한국군 국방지휘통제체계 AI 보안 요구사항 분석과 시사점

Security for AI Issues and Implications for the Korean Military Command and Control System

김진성[1] . 남승수[1] . 김찬수[2] . 장욱[2] . 최대선*,[1]

Jinsung Kim[1] . Seungsoo Nam[1] . Chansoo Kim[2] . Wook Jang[2] . Daeseon Choi*,[1]

[초 록]

KCCS는 전장 클라우드와 전군 C4I, 데이터·AI 플랫폼을 통합하는 데이터 중심 지휘통제체계로 발전하고 있으나, 기존 사이버 보안 프레임만으로는 AI 모델·데이터·프롬프트·운영정책의 취약점을 충분히 다루기 어렵다. 본 논문은 KCCS 관점에서 기존 사이버 보안과 AI 보안 간의 구조적 공백을 분석하고, 한국군 지휘통제체계에 필요한 AI 보안 통제 방향을 도출한다. 이를 위해 관련 연구와 국제 프레임워크를 바탕으로 C2 환경의 사이버 공격 유형과 보안 통제 항목을 정리하고, 이들이 주로 네트워크·시스템·데이터 중심으로 구성되어 AI 모델 보안과 보증 요구를 포괄하지 못함을 보인다. 이어서 데이터 포이즈닝, 적대적 예제, 프롬프트 인젝션, 모델 추출, AI 공급망 공격 등 AI 특화 취약점을 정리하고, 가드레일·정렬, AI 시험·보증, AI 공급망·프로비넌스가 부재한 KCCS 운용 시나리오를 통해 지휘결심상의 위험을 분석한다. 마지막으로 JADC2와의 비교를 통해 KCCS형 AI 보안 목록의 필요성과 우선 과제를 제안한다.

[ABSTRACT]

KCCS is evolving into a data-centric command and control system integrating the battlefield cloud, service-wide C4I systems, and data/AI platforms. However, existing cybersecurity frameworks are insufficient to address vulnerabilities in AI models, data, prompts, and operational policies. This paper analyzes the structural gap between conventional cybersecurity and security for AI from the perspective of KCCS and derives implications for AI security controls required in the Korean military command and control environment. Based on related studies and international frameworks, we categorize cyber attack types and cybersecurity controls in C2 environments and show that these controls remain focused on networks, systems, and data rather than AI model security and assurance. We then identify AI-specific vulnerabilities such as data poisoning, adversarial examples, prompt injection, model extraction, and AI supply chain attacks, and analyze operational risks in KCCS under scenarios where guardrails and alignment, AI testing and assurance, and AI supply chain and provenance are absent. Finally, by comparing KCCS with JADC2, we argue for the need for a KCCS-specific security-for-AI control catalogue and priority tasks.

Key Words : KCCS(차세대 지휘통제체계), JADC2(합동전영역지휘통제), Security for AI(AI 보안)

1. 서 론

1.1 연구배경

현대전은 정보 우위의 전쟁이다. 전장 곳곳의 센서가 만드는 데이터와 이를 실시간으로 처리하는 네트워크/클라우드 그리고 방대한 정보를 이해, 예측, 결심(Decision)으로 전환하는 AI 기술이 하나의 지휘통제체계 안에서 결합될 때 OODA(Observe-Orient-Decide-Act) 주기가 단축되고 Sensor-to-Shooter 흐름을 보장한다[1]. 데이터(센서)-네트워크(클라우드/엣지)-AI(분석/결심)의 결합이 곧 지휘통제(C2)의 전투력 자체가 될 수 있다. AI는 단순한 보조 수단이 아니라 분산된 전장 데이터를 신뢰 가능한 공통작전상황도(COP)와 행동 지침으로 바꾸는 핵심 도구이며, 전장 환경의 실시간성·불확실성·이기종성에 대응하도록 하는 핵심 기술 요소이다.

한국군은 공중·지상·해상·우주·사이버 합동 전 영역에서

1) 숭실대학교 AI 안전성 연구센터
(AI Safety Center, Soongsil University, Korea)
2) 국방기술진흥연구소
(Unmanned & Intelligent Technology Team, KRIT, Korea)
* Corresponding author, E-mail: sunchoi@ssu.ac.kr

Received : January 5, 2026 Revised : March 30, 2026
Accepted : March 30, 2026

신속·정확한 지휘 결심과 Sensor-to-Shooter 실행을 보장하기 위해 차세대 지휘통제체계 KCCS(Korea Command & Control System)를 단계적으로 전력화하고 있다[2]. 최근 연구에 따르면 KCCS를 전군 C4I(Command, Control, Communication, Computer)와 데이터·AI 플랫폼을 통합한 전장 클라우드 구조로 정의하고 있으며, 전장 데이터의 수집, 표준화, 공유, 활용을 통해 지휘관의 결심을 가속하는 것을 핵심 목적으로 명시하고 있다[3, 4]. 이러한 한국군 지휘통제체계의 발전 방향은 미국 국방부가 제시한 JADC2(Joint All-Domain Command & Control)를 한국군 환경에 맞게 구현하려는 연구와 같은 방향으로 발전하고 있다[1]. 특히 JADC2는 멀티센서 지각·융합, 상황 예측과 COA(Course of Action) 생성, 전술 엣지 배포 등 지휘 결심 전 과정에 AI 기술을 광범위하게 적용하며, 이를 신속 결심을 가능하게 하는 핵심 수단으로 규정하고 있다. 이에 발맞춰 KCCS도 전군 C4I와 데이터·AI 플랫폼을 통합하는 '전장 클라우드' 구상 아래, 상황 인식·예측·지휘 결심 지원과 엣지 분석 등 핵심 임무 영역에서 AI의 비중을 단계적으로 확대·고도화하는 방향으로 추진되고 있다[2].

