Vol. 2026-02
November 2026

동향 AI 안전 동향 분석 리포트 26-02

군사 AI의 인간 판단·통제: 국제 정책·연구 동향

Human Judgment and Control over Military AI:
International Policy and Research Trends

유고운



ETRI 한국전자통신연구원 www.etri.re.kr
인공지능안전연구소 AI Safety Institute

AI 안전 동향 분석 시리즈 26-02(2026년 11월)

「군사 AI의 인간 판단·통제: 국제 정책·연구 동향」

유고운(석사후연수연구원)

한국전자통신연구원 인공지능안전연구소 AI안전정책 및 대외협력실

본 보고서는 주요국 및 국제기구·다자협의체의 군사 분야 AI 적용 관련 정책·원칙 및 규범 동향을 바탕으로 분석·정리한 자료입니다. 보고서 내용은 정보 제공 및 연구 목적을 위한 것이며, 원문 자료와 해석 또는 적용상 차이가 있는 경우에는 원문이 우선합니다. 또한 본 보고서의 분석 및 해설에는 연구진의 견해가 포함될 수 있으며, 소속 기관의 공식 입장을 의미하지 않습니다.

AI Safety Trends Analysis Series 26-02 (November 2026)

「Human Judgment and Control over Military AI: International Policy and Research Trends」

Goun Yoo (Post-Master's Research Fellow)

AI Safety Policy and International Collaboration Section, Korea AI Safety Institute (AISI), Electronics and Telecommunications Research Institute (ETRI)

This report provides an analysis of policies, principles, and normative developments regarding the application of AI in the military domain across major countries, international organizations, and multilateral forums. It is intended as a reference material for informational and research purposes, and in the event of any discrepancy between this report and the original source materials, the original sources shall prevail. The analyses and interpretations presented in this report may reflect the views of the authors and do not necessarily represent the official position of their affiliated institutions.

Contents

Ⅰ. 논의 배경과 인간 통제 개념 ····· 1

1. 인간 통제 개념의 형성과 국제 논의 동향 ····· 1

2. 인간 통제 수단의 구분과 실효성 조건 ····· 9

3. 인간 통제의 대상과 구성요소 ····· 11

4. 개념 적용의 확장 쟁점 ····· 14

Ⅱ. 주요국·국제기구의 군사 AI 정책 문서에서의 인간 통제 ····· 16

5. 대한민국(Korea) ····· 16

6. 미국(US) ····· 18

7. 영국(UK) ····· 21

8. 유럽연합(EU) 및 NATO ····· 23

9. 프랑스(France) 및 독일(Germany) ····· 26

10. 일본(Japan) ····· 27

11. 중국(China) ····· 29

Ⅲ. 인간 통제 구현 수단별 연구동향 ····· 31

12. 인간 판단과 의사결정의 실효성 ····· 31

13. AI 행동범위의 사전 제한과 감독·개입 ····· 33

14. 시험·평가·검증 및 AI 보증 ····· 34

15. 인간-AI 상호작용과 다중 자율체계의 통제 ····· 36

16. 책임·추적성과 생애주기 통제 ····· 37

17. 연구동향 종합 ····· 38

Ⅳ. 종합 및 시사점 ····· 40

요 약 문

AI의 군사적 활용은 더 이상 무인 플랫폼이나 자율무기에 국한되지 않는다. 그 범위는 정보·감시·정찰(ISR), 지휘통제, 지휘관 의사결정 지원, 표적 식별 및 우선순위 결정, 유무인 복합운용, 사이버 작전, 군수 및 교육훈련으로 확장되었다. 본 보고서는 이처럼 확장된 영역에서 인간 통제 개념이 어떻게 변화해 왔는지, 주요국이 이를 어떻게 제도화했는지, 그리고 국제인도법 수호기관·정책연구기관·기술연구기관이 이 문제를 어떻게 다루고 있는지를 분석한다.

인간 통제의 개념: 초기 논의는 인간이 개별 행위를 하나하나 승인하는지를 기준으로 단일 구분을 통해 문제를 설정하였다. 최근의 정책과 연구는 이러한 운용상 분류에 더해, 이를 가로지르는 별도의 관점인 전 생애주기 인간 통제를 함께 제시한다. 이는 개발, 검증, 시험평가, 임무 범위 설정, 배치, 운용, 무효화, 사후 검토에 이르는 전 과정에 인간의 판단과 책임을 위치시키는 접근이다. 두 관점은 서로 경쟁하는 범주가 아니라 서로 다른 층위에서 작동한다. 형식적 승인 절차만으로는 충분하지 않다. 운용자가 권고의 근거를 이해할 수 없거나, 판단할 시간이 부족하거나, 수백 건의 권고를 동시에 처리해야 한다면 그 통제는 실질적 통제가 아닌 형식적 통제에 그친다. 본 보고서는 각국 및 기관의 정책에서 반복적으로 나타나는 요소를 확립된 분류가 아닌 분석 틀로서 여섯 가지로 묶는다. 인간의 판단 및 승인, 자율성 범위의 사전 제한, 감독·무효화·비활성화, 설명가능성 및 상황인식, 인간의 책임성, 개발 및 시험평가 단계의 통제가 그것이다.

주요국 정책: 각국의 용어는 서로 다르다. 미국은 국방부 훈령 3000.09와 책임 있는 AI 체계를 통해 "무력 사용에 대한 적절한 수준의 인간 판단"을 제도화하고 있으며, 2026년 6월 발표된 NSPM-11은 국가안보 AI 정책을 도입·적응·보증·책임성을 중심으로 재편하고 해당 훈령의 개정을 지시하였다. 영국은 "맥락 적합한 인간 관여"라는 표현을 사용하며, 이를 JSP 936으로 구체화하고 2026년 6월부터는 신속 AI 도입 태스크포스(Rapid AI Delivery Taskforce)를 통해 신속하게 추진하고 있다. 독일은 2026년 국방부 내부에 구속력 있는 AI 개념을 발표하였고, 프랑스는 완전자율 치명무기체계를 개발하지 않겠다는 입장을 유지하면서 AMIAD를 통해 역량을 구축하고 있다. 일본은 방위장비청(ATLA)을 통해 연구개발 단계의 책임 있는 AI 지침을 발표하였고, 중국은 무기체계가 인간의 통제하에 있으며 언제든 중단될 수 있다는 점을 강조한다. NATO는 인간 통제를 통제가능성(Governability)을 포함한 여섯 가지 책임 있는 사용 원칙과 연계하고 있으며, 한국은 REAIM 정상회의 공동주최를 통해 국제 논의를 주도하는 한편 국방혁신 4.0에 따라 역량을 강화하고 있다. 유의미한 인간 통제(Meaningful Human Control)는 여전히 중요한 담론적 개념이지만, 각국이 일률적으로 채택한 확정된 단일 법적 용어는 아니다.

연구 동향: 본 보고서는 검토한 기관별 연구를 여섯 가지 변화로 읽어내며, 이는 확립된 분류가 아닌 본 보고서의 자체적 종합이다. 무기의 자율성에서 의사결정의 자율성으로, 최종 승인 단계의 통제에서 전 생애주기 통제로, AI 성능에서 AI 보증으로, 단일 체계에서 다중 에이전트 및 군집 체계로, 기술적 인터페이스에서 조직 및 교육훈련 문제로의 전환, 그리고 AI와 전략적 안정성, 특히 AI-핵 연계가 독자적 연구 축으로 부상한 것이 그것이다. 한편 다자 논의도 지속되었다. 특정재래식무기금지협약(CCW) 정부전문가그룹(GGE)은 2026년 9월 향후 문서의 구성요소를 담은 최종보고서를 컨센서스로 채택하면서 인간 관여에 관한 표현을 수식어 없는 '통제와 인간 판단(control and human judgement)'으로 정리하였다. 제네바 현지시간 2026년 11월 21일 새벽 종료된 제7차 검토회의는 이 구성요소를 바탕으로 치명적 자율무기체계에 적용되는 국제인도법 원칙과 의무 및 이행 방안을 구체화하는 문서를 2027~2028년에 작성하도록 GGE에 권한을 부여하였으나, 문서의 법적 형식은 열어 두었으며 참관인 참여를 둘러싼 이견으로 협약에 대한 정식 검토는 완료하지 못하였다. 제3차 REAIM 정상회의는 2026년 2월 「행동으로의 경로(Pathways to Action)」를 채택하였으나 미국과 중국의 지지는 얻지 못하였다.

Executive Summary

Military applications of AI are no longer confined to uncrewed platforms or autonomous weapons. They now span ISR, command and control, commander decision support, target identification and prioritization, manned–unmanned teaming, cyber operations, logistics and training. This report analyzes how the concept of human control has evolved across this expanded domain, how major states have institutionalized it, and how the ICRC, policy institutes and technical research agencies are addressing it.

The Concept of Human Control: Early debate framed the question through a single distinction defined by whether a human approves each individual action. Recent policy and research pair that operational classification with a distinct, cross-cutting perspective: lifecycle human control, which locates human judgment and accountability across development, verification, testing and evaluation, mission scoping, deployment, operation, override and after-action review. The two operate at different levels rather than as competing categories. A formal approval step is not in itself sufficient: if the operator cannot understand the basis of a recommendation, has too little time, or must process hundreds of recommendations at once, the control is nominal rather than substantive. This report groups the elements that recur across national and institutional policy into six, as an analytical framework rather than an established classification: human judgment and approval; pre-set limits on the scope of autonomy; supervision, override and deactivation; explainability and situational awareness; human accountability; and controls at the development and test-and-evaluation stage.

National Policies: Terminology diverges. The United States institutionalizes "appropriate levels of human judgment over the use of force" through DoD Directive 3000.09 and its Responsible AI framework, while NSPM-11 of June 2026 reorganizes national security AI policy around adoption, adaptation, assurance and accountability and directs an update of that directive. The United Kingdom uses "context-appropriate human involvement," operationalized through JSP 936 and, since June 2026, pursued at speed through the Rapid AI Delivery Taskforce.

Germany issued a binding internal AI concept for the defense ministry in 2026; France maintains that it will not develop fully autonomous lethal systems while building capability through AMIAD; Japan has issued an R&D-stage responsible AI guideline through ATLA; China stresses that weapon systems remain under human control and can be suspended at any time; NATO ties human control to its six Principles of Responsible Use, including Governability; and Korea has helped lead international deliberations by co-hosting the REAIM Summit while advancing capability under Defense Innovation 4.0. Meaningful Human Control remains an important discourse concept rather than a single settled legal term adopted uniformly across states.

Research Trends: This report reads the institutional research surveyed here as showing six shifts, offered as the report's own synthesis rather than an established taxonomy: from autonomy in weapons to autonomy in decision-making; from final-approval control to lifecycle control; from AI performance to AI assurance; from single systems to multi-agent and swarm systems; from technical interfaces to organizational and training questions; and the emergence of AI and strategic stability, particularly the AI-nuclear nexus, as a distinct research axis. Multilateral deliberation has meanwhile continued: in September 2026 the CCW GGE adopted by consensus a final report setting out elements of a possible future instrument, settling on the unqualified formulation "control and human judgement," and the Seventh Review Conference, which closed in the early hours of 21 November 2026 (Geneva time), mandated the GGE to elaborate in 2027–2028 an instrument specifying the principles and obligations of international humanitarian law applicable to lethal autonomous weapons systems and measures for their implementation, leaving its legal form open, although a dispute over observer participation prevented a formal review of the Convention and its Protocols; and the third REAIM Summit adopted Pathways to Action in February 2026, though without the support of the United States or China.

Ⅰ 논의 배경과 인간 통제 개념

1 인간 통제 개념의 형성과 국제 논의 동향

1-1. 군사 AI 적용 범위의 확대

■ 군사 AI의 적용은 무인체계·자율무기에 한정되지 않고 군사 활동 전반으로 확대되고 있으며, 이에 따라 인간 통제 논의의 대상도 함께 확장됨

- 현재 적용 영역은 정보·감시·정찰(ISR), 지휘통제(C2), 지휘관 의사결정 지원, 표적 식별 및 우선순위 설정, 유무인 복합체계, 사이버 작전, 군수·정비, 교육·훈련 등을 포괄함
- 한국의 국방 AI 기술로드맵 역시 감시정찰, 지휘통제, 유무인 복합체계, 사이버의 4개 분야를 적용 대상으로 설정하고 있음
- 이에 따라 정책적 쟁점은 '인간이 AI가 수행하는 군사적 판단과 무력 사용을 어느 수준까지 어떤 방식으로 통제해야 하는가'로 정식화됨
- 다만 인간 통제 논의는 AI가 이용되는 군사적 맥락에 따라 그 근거와 요구되는 통제 수준이 달라지므로, 무력 사용에 관한 AI 이용과 그 밖의 군사적 맥락에서의 AI 이용을 구분할 필요가 있음
- (무력 사용 맥락) 국제관계에서의 무력 사용은 UN헌장 제2조 제4항 및 국제관습법상 원칙적으로 금지되며, 무력충돌 시에도 전투의 방법과 수단을 선택할 권리는 무제한적이지 않다는 국제인도법 원칙이 적용됨. 이에 따라 무력 사용에 관한 결정을 완전한 자율성을 지닌 AI에 위임해서는 안 된다는 논리에서 LAWS 관련 인간 통제 논의가 전개됨[1)]
- (그 밖의 군사적 맥락) 정보·감시·정찰, 군수·정비, 교육·훈련 등 무력 사용 결정을 직접 수반하지 않는 영역에서의 AI 이용은 위와 논의의 근거가 다르며, 요구되는 인간 통제의 수준도 달라짐
- 의사결정지원체계는 표적 선정 등 무력 사용 결정을 지원하는 경우 전자의 논의와 연결되고, 그 밖의 지휘·참모 업무를 지원하는 경우 후자에 해당함
- 이러한 구분에 따라 Ⅲ장의 각 절에는 해당 연구가 주로 다루는 맥락을 [무력 사용], [그 밖의 군사

1) Charter of the United Nations, Art. 2(4); CCW GGE on LAWS, Report of the 2024-2025-2026 Sessions of the Group of Governmental Experts on Emerging Technologies in the Area of LAWS, CCW/GGE.1/2026/3, para. 26, adopted 4 Sep 2026

맥락], [공통]으로 표기함

■ **Ⅱ장과 Ⅲ장의 분석 범위와 자료의 성격은 다음과 같음**

- Ⅱ장은 주요국·국제기구의 군사 AI 정책 문서에서 인간 통제가 어떻게 규정되어 있는지를 분석 대상으로 하며, CCW GGE·REAIM 등 국제 논의에서 각국이 보인 입장은 이를 보완하는 범위에서 함께 다룸
- Ⅲ장에서 다루는 ICRC·UNIDIR·SIPRI 등의 문헌은 국제규범을 형성하는 주체의 문서가 아니라 국제인도법 수호기관(ICRC)과 연구기관(UNIDIR·SIPRI 등)의 연구로서, CCW GGE와 REAIM에서 이루어지는 국가 간 논의를 보충하는 자료로 활용함

