하고 있다.

본론에서는 인공지능 적용 무기체계 평가와 관련된 우리 군의 문제점을 고찰하고, 선진국의 평가 시스템을 설명한 후 앞으로 우리 군이 어떻게 인공지능 평가를 수행해야 하는지에 대하여 논하고자 한다.

3. 우리 군의 AI 적용 무기체계 평가의 문제점

인공지능에 대한 관심이 지금처럼 뜨거웠던 적이 없었고, 한 시대의 거대한 패러다임이 변화하는 시점이다. 우리 군도 늦은 감이 없지 않지만 최근 본격적으로 AI 적용 무기체계를 확보하기 위해 노력하고 있다. 하지만 아직 인공지능을 적용하기에는 기술적, 인식적 부분에 어려움을 겪고 있으며 특히 무기체계 획득 분야에서 가장 어렵기로 정평이 나 있는 시험평가 방법이 모호한 상황으로 아래와 같은 문제들이 있다.

첫째, 인공지능 무기체계는 확률적 성격을 보유하고 있으며, 지속적인 학습을 통해 성능향상이 가능하다는 일반 무기체계와의 차이점에도 일반 무기체계에 적용하여 평가하던 확정적(deterministic) 평가방식을 유지하고 있다. 합참은 합동작전운용성능(JROC) 작성지침에서 인공지능 평가에 대한 내용을 일부 반영하고 있다. 주요 내용은 ① 운용개념 구현을 위한 AI기술 요구능력은 주요 ROC에 반영(표적식별 00% 이상, 탐지정확도 00% 이상으로 명시) ② AI기술 구현을 위한 컴퓨팅 형태는 기술적, 부수적 성능에 반영하는 것으로 명시되어 있다[2]. 주요 ROC에 수치화된 능력을 요구하는 것이며, 아직 결정론적(deterministic) 성능평가에서 벗어나지 못하고 있음을 의미한다.

두 번째는 인공지능 적용 무기체계의 필요 능력에 따라 학습에 대한 기회 부여 즉, 학습의 시간이 필요하나 개발 기간에 학습 시간이 구체적으로 포함되어 있지 않다. 일반적으로 판단되는 AI의 학습 기간은 Table 2에서 설명하는 것 같이 수주~수년까지 광범위한 형태를 보인다.

세 번째는 인명 살상 등 생명, 위험, 안전과 직결되는 무기체계의 경우 인공지능 자체의 환각(hallucination)에 의한 비윤리적 행위가 발생할 수 있고, 이는 전쟁의 성패를 결정할 수준의 심각한 부수적 피해

Table 2. Estimated learning periods by type of weapon system

Weapon System Type (example)	Key AI Functions	Estimated Learning / Maturation Period	Remarks
Intelligent Surveillance System	Identification of intruders/equipment in CCTV footage	6 months – 1 year	Easy data acquisition, relatively fast deployment
Autonomous Aerial Drone	Path planning, obstacle avoidance, formation flight	2 – 5 years	Requires 3D spatial awareness and aerodynamics considerations
Unmanned Ground Vehicle (UGV)	Off-road navigation, ambush detection	More than 5 years	Highly complex variables (mud, forest, buildings, etc.)
Strategic Decision AI	Wargame simulation, commander decision support	Continuous	No fixed endpoint, requires continuous data updates

* The values presented are approximate estimates; appropriate learning periods should be refined based on accumulated deployment cases of AI-enabled weapon systems.

** Even after functional maturity is achieved through training, continuous learning remains necessary.

(collateral damage)를 야기할 수 있다. 따라서 결정적인 순간에 인간이 개입하여 평가할 수 있는 **human-in-the-loop** 개념이 필수적으로 포함되어야 하나 이에 대한 개념이 부재하다.

네 번째, AI는 학습을 통해 그 성능을 향상시켜 가는 체계임을 고려하여 평가 후에도 계속해서 후속 관리가 필요하다. 하지만 전력화 후 어떻게 지속적인 학습을 통해 어떻게 무기체계의 성능을 향상시킬지에 대한 후속 관리 개념이 없다.

