Journal of the KNST

2025; Vol.8, No.3; pp.345-353

ISSN: 2635-4926 https://doi.org/10.31818/JKNST.2025.9.8.3.345

Received: 2025/08/24
Revised: 2025/09/03
Accepted: 2025/09/29
Published: 2025/09/30

*Corresponding Author:
Jinwoo Kim
Dept. of Naval Strategy and Force Planning, Republic of Korea Naval War College
271, Jaun-ro, Yuseong-gu, Daejeon, 34059, Republic of Korea
Tel: +82-42-975-2420
E-mail: pkm311@gmail.com

임무공학(ME)에 기반한 AI 무인체계 개발 방법론: 설명가능성 및 통제가능성의 핵심 기능 분석 중심으로

A Study on Mission Engineering-based Development Methodology for AI Unmanned Systems: Focusing on Explainability and Controllability

김진우[1*], 양정규[2]

[1]해군 대령/해군대학 해양전략·전력학처장/컴퓨터공학 박사·국제정치학 박사
[2]해군 중령/해군대학 획득관리 교관/조선해양공학 박사과정 수료

Jinwoo Kim[1*], Jungkyu Yang[2]

[1]CAPT, ROK Navy/Director of Maritime Strategy & Military Force Studies, ROK Naval War College/Ph.D. in computer science & engineering, Ph.D. in international political science
[2]CDR, ROK Navy/Instructor on Defense Acquisition, ROK Naval War College/Ph.D. Candidate, Dept. of Naval Architecture and Ocean Engineering, Inha University

Abstract

인공지능(AI) 기반 무인체계는 다양한 임무에서 전력 승수로 활용될 잠재력이 크지만, 단순한 성능 충족만으로는 불확실한 전장에서 운용 신뢰성을 확보하기 어렵다. 본 연구는 이러한 한계를 보완하기 위해 설명가능성과 통제가능성을 핵심 요인으로 설정하고, 이를 임무공학(mission engineering) 절차에 통합하는 개념적 틀을 제시하였다. 또한 임무 단계별 적용과 위험 수준별 요구, 소요-기술-체계 연계, 지속적 인증 개념을 논의함으로써 기술 효율성과 작전 신뢰성의 균형을 위한 정책적 시사점을 도출하였다.

AI-enabled unmanned systems possess strong potential as force multipliers, but performance metrics alone cannot ensure reliability in contested environments. This study frames explainability and controllability as core determinants and embeds them within the Mission Engineering process. By addressing risk-differentiated requirements, system–technology integration, and continuous certification, it offers policy insights for balancing technological efficiency with operational trustworthiness.

Keywords

AI 기반 무인체계(AI-enabled Unmanned Systems), 임무공학(Mission Engineering), 설명가능성(Explainability), 통제가능성(Controllability), 신뢰성 (Trustworthiness), 지속적 인증 (Continuous Certification)

1. 서론

인공지능(artificial intelligence, AI) 기반 무인체계는 감시·정찰, 자율 항행, 타격·교전 등 다양한 임무를 수행함으로써 미래 군사작전에서 전력 승수(force multiplier)로 작용할 잠재력을 지니고 있다. 그러나 높은 성능을 발휘하는 것만으로는 실전 운용의 신뢰성을 보장하기 어려운 것이 현실이다. 이는 실제 작전환경이 높은 불확실성과 빈번한 돌발 변수로 특징지어지며, 이때 운용의 안전성(safety)과 책임성(accountability)이 확보되지 않으면 해당 체계는 기대와 달리 오히려 위험을 가중하고 작전적 위험을 증대시킬 수 있기 때문이다.

국제적 논의에서도 이러한 문제의식이 지속하여 제기되고 있다. 미국 국방고등연구계획국(Defense Advanced Research Projects Agency, DARPA)의 XAI(explainable AI) 프로그램은 판단 근거가 불투명한 AI 체계가 운용자의 신뢰를 저해할 수 있음을 지적하였으며[1], Scharre(2018)는 인간 개입이 배제된 자율무기체계가 억제 전략과 책임성 구조를 약화시킬 수 있음을 경고하였다[2]. 또한

북대서양조약기구 과학기술기구(NATO Science and Technology Organization, NATO-STO)는 자율체계를 설계할 때 비상 정지 및 수동 전환 장치를 내장할 것을 권고하고 있으며[3], Horowitz(2019)는 최소한 human-in-the-loop 구조가 유지되어야 한다고 강조한다[4].

이러한 논의는 AI 무인체계가 단순한 성능 향상만으로는 충분하지 않으며, 운용자가 체계의 판단 근거를 이해할 수 있는 설명가능성(explainability)과 긴급 상황에서 개입할 수 있는 통제가능성(controllability)이 필수적임을 보여준다. 두 요소가 결여될 경우 잘못된 판단 수용, 임무 실패, 지휘·통제 구조의 붕괴, 국제적 비난 등 중대한 부정적 결과가 발생할 수 있다.