1-2. 국제 논의에서의 개념 형성

■ **국제사회는 자율무기체계(LAWS) 규율을 논의하는 과정에서 인간 통제를 국제인도법(IHL) 준수 여부를 판단하는 핵심 개념으로 설정함[2)]**

- 2013년 4월 유엔 특별보고관 Christof Heyns는 인권이사회 보고서에서 인간의 추가 개입 없이 표적을 선택·공격하는 치명적 자율로봇이 구별·예방·비례성 등 국제인도법 원칙을 준수할 수 있는지, 위법한 살상이 발생할 경우 누가 책임을 지는지의 문제를 제기하여 이를 유엔 차원의 공식 의제로 만듦[3)]
- 같은 해 ICRC는 CCW 당사국회의에서, 전장에서 본격적으로 사용되기 전에 실명 레이저 무기를 선제적으로 금지한 1995년 CCW 제4의정서(Protocol IV on Blinding Laser Weapons)를 자율무기 논의의 선례로 제시함[4)]
- 특정재래식무기금지협약(CCW, 2026년 11월 1일 기준 당사국 128개국) 체제는 2016년 자율무기체계 정부전문가그룹(GGE on LAWS)을 설치하였으며, 자율성이 증대될수록 무력 사용에 대한 인간의 관여를 어떻게 유지할 것인지가 논의의 중심축으로 형성됨
- 2019년 GGE 보고서는 국제인도법이 자율무기체계를 포함한 모든 무기체계에 전면 적용된다는 점(지침원칙 a)과, 무력 사용의 적법성 확보를 위해 '인간의 판단(human judgement)'이 필수적이라는 점(제17항 e)을 확인함

2) CCW, Report of the 2019 Session of the GGE on LAWS, Annex IV (Guiding Principles), CCW/GGE.1/2019/3, 25 Sep 2019
3) UN Human Rights Council, Report of the Special Rapporteur on Extrajudicial, Summary or Arbitrary Executions, Christof Heyns, A/HRC/23/47, 9 Apr 2013
4) ICRC, Statement by Kathleen Lawand, Meeting of the High Contracting Parties to the CCW, Summary Record of the Second Meeting, CCW/MSP/2013/SR.2, paras. 17-18, Nov 2013

■ **인간 통제를 지칭하는 용어는 주체별로 상이하게 발전하였으며, GGE 롤링 텍스트 논의에서는 '맥락 적합적 인간 판단 및 통제'라는 절충적 표현이 부상하였으나, 2026년 9월 채택된 최종보고서에는 수식어 없이 '통제와 인간 판단(control and human judgement)'이라는 표현으로 최종 반영됨[5)]**

- (유의미한 인간 통제, Meaningful Human Control) : 시민사회 및 다수 중견국이 제기한 개념으로 무력 사용에 대한 실질적·능동적 통제를 요구하며 통제의 '질'을 강조함
  - 영국 NGO Article 36이 2013년 처음 제시한 개념으로, 인간이 자율무기의 임무 파라미터를 설정하는 것만으로는 충분한 통제라 할 수 없으며 개별 공격마다 유의미한 인간 통제가 있어야 한다고 주장함[6)]
  - 그 요건으로 ① 정보(information): 운용자 등 공격 계획 책임자가 표적 영역의 맥락, 표적 선정 사유, 임무 목적, 예상되는 즉각적·장기적 무기 효과에 관한 정보를 보유할 것, ② 행동(action): 공격 개시에 인간 운용자의 적극적 행동이 요구될 것, ③ 책임(accountability): 정보 평가와 공격 실행에 책임 있는 자가 공격 결과에 책임을 질 것을 제시함
  - 2015년 CCW 비공식 전문가회의에서 인간 통제 논의를 진전시키기 위한 개념으로 다뤄졌으며, 의사결정 과정에 인간 개입이 존재하는지보다 통제의 질을 요구한다는 점에 특징이 있음
  - 다만 의미가 모호하고 국제인도법상 확립된 법적 개념이 아니라는 이유로 일부 국가가 용어 채택에 반대함

- (적절한 수준의 인간 판단, Appropriate Levels of Human Judgment) : 미국이 채택한 표현으로, 고정된 단일 기준이 아니라 무기체계·전쟁 영역·작전 맥락별로 달라지는 유연 기준임을 전제함
  - 미국은 2018년 GGE 논의부터 이 표현을 지지해 왔으며, MHC는 국제인도법상 개념화되지 않은 용어라는 입장을 취함
  - 또한 인간 통제(human control)와 인간 판단(human judgment)은 구분되어야 하고, 필요한 인간 관여의 수준은 고정될 수 없으며, 인간 개입이 많을수록 반드시 더 적절한 것은 아니라는 논리를 제시함[7)]

- (맥락 적합적 인간 판단 및 통제, Context-appropriate Human Judgement and Control) : 요구되는 통제의 정도를 표적 유형·사용 환경·기술적 성능 등 맥락 변수에 연동시키는 표현으로, 롤링 텍스트에서 상이한 입장을 연결하기 위한 절충안으로 등장함
  - 다수 대표단은 필요한 인간 통제의 수준이 개별 상황에 의존한다는 데 동의하였으며, 이는 절대적 통제를 요구하지 않는 대신 판단 기준의 불명확성을 남기는 결과를 초래함
  - 2025년 불가리아·덴마크·핀란드·프랑스·독일·이탈리아·룩셈부르크·네덜란드·노르웨이·스페인·

5) CCW GGE on LAWS, Report of the 2024-2025-2026 Sessions of the Group of Governmental Experts on Emerging Technologies in the Area of LAWS, CCW/GGE.1/2026/3, adopted 4 Sep 2026
6) Article 36, Killer Robots: UK Government Policy on Fully Autonomous Weapons, Apr 2013
7) U.S. Government, Human-Machine Interaction in the Development, Deployment and Use of Emerging Technologies in the Area of LAWS, CCW/GGE.2/2018/WP.4, 28 Aug 2018

스웨덴 등 유럽 11개국은 공동 작업문서에서 이 표현을 지지하며 개념을 구체화함[8)]

- 국제인도법에 '인간 판단 및 통제'를 규정한 독립된 명문 규칙은 없으나, 구별·비례성·예방 등의 의무를 준수하려면 현실적으로 인간의 판단과 통제가 필요하며, 그 수준은 맥락에 따라 달라진다고 봄

■ 논의는 무기 중심의 CCW 체제를 넘어 유엔 총회 차원의 포괄적 의제로 확장됨[9)]

- 2024년 12월 24일 유엔 총회는 「군사 영역의 인공지능과 국제 평화·안보에 대한 함의」 결의(A/RES/79/239)를 최초로 채택, AI의 전 생애주기에 국제법이 적용됨을 확인함
  - 동 결의는 한국과 네덜란드가 2024년 10월 유엔 총회 제1위원회에 공동 제출하여 11월 6일 위원회에서 채택된 후 총회에서 최종 채택되었으며, 사무총장에게 각국 견해를 수렴하도록 요청하면서 CCW에서 다루는 자율무기체계 이외의 영역에 특히 초점을 두도록 함[10)]
  - 이에 따라 사무총장은 2025년 6월 군사 영역의 AI 이용이 국제 평화·안보에 제기하는 기회와 도전에 관한 각국 견해를 종합한 보고서(A/80/78)를 총회에 제출함[11)]

1-3. 논의 경과 및 현재 지위

■ 자율무기체계 논의의 주된 장인 CCW의 정식 명칭은 '과도한 피해를 주거나 무차별적인 효과를 초래할 수 있는 특정 재래식 무기의 사용 금지 또는 제한에 관한 협약(Convention on Prohibitions or Restrictions on the Use of Certain Conventional Weapons Which May Be Deemed to Be Excessively Injurious or to Have Indiscriminate Effects)'으로, 무력충돌에서 과도한 상해나 불필요한 고통을 초래하거나 무차별적 효과를 갖는 특정 재래식 무기의 사용을 금지·제한하여 전투원과 민간인을 보호하는 것을 목적으로 함[12)]

- 동 협약은 1-1에서 언급한 '전투의 방법과 수단을 선택할 권리는 무제한적이지 않다'는 원칙과, 과도한 상해 또는 불필요한 고통을 초래하는 성격의 무기 사용을 금지하는 국제인도법 원칙에 기반함
- 프레임워크 협약과, 구체적으로 금지·제한되는 무기를 규율하는 5개 의정서로 구성되며, 의정서별로

8) Bulgaria et al., Working Paper on the "Human Element", CCW/GGE.1/2025/WP.4, Aug 2025
9) UN General Assembly, Artificial Intelligence in the Military Domain and Its Implications for International Peace and Security, A/RES/79/239, 24 Dec 2024; A/RES/80/58, 1 Dec 2025
10) UN General Assembly, Artificial Intelligence in the Military Domain and Its Implications for International Peace and Security, A/RES/79/239, 24 Dec 2024
11) UN Secretary-General, Artificial Intelligence in the Military Domain and Its Implications for International Peace and Security, A/80/78, 5 Jun 2025
12) Convention on Prohibitions or Restrictions on the Use of Certain Conventional Weapons Which May Be Deemed to Be Excessively Injurious or to Have Indiscriminate Effects, adopted 10 Oct 1980, entered into force 2 Dec 1983

당사국이 상이함

■ 2023년 CCW 당사국회의가 GGE에 부여한 현행 권한은 '문서(instrument)의 성격을 예단하지 않은 채 그 구성요소를 컨센서스로 도출하고 그 밖의 가능한 조치를 검토'하는 것이며, GGE는 그 결과를 제7차 검토회의에 최종보고서로 제출하도록 했음[13)]

- GGE는 2026년 3월 2~6일과 8월 31일~9월 4일 두 차례 회기를 거쳐 3년 권한의 마지막 해 작업을 마쳤으며, 의장이 작성한 '롤링 텍스트(rolling text)'를 기초로 논의한 끝에 2026년 9월 4일 최종보고서(CCW/GGE.1/2026/3)를 컨센서스로 채택함

- 최종보고서는 합의된 구성요소(set of elements)를 담고 있으며, 제7차 검토회의에 보고되어 검토회의가 향후 협상 결과를 예단하지 않은 채 이를 검토·승인(endorse)하도록 권고함. 다만 이는 조약이나 법적 구속력을 갖는 문서가 아니라 향후 논의의 기초에 해당함

- 특히 최종보고서 제34항은 국제인도법, 특히 구별·비례성·공격 시 예방 원칙의 준수를 위해 자율무기체계에 대한 '통제와 인간 판단(control and human judgement)'이 필요하다고 명시함[14)]
  - 그간 논의된 '유의미한', '적절한', '맥락 적합적' 등의 수식어 없이 채택되었으며, '인간'은 '판단'에만 결합되고 '통제'에는 붙지 않음
  - 통제는 인간이 사용 전 또는 사용 중에 취하는 조치를 통해 직접 또는 간접적으로 행사될 수 있으며, 통제 방식과 관계없이 국제인도법 준수에 필요한 조치를 결정·이행하는 데에는 인간의 판단이 필요하다고 규정함
  - 이 과정에서 작전 맥락과 무기체계의 특성·성능 등을 고려해야 한다고 명시하여, 맥락 의존성은 수식어가 아닌 본문 설명으로 반영됨

- 최종보고서가 병기한 '통제'와 '인간 판단'은 서로 다른 개념으로, 2025년 유럽 11개국 작업문서는 두 개념을 다음과 같이 구분함[15)]
  - (인간 판단) 무기체계의 개발·사용과 관련하여 인간이 자신의 판단을 행사하는 것으로, 특히 공격 개시 시 구별·예방·비례성 등 국제인도법 규칙과 원칙의 준수에 관한 판단을 포함함
  - (통제) 무기체계가 활성화된 이후에도 예측 가능한 방식으로 작동한다는 충분한 확신을 확보하는 것으로, 이를 통해 체계가 초래하는 효과가 국제인도법, 교전규칙 및 지휘관·운용자의 의도에 부합하는지 판단할 수 있어야 함

13) CCW GGE on LAWS, Report of the 2024-2025-2026 Sessions of the Group of Governmental Experts on Emerging Technologies in the Area of LAWS, CCW/GGE.1/2026/3, adopted 4 Sep 2026

14) CCW GGE on LAWS, Report of the 2024-2025-2026 Sessions of the Group of Governmental Experts on Emerging Technologies in the Area of LAWS, CCW/GGE.1/2026/3, para. 34, adopted 4 Sep 2026

15) Bulgaria et al., Working Paper on the "Human Element", CCW/GGE.1/2025/WP.4, para. 8(b)-(c), Aug 2025

- 다만 일부 국가는 통제가 인간이 아닌 소프트웨어를 통해서도 이루어질 수 있다는 점을 들어 'human control'이라는 표현에 우려를 나타냄

- 최종 회기에서 브라질이 47개국을 대표해 발표한 공동성명은 해당 구성요소가 GGE 권한을 충족하며 CCW 내 협상의 기초로 충분하다는 입장을 밝혔고, 아프리카그룹도 같은 취지의 성명을 발표함[16)]

- 2026년 11월 16일 제네바에서 개막한 제7차 검토회의는 현지시간 11월 21일 새벽 종료되었으며, 당사국들은 9월 GGE가 합의한 구성요소를 바탕으로 치명적 자율무기체계에 적용되는 국제인도법 원칙과 의무 및 이행 방안을 구체화하는 문서를 2027~2028년에 작성하도록 GGE에 권한을 부여함[17)]
  - 문서의 법적 형식은 열어 두었으며, GGE는 구속력 있는 의정서로 채택할지 여부를 포함한 권고안을 2028년 당사국회의에 제출할 예정임
  - 다만 참관인 참여를 둘러싼 이견으로 의제와 절차규칙에 합의하지 못해, 검토회의의 본래 임무인 협약과 기존 의정서에 대한 정식 검토는 완료하지 못함[18)]
  - GGE는 컨센서스 방식으로 운영되어 개별 국가가 사실상 거부권 행사가 가능하며 강한 금지·규제 언어가 합의 과정에서 완화되는 경향이 반복적으로 지적됨

■ 규제 방식으로는 일부 유형을 금지하고 나머지를 국제인도법에 따라 규율하는 '2단계 접근(two-tier approach)'에 대한 지지가 확대됨[19)]

- 최종보고서는 과도한 상해·불필요한 고통을 초래하거나 본질적으로 무차별적이거나 국제인도법에 따라 사용될 수 없는 체계의 사용 금지(제31항)와, 그 밖의 체계를 구별·비례성·예방 원칙 등 국제인도법에 따라서만 사용하도록 하는 규율(제33항) 등 향후 문서에 담을 금지 및 규율 요소를 제시함