하지만 AI 자체를 안전하게 운용하기 위한 AI 보안·보증(Security for AI)의 제도·기술 구성은 양 체계 간 괴리가 뚜렷하다. 미국은 JADC2 추진과 병행해 CDAO(Chief Digital and Artificial Intelligence Office)의 GenAI Toolkit(생성형 AI 가이드라인·가드레일), JATIC(Joint AI Test Infrastructure Capability)을 중심으로 한 AI 시험·보증(T&E/Assurance), DevSecOps v2.5와 Iron Bank/Platform One에 의거한 AIBOM(AI Bill of Materials)/SBOM(Software Bill of Materials)·서명·공급망 보증, Zero-Trust 참조 아키텍처 등을 정책-도구-운영 절차로 구체화해 운용한다[5-8]. 반면 KCCS에서 확인된 AI 보안 기술은 데이터·클라우드·AI 통합의 필요성과 전력화 청사진이 중심이며, 지휘 결심용 XAI 가드레일, AI 전용 T&E/Assurance, AIBOM(ML-BOM)/SBOM 등의 AI 보안 기술 규격·운영 절차는 향후 과제로 제시되는 단계에 머물러 있다[2,3]. 따라서, KCCS는 JADC2의 데이터·AI 통합 방향을 적극 수용하면서도, AI 보안 기술 레이어(XAI 가드레일, T&E, AIBOM)에 대한 제도화·표준화 체계는 미비하여 AI 모델에 대한 취약점 공백이 존재한다.

1.2 연구 목적 및 필요성

KCCS 도입되는 AI 기술은 전장 데이터의 이해·예측·결심을 가속하는 핵심 수단이지만, 동시에 AI 기술 취약점이 지휘 결심 오류로 직결될 수 있는 문제점을 가지고 있다. 예를 들어 AI 모델의 환각(Hallucination)과 근거 부재는 신뢰도 낮은 COA를 그럴듯한 명령으로 포장해 지휘 결심에 오판을 유도할 수 있고, 외부 지식에 의존하는 RAG 기반 참모는 자료 출처가 변조되거나 프롬프트 주입(Prompt Injection) 공격이 발생할 경우 보안 구역 노출, 좌표 변조, 규정 위반 지시처럼 실제 작전 위험으로 이어지는 명령을 산출할 수 있다. 또한 재밍(Jamming), 스푸핑(Spoofing), 통신 단절과 같은 학습 외 환경에서의 급격한 성능 저하, 적대적 공격과 데이터/모델 오염에 따른 공격 탐지 실패·허위 경보 증가는 시간 우위를 목표로 하는 C2의 장점을 상쇄하고, 우군 피해나 임무 실패로 이어질 수 있다. 공급망 측면에서도 모델, 데이터, 코드, 컨테이너 중 어느 하나라도 무결성이 깨지면 백도어 주입, 취약 라이브러리 악용, 파라미터 바꿔치기 등으로 이어져, 배포 후에는 원인 추적이 곤란해지는 구조적 리스크가 발생한다[5-8, 11-13].

따라서 본 논문에서는 지휘통제체계에 AI 기술이 도입되는 상황에서, 기존 사이버 보안으로는 정의할 수 없는 AI 모델의 보안 문제를 진단하고자 한다. 이를 위해 C2 환경에서의 사이버 공격·사이버 보안 목록을 살펴보고, AI 도입으로 인해 등장하는 새로운 취약점과 Security for AI 개념을 설명한다. 또한, JADC2에서 정책·도구·운영 절차 수준으로 구체화되고 있는 AI 보안·보증 사례를 정리하고, KCCS 관련 연구·정책과 비교·분석함으로써 KCCS에서 AI 보안이 사이버 보안 목록에 포함되지 않은 문제를 드러낸다. 마지막으로 한국군 지휘통제체계 관점에서 AI 보안 시사점과 향후 과제를 제안한다.

2. 지휘통제체계의 사이버 공격과 사이버 보안

지휘통제체계는 전장 클라우드, 합동망, 무기체계 인터페이스, 외부 정보망 등 다수의 구성요소가 복합적으로 연결된 시스템이므로, 일반 행정·업무망보다 넓은 공격 표면과 상호의존성을 가진다. 특히 KCCS와 같이 전군 C4I와 데이터·AI 플랫폼을 통합하는 구조에서는, 하나의 취약한 노드나 계정 탈취 사고가 COP 갱신 지연, 센서-슈터 간 데이터 흐름 차단, 잘못된 지휘명령 전파로 직결될 수 있다. 2장에서는 먼저 선행 연구와 공개 위협 보고서를 바탕으로, C2 환경에서 빈번하게 언급되는 대표적인 사이버 공격 유형을 정리하고, KCCS 관련 연구와 NIST(National Institute of Standards and Technology), ENISA(The European Union Agency for Cybersecurity), MITRE 등의 국제 프레임워크를 기반으로, 현재 논의되고 있는 사이버 보안 프레임과 통제 항목을 정리한다[2-4, 10, 12-14]. 마지막으로, 이러한 사이버 보안 목록이 주로 네트워크·시스템·데이터 중심으로 구성되어 있어, AI 모델·학습 데이터·프롬프트·정책을 별도로 다루지 못하고 있다는 한계를 분석함으로써 지휘통제 AI 보안 논의를 위한 시사점을 마련하고자 한다.

2.1 C2 환경에서의 사이버 공격 유형

지휘통제체계는 전장 클라우드, 합동망, 무기체계 인터페이스 등 다양한 요소가 결합된 복합 시스템으로, 일반 IT 시스템보다 넓은 공격 표면을 가진다[1-4]. 본 논문에서는 선행 연구와 위협 모델을 참고하여 C2 환경에서 빈번히 논의되는 공격을 피싱(phishing), 사회공학(social engineering), 내부자 위협(insider threat), 취약·레거시 시스템 방치(Unpatched / Legacy systems), 계정·권한 탈취(ATO), 서비스 거부(DDoS), 네트워크 침투와 수평 이동(lateral movement), 공급망(supply chain) 공격, 데이터 유출/변조

(Data tampering), 랜섬웨어(ransomware), 웹 취약점 공격(Web application attacks) 등으로 정리한다. 표 1은 공격들을 C2 관점에서 재구성한 것이다. 예를 들어, 피싱, 사회공학은 C2 운영자 계정 탈취와 악성 명령 주입으로 이어질 수 있으며, 취약·레거시 시스템 방치는 전장 클라우드의 특정 노드를 발판으로 삼는 침투 경로를 제공한다. 서비스 거부 공격은 COP 갱신 지연과 센서-슈터 간 정보 흐름 차단으로, 공급망 공격은 C2 소프트웨어 패치·업데이트 경로를 통한 악성 코드 유입으로, 데이터 유출/변조는 상황 인식 자체의 왜곡으로 연결된다. 이와 같이 C2 환경에서의 사이버 공격은 단순한 시스템 장애를 넘어, 지휘관의 상황 판단과 교전 규칙, 사격 통제에 직접적인 영향을 미칠 수 있다. 따라서 KCCS와 같은 차세대 지휘통제체계에서는 각 공격 유형을 명시적으로 식별하고, 이에 대응하는 사이버 보안 통제를 구조적으로 설계하는 것이 필수적이다.