다음 장에서는 우리 군이 추구해야 할 인공지능 평가에 대한 세부적인 방안을 도출하기 위해 해외의 인공지능 평가사례를 살펴보겠다.

4. 미국 사례

미 국방부는 '22년 2월 AI 업무의 전문성, 일관성, 효율성 향상을 위해 기관별 분산되어 있던 합동AI센터, 국방디지털서비스국, 데이터책임관실을 통합하여 CDAO(Chief Digital & AI Office)를 창설하였다 [3]. CDAO는 AI 무기체계 시험평가를 위한 **guide book**에서 기존 무기체계와의 비교를 통해 평가 중점을 명시하고 있으며, 이를 정리하면 Table 3와 같다.

미국의 경우 기존 확정적(deterministic) 무기체계의 평가와는 다르게 AI 적용 무기체계의 경우 평가 대상, 평가시점, 핵심기법, 성공기준에서 다른 형태를 보여준다[4].

첫째, 무기체계의 단순 성능을 평가하는 것이 아니라

Table 3. Comparison of evaluation focus: Conventional vs. AI-enabled weapon systems

Category	Conventional Weapon Systems (Deterministic)	AI-Enabled Weapon Systems (Probabilistic)
Evaluation Target	Hardware, performance, durability, etc.	Data quality, algorithm robustness, ethics
Evaluation Timing	After development completion (one-time)	Entire life cycle from development to operation (continuous)
Key Methods	Live-fire tests, physical testing	Digital twin, red teaming, ¹⁾ runtime monitoring ²⁾
Success Criteria	Compliance with specifications (e.g., accuracy, speed)	Trustworthiness, explainability, maintenance of human control

* Reconstructed based on DoD, Developmental Test and Evaluation of Artificial Intelligence-Enabled Systems Guidebook(Feb. 2025)

인공지능 학습의 수준을 판단한다. 따라서 평가 대상은 수치화된 것이 아니라 학습 알고리즘의 강건성, 데이터의 품질, 판단의 윤리성 등을 중심으로 평가한다.

둘째, 평가시점은 개발이 완료된 시점으로 하는 것이 아니라 평가 시점과 운용 주기간 지속적으로 실시한다.

셋째, 실사격·물리적 성능시험이 아닌 디지털트윈을 구성하여 인공지능을 평가하거나 악의적 학습 또는 비의도적, 우연에 의한 오판이 발생하지 않는가에 대해 집중하여 인공지능을 평가한다.

넷째, 성공 기준은 특정 수치화된 스펙 충족 여부가 아니라 인공지능이 얼마나 신뢰성을 갖는지, 운용자가 통제권을 유지할 수 있는지를 평가한다.

¹⁾ 시모델의 잠재적 취약성과 편향성을 테스트하여 AI의 잠재적 취약성 대응할 수 있는 방법

²⁾ 런타임 모니터링(Run-time Monitoring) : 작전수행 중인 AI가 안전

범위를 벗어난 행동을 하려 할 때, 즉시 차단하는 안전 가드레일

5. 영국 사례

영국의 경우는 미국의 사례와 다르게 인공지능 무기체계가 어떠한 성능을 발휘해야 한다는 측면보다는 법적, 윤리적 보증을 시험평가의 핵심으로 판단하고 있다. 전술한 바와 같이 인공지능이 적용된 무기체계의 경우 hallucination에 의한 부수적 피해를 발생시킬 가능성이 있으며, 통제되지 못할 경우 어떠한 결과를 초래할지 아무도 예상할 수 없다. 유럽은 이러한 측면을 보다 중요하게 판단하고 있다.

인공지능의 선진국인 미국과 차별화되는 유럽의 특징을 평가기준, 핵심기술, 검증 주체, 도입 철학 부분으로 나누어 Table 4와 같이 정리할 수 있다.

특히 평가기준에서 설명하는 바와 같이 성능 중심보다는 신뢰성과 법적 책임에 더 집중하는 것을 확인할 수 있다.