이에 본 연구는 이러한 문제의식을 바탕으로 AI 무인체계 개발 과정에 설명가능성과 통제가능성을 체계적으로 반영할 수 있는 개념적 분석틀(conceptual analytical framework)을 제시한다. 특히 임무공학(mission engineering, ME) 절차에 주목하여, 임무 분석-능력 도출-체계 설계-시험평가의 전 과정 속에서 설명가능성과 통제가능성이 고려될 수 있는 구조적 접근을 탐색한다. 이를 위해 두 변수를 투명(transparent)-부분적 해석 가능(semi-transparent)-불투명(opaque), 완전 통제 가능(fully controllable)-제한적 통제 가능(limited controllable)-통제 불가(uncontrollable)의 개념적·기술적 수준으로 구분하여, 임무 유형에 따라 차별화된 요구 수준을 도출하고자 한다.

연구 방법은 다음과 같은 절차로 전개된다. 첫째, 기존 연구와 국제적 논의를 검토하여 AI 무인체계 개발에서 설명가능성과 통제가능성이 요구되는 근거를 규명한다. 둘째, 두 요인의 수준별 속성과 영향 요인을 체계적으로 분류하여 분석의 준거틀을 마련한다. 셋째, 이를 임무공학 절차에 접목할 수 있는 구체적 적용 방법론을 설계한다. 넷째, 전력소요 반영 가능성, 시험평가의 한계, 지속적 검증 체계 구축 필요성 등을 종합하여 정책적 시사점을 도출한다.

본 논문은 2장에서 이론적 논의와 분석의 틀을 정립하고, 3장에서 임무공학 기반의 적용 방법론을 제시한다. 이어 4장은 제도적·기술적 정책 방향을 논의하며, 5장은 결론으로 연구의 성과와 한계, 향후 연구를 종합한다.

2. 이론적 논의와 분석의 틀

2.1 AI 기반 무기체계의 특징

AI 기반 무기체계는 기계학습과 딥러닝 등 인공지능 기술을 적용하여 탐지·판단·행동의 일련 과정을 수행하는 데 있어 인간의 개입을 최소화하고, 일정 수준의 자율성을 바탕으로 임무를 독립적으로 수행할 수 있는 무기체계를 의미한다. 이는 전통적 무기체계와 뚜렷한 차이를 가진다. 기존 무기체계가 사전에 정의된 알고리즘이나 고정된 규칙에 따라 예측 가능한 방식으로 작동한다면, AI 무기체계는 데이터 학습을 통해 성능이 지속적으로 변화하고 고도화된다는 점에서 본질적으로 다르다. 이러한 특성은 AI 무기체계가 단순한 자동화 수준을 넘어, 임무 환경의 불확실성에 대응하는 능력을 내재화하고 있음을 보여준다[5].

AI 기반 무기체계의 특징은 자율성, 적응성, 불확실성으로 요약될 수 있다. 자율성은 인간의 직접적인 지시 없이도 임무를 수행할 수 있는 능력을 의미하며, 적응성은 변화하는 전장 환경 속에서 학습과 경험을 통해 스스로 성능을 개선할 수 있는 능력을 뜻한다. 그러나 동시에 이러한 자율적·적응적 성격은 결과 예측을 어렵게 만들며, 체계가 오작동하거나 예상치 못한 방식으로 행동할 가능성을 내포한다는 점에서 불확실성을 수반한다[6].

실제 사례에서도 이러한 특징은 확인된다. 러시아-우크라이나 전쟁에서 활용된 Lancet 드론은 표적 인식에 AI 기술을 접목하여 기존 무인기보다 높은 정확도를 보여주었으나, 동시에 표적 오인식 가능성이라는 위험을 드러냈다[7]. 또한 자율 기동을 수행하는 무인수상정(USV)은 광범위한 해역에서 장시간 임무를 수행할 수 있는 장점이 있지만, 복잡한 해상 상황 속에서는 예기치 못한 충돌 가능성이 존재한다[8]. 더불어 우크라이나군이 활용한 GIS Arta나 Delta와 같은 AI 지원 지휘통제체계는 작전 결심 시간을 단축시키는 성과를 거두었으나, 잘못된 데이터가 투입될 경우 오류가 연쇄적으로 확산될 수 있다는 한계를 동시에 가지고 있다.

따라서 AI 기반 무기체계는 작전 효율성과 효과성을 비약적으로 향상시킬 수 있지만, 설명가능성과 통제가능성이 충분히 확보되지 않는다면 오히려 위험을 증대시킬 수 있다. 이는 설명가능성과 통제가능성

에 주목하여 임무공학적 접근을 채택한 본 연구의 논리적 기반을 형성한다.

2.2 임무공학(mission engineering)의 적용 개념

임무공학은 개별 체계의 성능 향상에 집중하는 기존 시스템공학(systems engineering)의 접근과 달리, 임무 달성(mission success) 자체를 목표로 하는 상위 개념적 방법론이다. 미국 국방부(Department of Defense, DoD)는 이를 "개별 체계의 성능이 아닌 다양한 체계 간 연계를 통해 임무 달성을 보장하는 방법론"으로 정의하며, 이는 체계 중심에서 임무 중심으로의 패러다임 전환을 의미한다[9].

임무공학의 특징은 첫째, 여러 체계의 연동을 전제로 한 다체계 통합(system of systems), 둘째, 설계·운용 과정 전반이 장비 성능이 아닌 임무 성공 보장을 지향한다는 점, 셋째, 실제 작전환경의 불확실성에 대응하기 위해 운용자의 개입과 상황 적응을 강조한다는 점으로 요약된다. 즉, 임무공학은 전장 환경과 운용자의 역할까지 포괄하여 전체 임무 체계를 설계·관리하는 접근이다.