표 1. 사이버 공격 유형 목록
Table 1. List of cyber attack categories

번호	분류	사이버 공격 유형
1	거버넌스	Phishing
2		Social engineering
3		Insider threat
4		Policy bypass / Policy evasion
5		Shadow IT abuse
6		내부자 이상 행위 (Insider anomaly / malicious insider)
7	엔드포인트	Unpatched / Legacy systems
8	공급망	Supply chain attack
9	IAM	Credential theft / Credential access
10		계정 탈취 (Account takeover, ATO)
11		권한 상승 (Privilege escalation)
12		세션 하이재킹 (Session hijacking)
13	네트워크	서비스 거부 공격 (DDoS, denial-of-service)
14		네트워크 침투 (Network intrusion)
15		수평 이동 (Lateral movement)
16	클라우드	클라우드 설정 오류 악용 (Cloud misconfiguration exploitation)
17	데이터	데이터 유출/침해 (Data breach / Data exfiltration)
18		기밀 정보 노출 (Confidential information exposure)
19		데이터 변조·무결성 훼손 (Data tampering / Integrity violation)
20	대응·복구	랜섬웨어 (Ransomware)
21		APT 공격 (Advanced persistent threat)
22		침해사고 확산 (Incident propagation)
23		대응 지연 (Delayed response)
24		재공격 노출 (Re-attack exposure)
25		복구 실패·영구 손실 (Recovery failure / Permanent loss)
26	애플리케이션	웹 취약점 공격 (Web application attacks)
27	공급망	취약 라이브러리·오픈소스 악용 (Vulnerable library / OSS exploitation)
28		소프트웨어 공급망 공격 (Software supply chain attack)

2.2 KCCS 관련 사이버 보안 프레임

KCCS 관련 연구와 자료를 보면, 지휘통제체계 보안은 주로 망·시스템·데이터 중심 사이버 보안 프레임에서 논의된다[2-4,10]. 망 분리·망 연동 정책을 통해 전장망, 행정망, 인터넷망을 층위별로 분리하고 필요 최소한의 연계를 허용하는 구조가 제시된다. 또한, 전장 클라우드와 엣지 노드에 대한 접근 통제, 암호화, 로깅·감사, 취약점 관리 등 클라우드/플랫폼 보안이 강조된다. 데이터 관리 측면에서는 전장 데이터의 표준화, 메타데이터 관리, 백업·복구 체계, 로그 감사·추적 등을 통해 데이터 기밀성, 무결성, 가용성을 보장하려는 노력이 논의된다. 국제 표준과의 정합성을 고려하여 NIST

표 2. 사이버 보안 유형 목록
Table 2. List of cybersecurity categories

번호	분류	사이버 보안 유형
1	거버넌스	보안 인식·훈련 (Security awareness and training)
2		정책 준수·강제 (Security policy enforcement)
3		지속적 개선·재발 방지 (Continuous improvement and lessons learned)
4	IAM	신원·접근관리 (Identity and access management)
5		강한 인증·MFA (Strong authentication and MFA)
6		계정 보호·이상 로그인 탐지 (Account protection and anomalous login detection)
7		최소 권한·권한 관리 (Least privilege and privilege management)
8	엔드포인트	자산·구성 관리 (Asset and configuration management)
9		취약점·패치 관리 (Vulnerability and patch management)
10		엔드포인트 보안 (Endpoint protection, AV/EDR)
11	네트워크	네트워크 경계보호·IDS/IPS (Network perimeter security and IDS/IPS)
12		네트워크 세분화·제로트러스트 (Network segmentation and zero trust)
13	클라우드	클라우드 설정 관리·CSPM (Cloud configuration hardening and CSPM)
14	데이터	데이터 보호·DLP·암호화 (Data protection, DLP and encryption)
15		데이터 분류·마스킹 (Data classification and masking)
16		무결성 보호·감사 로그 (Data integrity controls and audit logging)
17	애플리케이션	웹·API 보안·WAF·Secure SDLC (Web and API security)
18	공급망	공급망·SBOM·SCA·CI/CD 보안 (Supply chain security, SBOM and CI/CD security)
19	대응·복구	보안 모니터링·SIEM·UEBA·위협 헌팅 (Security monitoring, SIEM, UEBA and threat hunting)
20		침해사고 대응·격리 (Incident response and containment)
21		백업·재해복구·복원 테스트 (Backup, disaster recovery and restore testing)

CSF 기반의 Identify-Protect-Detect-Respond-Recover 기능, ENISA 위협·대응 카테고리, 국내 K-RMF와 연계된 DevSecOps, SBOM, 정적/동적 분석, 취약 오픈소스 감지 등 일반적인 사이버 보안 통제 항목이 KCCS·전장 클라우드 보안 논의의 주요 축으로 등장한다[12,13]. 이를 정리하면, 네트워크·클라우드 보안, 데이터 보안, 계정·권한 관리, 제로트러스트(Zero-Trust), DevSecOps, SBOM, 로그·감사·모니터링 등이 KCCS 관련 사이버 보안의 핵심축이라고 볼 수 있다. 표 2는 이러한 연구와 국제 프레임워크를 바탕으로, 사이버 보안 카테고리를 정리한 것이다. 그러나 표 2 에서 정리한 KCCS 관련 사이버 보안 목록을 자세히 살펴보면, 공통적으로 네트워크·시스템·데이터와 같은 전통적 자산을 중심으로 설계되어 있음을 확인할 수 있다.