- 다만 자율무기체계의 정의, 규율 범위, 표적 결정에 필요한 인간 관여 수준 등 근본 개념에 대한 국가 간 이견이 지속되고 있음

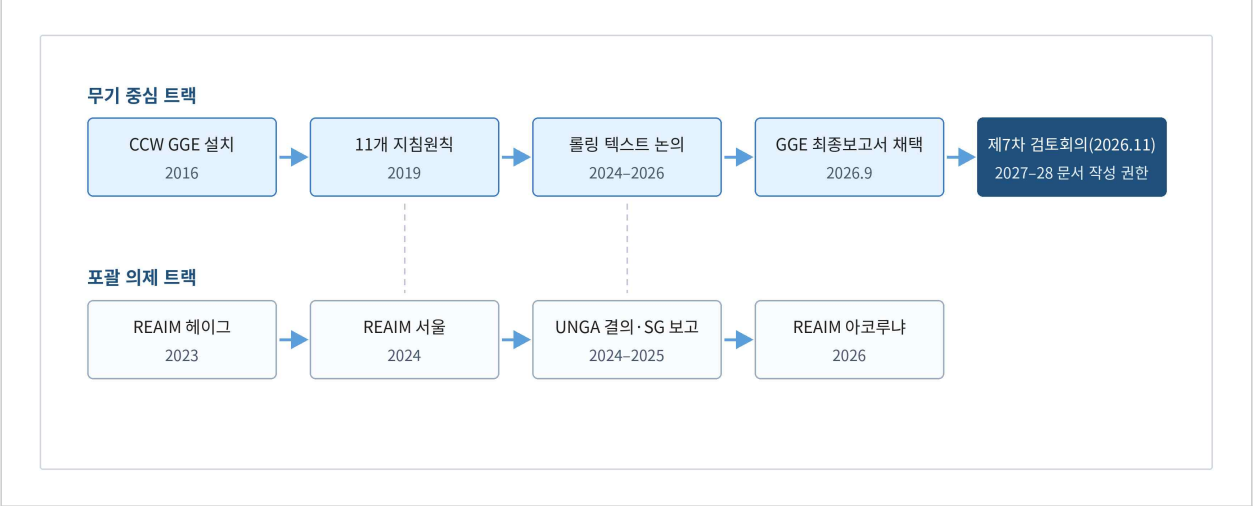
16) Automated Decision Research, Over 70 States Now Support Elements Work at CCW GGE as a Basis for Negotiations, Sep 2026; Stop Killer Robots, Joint Statement to the CCW GGE on LAWS, Sep 2026

17) Swiss Federal Department of Foreign Affairs, New Mandate on Lethal Autonomous Weapons Systems Adopted in Geneva: The FDFA Welcomes This Step Forward, 21 Nov 2026; Permanent Representation of Italy to the Conference on Disarmament, Autonomous Lethal Weapons: A New Two-Year Mandate Has Been Adopted, 25 Nov 2026

18) UN Web TV, 11th Meeting, Seventh Review Conference of the High Contracting Parties to the CCW, 21 Nov 2026; APILS, CCW Seventh Review Conference: A New Mandate on Autonomous Weapons, But No Review, 21 Nov 2026

19) CCW GGE on LAWS, Report of the 2024-2025-2026 Sessions of the Group of Governmental Experts on Emerging Technologies in the Area of LAWS, CCW/GGE.1/2026/3, paras. 31, 33, adopted 4 Sep 2026

| 그림 1. 군사 AI 이용 관련 국제 논의 트랙 |



출처 CCW GGE·REAIM 및 UNODA 공개 문서를 기초로 재구성

■ **유엔군축연구소(UNIDIR)는 2026년 9월 발간한 보고서에서, 군사 AI 규범 논의에서 회원국 간 수렴이 나타나는 영역과 이견이 지속되는 영역을 구분하여 제시함**[20]

- 수렴 영역: 군사 AI의 생애주기 전반에 대한 국제법 적용, 무력 사용에서의 통제와 인간 판단 유지, 인간의 책임과 책무성, 무력충돌 시 민간인 보호, 국제협력의 가치
- 이견 영역: 용어에 대한 공통 이해, 자발적 규범과 구속적 규범 중 어느 쪽이 바람직한지, 레드라인의 식별, 논의를 진행할 적절한 포럼과 방식

1-4. REAIM 프로세스와 2026년 결과문서

■ **REAIM은 CCW GGE와 같은 조약 당사국 간 논의체가 아니라 일부 국가의 이니셔티브로 출범한 정상회의 형식의 논의체로, 채택 문서에 대한 지지 여부도 국가별로 선택함**

- CCW GGE가 협약의 목적 범위 내에서 자율무기체계를 다루는 데 비해, REAIM은 자율무기체계 문제를 중복해 다루기보다 그 밖의 군사 AI 문제를 폭넓게 다루고, 그 논의를 유엔 총회 결의를 통해 유엔 차원으로 확대하는 방식으로 운영됨
- (제1차, 2023년 2월 헤이그) 군사 영역 AI의 책임 있는 개발·배치·사용 문제를 국제 의제로 제기하고 「행동촉구(Call to Action)」를 채택함(정상회의 당시 57개 국가 및 지역 지지, 후속 집계로는 네덜란드 정부가 2025년 4월 공개한 명단 기준 58개 국가 및 지역)[21]

20) B. Unal, From Cyberspace to AI: Translating Norms of Responsible State Behaviour in Cyberspace to AI in the Military Domain, UNIDIR, 18 Sep 2026

- (제2차, 2024년 9월 서울) 국제법 준수, 인간의 통제·판단, 책임성, 신뢰성, 위험관리 등의 원칙을 담은 「행동청사진(Blueprint for Action)」을 채택함(정상회의 당시 61개국 지지)[22)]

■ 2026년 2월 4~5일 스페인 아코루냐에서 제3차 REAIM 정상회의가 개최되어 결과문서 「REAIM Pathways to Action」이 채택됨[23)]

- 2023년 헤이그 「행동촉구(Call to Action)」와 2024년 서울 「행동청사진(Blueprint for Action)」의 원칙을 실행 단계로 옮기는 것을 목적으로 하며, 총 28개 항으로 구성됨
- 채택 당시 참석 85개국 중 35개국이 지지하였으며(후속 집계로는 스페인 외교부 공개 명단 기준 45개국), 미국과 중국은 지지국에 포함되지 않아 주요 군사국의 참여가 제한된 상태에서 채택됨
- 주요 내용은 생애주기 전반의 국제법 준수, '설계 단계부터의 책임(responsible by design)' 접근, 책임 공백(accountability gap) 방지, AI 보증(AI assurance) 분야 진전, TEVV의 정책·교리·조달 절차 통합, 신무기 법적 검토 수행 등임
- 특히 제12항은 AI 기반 의사결정지원체계가 인간의 판단 행사를 지원해야 하며 대체해서는 안 된다는 점과, 인간 관여의 성격과 정도가 작전 맥락·수행 기능·기술적 특성 및 훈련·피로 등 인적 요인을 고려하여 결정되어야 한다는 점을 명시함
- 2026년 6월 제네바에서는 유엔 총회 결의 80/58에 따른 비공식 다자 이해관계자 교류가 개최되었고, 케냐가 2027년 4월 제4차 정상회의 개최를 발표함

21) REAIM 2023, Call to Action, 16 Feb 2023 (57 endorsing countries and territories at the Summit); Government of the Netherlands, REAIM 2023 Endorsing Countries and Territories, published 4 Apr 2025 (58 countries and territories)

22) REAIM 2024, Blueprint for Action, 10 Sep 2024; Reuters, South Korea Summit Announces 'Blueprint' for Using AI in the Military, 10 Sep 2024 (61 endorsing countries at the Summit)

23) REAIM 2026, REAIM Pathways to Action, A Coruña, 5 Feb 2026; CFR, Military AI Adoption Is Outpacing Global Cooperation, Feb 2026 (35 endorsing countries at the Summit); Spain Ministry of Foreign Affairs, REAIM 2026 endorsing countries (45 countries, list checked 26 Nov 2026)

2 인간 통제 수단의 구분과 실효성 조건

2-1. 루프 기반 구분과 그 한계

| 참고. 루프 기반 3유형*|

Human-in-the-loop	행동 실행 전 인간의 승인이 필요한 형태
Human-on-the-loop	AI가 자율 작동하되 인간이 감독하고 중단·개입할 수 있는 형태
Human-out-of-the-loop	활성화 이후 실시간 승인·감독이 부재한 형태

*초기 논의에서는 운용 국면에서 인간이 개별 행동을 승인하는지를 기준으로 통제를 세 유형으로 구분하였음

■ 세 유형은 보조적 설명 도구로서 유용하나, 이것만으로는 통제의 실질을 판정하기 어렵다는 점이 공통적으로 지적됨

- 운용자가 최종 승인 절차를 수행하더라도 판단 근거를 이해할 수 없거나, 시간 압박과 정보 비대칭이 존재하거나, 다수의 추천을 동시에 처리해야 하는 경우 승인은 형식적 절차에 그칠 수 있음
- 2026년 GGE 논의와 REAIM 결과문서도 루프 위치가 아니라 표적·지역·시간 제한, 감독·개입·중단, 책임 소재, 추적성 등 실제 통제 수단을 중심으로 구성되어 있음
- 이에 따라 본 보고서는 루프 구분을 분석 틀로 사용하지 않고, 3-2의 6대 구성요소와 실제 통제 수단을 중심 분석 틀로 삼음

■ 생애주기 전반의 인간 통제 관점은 위 세 유형과 층위를 달리하는 접근으로, 별도의 확립된 개념 유형이 아니라 통제를 바라보는 시야의 확장에 해당함

- 루프 기반 구분이 운용 시점의 위치를 나누는 데 비해, 생애주기 관점은 개발·검증·임무설정·배치·운용·중단·사후평가 전 과정을 가로지르는 거버넌스 시각임
- REAIM 2026 결과문서가 '설계 단계부터의 책임' 접근을 통해 최초 기획부터 설계·개발·시험·배치·모니터링·폐기까지 책임 있는 AI 원칙을 통합하도록 권고한 것도 같은 층위의 접근에 해당함

2-2. 통제 수단의 구분 기준

■ **통제 수단은 단일 기준으로 나열되지 않으며, 통제가 행사되는 시점, 통제의 대상, 거버넌스 수단이라는 서로 다른 기준이 결합되어 적용됨[24)]**

- (시점 기준) 일부 GGE 문서 및 연구 문헌은 교전 시점에 행사되는 직접 통제와 배치 이전 파라미터 설정을 통한 간접 통제를 구분함. 다만 이는 확립된 표준 분류가 아니라 논의의 편의를 위한 분석 틀에 해당함
- (대상 기준) ICRC와 SIPRI는 무기 파라미터에 대한 통제, 사용 환경에 대한 통제, 인간 감독의 형태라는 세 가지 조치를 제시함
- (거버넌스 기준) 신무기 법적 검토, 고위급 승인, 시험평가 절차 등은 개발·획득 단계에서 작동하는 제도적 수단으로, 위 두 기준과 층위를 달리함

| 표 1. 기준별 인간 통제 수단 |

구분 기준	유형	주요 수단 및 한계
통제 시점	직접 통제 (교전 시점)	운용 중 감시, 중단·비활성화 메커니즘. 고속 교전 환경에서 개입 가능 시간이 부족하고 자동화 편향이 작용할 수 있음
통제 시점	간접 통제 (배치 이전)	표적 유형·교전 범위·작동 시간의 사전 파라미터 설정. 설정 시점과 공격 시점의 간극이 클수록 비례성 판단이 곤란
통제 대상	무기 파라미터	체계가 반응할 수 있는 표적 유형 자체를 한정
통제 대상	사용 환경	민간인 밀집 지역 배제, 지리적 범위 및 지속 시간 한정. 환경 변화 등 동적 요인의 사전 반영이 어려움
통제 대상	인간 감독의 형태	운용 중 감시 및 개입·중단 능력의 유지
거버넌스 수단	절차적 통제	신무기 법적 검토, 고위급 승인, 시험평가·변경관리. 내부 절차 의존 시 외부 검증이 어려움

출처 ICRC·SIPRI(2020), CCW GGE 문서 및 각국 정책문서를 기초로 재구성

2-3. 통제 실효성의 기술적 조건

■ **인간 통제가 형식적 요건에 그치지 않으려면 운용자가 체계의 작동을 예측하고 이해할 수 있어야 한다는 기술적 전제가 요구됨[25)]**

- 운용자는 체계의 특성을 숙지하고 해당 체계가 배치 환경에 적합하다는 점을 확인할 수 있어야 하며,

24) ICRC, ICRC Position on Autonomous Weapon Systems, 12 May 2021; ICRC & SIPRI, Limits on Autonomy in Weapon Systems, Jun 2020; Human Rights Watch, Statement to CCW GGE Consultation on LAWS, 28 May 2025

25) UNIDIR, The Black Box, Unlocked: Predictability and Understandability in Military AI, 2020

의식적 판단과 법적 준수를 위한 충분하고 신뢰할 수 있는 정보를 보유해야 함

- 예측가능성(predictability)과 이해가능성(understandability)은 AI 체계의 불투명성으로 인해 제약되며, 이는 통제의 형식적 유지와 실질적 유지 사이의 괴리를 발생시킴

■ **시험·평가·검증·확인(TEVV) 체계가 AI 기반 역량의 도입 속도를 따라가지 못하고 있다는 문제가 제기됨[26)]**

- AI 체계는 형식적 방법론과 시뮬레이션만으로는 부분적 평가만 가능하며, 불투명성으로 인해 신뢰할 수 있는 평가가 추가로 제약됨
- 이로 인해 실제 전장이 사실상의 시험 환경으로 기능하고 있다는 비판이 제기되며, 이는 신무기 법적 검토 의무(제1추가의정서 제36조)의 이행 실효성과 직결됨

3 인간 통제의 대상과 구성요소

3-1. 통제 대상의 두 층위

■ **군사 AI의 인간 통제는 통제의 대상을 기준으로 무력사용·무기 운용에 대한 통제와 AI 체계 자체에 대한 통제의 두 층위로 구분하여 볼 수 있음**

- 전자는 어떤 표적에 대해 언제·어디서 무력을 사용할 것인지를 인간이 최종 판단·승인하고 필요 시 개입·중단할 수 있는지를 중심으로 함
- 후자는 AI의 개발·검증·배치·운용 전 과정에서 AI의 행동범위, 성능과 한계, 변경 가능성 등을 인간이 이해·설정·감독할 수 있는지를 중심으로 함
- 두 층위는 상호 배타적인 구분이 아니라 서로 연결되어 있으며, 실질적 인간 통제는 두 층위가 함께 작동할 때 확보될 수 있음
  - 무기 중심의 CCW 논의가 전자에, AI 의사결정지원체계 및 AI 보증 논의가 후자에 각각 대응하며, 최근 논의의 확장은 후자의 비중이 커지는 흐름으로 나타남