실제 적용에서도 이러한 성격은 확인된다. 미 국방부는 자율 무인체계 운용 개념을 발전시키며 임무공학을 통해 요구 능력(required capability)을 정의하고, 체계 설계 및 시험평가를 연계하고 있다[10]. 이로써 개별 기술 개발이 아닌 임무 성공이라는 상위 목표에 기술과 설계가 정렬되도록 유도한다.

AI 기반 무기체계는 다양한 센서, 통신, 무인체계, 지휘통제체계가 결합된 복합 구조 속에서 운용된다. 따라서 설명가능성과 통제가능성을 확보하기 위해서는 개별 장비 분석만으로는 부족하며, 임무환경 전체를 고려하는 임무공학적 접근이 요구된다. 다시 말해, 임무공학은 AI 무기체계의 안전성과 신뢰성을 보장하기 위한 분석틀이자 실행 경로로서 중요한 의의를 지닌다.

2.3 설명가능성·통제가능성: AI 무인체계 운용의 핵심 기능

AI 기반 무인체계는 탐지거리, 명중률 등 전통적 성능 지표만으로는 운용의 신뢰성을 보장하기 어렵다. 자율성이 확대될수록 체계의 의사결정 과정을 운용자가 충분히 이해하지 못하거나 긴급 상황에서 개입할 수 없는 조건이 발생할 수 있으며, 이는 임무 실패와 오작동으로 인한 우군 피해, 지휘·통제 구조의 혼란, 국제적 규범 위반으로 이어질 수 있다. 따라서 운용자가 AI의 판단 근거를 이해할 수 있는 설명가능성과 필요 시 최종적 통제권을 행사할 수 있는 통제가능성을 확보하는 것이 성능 향상만큼 중요하다.

설명가능성은 운용자가 체계의 결정을 단순히 받아들이는 것이 아니라, 그 판단 과정과 근거를 해석하고 검증할 수 있도록 보장한다는 점에서 신뢰 형성의 기반이 된다. 운용자가 설명가능성을 확보하지 못하면 AI의 판단을 맹신하거나 반대로 전면 불신하는 양극단적 상황이 초래될 수 있다.

통제가능성은 지휘·통제 구조의 안정성과 직결된다. AI 무인체계는 학습과 환경 적응 과정에서 예기치 못한 행동을 보일 수 있는데, 이때 인간이 긴급 정지·임무 중단·대체 명령을 실행할 수 없다면 체계는 사실상 통제 불능 상태가 된다. 이는 억제 전략의 신뢰성과 책임성 구조를 약화시키며, 무력 충돌의 안정성을 위협할 수 있다. 따라서 통제가능성은 단순한 기술적 보완 장치가 아니라, 군사적 정당성과 윤리적 책임을 보장하는 최소 조건이다[11].

결국 설명가능성과 통제가능성은 독립적인 개념이 아니라 상호 보완적으로 작동한다. 설명가능성이 있어야 운용자는 AI의 판단을 합리적으로 수용할 수 있고, 통제가능성이 보장되어야 최종 책임을 인간이 행사할 수 있다. 두 요건이 함께 충족될 때 AI 무인체계는 전력 승수로 기능할 수 있으며, 그렇지 않을 경우 심각한 위험 증폭자가 될 수 있다.

2.4 분석의 틀

AI 기반 무인체계는 단순한 성능 지표 충족만으로 운용상의 안전성과 신뢰성을 확보할 수 없다. 특히 설명가능성과 통제가능성이 결여될 경우, 운용자는 AI 판단 근거를 이해하지 못해 체계에 대한 신뢰가 약화되거나 맹목적으로 의존하게 된다. 또한 돌발 상황에서 인간의 개입이 차단될 경우 지휘·통제 구조가 무력화될 수 있다. 이는 임무 실패, 오작동 확산, 우군·민간 피해, 국제적 비난으로 이어질 수 있다는 점에서 심각한 전략적 위험을 내포한다. 이러한 잠재적 위험은 Table 1과 같이 정리할 수 있다.

Table 1. Potential risks arising from the absence of explainability and controllability

Category	Absence of explainability	Absence of controllability
Operator dimension	Opaque reasoning → Weakened trust, excessive distrust/blind reliance	Inability to intervene in emergencies → Unclear operator responsibility
Operational dimension	Acceptance of false alarms/misjudgments → Risk of mission failure	Loss of control during malfunction → Escalation of damage, collateral harm
Institutional/ policy dimension	Lack of decision-making transparency → Decline in policy credibility	Weakening of command-and-control structures → Vulnerable deterrence strategy, international criticism

이러한 문제의식에 따라 본 연구는 설명가능성과 통제가능성을 AI 무인체계 개발 과정의 핵심 요건으로 정의하고, 이를 질적 수준에 따라 구분한다. 설명가능성(X)은 운용자가 AI 판단 근거를 이해·검증할 수 있는 능력이며, 통제가능성(C)은 운용자가 돌발 상황에서 체계를 개입·제어할 수 있는 능력을 뜻한다. 두 변수는 각각 Table 2와 같이 세 단계 수준으로 구분된다.