2.3 AI 보안의 부재

암호화, 접근 통제, 취약점 관리, 로그 감사, 네트워크 분리·모니터링과 같은 통제는 필수적이지만, 이들은 AI 모델·학습 데이터·프롬프트·정책을 독립적인 보안 대상으로는 설정하지 않는다. 특히 AI 모델의 환각, 데이터 포이즈닝, 적대적 예제, 프롬프트 인젝션, 모델 추출, 모델 인퍼런스, AI 공급망 공격과 같은 AI 특화 위협은 대부분 기존 사이버 보안 분류에는 포함되어 있지 않거나, 일반적인 소프트웨어 취약점의 하위 항목으로 간접적으로만 언급된다[5-8,11-13]. 다시 말해, KCCS 관련 사이버 보안 구성에서는 AI를 보안 기술로 활용하는 AI for Security 관점은 있지만, AI 모델 자체를 보호·검증·추적해야 하는 Security for AI 관점은 독립적인 축으로 자리 잡지 못한 상태이다. 따라서 KCCS의 사이버 보안 목록에는 AI 보안이 사실상 부재하며, 이는 향후 KCCS에 AI 모델이 본격 투입될수록 새로운 취약점을 방치하는 구조적 요인이 될 수 있다.

3. 지휘통제체계의 AI 도입과 AI 보안 위협

앞선 2장에서 살펴본 바와 같이, KCCS를 포함한 지휘통제체계는 피싱, 내부자 위협, 네트워크 침투, 공급망 공격, 데이터 유출·변조, 랜섬웨어 등 다양한 사이버 공격에 노출되어 있으며, 이에 대응하기 위한 망·시스템·데이터 중심 사이버 보안 통제 체계를 점진적으로 확립해 나가고 있다. 그러나 이러한 통제는 어디까지나 전통적 IT 자산을 보호 대상으로 삼고 있으며, KCCS 내에 도입될 AI 모델과 학습·추론 파이프라인을 독립된 보안 대상으로 취급하지는 않는다. 그래서 3장에서는 먼저 JADC2와 KCCS에서 AI 모델이 어떤 임무 영역에, 어떤 방식으로 활용되고 있는지 설명하고, AI 모델의 취약점과 공격 기법을 AI를 위한 보안 관점에서 정리한다. 나아가 이러한 취약점이 KCCS 운용 맥락에서 구체적으로 어떤 지휘결심 오류·상황 인식 왜곡·공급망 리스크로 이어질 수 있는지 시나리오 형태로 제시함으로써, 기존 사이버 보안 통제만으로는 충분히 커버되지 않는 AI 보안 공백을 입증하고자 한다.

3.1 JADC2·KCCS에서의 AI 활용 개요

3.1.1 JADC2 AI 기술

JADC2에서는 여러 임무 상황에 맞게 폭넓은 AI 기술을 사용한다. DoD는 생성형 AI 전력화를 담당하는 AI RCC(AI Rapid Capabilities Cell)를 출범시켜 C2·결심지원 등 15개 핵심 사용례를 중심으로 파일럿-확대 적용을 추진 중이다[9]. 구체적으로는 멀티센서 데이터를 상관·분류하는 지각·융합, 적 행동과 전장 양상을 예측하는 상황 이해·예측, COA 생성·비교·워게이밍 기반 결심 지원, 자율/유무인 복합(MUM-T) 운용, RAG 기반 지식·문서 AI, 군수/정비/네트워크 최적화, CLOPS·DevSecOps 기반 전술 배포·운영(cATO 지향) 등에서 AI가 활용된다[1,5-7]. 이처럼 JADC2에서는 AI가 COP 구성, 위협 평가, 교전 의사결정, 전력 배분, 네트워크 운용 등 C2 전 과정에 깊숙이 내장되어 있다.

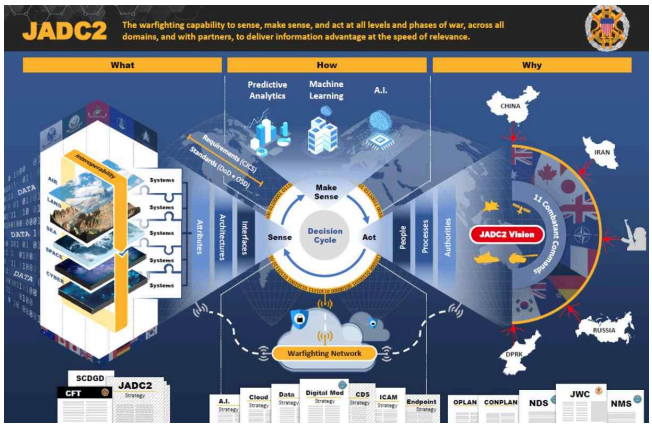

그림 1. JADC2 구조도[5]

Fig. 1. JADC2 placemat[5]

3.1.2 KCCS AI 기술

KCCS에서도 인공지능은 단순한 보조 도구가 아니라, 전장 인식에서 지휘결심, 전투 수행과 군수·통신 운용에 이르는 전 주기를 관통하는 핵심 기능으로 활용될 것으로 예상된다. 레이더·무인 정찰체계 등 다양한 센서에서 수집되는 이종 데이터를 AI가 자동으로 상관·분류해 중복 표적과 오경보를 줄이고, COP에 통합된 전장 그림을 제공하는 기능이 대표적이다. 나아가 과거 교전 사례, 적 활동 패턴, 우리군 배치 정보 등을 결합한 모델을 통해 위협도 평가, 적 행동·전개 양상 예측, 전장 양상 변화 전망과 같은 상황 이해·예측 기능을 제공함으로써 지휘관의 선제적 결심을 지원한다[2-4,10]. 향후에는 RAG 기반 지휘참모, 표적 추천, 교전 규칙 검토, 군수·정비 예측, 네트워크 자원 최적화 등으로 AI 활용 영역이 확대될 가능성이 크다.