Table 4. Characteristics of AI adoption in Europe

Category	Adoption Approach	Characteristics
Evaluation Criteria	Focus on trust and legal accountability	Even if performance is high, adoption is rejected if legal responsibility is unclear
Core Technology	Explainable AI	AI decision-making must be understandable to humans
Verification Authority	Multi-agency validation	Engineers, ethicists, and international law experts participate in evaluation
Adoption Philosophy	Human-in-the-loop centered	AI serves as a support tool; final decisions must be made by humans

* Reconstructed based on UK MoD, Defence Artificial Intelligence Strategy (June 2020), and related sources.

6. 중국 사례[5]

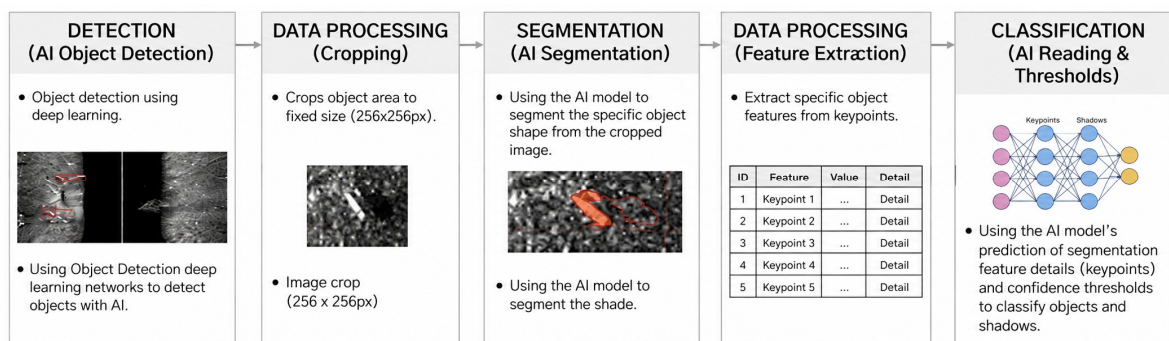
중국은 미국이나 유럽처럼 다양한 사례를 조사하는데 한계가 있었다. 현재까지 공개된 자료를 바탕으로 파악한 중국의 인공지능 적용 무기체계 평가 방법은 다음과 같다.

중국은 미국, 유럽과 확연히 다른 평가방식을 사용 중이다. 민간 대기업이 이미 확보한 방대한 데이터와 알고리즘 평가체계를 그대로 군사 분야에 이식하여 초기성능을 보장하고, 정부 주도의 국방 데이터 셋을 구축한 후 이를 통해 알고리즘의 정확도를 정량적으로 측정하고 있다. 또한 미국, 유럽은 안정성 검증을 중심으로 인공지능을 평가하기 위해 노력 중이나 중국은 우선 배치 후 반복해서 개선하는 방식을 사용하는 등 성능 발전을 중심으로 위험을 크게 감수하는 성향을 보이고 있다.

7. 우리 군사례(신속시범사업 실사례 중심)

우리 군에서는 민간의 선진기술을 신속하게 받아들이고 이를 사업화하여 군의 소요와 연계시키기 위한 신속시범사업을 추진 중이다. 민간의 기술을 기존의 소요 절차 대비 간소한 형태로 추진할 수 있는 방위사업으로 민간 기업들의 반응이 뜨거우며, 특히 기존의 방식으로 추진이 어려웠던 AI, 무인기 등 선진기술을 접목하기 위한 다양한 시도들을 추진 중이다.

최근 해군은 최초로 AI를 적용한 무기체계인 big data에 기반한 AI 자동기뢰탐지체계를 신속시범사업으로 추진 중이다. AI 자동기뢰탐지체계는 해저에 있는 기뢰의 모양을 학습하여 해저 영상촬영이 가능한 side scan SONAR, AUV(autonomous underwater



Source: Pilot Operational Plan for AI-Based Mine Detection System

Fig. 1. Automatic mine detection procedure

vehicle) 등의 장비에서 획득한 이미지 정보를 판독하여 실제 기뢰인지 여부를 판단, 소해작전을 지원하기 위해 운용하는 전력이다.