Table 2. Concepts and levels of explainability and controllability

Category	Concept definition	Level classification
Explainability (X)	The ability of the operator to understand and verify the AI's reasoning process	Transparent/ Semi-transparent/ Opaque
Controllability (C)	The ability of the operator to intervene and control the system in unexpected situations	Fully controllable/ Limited controllable/ Uncontrollable

이 분석틀은 임무공학의 절차 내에서 설명가능성과 통제가능성을 구조적으로 반영하기 위한 기초를 제공한다. 즉, 임무분석 단계에서는 목표 수준을 설정하고, 능력 도출 단계에서는 이를 요구조건에 포함하며, 체계 설계 단계에서는 XAI 기법과 human-in-the-loop 구조를 적용하고, 시험평가 단계에서는 운용자의 이해도와 개입 가능성을 질적으로 검증하는 방식으로 연계된다.

3. 임무공학에 기반한 AI 무인체계 개발 방법론 제안

AI 무인체계의 개발과 운용은 단순히 성능 지표의 향상만으로는 충분히 설명되지 않으며, 실제 전장 환경에서는 상황의 불확실성과 복잡성이 높고, 긴급 돌발 상황과 적의 의도적 교란이 빈번하게 발생한다. 이러한 조건에서 운용자의 신뢰와 개입 가능성이 확보되지 않는다면 체계의 자율성은 곧바로 위험으로 전환될 수 있다. 따라서 본 장은 임무공학 절차에 설명가능성과 통제가능성의 핵심 기능을 구조적으로 내재화하는 적용 방법론을 제시한다. 이를 통해 AI 무인체계가 단순한 기술적 진보를 넘어 실제 작전 운용 환경에서 신뢰할 수 있는 전력 자산으로 기능할 수 있는 방법론을 제안하고자 한다.

3.1 임무공학 기반 적용 방법론

임무공학은 임무 성취(mission accomplishment)를 보장하기 위해 설계·개발 과정을 통합하는 절차적 접근으로, 임무분석-능력도출-체계설계-시험평가의 네 단계로 구분된다. 본 연구는 각 단계에 설명가능성과 통제가능성의 핵심 기능을 반영할 수 있는 통합 메커니즘을 Fig. 1과 같이 제시한다.

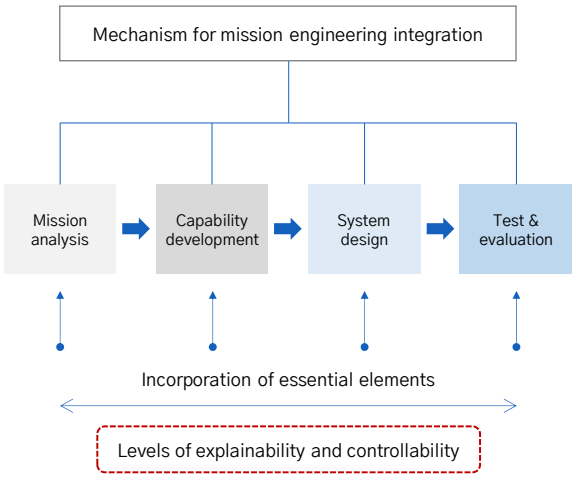

Fig. 1. Mission engineering-driven application methodology

3.1.1 임무분석(mission analysis) 단계

임무분석 단계는 설명가능성과 통제가능성을 구체적인 임무 요구조건으로 전환하기 위한 출발점이 되며, 이 과정에서 먼저 임무 목적과 작전환경을 규정하고, 정상·비정상·경계 상황을 포함한 잠재적 위험 시나리오를 도출하는 단계이다.

예컨대 해양작전에서는 기상 변화, 통신 두절 및 교란, 다중 위협 상황 등을 고려할 수 있으며, 이를 바탕으로 설명가능성과 통제가능성의 목표 수준을 설정한다. 감시 정찰 임무의 경우, 설명가능성은 "AI가 제시하는 탐지 근거를 운용자가 이해할 수 있는 수준 이상"으로, 통제가능성은 "긴급 상황에서 운용자가 항로 변경이나 탐지 임계치 조정 등 제한적 개입을 할 수 있는 수준 이상"으로 규정되어야 한다. 이와 반대로 교전 임무의 경우 설명가능성은 부분적 수준의 규정만으로도 충분히 확보될 수 있으나, 통제가능성은 반드시 완전 통제가 가능한 구조로 요구되어야 한다.

이 단계에서는 운용 개념(CONOPS)을 정교화하여 인간과 기계 간 역할 분담, 설명이 제공되는 시점과 형식, 개입 권한 및 절차를 명확히 해야 하며, 예를 들어 human-in-the-loop(HITL, 인간에 의한 최종 승인), human-on-the-loop(HOTL, 인간에 의한 감시·승인), human-out-of-the-loop(HOOTL, 인간의 통제 배제/완전 자율) 구조 중 어느 형태가 해당 임무에 적합한지, 그리고 긴급 상황 시 개입 권한이 어느 수준까지 보장되어야 하는지를 결정해야 한다. 또한 임무 시나리오와 위협 모델링을 통해 설명가능성과 통제가능성이 약화될 수 있는 취약 지점을 사전에 식별해야 한다. 예를 들어, 표적 인식 과정에서 발생하는 오경보·오인식은 설명가능성의 한계를 드러내며, 통신 두절이나 전자전 상황에서의 통제권 상실은 통제가능성 저하의 대표적 사례가 될 수 있다. 이를 통해 개발 단계에서 최우선적으로 보완해야 할 방안을 도출할 수 있다.