3.2 AI 보안의 개념과 취약점

기존 사이버 보안에서는 AI를 활용하여 보안 기능을 강화하는 AI for Security와, AI 자체를 보호·검증·통제하는 Security for AI가 종종 혼용되지만, 두 개념은 목적과 보호 대상이 다르다. 전자는 침입 탐지, 이상 행위 탐지, 위협 분석 등 보안 기능을 강화하기 위한 AI 활용을 의미하는 반면, 후자는 AI 모델, 데이터, 프롬프트, 운영정책을 공격과 오동작으로부터 보호하고 그 신뢰성을 보증하는 것을 의미한다. 본 논문은 Security for AI에 초점을 맞춘다. AI 보안 관점에서 대표적인 취약점은 다음과 같이 정리할 수 있다. 첫째, 데이터 포이즈닝(data poisoning)과 백도어(backdoor) 삽입은 학습 단계에서 데이터나 레이블을 조작하여 특정 입력에서 의도된 오판을 유도하거나 전체 성능을 은밀히 열화시키는 공격이다. 둘째, 적대적 예제(adversarial examples)는 추론 단계에서 입력을 미세하게 교란하여 모델의 오분류를 유도하는 기법으로, 센서 노이즈·위장·환경 교란과 결합될 경우 전장 상황 인식과 표적 분류의 신뢰성을 훼손할 수 있다. 셋째, 프롬프트 인젝션(prompt injection)과 탈옥(jailbreak)은 LLM·RAG 기반 시스템에서 모델의 정책과 역할을 우회하여 금지된 정보 접근이나 규정 위반 출력을 유도하는 공격이다. 넷째, 모델 추출(model extraction)과 모델 역추론(model inversion, membership inference)은 반복 질의와 출력 분석을 통해 모델의 내부 행태나 학습 데이터의 일부 민감 정보를 추론하는 공격이다. 다섯째, AI 공급망 공격(AI supply chain attack)은 학습 데이터셋, 파인튜닝 파이프라인, 모델 가중치, 배포 아티팩트, 의존 라이브러리 등 공급망의 어느 한 지점을 겨냥해 악성 요소를 삽입하거나 무결성을 훼손하는 공격을 의미한다. 이들 취약점은 KCCS에서 아직 공개된 직접 사고 사례가 충분히 축적되지 않았더라도, 유사한 AI 시스템에서는 이미 반복적으로 실증되어 왔다. 예를 들어 BadNets 연구는 학습 공급망이 오염될 경우 특정 트리거가 부착된 표지판을 오인식하도록 만드는 백도어 모델이 가능함을 보였고, 물리 세계 적대적 예제 연구는 카메라를 통해 취득한 실제 영상에서도 모델 오분류가 재현될 수 있음을 입증하였다. 또한 Hidden Voice Commands 연구는 사람이 인지하기 어려운 음성 신호가 기계에는 명령으로 해석될 수 있음을 보였으며, 최근의 간접 프롬프트 인젝션 연구들은 외부 문서나 검색 결과에 삽입된 지시가 LLM 통합형 응용과 RAG 시스템의 응답을 왜곡할 수 있음을 보여준다. OWASP 역시 2025년 LLM Top 10에서 프롬프트 인젝션, 학습 데이터 포이즈닝, 공급망 취약성 등을 핵심 위험으로 제시하고 있다[15-21]. 따라서 KCCS와 같이 센서-네트워크-AI-결심이 긴밀히 연결된 지휘통제체계에서는, 이러한 유사 사례를 단순한 연구실 수준의 문제가 아니라 실제 작전 보안과 결심 보증의 문제로 해석할 필요가 있다. 표 3은 이러한 AI 취약점을 학습·추론·운영·공급망 단계별로 정리한 것이다. 이 표는 이후 JADC2와 KCCS의 AI 보안 체계를 비교할 때, 어떤 취약점에 대해 어떤 통제와 보증 수단이 필요한지를 판단하는 기준으로 활용된다.

표 3. AI 모델 취약점 유형 목록
Table 3. List of AI model vulnerability categories

번호	AI 모델 취약점 유형
1	데이터 포이즈닝·백도어 (Data poisoning / Backdoor)
2	적대적 예시·회피 공격 (Adversarial examples / Evasion)
3	모델 도난·추출 (Model stealing / Extraction)
4	프라이버시·멤버십 추론 (Privacy / Membership inference)
5	환각·불신뢰 출력 (Hallucination / Unreliable outputs)
6	편향·불공정성 (Bias / Unfairness)
7	드리프트·성능 열화 (Data/Concept drift, Model degradation)
8	AI 공급망·모델 변조 (AI supply chain and Model tampering)

3.3 KCCS 관점에서의 AI 공격 시나리오

KCCS가 전장 클라우드 위에서 다양한 AI 서비스를 함께 구동하는 구조라는 점을 감안하면, 앞서 정리한 AI 보안 기능이 없을 때 다음과 같은 구체적인 위험이 발생할 수 있다. 첫째, 가드레일·정렬 부재 시나리오이다. KCCS 내부에 LLM 기반 참모 서비스나 표적 추천 AI가 탑재되어 있음에도, 역할·임무·보안등급에 따른 출력 제한, OPSEC/PII 자동 필터링, 고위험 명령에 대한 Human-in/On-the-Loop 및 2인 승인 절차가 명확히 정의·도구화되어 있지 않은 상황을 가정한다. 이 경우 프롬프트 인젝션이나 잘못 설계된 RAG 문맥으로 인해 노출되면 안 되는 민감 좌표, 구체적인 작전 계획, 암호키 유사 정보와 같은 기밀성이 높은 정보가 질의·응답 과정에서 그대로 노출될 수 있다. 또한, 교전 규칙 완화, 특정 좌표에 대한 타격 허용과 같은 고위험 결심을 요구하는 질의가 단순 시나리오 검토 차원으로 제기되었더라도, 이를 자동으로 걸러내거나 사람 승인 단계로 전환하는 장치가 없다면, 해당 결과가 실제 작전 권고나 암묵적 기준으로 오인·내재화될 가능성이 존재한다. 둘째, AI 시험·보증 부재 시나리오이다. 전장 클라우드 위에서 위협도 예측 모델, 표적 분류 모델 등이 정기적으로 갱신·배포되지만, 이들에 대해 AI 전용 T&E 체계 없이 제한된 기능 시험 위주로만 검증이 이루어지는 상황에서는 센서 고장, GPS 스푸핑, 통신 지연과 같이 KCCS 운용 환경에서 빈번히 발생할 수 있는 비정상 조건에 대해, 모델이 어떻게 성능 열화·오작동을 보이는지 평가하는 표준 시험지표와 합격 임계값(KPI)이 정의되어 있지 않다면, 특정 조건에서만 반복적으로 발생하는 오판 패턴을 사전에 식별하기 어렵다. 더 나아가 모델 업데이트 이