AI 자동기뢰탐지체계는 해군 최초로 AI를 탑재하여 신속시범사업에서 규정하고 있는 방식에 따라 평가할 예정인 전력이며, 인공지능에 대한 평가는 Fig. 1과 같은 절차를 통해 진행되었다. 신속시범사업을 진행하면서 가장 어려웠던 부분은 인공지능에 대한 평가를 어떻게 진행할 것인가였다.

최초 다양한 방법이 논의되었으나 앞서 언급한 JROC의 기준에 의거하여 Table 5와 같이 재현율(accuracy)을 적용, 탐지율 80% 이상을 달성하는 것을 목표로 정하였다.

Table 5. Evaluation method for AI-based mine detection system

Category	Content
Detection Rate (Accuracy)	$\frac{TP+TN}{TP+FN+FP+TN}$
TP (True Positive)	AI correctly detects a deployed mock mine as a mine.
FN (False Negative)	AI fails to detect a deployed mock mine (Missed).
FP (False Positive)	AI incorrectly identifies a non-mine underwater obstacle as a mine (False Alarm).
TN (True Negative)	AI correctly identifies a non-target and different shaped object as a non-mine.

해당 전력은 탐지 성능을 향상시키기 위해 AI를 적용시키는 것임에도 불구하고, 다양한 변수(조류, 파도, 온도 등)에 종속되는 탐지능력 자체에만 초점이 맞춰지는 평가방식을 적용하게 되었다.

특히 side scan SONAR, 카메라의 성능에 의해 좌우되는 수중영상 해상도는 탐지율에 직접적으로 영향을 미치지만, AI는 탐지율과 무관하게 학습을 통해 정확도를 점차 향상시키는 것이 가능하다. 실제로 Fig. 2는 수중 탐색기인 Klein과 GAVIA를 같은 구역에서 같은 진행 방향에 따라 운용하고, 기뢰를 탐지한 이미지에 대해서도 자료마다 수집 수단에 의한 차이가 있음을 보여준다.

이외에도 음탐기의 탐색 진행 방향, 만조, 간조, 파도, 수온 등 해양 환경적 요인에 의한 영상촬영의 어려움 등 발생 가능한 변수들을 고려 시 단순히 탐지율만을 달성하는 조건으로 평가를 진행하는 것은 해당

전력에 적용된 AI의 능력을 정확히 판단하기에는 부적합하다는 판단에 이르게 되었다.

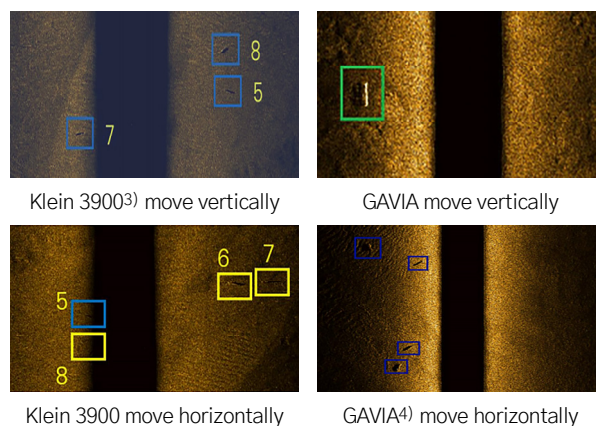


Fig. 2. Sensor-based mine detection imagery

하지만 현재 합참에서 규정하고 있는 JROC 작성 지침 기준 외 다른 대안이 없기에 Table 5의 내용으로 평가를 진행할 예정이다. 이러한 방식은 AI의 학습을 통한 능력향상 정도를 평가 결과에 반영할 수 없기 때문에 성능이 우수한 AI임에도 평가 과정을 통과하지 못하여 해당 전력을 사장시킬 가능성이 있으며, 이러한 문제점을 해결할 필요가 있다.

다음 장에서는 국외 사례와 우리 군 사례를 바탕으로 인공지능 적용 무기체계 확보 시 어떤 분야를 반드시 평가하고, 구체화되어야 하는지를 제시하였다.