26) Human Rights Watch, Addressing Artificial Intelligence in the Military Domain, 14 Jun 2026

| 표 2. 통제 대상의 두 층위 |

구분	핵심 질문	주요 통제 수단
무력사용·무기 운용에 대한 통제	인간이 무력사용의 범위와 조건을 실질적으로 판단·결정할 수 있는가	표적·지역·시간·임무 범위의 제한, 인간의 승인, 감독·개입·중단, 책임 소재의 유지
AI 체계 자체에 대한 통제	인간이 AI의 능력·한계와 행동을 개발부터 운용 이후까지 이해·검증·관리할 수 있는가	데이터·모델 검증, 설명가능성·추적성, V&V 및 시험평가(T&E), 변경관리, 모니터링·사후평가

출처 ICRC·SIPRI(2020), CCW GGE 문서 및 각국 정책문서를 기초로 재구성

3-2. 인간 통제의 6대 구성요소

■ 위 두 층위에서 실제로 작동하는 통제 요소로, 본 보고서는 공개된 각국·국제기관 정책문서에서 반복적으로 나타나는 항목을 다음 여섯 가지로 분석·정리함[27)]

- 선별 기준은 ① CCW GGE 최종보고서 등 국가 간 합의 문서와 주요국 정책문서에 공통적으로 나타나고, ② 개발부터 사용·사후 책임까지 생애주기의 서로 다른 단계를 포괄하며, ③ 서로 중복되지 않는 통제 조치일 것의 세 가지임

- 이에 따라 여섯 요소는 GGE 최종보고서의 통제와 인간 판단(제34항), 운용 범위 제한 및 개입·비활성화 조치(제35항), 능력·한계의 이해와 시험평가(제38항), 책임성(제40~41항) 규정과도 대응함[28)]

■ (제1요소) 인간의 판단과 승인

- 무력 사용이나 중대한 결과를 초래할 수 있는 결정에 대해 인간의 판단이 실질적으로 유지되어야 한다는 원칙으로, 미국은 이를 '무력 사용에 대한 적절한 수준의 인간 판단'으로 제도화함

■ (제2요소) AI 행동범위의 사전 제한

- 표적의 유형, 작동 지역, 작동 시간, 임무 목적 및 교전조건 등을 인간이 사전에 설정함으로써 자율성이 행사될 수 있는 범위를 한정함

- ICRC가 제시하는 자율무기 통제 방안에서도 시간·공간·표적 제한이 핵심 요소로 다루어짐

27) U.S. DoD, DoD Directive 3000.09, Autonomy in Weapon Systems, 25 Jan 2023; NATO, Summary of the NATO Artificial Intelligence Strategy, 22 Oct 2021; ICRC, FAQ on Artificial Intelligence in the Military Domain, June 2026

28) CCW GGE on LAWS, Report of the 2024-2025-2026 Sessions of the Group of Governmental Experts on Emerging Technologies in the Area of LAWS, CCW/GGE.1/2026/3, paras. 34-41, adopted 4 Sep 2026

■ (제3요소) 감독·개입·중단 가능성

- 인간이 체계의 이상 행동을 식별하고 AI의 행동을 무효화하거나 임무를 중단·비활성화할 수 있는 실질적 능력을 보유해야 함
- NATO의 책임 있는 사용 원칙 중 통제가능성(Governability)은 의도하지 않은 결과를 탐지·회피하고 필요 시 체계를 작동 해제 또는 비활성화할 수 있어야 한다고 규정함

■ (제4요소) 설명가능성과 상황인식

- AI의 능력과 한계를 이해하고 AI의 추천을 독립적으로 검토할 수 있어야 하며, 설명가능성·추적가능성·데이터 품질·불확실성 정보가 인간 통제를 가능하게 하는 기술적 전제조건으로 기능함

■ (제5요소) 인간의 책임성

- AI가 판단을 보조하더라도 군사행동에 대한 책임을 알고리즘에 귀속시킬 수 없다는 입장으로, 미국·NATO·영국·중국 등이 표현의 차이에도 불구하고 공통적으로 강조함

■ (제6요소) 개발 및 시험평가 단계의 통제

- 전장에서의 최종 승인뿐 아니라 개발 단계의 데이터 관리, 검증·확인(V&V), 시험평가(T&E), 레드팀, 운용자 교육과 사후평가까지 인간 통제의 범위가 확대되는 추세

3-3. 주체별 인간 통제 용어

■ 동일한 문제의식에도 불구하고 주체별로 채택한 용어와 강조점이 상이하여, 개념의 국제적 단일화에는 이르지 못함[29)]

29) REAIM 2024, Blueprint for Action, 10 Sep 2024; Ministry of Foreign Affairs of the PRC, Position Paper of the PRC on Regulating Military Applications of Artificial Intelligence, 13 Dec 2021

| 표 3. 주요 주체별 인간 통제 관련 용어 |

주체	주요 표현	특징
미국 국방부(DoD)	Appropriate levels of human judgment over the use of force	무력사용에 대한 인간 판단을 핵심 기준으로 설정
영국 국방부(MoD)	Context-appropriate human involvement	임무·위험·운용환경에 적합한 수준의 인간 관여
NATO(국제기구)	Appropriate human-machine interaction / human responsibility	책임성·설명가능성·신뢰성·통제가능성과 연계
중국 외교부	Under human control / human suspension at any time	인간 통제와 언제든 중단할 수 있는 능력, 인간의 최종 책임 강조
ICRC(국제인도법 수호기관)	Sufficient human control and judgment	국제인도법 준수 및 인간의 도덕적·법적 판단 보장
REAIM(국가 간 정상회의)	Appropriate levels of human involvement / control	군사 AI의 책임 있는 개발·배치·사용이라는 포괄적 접근

출처 각 주체의 공개 정책문서를 기초로 재구성

■ '유의미한 인간 통제(MHC)'는 중요한 국제 논의 개념이나, 모든 국가가 동일한 정의를 채택한 확립된 단일 법률용어로 보기는 어려움

- 실제 정책 문서에서는 human judgment, human involvement, human-machine interaction, human control 등의 표현이 병존하고 있음
- 용어의 차이는 요구되는 통제의 강도와 판단 주체의 재량 범위에 관한 입장 차이와 연동되어 있음 – MHC는 통제의 질적 충분성을 요구하는 방향으로, appropriate 계열 용어는 맥락별 재량을 허용하는 방향으로 각각 기능함

■ 주요 주체 간에 실질적으로 수렴되는 공통 요소도 확인됨

- 국제법 준수, 인간의 책임 유지, AI의 능력과 한계에 대한 운용자의 이해, 검증·시험평가, 운용범위 제한, 비정상 행동에 대한 개입 가능성이 공통 요소로 도출됨
- 반면 구체적 제도화 방식, 법적 구속력의 수준, 외부 검증 수단의 유무에서는 국가 간 차이가 뚜렷함

4 개념 적용의 확장 쟁점

4-1. AI 의사결정지원체계로의 확장

■ **논의가 자율무기체계(LAWS) 중심에서 AI 의사결정지원체계(AI-DSS)를 포함하는 포괄적 군사 AI 논의로 확대되면서, 인간 통제 개념의 적용 범위도 함께 넓어짐[30)]**

- 의사결정지원체계는 인간이 최종 결정을 내린다는 점에서 형식적으로는 통제가 유지되나, 제시된 선택지와 분석 결과가 인간의 판단 범위를 사실상 규정한다는 점에서 통제의 실질이 문제됨

■ **의사결정지원체계로의 확장은 자동화 편향(automation bias)이라는 새로운 통제 쟁점을 수반함[31)]**

- 운용자가 체계 산출물을 과신하여 독자적 검증을 축소하는 경향을 의미하며, 의사결정 주기가 단축될수록 영향이 증대됨

4-2. 핵지휘통제 및 전략적 안정성

■ **핵무기 사용 결정에 대한 인간 통제 유지는 여러 다자 문서와 주요국 성명에서 폭넓게 지지되는 원칙이나, 구속력 있는 합의에는 이르지 못함[32)]**

- REAIM 2024「행동청사진」제5항은 핵무기 사용에 관한 주권적 결정을 알리고 집행하는 모든 행위에 인간의 통제와 관여를 유지하는 것이 특히 중요하다고 명시함
- 2024년 11월 미·중 정상회담에서 양국 정상은 핵무기 사용 결정에 대한 인간 통제 유지의 필요성을 확인함
- 미국 주도「책임 있는 군사 AI 및 자율성 이용에 관한 정치적 선언」은 핵무기나 핵 지휘통제에 관한 명시적 언급을 두지 않으며, IISS는 초안에 있던 관련 문언이 공표 전 삭제된 것으로 분석함

■ **AI가 조기경보·정보분석·지휘통제에 결합될 경우 의사결정 시간 압축으로 오판·확전 가능성이 증대될 수 있다는 분석이 제기되며, 인간 통제 논의가 전략적 안정성 영역으로 확장됨[33)]**

30) UN Secretary-General, Artificial Intelligence in the Military Domain and Its Implications for International Peace and Security, A/80/78, 5 Jun 2025
31) L. Bruun & M. Bo, Bias in Military Artificial Intelligence and Compliance with International Humanitarian Law, SIPRI, Aug 2025
32) REAIM 2024, Blueprint for Action, 10 Sep 2024; IISS, Military AI Governance under Strain, Jun 2026
33) V. Chernavskikh & J. Palayer, Impact of Military Artificial Intelligence on Nuclear Escalation Risk, SIPRI Insights on Peace and Security No. 2025/06, Jun 2025; J.-P. Rivera et al., Escalation Risks from Language Models in Military and Diplomatic Decision-Making, ACM FAccT 2024, Jun 2024

Ⅱ 주요국·국제기구의 군사 AI 정책 문서에서의 인간 통제

5 대한민국(Korea)

5-1. 국방 AI 추진 체계

■ **한국의 군사 AI 정책은 「국방혁신 4.0 기본계획」을 상위 틀로 하여 전개됨[34)]**

- 2023년 3월 발표된 기본계획은 AI 과학기술 강군 육성을 목표로 하며, 국방기획지침 및 합동군사전략서, 국방중기계획 작성의 기준으로 활용됨
- AI·유무인 복합체계를 중심으로 전력 구조를 전환하는 방향을 제시하며, 세부 적용 영역은 기술로드맵과 전담 조직을 통해 구체화됨

■ **국방기술진흥연구소(국기연)가 2022년 1월 발간한 「미래국방 2030 기술전략: 국방 AI 기술로드맵」은 무기체계별 핵심기술을 구체화한 국내 최초의 국방 AI 기술로드맵임[35)]**

- 감시정찰·지휘통제·유무인 복합·사이버 4개 분야 무기체계 30종에 대해 272개 핵심기술을 도출, 이 가운데 AI 판단 결과의 신뢰성을 확보하기 위한 '신뢰지능'과 지휘관의 상황판단을 지원하는 '전장환경 및 상황인지'를 중점 분야로 제시함

■ **2024년 4월 국방과학연구소(ADD) 내부 조직으로 국방AI센터가 창설되어 기술개발 및 정책지원 기능을 수행함[36)]**

- 총 110명 규모로 구성, AI 기반 유무인 복합체계 및 전장상황인식 등 핵심기술 개발과 군 AI 소요기획·기술기획 지원을 주요 기능으로 함

5-2. 안전성·신뢰성 확보 노력

34) 국방부, 「국방혁신 4.0 기본계획」, 2023. 3.
35) 국방기술진흥연구소, 「미래국방 2030 기술전략: 국방 AI 기술로드맵」, 2022. 1.
36) 국방부, 「국방AI센터 창설」, 2024. 4.

■ **2025년 7월 14일 국방기술품질원(기품원)은 국방 AI 표준화 워킹그룹을 출범시켜 AI 적용 무기체계의 안전성·신뢰성 확보를 위한 표준 개발에 착수함[37)]**

- 국방부, 방위사업청, 각 군, 국방과학연구소, 한국국방연구원 등 국방 관련 기관과 민간 AI 전문기관 및 방산업체에서 약 30명의 전문가가 참여함
- 객체 탐지·추적 분야에서 AI 시험·인증을 위한 데이터 품질 및 모델 성능 점검 표준 2종을 우선 개발하고, 이후 의사결정 지원, 자율형 시스템, 멀티모달리티 등 기술 수요가 높은 분야로 확대할 계획
- 해외 선진국의 AI 표준을 분석하여 국방 환경에 맞게 재설계함으로써 '한국형 국방 AI 표준체계'를 구축하는 방향을 제시함

■ **2026년 국방부는 전국 5개 AX(AI Transformation) 거점에 군·산·학 협력센터를 구축하여 민·군 AI 협력 기반을 조성하고 있음[38)]**

- 합참의 AI 기반 전략적 의사결정 지원모델과 공군의 표적 자동식별 기술 등이 과제에 포함되어, 인간 통제의 실질이 쟁점화되는 의사결정 지원 및 표적화 자동화 영역이 국내에서도 개발 단계에 진입함

5-3. 국제 규범 참여 및 제도화 과제

■ **한국은 REAIM 정상회의 공동주최를 통해 군사 AI 국제 논의를 주도함[39)]**

- 2024년 9월 9~10일 서울에서 제2차 REAIM 정상회의를 개최하였으며, 네덜란드·싱가포르·케냐·영국이 공동 주최국으로 참여함
- 96개국에서 1,952명이 참가하고 38명의 장관급 인사가 참석하였으며, 결과문서로 「행동청사진(Blueprint for Action)」이 채택되어 61개국이 지지함
- 청사진은 국제법 준수, 인간의 적절한 통제·관여, AI 체계의 신뢰성, 설명가능성을 주요 원칙으로 제시함
- 한국은 미국이 주도한 「책임 있는 군사 AI 및 자율성 이용에 관한 정치적 선언」에도 참여하고 있음

■ **국내 법제에서 국방·국가안보 목적 AI는 일반 AI 규율의 적용 대상에서 제외되어 있음[40)]**

37) 국방기술품질원, 「국방AI 표준화 워킹그룹 출범」 보도자료, 2025. 7. 14.
38) 국방부, 「제2차 국방데이터 정책포럼 – 5개 AX 거점 내 '군·산·학 협력센터' 구축 착수」 보도자료, 2026. 6. 18.
39) 대한민국 외교부, 「REAIM Summit 2024 결과」, 2024. 9.
40) 「인공지능 발전과 신뢰 기반 조성 등에 관한 기본법」(법률 제20676호, 2025. 1. 21.) 제4조 제2항; Regulation (EU) 2024/1689 (Artificial Intelligence Act), Article 2(3), 13 Jun 2024