후 특정 표적 유형만 지속적으로 과소·과대 평가된다거나, 특정 지역·부대에 편향된 예측을 내는 등 성능 저하·편향 징후가 현실에서 관찰되더라도, 카나리 배포·롤백 기준이 없으면 이를 신속히 되돌리지 못한 채 운용할 수밖에 없다. 이러한 상황은 평시 기능 시험에서는 정상으로 보였던 AI 모듈이, 실제 전장·위기 환경에서만 일관된 오판을 야기하는 블랙스완형 실패 모드로 이어질 수 있으며, KCCS 전체의 신뢰성을 심각하게 저하시킬 수 있다. 셋째, AI 공급망·프로비넌스 부재 시나리오이다. KCCS용 AI 모델, 학습 데이터, 전장 로그, 프롬프트 템플릿 등이 복수의 방산 업체, 연구기관, 운용부대 사이를 오가며 개발·개선·재배포되는 구조를 상정할 때, 어떤 데이터로 어떤 버전의 모델이 학습되었는지, 누가 언제 어떤 설정으로 이를 빌드·배포했는지를 일관되게 추적할 수 있는 AIBOM/SBOM 및 서명 체계가 없다면, 공급망 전 과정에 대한 신뢰 사슬이 형성되지 않는다. 이 경우 백도어가 삽입된 모델이 정상 패치로 위장하여 유입되는 악의적인 모델 바꿔치기, 특정 표적을 항상 비위협으로 학습시키도록 설계된 의도적 데이터 오염, 승인되지 않은 비공식 모델·프롬프트의 무단 사용 등이 발생하더라도, 그 출처와 영향 범위를 신속히 파악·차단하기 어렵다. 결국, 망 분리·체계 인증과 같은 전통적 보안 조치만으로는 충분하지 않으며, AI 자산(모델·데이터·정책) 단위의 계보 관리, 전자 서명, 정책 게이트 없이는 내부 공급망을 통한 공격과 오류가 지휘결심 전 과정에 서서히 스며들 수 있다는 점이 문제로 남는다. 이러한 시나리오들은 기존의 사이버 보안 통제만으로는 AI 모델 특유의 취약점을 충분히 커버할 수 없음을 보여주며, KCCS에 별도의 AI 보안 사항이 필요함을 시사한다.

위 시나리오들은 단순한 가정적 서술이 아니라, 공개된 AI 보안 연구와 산업 가이드라인에서 반복적으로 확인된 위험 양상을 KCCS 운용 맥락으로 변환한 것이다. 즉, 학습 데이터와 모델 공급망의 오염 가능성, 물리 환경에서의 적대적 교란, 외부 문서를 통한 간접 프롬프트 인젝션, 반복 질의를 통한 모델 정보 추론 등은 이미 일반 AI 시스템과 LLM 응용에서 실증되었으며, 지휘통제체계에서는 그 결과가 단순 서비스 오류가 아니라 상황 인식 왜곡, 지휘결심 오판, 규정 위반 명령 생성으로 연결될 수 있다는 점에서 위험의 수준이 더 높다. 따라서 KCCS에서 AI 보안은 기존 사이버 보안의 세부 항목이 아니라, 지휘결심의 신뢰성과 작전 통제 가능성을 보장하기 위한 독립적 관리 영역으로 다루어져야 한다.

4. JADC2와 KCCS의 AI 보안 비교 및 시사점

2장과 3장에서 살펴본 바와 같이, KCCS는 전장 클라우드와 데이터·AI 플랫폼을 중심으로 한 데이터 중심 C2를 지향하면서도, 사이버 보안 측면에서는 여전히 전통적인 네트워크·시스템·데이터 보호에 초점을 두고 있고, AI 모델 자체를 보호·보증 대상으로 삼는 전용 AI 보안은 부재하다는 것을 확인하였다. 반면 미군의 JADC2 생태계에서는 CDAO의 GenAI Toolkit, JATIC, DevSecOps v2.5 및 AIBOM/SBOM 체계 등, AI 보안·보증을 위한 정책-도구-운영 절차가 점진적으로 구체화되고 있다[5-8,10-13]. 4장은 이러한 차이를 보다 구조적으로 비교·정리하고, KCCS에 요구되는 AI 보안 체계 설계 방향을 도출하는 것을 목표로 한다. 먼저 JADC2 사례를 기반으로, 지휘통제 AI를 위한 보안·보증을 가드레일·정렬, AI 시험·보증, AI 공급망·프로비넌스의 세 축으로 정리하고, 이어서 KCCS 관련 문헌과 정책 자료를 분석하여 이 세 축이 한국군 지휘통제체계에서는 어떤 부분이 부재한지를 도출한다. 마지막으로, 이러한 비교 결과를 바탕으로 KCCS 사업·운영·연구개발 측면에서의 정책적·기술적 시사점과 우선 과제를 제안함으로써, 한국군 지휘통제체계 AI 보안 이슈 및 시사점을 도출한다.

4.1 JADC2의 AI 보안

미국은 JADC2 추진과 병행하여 AI 보안을 독립적인 계층으로 취급하고, 다음의 세 축을 중심으로 정책·도구·운영 절차를 정교화하고 있다[5-8]. 첫째, 가드레일·정렬이다. CDAO는 생성형 AI 가이드라인·가드레일을 공표하고 이를 GenAI Toolkit으로 구현하여, 체크리스트·평가·승인 흐름을 통해 C2-AI에 대한 정책-as-code를 적용한다[5]. 허용된 데이터 소스만 연결하는 RAG 도메인 제한, 인용·버전 강제, OPSEC·PII 자동 필터링, 고위험 명령에 대한 Human-in/On-the-Loop 및 2인 승인 절차, 출력 레드액션(redaction)과 같은 원칙이 여기에 포함된다. 둘째, AI 시험·보증이다. JATIC는 재현성·추적성·강건성을 중심으로 하는 AI 시험·평가/보증 도구와 프레임워크를 공개 제공하며, 프롬프트/데이터 주입, 환각, 적대적 입력, OOD 환경에 대한 강건성 시험카드와 KPI를 정의한다[6]. 또한, 배포 전·후에는 카나리/롤백 게이트와 연동하여, 성능 저하·편향·드리프트가 감지되면 자동으로 등급을 낮추거나 이전 버전으로 복귀할 수 있는 구조를 지향한다. 셋째, AI 공급망·프로비넌스이다. DevSecOps v2.5, Iron Bank, Platform One 등을 통해 소프트웨어·AI 아티팩트에 대한 SBOM·서명·정책 게이트를 적용하고, AIBOM/ML-BOM을 활용해 모델 가중치, 학습·평가 데이터, 프롬프트·정책·설명 규칙까지 메타데이터로 묶어 관리한다[7,8]. 이를 통해 빌드-검증-배포 전 과정을 추적·검증하고, 누가·언제·무엇을 사용했는지를 계보 형태로 남겨 바꿔치기·오염·스푸핑 위험을 줄인다. 요약하면, JADC2는 AI 모델을 지휘통제체계 내부의 독립 자산으로 보고, 가드레일·정렬, 시험·보증, 공급망·프로비넌스라는 세 축을 통해 AI 보안을 구조적으로 관리하고 있다.