8. 우리 군 발전방향 제언

AI 적용 무기체계는 모델 고유의 비예측성, 작은 변화에도 민감하게 반응하는 모델 민감도, 모델의 복잡성과 모호성, 변수의 다차원성, 학습 데이터에 기반한 모델 의존성 등의 어려움으로 인해 성능 예측이 어렵고, 도출 결과 역시 확률적이다.

그럼에도 전쟁의 승패, 인간의 생명과 직결되는 무기체계는 충분한 성능과 신뢰성이 보장되어야 하며, 이를 위해서는 특성에 부합한 시험평가를 통해 군이 요구하는 시기에 능력을 갖춘 무기체계를 확보하는

3) 최대 200 m 수심까지 관측 가능한 음향 표적탐지 장치(Side Scan SONAR), 진행방향 양쪽 해저면의 고해상도 음향 이미지를 획득

4) 최대 1,000 m까지 잠수하여 해저 영상을 탐사할 수 있는 AUV (Autonomous Underwater Vehicle) 진행방향 양쪽 해저면의 고해상도 음향 이미지를 획득

것이 중요하다. 그러면서도 동시에 빠르게 발전하는 AI 등의 첨단 기술이 적용된 성능이 충분히 보장된 무기체계를 확보하는 것 역시 중요한 상황이다.

이러한 운용자의 요구를 만족하기 위한 결론을 도출하기 위해 본론에서 제시한 미국, 유럽, 중국, 우리 군의 사례에 대한 고찰을 통해 4가지 발전 방향을 제시하였다.

첫째, AI가 적용된 무기체계는 정량화된 수치 정보를 바탕으로 주요 ROC에 포함하여 시험평가하는 관행을 극복할 필요가 있다. 예를 들어 인공지능이 보여줘야 할 성능을 기술적 부수적 성능 파트에 포함하는 것이다. 기술적 부수적 성능은 단위전력의 운용개념·계획을 변경시키지 않는 사항, 작전운용성능 및 운용에 간접적으로 영향을 미치는 사항 등에 대해 방위사업청의 실무위원회를 통해 수정이 가능한[6] 제도적 장치이다. 기술적 부수적 성능은 ROC에 대비하여 수정이 용이하며, 사업추진 간 AI의 성능 발휘 정도, 학습의 효과성을 보다 유연성 있게 관리할 수 있다.

둘째, AI 적용 무기체계의 평가 시 인공지능이 학습하는 데이터의 양과 품질을 평가해야 한다는 것이다. 국방 분야는 일반분야 대비 보안의 중요성이 그 무엇보다도 강조되는 분야이다. 따라서 인공지능이 학습해야 할 충분한 데이터 확보가 어려운 경우가 대부분이다. 이렇듯 데이터의 양이 부족한 상황에서는 데이터의 양보다 품질에 집중(data-centric AI)하는 것이 중요하다. 데이터의 양이 부족할 경우 잘못된 데이터 학습으로 인해 발생하는 인공지능의 편향(bias)이 다량의 데이터를 학습한 경우보다 크게 발생하게 된다. 삼성 SDS의 연구에 의하면 클래스 불균형(class imbalance)을 지능적으로 해결하고, 노이즈를 제거하는 데이터 중심의 전처리 수행이 모델의 구조를 바꾸는 것보다 20% 이상의 성능개선이 이뤄졌다는 결과를 보여줬으며[7], 이는 data 평가의 중요성을 나타내는 것이라 할 수 있다. IBM 역시 데이터 라벨링(label) 오류를 10% 줄이는 것이 데이터의 양을 10배 늘리는 것보다 모델의 성능향상에 더욱 기여한다는 연구결과[8]를 도출하였고, 데이터를 정확하게 분류하는 등 품질 평가의 중요성을 역설하고 있다.