따라서, 최종적으로 도출되는 산출물에는 임무분석 보고서, 설명가능성과 통제가능성 목표 수준 표, 위험·가정 목록(assumption log), 초기 인증·윤리 고려사항 등이 포함되어야 하며, 그 예시로 "정찰 임무에서 운용자는 표적 분류 시 AI가 제시하는 주요 근거와 신뢰도를 확인할 수 있어야 하며, 필요 시 항로 변경 및 임무 중지를 즉시 명령할 수 있어야 한다"라는 형태의 요구조건이 산출되어야 한다.

3.1.2 능력 도출(capability development) 단계

능력 도출 단계는 임무분석에서 설정된 목표 수준과 운용개념을 바탕으로, 이를 구체적인 임무 수행 능력 요구조건으로 전환하는 과정이다. 이 단계에서는 설명가능성과 통제가능성의 핵심기능을 단순히 부수적 항목으로 규정하는 이상으로 각 임무 능력 속에 내재화하는 것이 중요하다. 예를 들어 탐지 능력을 정의할 때는 탐지율·탐지거리와 같은 전통적 지표에 부가하여, "탐지 결과와 그 근거를 운용자가 이해할 수 있는 형태로 제시할 것"이라는 설명가능성 요구가 반드시 포함되어야 한다. 교전 능력에서는 "목표물 식별 이후 무력 사용 결정 이전에 운용자가 개입하거나 임무를 중단할 수 있는 권한을 보장할 것"이라는 통제가능성 요구가 함께 반영되어야 한다.

이 과정은 설명가능성과 통제가능성을 단순한 '보조 기능'이 아닌 임무 수행의 필수적 능력 요소로 격상시키는 효과를 가지며, 더 나아가 설명가능성과 통제가능성은 기술적 차원뿐 아니라 운용자 교육·훈련, 지휘체계 절차, 데이터 신뢰성 관리 등 제도적 차원과도 연계되어야 한다. 예를 들어, 설명가능성은 AI 모델이 제공하는 근거의 품질과 더불어 운용자가 이를 이해·활용할 수 있는 교육 체계가 함께 마련될 때 실효성을 가지며, 통제가능성 역시 단순히 물리적 긴급정지 버튼의 존재에 그치지 않고, 실제 작전 상황에서 언제, 누구에 의해, 어떤 절차로 작동하는지 명확히 규정될 때 의미를 갖는다. 능력 도출 단계의 최종 산출물은 설명가능성과 통제가능성이 내재화된 임무 능력 요구조건 정의 문서이며, 이는 이후 설계와 시험평가 단계에서 준거로 기능하여야 한다.

3.1.3 체계 설계(system design) 단계

체계 설계 단계는 앞선 단계에서 정의된 설명가능성과 통제가능성의 핵심기능 요구조건을 실제 아키텍처 설계에 반영하는 과정이다.

설명가능성을 확보하기 위해서는 XAI 기법을 적용하여, 표적 탐지·분류 결과의 판단 근거를 시각적으로 제시하거나 설명할 수 있는 모듈을 내장해야 한다. 또한 운용자가 실전 상황에서 즉시 이해할 수 있도록

설명은 복잡한 기술 용어 대신 직관적인 형태로 제공되어야 하며, 필요할 경우 요약·세부 설명을 선택적으로 확인할 수 있는 사용자 인터페이스(UI) 설계가 반영되어야 한다. 통제가능성을 보장하기 위해서는 HITL 구조를 반영하여 운용자가 주요 의사결정 단계에 개입할 수 있도록 해야 한다. 자동-반자동-수동 모드 전환 기능과 긴급 정지 장치(비상 정지 버튼, 원격 제어 체계 등)는 초기 설계 단계에서부터 통합적으로 고려되어야 하며, 단순한 보조 기능이 아니라 안전 운용의 핵심 요소로 기능해야 한다.

또한 설명 로그와 개입 기록을 표준화된 포맷으로 저장·관리하여, 사후 검증과 책임성 확보가 가능하도록 하는 데이터 관리 구조가 필요하다. 통제권 전환 명령이 지연·왜곡 없이 체계에 반영될 수 있도록 보안적·통신적 설계 역시 필수적으로 고려되어야 한다.

3.1.4 시험평가(test & evaluation) 단계

시험평가 단계는 설명가능성과 통제가능성이 실제 운용에서 구현되는지 검증하는 절차이다. 기존 시험평가가 주로 성능 지표 충족 여부에 집중했다면, AI 무인체계 시험평가는 운용자의 신뢰와 개입 가능성, 그리고 체계의 안정성을 종합적으로 평가해야 한다.

설명가능성 측면에서는 운용자가 AI의 판단 근거를 얼마나 명확히 이해할 수 있는지, 그리고 그 설명이 실제 의사결정에 기여하는지 검증해야 한다. 지나치게 기술적이거나 복잡한 설명은 실전 활용이 어려우므로 설명의 명확성·일관성·속도가 주요 평가 기준이 된다.