- 「인공지능 발전과 신뢰 기반 조성 등에 관한 기본법」 제4조 제2항은 국방 또는 국가안보 목적으로만 개발·이용되는 인공지능에 대해 동 법을 적용하지 않는다고 규정함
- EU 인공지능법 제2조 제3항이 군사·국방·국가안보 목적 전용 AI를 적용 범위에서 배제하는 것과 같은 구조로, 양자 모두 해당 영역의 규율을 일반 AI 법제 외부에 두고 있음
- 2026년 9월 기준 공개 자료 범위에서는 미국 DoDD 3000.09나 영국 JSP 936에 상응하여 군사 AI의 인간 통제를 독립적으로 규정한 국방부의 공개 단일 지침은 확인되지 않으며, 관련 내용은 AI 전력화 로드맵, 안전성·신뢰성 표준화, 국제 책임규범 참여에 분산되어 있음

| 표 4. 한국 국방 AI 정책의 구성 요소 |

영역	주요 수단	인간 통제 관련 함의
전력화 전략	국방혁신 4.0 기본계획 국방AI 발전모델	원격통제형–반자율형–자율형 단계 구분을 통해 자율성 수준별 통제 설계 가능
기술 기획	국방AI 기술로드맵(30종 무기체계, 272개 기술)	표적식별·군집 지휘통제 등 통제 쟁점이 집중되는 기술군 포함
전담 조직	국방AI센터(ADD, 2024)	유무인 복합체계·전장상황인식 기술개발 및 소요기획 지원
안전성·신뢰성	국방AI 표준화 워킹그룹(2025)	AI 적용 무기체계의 품질·신뢰성 표준 개발 착수
국제 규범	REAIM 공동주최(2024) 정치적 선언 참여	적절한 인간 통제·관여, 신뢰성·설명가능성 원칙 수용

출처 국방부·방위사업청·외교부 공개 자료를 기초로 재구성

6 미국(US)

6-1. 자율무기체계 훈령과 인간 판단 요건

■ 미국의 핵심 문서는 「국방부 훈령 3000.09(DoDD 3000.09, Autonomy in Weapon Systems)」임[41)]

- 2012년 11월 제정 후 2023년 1월 25일 개정되었으며, 자율·반자율 무기체계는 지휘관과 운용자가 무력 사용에 대해 '적절한 수준의 인간 판단'을 행사할 수 있도록 설계되어야 한다고 규정함
- 현실적 운용조건 내 검증·확인(V&V) 및 시험평가(T&E), 운용자의 적절한 주의, 체계 실패 가능성에

41) U.S. DoD, DoD Directive 3000.09, Autonomy in Weapon Systems, 25 Jan 2023

대한 관리를 함께 요구함

- '적절한(appropriate)'은 의도적으로 유연하게 설정된 개념으로, 모든 맥락에 일률 적용되는 고정된 인간 판단 수준은 존재하지 않으며 무기체계·전쟁 영역·작전 맥락·체계 내 기능별로 달라질 수 있음을 전제함[42)]

- 동 훈령은 높은 수준의 자율성을 일률적으로 금지하지 않으며, 완전자율 살상무기의 금지나 모든 상황에서의 실시간 인간 개입을 규정하지도 않음. 대신 적용 대상 체계는 동 훈령이 정한 설계·시험·검토·승인 절차를 거치도록 함

■ **인간 통제의 이행은 고위급 검토(senior review)와 인간-기계 인터페이스 요건을 통해 절차적으로 담보됨[43)]**

- 면제 대상에 해당하지 않는 자율무기체계는 정식 개발 착수 전과 배치 전의 2단계 고위급 검토를 거쳐야 하며, 긴급한 군사적 필요가 있는 경우 국방부 부장관이 이를 면제할 수 있음

- 기존 승인 체계의 변형이라도 알고리즘, 임무 범위, 운용 환경, 표적군, 예상 적 대응수단이 기존 승인 범위를 벗어날 정도로 달라지는 경우 새로운 고위급 검토가 요구됨

- 인간-기계 인터페이스는 훈련된 운용자가 용이하게 이해할 수 있도록 설계되어야 하며, 기능의 활성화·비활성화 절차가 명확하고 체계 상태에 관한 투명한 피드백을 제공해야 함
  - 자율무기체계 작업반(AWS Working Group)이 고위급 검토 절차를 지원·자문하며, 정책·연구공학·획득·법무·CDAO·시험평가·합참 대표로 구성됨

■ **사이버 역량, 비무장 플랫폼, 무유도탄, 운용자가 수동 유도하는 탄약, 지뢰, 불발탄 및 무기체계가 아닌 자율체계는 적용 대상에서 제외됨[44)]**

6-2. 책임 있는 AI 체계 및 최근 정책 변화

■ **훈령 3000.09는 2020년 채택된 국방부 AI 윤리원칙 및 「책임 있는 AI 전략·이행 경로」와 연결되는 구조로 운영됨[45)]**

- AI 윤리원칙은 책임성(Responsible), 형평성(Equitable), 추적가능성(Traceable), 신뢰성(Reliable), 통제가능성(Governable)의 5개 원칙으로 구성됨

42) U.S. Government, Human-Machine Interaction in the Development, Deployment and Use of Emerging Technologies in the Area of LAWS, CCW/GGE.2/2018/WP.4, 28 Aug 2018
43) U.S. DoD, DoD Directive 3000.09, Autonomy in Weapon Systems, Sections 1.2, 4 and Glossary, 25 Jan 2023
44) K. M. Sayler, Defense Primer: U.S. Policy on Lethal Autonomous Weapon Systems, CRS IF11150, updated 26 Mar 2026
45) U.S. DoD, AI Ethical Principles, 24 Feb 2020; Responsible AI Strategy and Implementation Pathway, Jun 2022

- 통제가능성 원칙은 AI의 의도하지 않은 결과를 탐지·방지하고 필요한 경우 체계를 중단하거나 비활성화하는 능력을 강조함

■ 2026년 6월 5일 「국가안보 영역의 인공지능에 관한 대통령 각서(NSPM-11)」가 서명되어, 도입 가속을 전면화하면서 보증·책임성을 병렬적으로 요구하는 방향으로 국가안보 AI 정책이 재편됨[46)]

- 바이든 행정부의 국가안보각서(NSM-25)를 폐지·대체하였으며, 정책을 채택(Adoption)·적응(Adaptation)·보증(Assurance)·책임성(Accountability)의 4개 축으로 구성함
  - (채택) 임무 영역 식별과 도입 장벽 제거, 산업계와 협력 통한 프런티어 모델 신속한 제공
  - (적응) 복수 공급자의 상용·오픈소스 AI를 국가안보 전반에서 공유·활용
  - (보증) 임무 핵심 AI 체계에 대한 통제 유지, 정부의 승인 없는 실질적 변경 방지
  - (책임성) 지휘 계통과 작전 권한을 존중하는 도입 및 지휘관·기관장의 책임 유지

- 신뢰성(reliability), 강건성(robustness), 행동 유도 가능성(steerability), 통제가능성(controllability)을 가능하게 하는 기술의 연구개발을 우선하도록 하고, 표준화된 시험·평가·검증·확인(TEVV) 방법론을 120일 내 제출하도록 지시함

- 계약 측면에서는 각서 제2절의 정책에 배치되는 행위를 반복적으로 보인(repeated pattern of conduct) 기업의 계약을 법이 허용하는 범위에서 해지하도록 하되, 임무상 필요에 따라 최대 1년의 유예를 허용함
  - 이는 특정 기업의 행태를 대상으로 한 조치로서, 공급기업의 사용정책 자체를 일반적으로 금지한 것은 아님

- 자율무기체계에 관한 훈령 3000.09를 90일 내에 갱신하되 지휘 계통을 보존하도록 지시하였으나, 2026년 9월 20일 현재 공개된 국방부 지침 목록에서는 2023년 판이 여전히 최신 공개본으로 표시되고 있음
  - 도입 가속과 보증·책임성 요건이 동시에 강화되는 구조이므로, 단일 방향의 규제 완화 또는 강화로 평가하기는 어려움

■ 2026년 1월 공표된 「국방부 인공지능 전략」은 속도를 우선하는 'AI-first' 기조를 공식화함

46) The White House, National Security Presidential Memorandum/NSPM-11: Artificial Intelligence in the National Security Enterprise, 5 Jun 2026

(부처 발표 문서이며 법령은 아님)[47]

- 7개 선도 과제(Pace-Setting Projects) 지정, 최신 상용 모델의 30일 내 배치, 비법정 요건을 면제할 수 있는 장벽제거위원회(Barrier Removal Board), 데이터 접근 의무화 등을 규정함
- CDAO를 '전시 CDAO'로 규정하여 운영승인(ATO) 지연 및 데이터 공유 장애를 우선 해소 대상으로 설정하고, 부처 내 기존 AI 정책 지침의 정합성 확보를 지시함
- FY2026 국방수권법(NDAA) 제1513조는 AI 체계에 대한 위험 기반 보안 프레임워크를, 제1533조는 AI 모델의 개발·조달 단계 윤리원칙 준수 프레임워크 마련을 요구함
  - 동 전략이 속도와 규제 완화를 우선하는 방향으로 해석될 경우, 훈령 3000.09의 절차 통제와 긴장 관계를 형성할 수 있다는 분석이 제기됨

■ **국제 규범 측면에서 미국은 자국 주도의 비구속적 선언을 선호하며, 유엔 차원의 신규 규제 절차 창설에는 반대 입장을 표명함[48]**

- 2023년 2월 REAIM 헤이그 정상회의에서 「책임 있는 군사 AI 및 자율성 이용에 관한 정치적 선언」을 출범시켰으며, 2025년 3월 기준 58개국이 지지함
- 유엔 결의 79/239에 대한 제출 문서에서 유엔 주도의 새로운 규제 절차 창설은 부적절하며 국제 평화·안보 강화 노력을 저해할 수 있다는 입장을 밝힘
- CCW GGE에서는 2018년부터 '적절한 수준의 인간 판단' 표현을 지지해 왔으며, REAIM에서는 2023년 「행동촉구」와 2024년 「행동청사진」을 모두 지지하였으나 2026년 「Pathways to Action」에는 참여하지 않음[49]

7 영국(UK)

7-1. 맥락 적합적 인간 관여

■ **영국 국방부의 2022년 「야심적·안전·책임 있는(Ambitious, Safe, Responsible)」 정책은 군사 AI에 대해 '맥락 적합적 인간 관여(context-appropriate human involvement)' 개념을 사용함[50]**

47) U.S. Department of War, Artificial Intelligence Strategy for the Department of War, 9 Jan 2026
48) U.S. Department of State, Political Declaration on Responsible Military Use of Artificial Intelligence and Autonomy, Nov 2023
49) RSIS, The Uncertain Future of the REAIM Summit, IDSS Paper IP26029, 26 Feb 2026

- AI가 표적을 식별·선택·공격하는 과정에서 요구되는 인간 개입의 형태는 임무 및 상황에 따라 달라질 수 있으며, 실시간 감독뿐 아니라 체계의 임무와 파라미터를 사전에 제한하는 방식도 인간 통제의 일부로 봄
- 작전 성과가 판단과 행동의 속도·정확성에 좌우되더라도, 문제가 되는 수준의 재량과 판단을 AI에 위임하지 않고도 높은 수준의 역량 구현이 가능하다는 논리를 전제함
- 인간은 군사 행위에 대해 계속 책임을 부담하며, 인간의 적절한 관여 없이 완전히 자율적으로 표적을 식별·선택·공격하는 무기체계를 개발할 의도가 없다는 입장을 밝힘
- 국방 AI 전략의 윤리원칙은 인간중심성(human-centricity), 책임(responsibility), 이해(understanding), 편향 및 피해 완화(bias and harm mitigation), 신뢰성(reliability)의 5개로 구성됨

7-2. JSP 936 및 보증 체계

■ 2024년 11월 발간된 「JSP 936 신뢰가능한 국방 인공지능(Dependable AI in Defence)」은 윤리원칙을 실무적 AI 보증(assurance) 체계로 발전시킴[51)]

- 단순 지침이 아닌 '지시(Directive)'로 규정되어 국방부 전 직원이 AI 생애주기 전 단계에서 적용해야 하며, 로봇·자율체계(RAS) 및 디지털 체계를 포함한 국방 전반의 AI 응용을 대상으로 함
- AI 생애주기 전반에 걸쳐 책임성·안전성·보안·설명가능성 및 적절한 인간 감독을 확보하는 것이 핵심 목표로 설정됨
- 책임 원칙은 AI 산출물에 대한 명확한 인간 책임 소재를 확립하고 유의미한 인간 통제를 유지할 것을, 인간중심성 원칙은 AI가 인간에게 미치는 영향을 평가하여 긍정적 효과가 부정적 효과를 상회하도록 할 것을 요구함
- Part 1(지시)은 2024년 11월 13일 공표되어 적용 중이며, 이행 지원을 위한 Part 2 지침은 「AI 실무자 핸드북(AI Practitioner's Handbook, AIPH)」이라는 온라인 형태로 제공됨
  - AIPH는 2026년 9월 기준 베타(BETA) 서비스로 운영되며 개별 지침이 순차 공개되고 있고, 일부 자료는 국방부 내부망 이용자에게만 제공됨

■ 2025년 도입된 책임 있는 AI 선임책임자(Responsible AI Senior Officers) 제도는 JSP

50) UK Ministry of Defence, Ambitious, Safe, Responsible: Our Approach to the Delivery of AI-enabled Capability in Defence, Jun 2022; Defence Artificial Intelligence Strategy, Jun 2022

51) UK Ministry of Defence, JSP 936 V1.1 Dependable AI in Defence, Part 1: Directive, Nov 2024

936을 조직 차원에서 구현하기 위한 조치임[52)]

- 해당 책임자는 윤리적 감독을 유지하고 체계 위험을 추적하며 국방 AI·자율성 부서와 연계하는 역할을 수행함

7-3. 2026년 도입 가속 정책

■ **2026년 6월 10일 국방부는 「국방의 중심에 AI를 두기(Putting AI at the Heart of UK Defence)」를 발표하고, 신속 AI 도입 태스크포스(TF RAID)를 신설함[53)]**

- 2025년 국방전략검토(SDR)가 제시한 자율성·AI 전환 방향을 즉시 이행 지시로 전환한 문서로 평가되며, 국방참모총장 직속 조직으로 설치되어 국방투자계획을 통해 1억 파운드가 배정됨
- 정책 문서는 AI 보안연구소(AISI)와 협력하여 공급자·과업에 관계없이 모델을 신속히 평가할 수 있는 AI 보증 역량 개발을 모색하도록 함