셋째, 최종 결과를 판단하기보다 알고리즘의 효과성과 강건성(robust)을 평가해야 한다. 즉, AI 모델의 알고리즘이 우수하여 학습을 통해 성능이 향상되는 정도를 측정하고 어느 정도의 성능 향상이 이뤄지고

있다면 조금 능력이 부족하더라도 전력화를 시켜야 한다. AI는 현장에서 운용하며 이뤄지는 지속적인 학습을 통해 부족한 부분을 해결해 나가고, Table 2에서 제시한 바와 같이 특정 학습기간을 부여하여야 한다. 부여된 기간을 기준으로 특정 기간이 지났을 때 최초 대비 어느정도의 성능이 향상되었는지를 평가해야 한다.

넷째, 생명·안전과 직결되는 AI적용 무기체계의 경우는 반드시 사람이 최종 개입할 수 있는 장치를 마련해야 하며, 이를 주요 평가 요소로 반영해야 한다. 예를 들어 기뢰탐지의 경우 단 1발만 잘못 찾아도 배 1척을 침몰시킬 수 있는 최악의 상황을 야기할 수 있다. 이러한 우려를 불식하기 위해 인공지능은 운용자에게 위험, 가능성 등을 전파하고, 최종적으로 숙련된 사람이 판단할 수 있도록 하는 human-in-the-loop 개념이 잘 작동하고 있는지를 평가하여야 한다.

9. 결론 및 향후 연구방향

본 연구에서는 선진국의 인공지능 적용 무기체계 평가방법, 우리 군의 인공지능 평가방법 사례를 제시하였고, 다양한 사례를 통해 얻은 교훈을 바탕으로 우리 군에서 추구해야 할 인공지능 평가의 기본 개념과 방법에 대한 제언을 하였다.

하지만 실제 전력건설 현장에서 활용하기 위해서는 보다 구체적이고, 수치화된 기준의 제시가 필요하며, 따라서 다음과 같은 후속연구를 진행할 예정이다.

첫째, 무기체계별 AI 학습을 위한 데이터의 품질과 양을 수치화하여 제시할 수 있는 연구가 필요하다. 전술한 바와 같이 군의 특수성을 고려하여 품질과 양에 대한 적정 수준을 제시함으로써 사업 추진의 기준을 제시할 수 있을 것이다.

둘째, 인공지능 적용 무기체계의 학습기간과 학습율에 대한 기준을 제시할 수 있는 연구가 필요하다. 무기체계별 학습의 기간은 다르나 각기 다른 기간의 학습을 통해 어느정도의 학습률 향상이 이뤄졌을 경우 무기체계로서 가치를 가지고 있는가에 대한 기준 제시를 통해 보다 뛰어난 성능의 인공지능 적용 무기체계 확보가 가능할 것으로 판단된다.

추가 연구를 통해 전력건설 현장에서 활용 가능성과 동시에 인공지능의 특성과 능력을 더 잘 사용하기 위한 구체화된 지침과 기준 제시가 필요하다.

참고문헌

- [1] 국방기술진흥연구소, 미래국방 2030기술전략(AI), 2022. 1., p. 16.
- [2] 합참, 2024 합동작전운용성능(JROC) 작성지침, 2025. 1., p. 7.
- [3] Kangmin Kim & Yong-Bok Lee, 'AIEC T&E Strategy Proposal for AI-Based Video Surveillance System,' Journal of the Korea Academia-Industrial Cooperation Society, VOL. 26, NO. 3, 2025, pp. 500-507.
- [4] DoD, DoD Manual 5000, 101 Operationl Test and Evaluation and Live-Fire Test and Evaluation of Artificial Intelligence-Enabled and Autonomous System, 2024. 11., pp. 6-9.
- [5] DoD, Military and security developments involving the people's Republic of China, 2024. 12., pp. 33-35.
- [6] 방위사업청, 방위사업관리규정 제16조(기술적부수적 성능의 결정 및 수정), 2025. 9., p. 5.
- [7] 주민식, 인공지능의 한계 3-AI 응용 상용화 85% 실패, 그리고 Data-Centric AI로의 이동(삼성 SDS 인터넷 게시자료). <https://www.samsungsd.com/kr/insight/data-centric-ai.html> (accessed 2026.01.)
- [8] Michael Desmond, Catherine Finegan-Dollak, Jeff Boston, & Matt Arnold, 'Label Noise Context,' in Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Online, July 2020, pp. 157-186.