통제가능성 측면에서는 비상 정지, 수동 전환, 임무 중단 등의 조치가 실제 작전 시나리오 속에서 안정적으로 작동하는지 평가해야 하며, 통신 불안정 상황에서 긴급 중단 명령이 제대로 전달되는지, 예기치 못한 표적 식별 시 운용자가 즉시 개입할 수 있는지 등을 검증해야 한다.

시험평가는 단순히 합격·불합격으로 끝나는 절차가 아니라, 설명가능성과 통제가능성 수준을 질적으로 진단하고, 지속적 개선 방향을 제시하는 과정으로 이해되어야 한다. 특히 AI 무인체계는 운용 중에도 학습과 데이터 축적을 통해 지수적으로 성능이 변화하므로, 설명가능성과 통제가능성 또한 주기적 검증과 보완이 필요하다.

3.2 임무 유형의 개념적 시나리오(예시: 무인수상정)

설명가능성과 통제가능성의 요구 수준은 임무의 성격과 위험 정도에 따라 달라진다. 감시·정찰 임무에서는 운용자가 AI가 제시하는 탐지 및 식별 근거를 이해할 수 있는 수준의 설명가능성이 필요하며, 동시에 긴급 상황에서 항로 변경이나 탐지 임계치 조정 등 제한적 개입이 가능해야 한다. 반면 타격·교전 임무와 같이 치명적 결과를 초래할 수 있는 임무에서는 부분적 설명가능성만으로도 충분할 수 있으나, 통제가능성은 반드시 완전한 수준으로 확보되어야 한다. 운용자가 체계의 행동을 즉시 중단하거나 수정할 수 있는 권한을 갖지 못한다면, 오작동은 곧바로 돌이킬 수 없는 피해로 이어질 수 있기 때문이다. 자율 경계·항행 임무의 경우에는 장시간 운용과 복합적 환경 대응이 필수적이므로, 설명가능성은 투명한 수준에서 보장되어야 하며, 통제가능성은 최소한 제한적 개입이 가능하도록 설계되어야 한다.

이러한 차별성은 개념적 시나리오를 통해 충분히 도출할 수 있다. 예컨대 무인수상정(USV)이 해상 경계 임무를 수행한다고 가정할 경우, AI의 표적 판별 근거가 투명하게 제공된다면 운용자는 오경보 가능성을 줄일 수 있으며, 제한적 개입 권한을 통해 항로 수정이나 임무 중단이 가능하다. 반면 교전 상황에서 운용자가 즉각적 개입 권한을 갖지 못한다면, 체계의 오작동은 곧 치명적 피해로 이어질 수 있다. 이는 설명가능성과 통제가능성이 단순한 보조적 속성이 아니라, 임무 수행의 성패를 좌우하는 핵심 요건임을 시사한다.

따라서 설명가능성과 통제가능성은 임무 유형에 따라 차별화된 수준으로 설정되어야 하며, 이를 임무공학 절차 속에 구조적으로 반영해야 한다. 이는 다양한 임무 환경에서 안전성·신뢰성 보장을 위한 방향성 정립에 중요하며, 이러한 분석은 향후 제도적 개선과 정책적 논의 과정에서 검토될 수 있는 기초 자료와 AI 무인체계 개발에서 기술적 효율성과 작전적 신뢰성 간 연계를 위한 중요한 시사점을 제공한다.

4. AI 무인체계 개발을 위한 정책적 제언

AI 무인체계의 개발은 단순한 기술적 진보만으로는 실질적 전력화에 이르기 어렵다는 한계가 지적된

다. 특히 설명가능성과 통제가능성은 체계의 안전성과 신뢰성을 보장하는 핵심 기능으로 단순한 기술 설계 수준을 넘어 제도적·정책적 차원에서 제도화될 필요가 있다. 이에 본 장에서는 임무공학적 관점에서 AI 무인체계 개발을 지원하기 위한 정책적 제언을 첫째 제도적·운용적 연계 강화, 둘째 시험평가 체계 개선이라는 두 가지 측면에서 제시하고자 한다.

4.1 제도적·운용적 연계 강화 및 시험평가 체계 개선

첫째, 소요기획-기술기획-체계개발 간 연계성 강화의 필요성이 지적된다. 현재 국방 연구개발 과정은 소요기획, 기술기획, 체계개발이 비교적 독립적으로 운영되는 경우가 많아, 임무분석 단계에서 도출된 설명가능성과 통제가능성 요구가 실제 개발 과정에 충분히 반영되지 못하는 한계가 존재한다. 이를 보완하려면 소요기획 초기 단계에서 설명가능성과 통제가능성을 주요 성능 속성(key system attribute, KSA)으로 설정하고, 이후 기술기획과 체계개발의 전 과정에서 이를 일관되게 추적할 수 있는 제도적 구조가 필요하다. 아울러 통합개념팀(integrated concept team, ICT)을 활용하여 소요군·연구개발기관·시험평가기관이 초기 단계부터 협력하는 방식을 고려할 수 있다.

이러한 접근은 임무분석 결과가 운용요구성능(required operational capability, ROC) 및 기술 기획 목표와의 정합성을 확보하도록 함으로써, 소요-기술-체계 설계 간 연속성과 체계성을 보장하는 데 기여할 수 있다.