■ **시민사회 일각에서는 보증 체계가 자기보고 및 내부 검토에 의존하여 외부 검증 수단이 미비하다는 비판을 제기함[54)]**

- AOAV는 의회 감독·외부 감사·시민사회 참여에 관한 언급이 포함되어 있지 않다는 점과, 규정된 절차에 따라 개발될 경우 자율 표적 체계도 허용될 수 있는 구조라는 점을 지적함
  - JSP 936 자체는 국방부의 기본 정책 프레임워크로서 생애주기 보증과 적절한 인간 감독을 명시하고 있으므로, 위 평가는 해당 단체의 비판적 해석에 해당함
- 정책 요건을 기술적·운용적 요건으로 전환하는 과정에서 증거 충분성, 인간-AI 상호작용, 운용환경 정의, 체계 간 통합, 성능 관리, 안전·보안 분석, 윤리성 측정, AI의 내재적 복잡성 완화 등 여덟 개 영역의 이행 과제가 제기됨

8 유럽연합(EU) 및 NATO

8-1. 유럽연합의 정책

■ **EU 인공지능법(AI Act)은 군사·국방·국가안보 목적에 전용되는 AI 체계를 적용 범위에서**

52) UK Ministry of Defence, Laying the Groundwork: Responsible AI Senior Officers' Report 2025, 3 Oct 2025
53) UK Ministry of Defence, Putting Artificial Intelligence (AI) at the Heart of UK Defence, 10 Jun 2026; UK Ministry of Defence, Rapid AI Delivery Taskforce (TF RAID), GOV.UK
54) AOAV, War by Algorithm: A Review and Critique of JSP 936, Jul 2025

배제하고 있어, 군사 AI에 대한 직접적 규율 수단으로 기능하지 않음[55)]

- 제2조 제3항은 군사·국방·국가안보 목적으로 전용(exclusively)하여 시장 출시·서비스 개시·사용되는 AI 체계에 대해 수행 주체의 유형과 무관하게 동 규정이 적용되지 않는다고 규정함
- 전문 제24항은 배제의 근거로 EU 조약 제4조 제2항에 따라 국가안보가 회원국의 전속 책임이라는 점과 국가안보 활동의 특수성을 제시함
- 다만 '전용' 요건이 판단 기준이므로, 민간 또는 법 집행 목적으로 출시된 체계가 군사 목적으로 전용되는 경우 등 이중용도 영역에서는 적용 경계가 불명확하다는 문제가 제기됨

■ **인간 통제 요건은 인공지능법 대신 방위 재정지원 수단과 기관 지침을 통해 부분적으로 규율됨[56)]**

- 유럽방위기금(EDF) 설립 규정(Regulation (EU) 2021/697)은 '인간에 대한 공격 수행 시 표적 선정 및 교전 결정에 유의미한 인간 통제의 가능성이 없는 자율살상무기'의 개발 활동을 자금 지원 대상에서 배제함
  - '유의미한 인간 통제'를 자금 지원 적격성 판단 기준으로 채택한 최초의 EU 법적 문서로 평가되나, 인간을 대상으로 한 공격에 한정된 문언이어서 적용 범위가 좁게 해석될 수 있다는 지적도 제기됨

- 단기 공동조달 수단인 EDIRPA도 유사한 제한을 두었으나 2025년 말 만료되었으며, 후속 수단인 SAFE는 AI 등 첨단 역량의 공동조달을 명시적으로 지원하면서도 유의미한 인간 통제 배제 조항을 반복하지 않음
  - 벨기에는 이사회 채택 시 SAFE가 오타와 협약이나 CCW 지침원칙 등 군축·비확산 문서에서 금지되는 방산품의 공동조달을 보다 명시적으로 배제하지 않은 점에 대해 유감을 표명하였으며, 이는 인간 통제 조항만을 특정한 것은 아님

- 유럽방위청(EDA)은 2025년 5월 「국방 분야 AI의 신뢰성(Trustworthiness for AI in the Defence Sector)」 백서를 발간하여 인공지능법과의 접점 및 방위 분야 고유의 제공자 역할 분류를 제시함
  - 회원국이 국방 주권을 보유하고 EDA는 조정 기능만을 수행하므로 EU 차원의 단일 횡단적 군사 AI 감독기관은 존재하지 않으나, 회원국의 신무기 법적 검토와 국내 감독 구조는 별도로 적용됨

8-2. NATO의 책임 있는 사용 원칙

55) Regulation (EU) 2024/1689 (Artificial Intelligence Act), Article 2(3) and Recital 24, 13 Jun 2024

56) European Defence Agency, Trustworthiness for AI in the Defence Sector, May 2025; S. Clapp, Defence and Artificial Intelligence, EPRS Briefing, European Parliament, Apr 2025

■ **NATO는 2021년 AI 전략에서 군사 AI의 책임 있는 사용을 위한 6개 원칙(PRU)을 채택함[57)]**

- 합법성(Lawfulness), 책임성 및 설명책임(Responsibility and Accountability), 설명가능성 및 추적가능성(Explainability and Traceability), 신뢰성(Reliability), 통제가능성(Governability), 편향 완화(Bias Mitigation)로 구성됨
- 책임성 및 설명책임 원칙은 인간의 책임을, 통제가능성 원칙은 적절한 인간-기계 상호작용과 AI 체계에 대한 개입·중단 능력을 각각 강조함
- 합법성 원칙은 국제인도법 및 국제인권법 준수를 명시적으로 포함하며, 동 원칙은 기존의 국가적·국제적 의무를 대체하지 않고 AI 영역에 구체화한 것으로 위치 지어짐

■ **2024년 7월 10일 개정 AI 전략은 생성형 AI 등 기술 변화를 반영하고 책임 있는 사용 원칙의 실제 구현을 우선순위로 설정함[58)]**

- 주요 우선순위는 PRU 이행 진전, 동맹 전반의 AI 체계 상호운용성 제고, 여타 신흥·파괴적 기술과의 결합, 산업계·학계 및 방위혁신 가속기(DIANA)·NATO 혁신기금을 통한 생태계 확장으로 제시됨
- 목표로는 책임 있는 AI 개발·사용의 모범 제시, 역량 개발 단계의 AI 도입 가속 및 상호운용성 제고, AI 기술 보호와 위험 관리, 적대적 AI 활용에 대한 대응이 설정됨
- 데이터·AI 검토위원회(DARB)가 원칙을 동맹 차원에서 운용하기 위한 제도적 장치로 기능하며, NATO 과학기술기구(STO)에서 PRU 충족 여부를 체계적으로 평가하기 위한 프레임워크 연구가 진행 중
  - 평가 프레임워크 연구는 신뢰(trust)와 불확실성 개념을 매개로 AI 체계가 각 원칙을 어느 정도, 어떤 조건에서 충족하는지를 판정하는 방식을 제안함

■ **일부 분석은 EU와 NATO의 역할 분담 구조상 인간 통제의 이행이 자금 지원 단계 이후의 조달·배치 국면에서 상호 검증되기 어렵다고 지적함[59)]**

- EU 자금 지원 프로그램을 벗어난 체계가 조달·소프트웨어 갱신·동맹 네트워크 편입 단계를 거치는 과정에서, 참여국 간에 인간 통제 요건의 이행을 확인할 공통 절차가 부족함이 주된 논거임
- 해당 분석은 공동작전에 AI 기반 역량을 투입하기 전 참여국이 상호 제시해야 할 최소 증빙 요건을

57) NATO, Summary of the NATO Artificial Intelligence Strategy, 22 Oct 2021
58) NATO, Summary of NATO's Revised Artificial Intelligence (AI) Strategy, 10 Jul 2024
59) Geopolitical Monitor, Europe's Military AI Gap Begins at Fielding, Aug 2026

정하는 '공통 보증 프로파일(common assurance profile)' 도입을 제안하나, 이는 현재 채택된 정책이 아니라 논평 성격의 제안임

| 표 5. 주요국·국제기구의 인간 통제 관련 규율 수단 비교 |

미국	DoDD 3000.09(2023) AI 윤리원칙(2020)	국방부 훈령(내부 구속)	무력사용에 대한 '적절한 수준의 인간 판단' 행사 가능하도록 설계, 통제가능성 원칙
영국	JSP 936 V1.1(2024) ASR(2022)	국방부 지시(내부 구속)	맥락 적합적 인간 관여, 인간 책임 소재 확립 및 유의미한 인간 통제 유지
EU	AI Act 제2조(3) EDF 규정	법률(적용 배제) + 자금지원 조건	군사 전용 체계 적용 배제, 유의미한 인간 통제 결여 LAWS 자금지원 금지
NATO	AI 전략 및 PRU (2021/2024)	동맹 정책(비구속)	통제가능성·책임성·설명가능성 등 6개 책임 있는 사용 원칙
독일	KI-Konzeption (2026)	국방부 구속 지침(내부 구속)	AI 체계의 승인·시험·운용에 관한 구속적 규칙, AI 이용 시 인간의 영향력 설정
일본	ATLA 책임 있는 AI 지침(2025)	연구개발 지침(R&D 단계)	국제법 준수, 인간 관여 확보, 위험 식별 및 생애주기 관리
중국	군사 AI 규제 입장문서 글로벌 AI 거버넌스 이니셔티브	정책 선언(비구속)	인간 통제 유지, 언제든 작동 중단 가능, 인간의 최종 책임

출처 각국·기구 공개 정책문서를 기초로 재구성

9 프랑스(France) 및 독일(Germany)

9-1. 프랑스

■ 프랑스는 AI의 군사적 활용 확대와 인간의 지휘책임 유지라는 두 방향을 병행하고 있음[60)]

- 2024년 5월 1일 국방 AI 전담기관인 AMIAD(Agence ministérielle pour l'intelligence artificielle de défense)가 창설되었으며, 2025년 국방부 AI 전략이 발표됨
- 무인 항공·지상·해상 플랫폼을 조종할 수 있는 AI 개발이 중점 방향으로 제시되어, 자율성 확대와 지휘책임 유지가 동일 기관 안에서 병행 추진되는 구조임

■ 프랑스 국방윤리위원회는 2025년 군사 AI에 관한 의견서를 발표하여 인간 책임과 통제의 지속을 강조함[61)]

60) Ministère des Armées, Lancement de la stratégie ministérielle pour l'intelligence artificielle, 2025

61) Comité d'éthique de la défense, Avis sur l'intelligence artificielle, 2025; France Diplomatie, Systèmes d'armes létales autonomes,

- AI는 인간의 지휘를 대체하기보다 지원해야 한다는 것이 기본 접근으로 제시됨
- 2026년 갱신된 외교부의 LAWS 정책에서도 프랑스가 정의하는 완전자율치명무기체계(SALA), 즉 인간의 개입 없이 치명적 무력을 사용하는 체계를 개발하지 않는다는 입장을 유지함

9-2. 독일

■ 독일은 2026년 국방부(BMVg) 업무영역 전반에 적용되는 「KI-Konzeption(AI 구상)」을 통해 군사 AI에 관한 구속적 내부 지침을 마련함[62)]

- 2026년 4월 9일 규정번호 K-9000/102로 발효되었으며 당초 비밀문서로 분류되었다가, 2026년 8월 공개본이 발간됨
- AI 체계의 승인(Zulassung)·시험(Erprobung)·운용(Einsatz)에 관한 구속적 규칙을 정하고, 국내외 법 요건의 준수와 책임 소재를 규정함
- AI 이용 시 인간의 영향력을 어떻게 설정할 것인지가 핵심 내용으로 제시되며, 무기체계와 행정 효율화 영역 모두에 적용됨
- NATO 동맹국 및 파트너군과의 호환성을 명시하여 다국적 작전에서의 상호운용성을 전제로 설계됨
  - 연방군은 국제법에 기초한 윤리 원칙과 NATO의 법적 요건을 AI 체계의 전 생애주기에 통합하는 접근을 취한다고 설명함

■ 유럽 주요국에서 인간 통제의 구체적 수준과 절차는 주로 정책·내부 지침으로 규정됨

- EU 차원에서는 회원국이 국방 주권을 보유하고 EDA가 감독 권한을 갖지 않으며, 인공지능법의 국가별 감독기관도 군사 전용 영역에는 관할이 미치지 않음

10 일본(Japan)

10-1. AI 활용 기본방침

■ 일본 방위성은 2024년「AI 활용 추진에 관한 기본방침(Basic Policy on Promoting the

2026

62) BMVg, KI-Konzeption für den Geschäftsbereich des BMVg (K-9000/102), 9 Apr 2026; Bundeswehr, Rechtliche Grundlagen für Künstliche Intelligenz, 2026

Utilization of AI)」을 발표함[63]

- AI의 국방 활용 분야, 위험관리, 데이터, 연구개발, 인재육성 등에 대한 기본 방향을 제시함
- 단일 포괄 법제보다 기존 분야별 법령과 정책 지침을 통해 대응하는 접근을 취하고 있음

10-2. 연구개발 단계의 책임 있는 AI 지침

■ 방위장비청(ATLA)은 2025년 6월 6일「장비품 등의 연구개발에서의 책임 있는 AI 적용 지침(제1판)」을 발표함[64]

- 2024년 기본방침에서 별도 지침 수립을 명시한 데 따른 후속 조치로, 방위성·자위대 고유의 지침으로 위치 지어짐
- 원칙적으로 AI 기술이 적용된 장비품(AI 장비품)의 시제 사업을 대상으로 하며, 위험 관리를 3단계로 수행하도록 규정함

  ① AI 장비품의 분류
  ② 저위험으로 분류되지 않은 경우 법적·정책적 심사
  ③ 적격 판정 시 기술적 심사

- 위 순서로 진행되며, 파괴 능력에 영향을 주는 경우 법적·정책적 심사로 이행됨
- 일본이 CCW에 제출한 LAWS 작업 문서에서 표명한 인간 중심 원칙, 신뢰성·예견가능성 확보, 국제법상 허용되지 않는 무기체계의 연구개발·운용 배제 입장을 연구개발 맥락에서 구체화한 것으로 설명됨
- 적용 대상은 연구개발 단계이며, 작전 운용 절차 전반을 규율하는 지침은 아님

■ (참고) 일본 정부 전체의 AI 규제 기조는 기업의 자율적 조치를 중시하는 라이트 터치(light touch) 접근이나, 이는 군사 AI 거버넌스와는 별개의 층위에 해당함[65]

- 내각관방의 AI 정책 논의는 기술 중립성 원칙에 따라 기존 분야별 법률에 기반한 규제 방식을 선호하며, 전략적 리더십 기구를 중심으로 위험 모니터링 및 정보 수집 체계를 구축하는 방향을 제시함