4.2 KCCS AI 보안 이슈

이에 비해 KCCS 관련 연구에서는 데이터·클라우드·망 통합, 제로 트러스트, DevSecOps 등 사이버 보안 관점의 논의가 두드러지지만, AI를 보안·보증 대상으로 다루는 제도·도구·운영 절차는 거의 찾아보기 어렵다[2-4,10,12,13]. 지휘결심용 XAI 가드레일, AI 전용 T&E/Assurance, AIBOM/SBOM 등은 대체로 향후 검토 과제 수준에서 언급되거나, 아예 등장하지 않는 경우가 많다.

KCCS는 사이버 보안 측면에서는 네트워크·클라우드·데이터를 보호하기 위한 다양한 통제 항목을 갖추고 있지만, AI 보안 목록과 도구는 비어 있는 상태에 가깝다. 즉, KCCS는

AI를 활용하는 지휘통제체계로 설계되면서도, AI 모델·데이터·정책을 별도의 보안 계층으로 다루는 명시적 프레임은 마련되어 있지 않다.

표 4. JADC2와 KCCS의 AI 보안 기술 비교
Table 4. Current status of security for AI tools in JADC2 and KCCS

구분	JACD2 AI 보안 기술	KCCS AI 보안 기술
가드레일·정렬	GenAI Toolkit : 존재	AI 특화 가드레일 : 미정립
시험·보증 기술	JATIC 등 전용 체계 : 존재	AI 전용 시험체계 : 미정립
AI 공급망·프로비넌스	AIBOM/SBOM 체계 : 존재	AI 공급망 개념 : 미정립

4.3 한국군 지휘통제체계의 AI 보안 시사점

이상의 분석이 시사하는 바는 단순히 KCCS에도 AI 보안이 필요하다"는 선언적 수준에 머물지 않는다. 오히려 2장과 3장의 분석을 종합하면, KCCS에 AI가 본격적으로 투입되는 순간 기존의 네트워크·시스템·데이터 중심 사이버 보안만으로는 지휘결심의 신뢰성과 통제 가능성을 충분히 보장하기 어렵다는 점이 보다 분명해진다. 즉, 기존 사이버 보안은 계정 탈취, 네트워크 침투, 데이터 유출·변조와 같은 전통적 위협에 효과적이지만, AI 모델의 환각, 프롬프트 인젝션, OOD 환경에서의 급격한 성능 열화, 모델 업데이트 이후의 편향·드리프트, 공급망상의 모델·정책 바꿔치기와 같은 AI 특화 실패 양상은 직접 통제하지 못한다. 따라서 KCCS에는 기존 사이버 보안 목록과 별도로, AI 자산을 독립적인 보호·보증 대상으로 설정하는 KCCS형 AI 보안 목록이 필요하다. 이러한 AI 보안 목록은 본 논문이 제안한 세 축, 즉 가드레일·정렬, AI 시험·보증, AI 공급망·프로비넌스를 중심으로 구성될 수 있다. 첫째, 가드레일·정렬은 지휘결심 지원 AI가 어떤 근거를 사용할 수 있는지, 어떤 수준의 출력을 허용하는지, 어떤 질의와 행동을 사람 승인 단계로 전환해야 하는지를 정책-as-code 형태로 규정하는 계층이다. 둘째, AI 시험·보증은 기존 기능 시험을 넘어 적대적 입력, 센서 불일치, 통신 지연, GPS 스푸핑, 데이터 드리프트와 같은 전장형 비정상 조건에서 모델이 어느 수준까지 성능을 유지하는지 검증하는 계층이다. 셋째, AI 공급망·프로비넌스는 모델, 학습 데이터, 프롬프트, 정책 규칙, 컨테이너, 라이브러리의 출처와 변경 이력을 추적하여, 지휘통제 AI의 신뢰 사슬을 형성하는 계층이다. 이 세 축은 기존 인프라 보안을 대체하는 것이 아니라, 그 위에 추가되어 AI의 특수한 실패 양상을 관리하는 보완적 통제로 이해되어야 한다. 정책적 우선순위도 명확하다. 첫째, 지휘결심과 직접 연결되는 LLM/RAG 기반 참모 기능, 위협도 평가, 표적 추천과 같은 고위험 사용례부터 AI 보안 요구사항을 명시해야 한다. 둘째, 획득·개발 단계에서는 AI 전용 시험평가 기준과 합격 임계값, 카나리 배포 및 롤백 기준을 제도화해야 한다. 셋째, 운영 단계에서는 모델·데이터·정책에 대한 계보 관리와 변경 통제를 DevSecOps 및 제로 트러스트 체계와 연동해야 한다. 결국 KCCS에서 AI 보안의 핵심은 AI를 쓸 것인가의 문제가 아니라, 어떤 조건과 통제 아래에서 AI를 작전 가능한 형태로 신뢰할 것인가의 문제라고 할 수 있다. 이러한 관점 전환이 이루어질 때 비로소 KCCS의 AI 전력화는 기술 시범 수준을 넘어 지휘통제체계의 제도화된 능력으로 정착할 수 있다.