둘째, 시험평가(test and evaluation, T&E) 체계의 개선 필요성도 강조된다. 기존 시험평가는 특정 시점에서 성능 충족 여부를 판정하는 일회적 방식에 머무르는 경향이 있었으나, AI 무기체계는 운용 중 학습과 데이터 축적에 따라 성능이 변동될 수 있다. 따라서 단순한 성능 확인만으로는 충분하지 않으며, 기존의 시험평가는 운용자의 이해 및 신뢰 수준, 긴급 상황에서의 개입 가능성, 환경 변화에 대한 적응 능력 등을 포함하는 질적 평가로 확장될 필요가 있다. 더 나아가 체계 운용 이후에도 정기적이고 반복적인 검증을 수행하는 '지속적 인증(continuous certification)' 개념을 도입한다면, 무기체계의 신뢰성과 안정성을 장기간에 걸쳐 유지·보완할 수 있을 것이다[12].

시험평가는 단일 실물시험에 의존하기보다는 HIL (hardware-in-the-loop, 실제 하드웨어 포함 시험), SIL(software-in-the-loop, 소프트웨어 모의 환경 검증), LVC(live-virtual-constructive, 실시간·가상·모의 통합 시험) 시뮬레이션과 디지털 트윈을 결합하는 방향으로 발전해야 한다. 이러한 다층적 접근은 AI 무인체계의 동적 특성과 불확실성을 반영함으로써 단순 성능 검증을 넘어 안전성과 신뢰성 확보에 기여할 수 있다.

4.2 정책적 제도화 및 국제적 정합성 확보

AI 무인체계의 설명가능성과 통제가능성은 단순한 기술적 과제가 아닌 정책적·제도적 기반 위에서 제도화되어야 할 목표라는 요구가 꾸준히 제기되고 있다. 이를 위해 다음의 세 가지 정책적 방향을 모색할 수 있다.

첫째, '제도적·조직적 기반의 정비'는 AI 무인체계 개발 과정에서 중요한 과제로 논의된다. 현재 소요군, 개발기관, 시험평가기관 간의 역할과 책임이 반드시 일관되게 규정되지 않고 있다는 지적이 제기되어 왔으며, 이로 인해 사업 추진 과정에서 개발자와 평가자가 서로 다른 기준을 적용하거나 상이한 해석을 내릴 가능성이 존재한다. 이러한 불일치는 설명가능성과 통제가능성과 같은 핵심 속성이 실제 사업 절차에 체계적으로 반영되는 것을 저해할 수 있다는 우려로 이어진다. 따라서 방위사업 절차서와 관련 지침에 해당 항목을 정식 평가 기준으로 명시하고, 데이터 기록 방식, 긴급 정지 기능 구현 여부, 권한 전환 절차 등 구체적 증빙 요건을 규정할 필요가 있다. 또한 운용자 교육·훈련 과정 역시 단순한 장비 조작 능력에 국한되지 않고, AI 체계의 판단 근거를 해석·검증하는 분석 능력, 오류 발생 시 신속히 개입할 수 있는 실무 역량, 상황 변화에 따라 권한을 전환하는 절차적 숙련도를 포함해야 하며, 이러한 제도적 정비와 운용자 역량 강화가 병행될 경우, 기술적 안전장치의 한계를 보완하고 인간 운용자의 개입 가능성을 제도화함으로써 AI 무인체계의 안전성과 신뢰성을 장기간 확보할 수 있을 것은 판단한다.

둘째, '위험 기반 접근(risk-based approach)'의 도입을 논의할 수 있다. 이는 모든 임무에 동일한 기준을 일률적으로 적용하는 대신, 임무의 위험 수준과 성격에 따라 요구조건을 달리 설정하는 방식을 의미한

다. 다시 말해, 위험도가 낮은 임무에는 완화된 기준을 적용하고, 위험도가 높은 임무에는 보다 엄격한 기준을 적용함으로써 효율성과 안전성을 동시에 확보하려는 접근이다[13]. 예를 들어 정찰 임무와 같이 위험도가 낮은 활동에서는 운용자가 판단 근거를 이해할 수 있는 설명가능성이 상대적으로 강조되며, 교전 임무와 같이 위험도가 높은 활동에서는 완전한 수준의 통제가능성이 핵심적으로 요구된다. 이러한 차등적 기준 설정은 불필요한 규제나 과도한 부담을 줄이는 동시에, 임무 특성에 적합한 실효적 기준을 제공할 수 있다는 장점이 있다.

셋째, 국제적 협력과 정합성 확보가 중요한 과제로 제기된다. NATO, UN CCW(Convention on Certain Conventional Weapons, 특정 재래식무기 금지협약) 등 국제기구에서 논의되는 human control 원칙과 설명가능성·통제가능성 권고는 점차 국제적 기준으로 정착하고 있다[14,15]. 한국 또한 이러한 국제 논의에 보조를 맞추어 국내 규범과 개발 지침을 정비할 필요가 있다. 이는 단순한 기술 표준 차원을 넘어, 연합작전에서의 상호운용성(interoperability), 무기체계 수출에서의 신뢰 확보, 국제사회에서의 윤리적 정당성 확보와 직결된다. 따라서 용어 정의, 데이터 포맷, 시험평가 시나리오를 국제 표준과 정합적으로 설계하고, 상호 인증 체계를 마련하는 노력이 병행되어야 한다.