63) 防衛省,「防衛省AI活用推進基本方針」, 2024. 7. 2.
64) 防衛装備庁,「装備品等の研究開発における責任あるAI適用ガイドライン(第1版)」, 2025. 6. 6.
65) Government of Japan, Interim Report of the AI Policy Study Group, Cabinet Secretariat, Feb 2025

- 방위성·ATLA의 군사 AI 연구개발 거버넌스는 위 일반 규제 기조와 구분되는 별도의 지침 체계로 운영되므로, 양자를 직접 연결하여 평가하기는 어려움

11 중국(China)

11-1. 정책 기조

■ 중국의 공개 정책에서 가장 직접적인 자료는 외교부가 발표한 「군사 분야 인공지능 규제에 관한 중화인민공화국 입장문서」임[66)]

- AI의 군사적 활용이 국제인도법을 포함한 국제법을 준수해야 하고, 무기체계가 인간의 통제하에 있어야 하며, 인간이 언제든 체계의 작동을 중단할 수 있어야 한다는 입장을 제시함
- 무기체계의 전체 생애주기에서 필요한 인간-기계 상호작용을 보장하고 인간이 책임의 최종 주체가 되어야 한다는 점을 명시함

■ 2023년 「글로벌 AI 거버넌스 이니셔티브」 및 2025년 유엔 제출 문서를 통해 기조가 재확인됨[67)]

- 2025년 4월 유엔 총회 결의 79/239에 따라 제출한 문서에서 '군사 AI에 대한 인간 중심 접근' 견지를 표명하고, 민첩한 거버넌스, 전향적 위험평가, 확산 위험 감소, 다자주의 견지 등을 제안함
- 2025년 11월 군축·비확산 백서에서도 다자주의·개방성·포용성을 옹호하고, 광범위한 합의에 기초한 보편적 국제 메커니즘 수립을 지지한다는 입장을 제시함

11-2. 협상 태도 및 평가

■ 중국은 CCW를 가장 적절한 논의 포럼으로 지지하면서도, 법적 구속력 있는 문서의 협상 개시는 '조건이 성숙할 때'를 전제로 함[68)]

- 협상 개시에 앞서 관련 작업 정의(working characterisations)와 규율 범위에 대한 합의가 선행되어야 한다는 입장을 2025년 GGE 회기까지 일관되게 유지함

66) Ministry of Foreign Affairs of the PRC, Position Paper of the PRC on Regulating Military Applications of Artificial Intelligence, 13 Dec 2021
67) Ministry of Foreign Affairs of the PRC, Global AI Governance Initiative, Oct 2023; PRC, Submission to the UN Secretary-General pursuant to Resolution 79/239, Apr 2025
68) Lieber Institute West Point, Human Oversight with Chinese Characteristics, May 2026

- 이러한 누적적 요건 설정은 현재 개발 중인 다수의 자율 역량을 규율 대상에서 사실상 제외하는 결과를 낳는다는 분석이 제시됨
- 다만 2026년 9월 GGE 최종보고서는 컨센서스로 채택되어 중국도 채택에 참여함

■ 중국은 REAIM 프로세스에 참여해 왔으나 2024년 이후의 결과문서는 지지하지 않음[69)]

- 2023년 헤이그 「행동촉구」에는 미국과 함께 지지국으로 참여함[70)]
- 2024년 서울 정상회의에 정부 대표단을 파견하고 장관급 대화에 참석하였으나 「행동청사진」은 지지하지 않았으며, 핵 관련 인간 통제 문언이 주요 배경이라는 분석이 제기됨
- 2025년 11월 백서는 핵무기를 특정하지 않고 '관련 무기체계(all relevant weapons systems)'가 인간의 통제하에 있어야 한다는 일반적인 표현을 사용함
- 2026년 2월 아코루냐 제3차 정상회의의 결과문서에도 중국은 서명하지 않았으며, 미국·이스라엘 등 주요 군사국도 지지국에 포함되지 않음

69) K. V. Hiebert, Prospects Have Dimmed for Autonomous Weapons Guardrails, CIGI, 30 Apr 2026
70) RSIS, The Uncertain Future of the REAIM Summit, IDSS Paper IP26029, 26 Feb 2026

III 인간 통제 구현 수단별 연구동향

12 인간 판단과 의사결정의 실효성

12-1. 의사결정지원체계와 인간 판단 [무력 사용]

■ **국제적십자위원회(ICRC)는 인간 통제의 적용 대상을 무기체계를 넘어 AI 의사결정지원체계(AI-DSS)로 확대하여 다루고 있음[71)]**

- AI-DSS가 표적 선정이나 작전계획을 지원할 경우 인간이 형식적으로 최종 결정을 내리더라도 데이터 오류, 자동화 편향, 설명가능성 부족으로 인해 인간 판단 자체가 AI에 과의존할 수 있다는 문제가 제기됨
- 이에 따라 AI 출력에 이의를 제기할 수 있는 운용자의 능력, 데이터 품질, 훈련, 시험 및 사후검토를 인간 통제의 구성요소로 다룸
- 제네바 아카데미와 공동으로 무력 사용에 관한 군사적 의사결정에서의 AI 활용을 분석하였으며, 2025년 4월 유엔 사무총장에게 결의 79/239에 따른 권고를 제출함

■ **SIPRI는 자율무기체계와 AI 기반 표적 의사결정지원체계를 비교하여 양자의 기능적 차이와 공통된 거버넌스 문제를 분석함[72)]**

- '인간 개입(human in the loop)'이라는 정태적 구분 대신 맥락 의존적 인간 감독 개념을 제시하며, 2단계 규제 접근에 대한 지지 확대를 관측함
- 설계·데이터 단계의 편향이 구별 원칙 적용과 취약 집단 보호에 미치는 영향을 분석하여, 편향이 국제인도법 준수 판단을 왜곡할 수 있다는 점을 제시함

■ **자동화 편향은 인간 판단의 실효성을 저해하는 핵심 위험 요인으로 지목되며, 이를 완화하기 위한 인간 역량 강화 연구가 진행 중[73)]**

71) ICRC, FAQ on Artificial Intelligence in the Military Domain, June 2026; ICRC & Geneva Academy, Expert Consultation Report: Artificial Intelligence and Related Technologies in Military Decision-Making on the Use of Force in Armed Conflicts, May 2024

72) A. Blanchard & L. Bruun, Autonomous Weapon Systems and AI-enabled Decision Support Systems in Military Targeting: A Comparison and Recommended Policy Responses, SIPRI, Jun 2025

73) Y.-K. Heng, Upskilling Human Actors against AI Automation Bias in Strategic Decision Making on the Resort to Force, Cambridge Forum on AI: Law and Governance, Jan 2026

- 무력 사용 결정 국면에서 인간 행위자의 자동화 편향을 완화하기 위한 교육·훈련 방안과 AI 지원 의사결정에서의 중개자 역할에 관한 연구가 제시됨

12-2. 대규모 언어모델 활용과 확전 위험 [무력 사용]

■ 대규모 언어모델의 군사적 의사결정 활용에 관한 실증 연구가 확대되고 있음[74)]

- 8개 가상 국가를 설정한 다중 에이전트 워게임 실험에서 중립적 초기 조건에서도 군비경쟁 역학과 공세적 전략이 빠르게 형성되었으며, 시험된 5개 모델 모두에서 통계적으로 유의한 초기 확전 경향이 관측됨
  - 다만 이는 특정 행동 공간과 시나리오로 구성된 실험 환경의 결과이므로, 실제 지휘결심 상황으로 일반화하는 데에는 외적 타당도의 한계가 있음

- 별도의 연구는 미·중 가상 위기 시나리오 워게임에 국가안보 전문가를 참여시켜 인간 팀과 모델의 행동을 비교하였으며, 모델이 지시 문구에 민감하고 편향을 보이며 인간 팀과 상이한 선택 양상을 나타낸다는 점을 보고함[75)]
  - 연구진은 고위험 영역에 배치하기에 앞서 신중한 검토가 필요하다는 결론을 제시하였으나, 모델이 인간보다 일률적으로 더 공세적이라고 단정하지는 않음

- 해당 연구들은 자동화 의존도가 높은 워게임에서 핵 사용 선택이 늘어난다는 선행 연구를 인용하며, 전산 모형이 전쟁의 비물질적 비용을 충분히 반영하지 못한다는 점을 논거로 제시함

■ AI가 핵 조기경보·정보분석·지휘통제에 결합될 경우의 의사결정 시간 압축과 오판 위험이 별도의 연구 축으로 형성됨[76)]

- SIPRI는 불투명한 AI 추천과 자동화 편향이 결합될 경우 핵 위기 안정성에 부정적 영향을 미칠 수 있다는 점을 분석하고, 「AI와 핵무기의 접점에서의 거버넌스 진전」 등을 통해 국가별 비교를 수행함

74) J.-P. Rivera et al., Escalation Risks from Language Models in Military and Diplomatic Decision-Making, ACM FAccT 2024, Jun 2024
75) M. Lamparth et al., Human vs. Machine: Behavioral Differences Between Expert Humans and Language Models in Wargame Simulations, AAAI/ACM AIES 2024, Oct 2024
76) V. Chernavskikh & J. Palayer, Impact of Military Artificial Intelligence on Nuclear Escalation Risk, SIPRI, Jun 2025

13 AI 행동범위의 사전 제한과 감독·개입

13-1. 운용범위의 사전 제한 [무력 사용]

■ **ICRC와 SIPRI가 제시한 세 유형의 통제 조치는 자율무기체계의 예측불가능성을 완화하기 위한 실무 수단으로 활용되고 있음[77)]**

- 무기 파라미터에 대한 통제는 체계가 반응할 수 있는 표적 유형을 한정하고, 사용 환경에 대한 통제는 지리적 범위와 지속 시간을 한정하며, 인간 감독의 형태는 운용 중 감시와 개입·중단 능력을 유지하는 것을 의미함

- 유럽방위기금 설립 규정이 '인간에 대한 공격 수행 시 표적 선정 및 교전 결정에 유의미한 인간 통제의 가능성이 없는' 체계를 자금 지원에서 배제한 것은 이러한 수단을 적격성 기준으로 제도화한 사례에 해당함

13-2. 감독형 통제와 개입 능력 [공통]

■ **미국 방위고등연구계획국(DARPA)의 연구는 자율성 확대와 통제 가능성을 양립시키는 기술적 수단을 구체화하고 있음[78)]**

- (ACE, Air Combat Evolution) 개량 F-16(X-62A VISTA)을 이용해 AI의 시계 내 공중전 비행시험이 2023~2024년 수행됨

- (AIR, Artificial Intelligence Reinforcements) ACE의 성과를 시계 외(BVR) 영역으로 확장하여 다수 항공기 편대의 전술적 자율성을 개발하며, 유인 시험기에서 '인간 온 더 루프' 방식으로 먼저 검증한 뒤 무인 전투기로 이전하는 경로를 취함

- (DISCORD) 실시간 센서 데이터와 고충실도 시뮬레이션을 결합해 전술을 지속적으로 생성·적응시키되, 단일 최적해가 아니라 다양한 전략 포트폴리오를 지휘관에게 제공하는 것을 목표로 함

- (ITM, In the Moment) 전문가 간 의견이 갈리고 정답이 합의되지 않은 고위험 영역에서 AI를 개별 인간의 판단 성향에 정렬시키는 것이 위임 의사에 영향을 주는지를 연구함
  - 이들 연구의 공통 지향점은 개별 플랫폼에 대한 직접 조작이 아니라 임무·행동범위·제약조건을 설정하고 예외 상황에 개입하는 감독형 통제(supervisory control)를 기술적으로 구현하는 데 있음

77) ICRC & SIPRI, Limits on Autonomy in Weapon Systems, Jun 2020; Regulation (EU) 2021/697 (European Defence Fund), 29 Apr 2021
78) DARPA, Air Combat Evolution (ACE); Artificial Intelligence Reinforcements (AIR), Program Pages

■ **NATO 변혁사령부의 차세대 표적화(NGT)는 표적화 주기 단축과 인간 통제 인터페이스를 동일 사업 안에서 함께 제시함[79)]**

- 회원국 체계를 모듈형 AI와 데이터 관리로 연결하여 '정보에서 결정까지의 지연'을 50% 이상 단축하는 것을 전략 목표로 하며, 1만 개 수준의 잠재 표적을 1천 개의 우선순위 교전 목록으로 압축하는 규모 확대를 상정함
- 그 중심에 정교한 인간 통제 인터페이스(human control interface)를 둔다고 설명하며, '기술은 지휘관을 지원해야지 대체해서는 안 된다'는 원칙과 설명가능성·추적가능성·신뢰성 및 유의미한 인간 통제를 반영한다고 밝힘

14 시험·평가·검증 및 AI 보증

14-1. 시험평가 체계의 재설계 [공통]

■ **신미국안보센터(CNAS)는 AI·자율체계의 특성에 맞는 시험평가 절차의 재설계를 핵심 연구 주제로 다룸[80)]**

- 2025년 3월 발간된 「안전하고 효과적인(Safe and Effective)」은 AI·자율체계가 전통적 무기체계와 달리 데이터와 운용환경 변화에 민감하고 행동 설명이 어렵다는 점을 고려하여, 기술 발전 속도에 부합하는 민첩하고 반복적인 시험평가가 필요하다고 분석함

14-2. AI 보증 개념의 제도화 [공통]

■ **유엔군축연구소(UNIDIR)는 규범적 언어를 기술적 검증 요건으로 전환하는 문제를 중심으로 연구를 축적하고 있음[81)]**

- 「군사 AI 거버넌스의 글로벌 만화경」(2024)과 「증거 기반 로드맵」(2025)을 통해 지역별 접근의 공통점과 차이를 분석하고 다자규범 논의의 경로를 제시함
- 2026년 2월 「군사 AI 거버넌스의 글로벌 프리즘」은 2025년 5개 지역 협의에서 제시된 각국 정책을 종합하고 조달 기준, 보증 요건, 사고 대응 및 위험 관리에 관한 논의를 정리하였으며, 참가국의 조달 파라미터 및 보증 요건 관련 입장을 정량 데이터로 함께 제시함

79) NATO Allied Command Transformation, Next Generation Targeting Project

80) J. Wallin, Safe and Effective: Advancing Department of Defense Test and Evaluation for Artificial Intelligence and Autonomous Systems, CNAS, 13 Mar 2025

81) UNIDIR, The Global Kaleidoscope of Military AI Governance, 2024; An Evidence-Based Road Map for Future Policy Action, Jul 2025; The Global Prism of Military AI Governance, Feb 2026

- 안보·국방 분야 국가 AI 전략 수립 지침 초안, 신뢰구축조치(CBM) 설계, AI 체계의 예측가능성·이해가능성 분석이 주요 연구 축을 구성함

■ 국가 차원에서도 보증 역량을 통해 통제의 기술적 전제조건을 확보하려는 시도가 확인됨[82)]