5. 결 론

본 논문은 데이터 중심 지휘통제체계를 지향하는 KCCS에 AI 기술이 본격 도입되는 상황에서, 기존 사이버 보안 프레임만으로는 관리하기 어려운 AI 보안 공백이 존재함을 분석하고, 이에 대한 정책적·기술적 함의를 도출하고자 하였다. 분석 결과, KCCS 관련 사이버 보안 논의는 네트워크·시스템·데이터 보호에 초점을 두고 있으며, AI 모델·학습 데이터·프롬프트·운영 정책을 독립적인 보호 및 보증 대상으로 설정하지 못하고 있음을 확인하였다. 반면 AI 시스템에서는 데이터 포이즈닝, 백도어, 적대적 예제, 프롬프트 인젝션, 모델 추출, 공급망 오염과 같은 취약점이 공개 연구와 산업 가이드라인을 통해 반복적으로 제시되고 있으며, 이러한 위험은 지휘통제 환경에서 상황 인식 왜곡과 결심 오류로 이어질 가능성이 높다. 따라서 KCCS의 AI 보안 문제는 단순한 기능 개선이 아니라, 작전 신뢰성과 통제 가능성을 보장하기 위한 체계 설계의 문제로 이해되어야 한다. 또한 JADC2 사례 분석을 통해, AI 보안은 단순한 원칙 수준이 아니라 가드레일·정렬, AI 시험·보증, AI 공급망·프로비넌스의 세 축을 중심으로 정책-도구-운영 절차의 형태로 구체화될 수 있음을 확인하였다. 이는 KCCS가 동일한 제도를 그대로 복제해야 한다는 뜻이 아니라, 최소한 AI를 독립적인 관리 대상으로 식별하고 지휘결심과 직접 연결되는 사용례부터 통제 요구사항을 구조화해야 함을 의미한다. 다시 말해, KCCS에는 기존 사이버 보안 카탈로그를 보완하는 별도의 "KCCS형 AI 보안 카탈로그"가 필요하며, 이 카탈로그는 AI의 입력, 근거, 출력, 행동, 업데이트, 공급망을 포괄해야 한다. 본 논문의 최종 주장은 다음과 같다. 첫째, 지휘통제체계에서 AI 보안은 기존 사이버 보안의 부속 항목이 아니라 독립된 보안 계층으로 제도화되어야 한다. 둘째, 그 구조는 가드레일·정렬, AI 시험·보증, AI 공급망·프로비넌스의 세 축을 중심으로 설계되는 것이 타당하다. 셋째, 한국군은 KCCS의 AI 전력화 이전 또는 최소한 병행 단계에서 고위험 사용례에 대한 AI 보안 요구사항, 시험평가 기준, 공급망 계보 관리 체계를 조기에 마련해야 한다. 이러한 조치가 수반되지 않을 경우, KCCS의 AI 도입은 결심 속도 향상이라는 이점을 제공하는 동시에 새로운 구조적 취약점을 지휘결심 체계 내부에 내재화할 가능성이 있다. 향후 연구에서는 본 논문이 제시한 세 축을 보다 구체적인 실증 과제로 전환할 필요가 있다. 예를 들어 KCCS용 LLM/RAG 참모 기능에 대한 프롬프트 인젝션·근거 오염 시험, 센서 융합 모델에 대한 적대·환경 강건성 시험, 모델·데이터·정책에 대한 KCCS형 AIBOM/ML-BOM 스키마 설계와 같은 후속 연구가 요구된다. 이러한 실증이 축적될 때, KCCS는 단순히 JADC2 개념을 수용하는 수준을 넘어 한국군 작전 환경에 적합한 지휘통제 AI 보안·보증 기준을 자체적으로 정립할 수 있을 것이다.

References

[1] J. Park and J. Kim, “Strategic Framework for Adapting AI Foundation Models to JADC2: Focused on the Korean Command and Control System,” Journal of Information Technology Services, Vol. 24, No. 3, pp. 1-16, 2025.

[2] G. Kim and T. Park, “Development Direction of the ROK Military Command and Control through the U.S. JADC2 Initiative: Focusing on the U.S. Air Force ABMS Program and its Approach to AI,” KIDA Defense Forum, No. 2057, Korea Institute for Defense Analyses, Seoul, Republic of Korea, Sept. 1, 2025.

[3] S. Park, J. Kang, Y. Lee, and J. Kim, “Research on functional area-specific technologies application of future C4I system for efficient battlefield visualization,” Journal of Information and Security, vol. 23, no. 4. Korea Convergence Security Association, pp. 109-119, 31-Oct-2023.

[4] S. Cho, Y. Kim, C. Jung and S. Hong, “Data Integration Plan for the AI-based Next-Generation Command and Control System (KCCS),” Korea Institute for Defense Analyses (KIDA), 2024.

[5] U.S. Department of Defense, “Summary of the Joint All-Domain Command and Control Strategy,” U.S. Department of Defense, Washington, DC, pp. 1-12, March, 2022.

[6] Chief Digital and Artificial Intelligence Office, “GenAI Toolkit: Generative AI Version of the Responsible AI Toolkit,” U.S. Department of Defense, Washington, DC, pp. 1-105, December, 2024.

[7] Chief Digital and Artificial Intelligence Office Public Affairs, “Responsible AI Toolkit,” U.S. Department of Defense, Washington, DC, November, 2023.

[8] Chief Digital and Artificial Intelligence Office, “Testing and Evaluation,” Pathway to AI Readiness, U.S. Department of Defense, 2025.

[9] Chief Digital and Artificial Intelligence Office, “Artificial Intelligence Rapid Capabilities Cell,” U.S. Department of Defense, Washington, DC, pp. 1-2, December, 2024.

[10] U.S. Department of Defense, “DoD Enterprise DevSecOps Fundamentals,” Office of the DoD Chief Information Officer, Washington, DC, pp. 1-44, October, 2024.

[11] U.S. Air Force Platform One, “Iron Bank: Hardened Container Repository,” U.S. Department of Defense, 2021.

[12] National Institute of Standards and Technology, “The NIST Cybersecurity Framework (CSF) 2.0,” U.S. Department of Commerce, Gaithersburg, MD, February 26, 2024.

[13] European Union Agency for Cybersecurity (ENISA), “ENISA Threat Landscape 2023/2024,” European Union Agency for Cybersecurity, Heraklion, Greece, 2023-2024.

[14] MITRE Corporation, “MITRE ATT&CK®: A Knowledge Base of Adversary Tactics and Techniques,” MITRE, McLean, VA, accessed 2025.