결론적으로, AI 무인체계 개발에서 설명가능성과 통제가능성은 선택적 기능이 아닌 임무 성공과 안전 운용을 담보하는 필수 핵심기능으로 주목받고 있다. 이를 정책적으로 제도화하기 위한 방향은 첫째, 소요-기술-체계 간 연계 강화와 시험평가 체계 개선을 통한 기술적 특성과 운용 신뢰성의 동시 확보, 둘째, 제도적 기반 정비와 위험 기반 접근을 통한 국내 제도화, 셋째, 국제적 정합성 확보를 통한 글로벌 신뢰와 상호운용성 보장으로 요약할 수 있다. 다만 본 연구가 제시하는 정책 방향은 확정적 해법이라기보다, 향후 제도개선과 정책 논의 과정에서 지속 검토될 수 있는 정책적 시사점으로서 의미가 있다.

5. 결론

본 연구는 AI 기반 무인체계의 개발 과정에서 설명가능성과 통제가능성이 단순한 부수적 기능이 아니라 실제 운용상의 안전성과 신뢰성을 담보하는 핵심요인으로 작동함을 논의하였다. 이를 위해 임무공학(mission engineering) 절차를 분석의 틀로 설정하고, 임무분석-능력 도출-체계 설계-시험평가의 각 단계에서 설명가능성과 통제가능성이 반영될 수 있는 가능성을 고찰하였다. 또한 임무의 성격과 위험 수준에 따라 두 요소의 요구 수준이 달라질 수 있음을 제시함으로써, 일률적 기준 적용보다는 차등적 기준 설정의 필요성을 강조하였다. 더 나아가 소요-기술-체계 간 연계 강화, 시험평가 체계의 전환, 위험 기반 접근, 국제적 정합성 확보와 같은 정책적 방향을 함께 논의함으로써, 기술적 차원에서의 논의와 제도적 차원에서의 논의를 연계하였다.

다만 본 연구는 아직 전력화되지 않은 무인체계의 실제 사례보다는 개념적 시나리오와 국제적 논의를 중심으로 전개되었기 때문에 실증적 검증의 한계가 존재하며, 설명가능성과 통제가능성을 질적 수준에서 구분하였으나 이를 정량적으로 계량화하는 단계까지는 나아가지 못했다는 제약이 있다. 그럼에도 불구하고 본 연구는 AI 무인체계의 안전성과 신뢰성 확보를 위해 설명가능성과 통제가능성을 임무공학 절차와 정책 제도화 과정 속에 위치시키려는 기반적 접근을 시도했다는 점에서 의의를 갖는다. 향후 연구에서는 디지털 트윈 기반 시뮬레이션과 실제 시험평가 자료 축적을 통해 보다 실증적인 분석이 이루어져야 하며, 나아가 운용자 교육 및 조직 제도와의 연계, 국제 규범과 국내 제도의 상호작용을 구체적으로 검토하는 접근이 필요하다.

참고문헌

[1] DARPA, Explainable Artificial Intelligence (XAI) Program, Defense Advanced Research Projects Agency, 2017.
[2] Scharre, P., Army of None: Autonomous Weapons and the Future of War, New York: W. W. Norton, 2018.
[3] NATO Science and Technology Organization, Science & Technology Trends 2020-2040: Exploring the S&T Edge, Brussels: NATO STO, 2020.
[4] Horowitz, M. C., When Speed Kills: Autonomous Weapon Systems, Deterrence, and Stability, Washington, D.C.: Center for a New American Security, 2019.
[5] Boulanin, V., & Verbruggen, M., Mapping the Development of Autonomy in Weapon Systems, Stockholm International Peace Research Institute, 2017.
[6] Cummings, M. L., Artificial Intelligence and the Future of

Warfare, London: Chatham House, 2017.
[7] 김진우, AI 무기체계 확산이 전략적 안정성에 미치는 구조적 영향 분석에 관한 연구: 러시아–우크라이나 전쟁 사례를 중심으로, 경기대학교 정치전문대학원 국제정치학 박사 학위 논문, 2025.
[8] 김진우, 김성훈, AI AT WAR: 전쟁의 게임 체인저, AI, 박영사, 2025.
[9] U.S. Department of Defense, Mission Engineering Guide, Office of the Deputy Director for Engineering, Washington, D.C., 2020.
[10] McEver, J., Dahmann, J., & Lowry, R., "Mission Engineering for System of Systems," Insight, Vol. 22, No. 2, 2019, pp. 45–50.
[11] Boulanin, V., & Verbruggen, M., Mapping the Development of Autonomy in Weapon Systems, Stockholm International Peace Research Institute, 2017.
[12] U.S. Department of Defense, Department of Defense Instruction 5000.87: Operation of the Software Acquisition Pathway, Washington, D.C., 2020.
[13] ISO, ISO 31000: Risk Management – Guidelines, Geneva: International Organization for Standardization, 2018.
[14] United Nations, Report of the 2019 Session of the Group of Governmental Experts on Lethal Autonomous Weapons Systems (LAWS), CCW/GGE, Geneva: United Nations, 2019.
[15] NATO, NATO's Approach to Autonomy in Weapon Systems, Brussels: NATO Policy Paper, 2021.