- 영국 국방부는 2026년 정책 문서에서 AI 보안연구소(AISI)와 협력하여 공급자·과업에 관계없이 모델을 신속히 평가할 수 있는 AI 보증 역량 개발을 모색하도록 함
- 국방 AI 보증 체계의 이행 과제를 데이터 적합성, 자율성 수준 설정, 신뢰 보정(trust calibration), 운용설계영역(ODD) 정의 등으로 구조화하는 학계 연구가 수행됨

■ 군사 영역 특화 평가체계(벤치마크) 개발이 새로운 연구 축으로 형성되고 있으나, 아직 초기 단계의 연구가 대부분임[83)]

- 최근 일부 벤치마크 연구는 기존 평가체계가 역사적 전투 사례 기반 인과추론 평가와 실시간 전략게임 환경을 활용한 전략추론 평가의 두 갈래로 발전해 왔으며, 국제인도법 준수를 강제 조건으로 검증하지 않는다는 한계가 있다고 지적함
- 표적 결정, 교전규칙 해석, 비례성 판단이 국제인도법 및 무력 충돌법 준수와 직결되므로 해당 차원을 시험하지 않는 평가체계는 근거 없는 신뢰를 형성할 수 있다는 문제가 제기되며, 법적 위험·도덕적 피해·지역적 편향을 함께 측정하는 평가체계가 제안되고 있음
  - 해당 연구들은 소수 모델과 제한된 수의 모의 분쟁 시나리오를 대상으로 한 사전공개(preprint) 단계 연구가 다수이므로, 확립된 연구 분야의 합의로 보기는 어려움

■ 시험평가 체계가 도입 속도를 따라가지 못한다는 지적은 규범 논의에서도 반복적으로 제기됨[84)]

- REAIM 2026 결과문서는 AI 보증 분야의 진전과 TEVV의 정책·교리·조달 절차 통합을 권고 사항으로 포함함

82) UK Ministry of Defence, Putting Artificial Intelligence (AI) at the Heart of UK Defence, 10 Jun 2026; C. Cockburn & S. Farrow, AI Assurance in UK Defence: Challenges in Operationalising JSP 936, arXiv:2606.09414, 8 Jun 2026

83) Z. Li et al., WARBENCH: A Comprehensive Benchmark for Evaluating LLMs in Military Decision-Making, arXiv:2603.21280, 22 Mar 2026

84) REAIM 2026, REAIM Pathways to Action, 5 Feb 2026

15 인간-AI 상호작용과 다중 자율체계의 통제

15-1. 인간-기계 통합의 조직적 조건 [그 밖의 군사 맥락]

■ RAND는 AI의 성능뿐 아니라 AI가 군사조직 및 인간 운용자와 결합되는 방식에 연구의 비중을 두고 있음[85)]

- 「전투원을 위한 인공지능의 한계 이해」(2024, 전5권)는 공군 자재사령부의 의뢰로 수행된 연구로, 사이버보안·예측정비·워게임·임무계획 등 사례를 통해 AI의 한계를 분석함

- 「하나의 팀, 하나의 전투」(2025)는 미 육군의 인간-기계 통합을 장비 도입 문제가 아니라 인간과 AI가 교류기억체계(transactive memory system)와 공유정신모델을 구축하는 사회기술적(sociotechnical) 통합 문제로 접근함
  - 인간이 다양하므로 동일 유형의 부대 간에도 AI의 특성·성능·가치가 달라질 수 있다는 점을 전제하고, 기계가 인간의 신뢰를 얻고 유지하는 과정을 모니터링할 것을 권고함

15-2. 다중 자율체계의 통제 [공통]

■ 단일 체계를 넘어 다수의 AI·자율체계가 함께 운용될 때의 통제 문제가 별도의 연구 축으로 형성됨[86)]

- CNAS의 「학습의 교훈」(2025)은 유인 체계가 인간의 직관·훈련·의사소통에 의존할 수 있는 반면 자율체계는 서로 다른 공급업체가 개발한 경우에도 상호 협조하도록 사전에 프로그래밍되어야 한다는 구조적 차이를 제시하고, 체계 간 호환성 확보를 위한 시험평가 조정과 공통 모델링·시뮬레이션 도구 활용을 권고함

- DARPA의 DICE는 다수 이종 AI 에이전트가 팀을 구성해 장기 임무를 수행하되 통제 범위 내에 머물도록 하는 분산형 AI 아키텍처를 연구하며, 개별 에이전트의 국소 추론 통제를 통해 집단의 창발적 행동을 지휘관 의도에 정렬시키고 이탈 에이전트에 대한 복원력을 확보하는 것을 목표로 함

- 다중 체계 환경에서는 인간이 개별 플랫폼을 직접 통제하는 방식보다 임무·행동범위와 제약조건을 설정하고 예외 상황에 개입하는 감독형 통제의 비중이 커지는 것으로 정리됨

85) RAND, Understanding the Limits of Artificial Intelligence for Warfighters, Vol. 1-5, RR-A1722, 2024; One Team, One Fight: Insights on Human-Machine Integration for the U.S. Army, RR-A2764, 2025

86) CNAS, Lessons in Learning: Ensuring Interoperability for Autonomous Systems in the Department of Defense, May 2025

16 책임·추적성과 생애주기 통제

16-1. 조달·법적 검토 단계의 통제 [공통]

■ SIPRI는 책임 있는 AI 공약이 실제로 확보되는 지점으로 조달 단계를 제시하고 있음[87)]

- 2026년 2월「군사 인공지능의 책임 있는 조달」은 법적 검토와 조달 절차의 연계 방안을 분석하며, 2026년 3월 캐나다 국방부와 공동으로 CCW GGE 부대행사를 개최하여 관련 권고를 협상 과정에 투입함
- 이는 개발·획득 단계에서 작동하는 절차적 통제가 운용 국면의 통제와 별개로 설계되어야 한다는 문제의식과 연결됨

■ 규범 문서 차원에서도 책임 소재와 추적성 확보가 구체적 이행 수단으로 제시됨[88)]

- REAIM 2026 결과문서는 책임 공백(accountability gap) 방지, 국가 책임자 지정, 감사 추적 및 사고 보고 체계 유지, 지휘·통제 계통의 명확화, 신무기 법적 검토 수행을 항목별로 권고함
- 유엔군축사무국은 결의 80/58에 따라 2026년 6월 제네바에서 국가·시민사회·산업계·학계가 참여하는 비공식 다자 이해관계자 교류를 개최하였으며, UNIDIR은 인권최고대표사무소와 공동으로 설계·데이터·시험·배치·사후지원 전반에서 산업계의 역할을 다루는「군사 영역 AI에 관한 책임 있는 산업 행동 프레임워크」를 발표함

■ UNIDIR은 교전 시점에 인간 통제가 유지되더라도 생애주기 전반에서 책임 소재가 모호하게 남을 수 있다고 지적함[89)]

- AI 체계가 하류에서 내리는 판단이 상류의 인간 운용자가 표적을 선정하거나 교전을 결정하는 판단에 영향을 미치기 때문임
- 이에 따라 설계·조달·배치·폐기에 이르는 전 단계에서 책임 위험이 발생할 수 있으며, 완화 수단으로 감사가능성, 추적성, 데이터 로그 보관, 시험·평가·검증(TEVV)이 제시됨
- 아울러 사람·절차·기술을 결합한 통합적 접근이 필요하다고 보았는데, 이는 ICT 보안 분야에서 확립된

87) SIPRI, Responsible Procurement of Military Artificial Intelligence, Feb 2026
88) REAIM 2026, REAIM Pathways to Action, 5 Feb 2026; UN General Assembly, A/RES/80/58, 1 Dec 2025; UNIDIR & OHCHR, Framework of Responsible Industry Behaviour for AI in the Military Domain, Jun 2026
89) B. Unal, From Cyberspace to AI: Translating Norms of Responsible State Behaviour in Cyberspace to AI in the Military Domain, UNIDIR, 18 Sep 2026

개념을 군사 AI에 적용한 것임

16-2. 규범 이행에 관한 평가 [공통]

■ **정책 싱크탱크와 시민사회는 다자 논의와 실제 도입 속도의 관계 및 규범 해석을 주된 분석 대상으로 삼고 있음[90)]**

- 미국외교협회(CFR) 및 조지타운대 안보·신흥기술센터(CSET) 연구진은 다자 협상이 국제 관료 절차의 속도로 진행되는 반면 각국 군의 AI 통합은 가속되고 있어, 규범 논의가 현실과 괴리될 위험을 지적함

- 국제거버넌스혁신센터(CIGI)는 주요 군사국의 불참 상황에서 중견국 중심의 행동규범 및 책임성 프레임워크 선행 구축을 대안으로 제시하고, 카네기 윤리위원회는 2024년 서울 「행동청사진」이 비교적 논쟁의 여지가 적은 내용을 담았음에도 약 30개국이 지지하지 않았다고 보고 '책임 있는 AI' 개념 자체가 경합 중이라는 해석을 제시함

- 웨스트포인트 리버연구소는 중국의 인간 감독 개념이 국제인도법 수호 입장으로 해석되기 어려운 구조적 이유를 분석하며, 휴먼라이츠워치 및 킬러로봇 중단 운동은 2단계 접근과 교전 시점의 인간 통제를 옹호하고 TEVV 미비 문제를 제기함

  – 위 평가들은 각 연구기관의 분석·규범적 입장으로, 객관적으로 확립된 사실과는 구분됨

17 연구동향 종합

■ **본 보고서는 앞서 검토한 통제 수단별 연구동향을 다음 여섯 가지 변화로 정리하였으며, 이는 공인된 분류가 아니라 본 보고서의 종합임**

- (제1변화) '무기의 자율성'에서 '의사결정의 자율성'으로 연구 대상이 확대됨.
  과거 논의가 자율무기의 무승인 공격 가능성에 집중되었다면, 최근에는 AI가 정보분석·표적화·작전계획·지휘결심에 어떻게 개입하는지가 중심 주제가 됨

- (제2변화) 인간 통제의 중심이 최종 공격 승인에서 생애주기 전반의 통제로 확대됨.
  데이터 설계, AI 훈련, V&V, T&E, 임무범위 설정, 운용자 훈련, 사후검토가 모두 통제의 구성요소로

90) CFR, Military AI Adoption Is Outpacing Global Cooperation, Feb 2026; GCSP, The Third Summit on REAIM, Feb 2026; K. V. Hiebert, Prospects Have Dimmed for Autonomous Weapons Guardrails, CIGI, 30 Apr 2026; Carnegie Council, From Principles to Action, 2024; Lieber Institute West Point, Human Oversight with Chinese Characteristics, May 2026; Human Rights Watch, Addressing Artificial Intelligence in the Military Domain, 14 Jun 2026

다루어짐

- (제3변화) AI 성능 자체보다 AI 보증(assurance)이 중요한 연구영역으로 등장함.
  정확도뿐 아니라 강건성, 신뢰성, 설명가능성, 추적가능성, 사이버보안, 적대적 강건성이 전력화의 핵심조건으로 연구됨

- (제4변화) 단일 AI 체계에서 다중 에이전트·군집체계로 연구가 이동함.
  이에 따라 개별 플랫폼 직접 통제보다 임무·행동범위와 제약조건을 설정하고 예외상황에 개입하는 감독형 통제의 중요성이 커짐

- (제5변화) 인간-기계 관계가 기술적 인터페이스를 넘어 조직·교육의 문제로 확대됨.
  자동화 편향, 신뢰의 과잉·부족, 운용자의 AI 이해능력, 책임구조, 인간-AI 팀 훈련이 별도 연구분야로 발전함

- (제6변화) AI와 전략적 안정성 문제가 새로운 연구 축으로 등장함.
  AI가 핵 조기경보, 정보분석, 지휘통제에 결합될 경우의 의사결정 시간 압축과 오판·확전 가능성이 주요 연구주제가 됨

Ⅳ 종합 및 시사점

■ 군사 AI에서 인간 통제의 핵심 질문은 '인간이 버튼을 누르는가'에서 '어떤 군사적 판단을 AI에 위임할 수 있으며, 그 위임의 범위와 조건을 인간이 실질적으로 이해·설정·감독하고 필요 시 철회할 수 있는가'로 이동하고 있음

- 루프 기반 구분은 여전히 유용한 분석 도구이나, 통제의 실질을 판정하기에는 불충분한 것으로 평가됨
- 이에 따라 개발·검증·임무설정·배치·운용·중단·사후평가를 포괄하는 생애주기 관점이 여러 정책·연구에서 공통적으로 강조되고 있으나, 구체적 용어와 제도화 방식은 국가별로 상이함

■ 주요국 정책에서 공통적으로 확인되는 요소와 차이점이 병존함

- (공통 요소) 국제법 준수, 인간의 책임 유지, AI의 능력과 한계에 대한 이해, 검증·시험평가, 운용범위 제한, 비정상 행동에 대한 인간의 개입 가능성
- (차이점) 미국은 훈령 3000.09와 책임 있는 AI 체계, 영국은 맥락 적합적 인간 관여와 JSP 936, 일본은 방위장비 연구개발 단계의 책임 있는 AI 지침, 프랑스는 인간 지휘책임과 SALA 개발 제한, 중국은 인간 통제·중단 가능성과 최종 인간 책임을 각각 강조함
- 법적 구속력의 수준과 외부 검증 수단의 유무에서 국가 간 편차가 크며, 다수 국가에서 인간 통제 요건은 법률이 아닌 내부 지침·정책 수준에 머물러 있음

■ 연구 측면에서는 개념 정의보다 통제 수단의 구현과 검증이 중심 주제가 되고 있으며, 인간 판단의 실효성, 행동범위의 사전 제한과 감독·개입, 시험·평가·검증과 AI 보증, 인간–AI 상호작용과 다중 자율체계, 책임·추적성이 주요 축을 구성함

- 기관 유형별로 상대적으로 강조하는 수단에 차이가 나타나나(ICRC·UNIDIR 등 국제인도법 수호기관·연구기관은 인도법 요건과 산업 행동 기준, 정책연구기관은 조달·편향·이행 수단, 기술연구기관은 감독형 통제와 신뢰 가능한 자율성), 실제 연구 영역은 상당 부분 중첩됨
- 2026년 GGE 최종보고서가 수식어 없는 '통제와 인간 판단'을 채택하고 통제가 사용 전·사용 중 조치를 통해 간접적으로도 행사될 수 있음을 인정한 것은, 본 보고서가 제시한 생애주기 관점과 같은 방향의

국제적 정리에 해당함

- 이어 2026년 11월 제7차 검토회의가 이 구성요소를 바탕으로 2027~2028년 문서 작성 권한을 GGE에 부여하면서 논의는 문서 작성 단계로 넘어갔으나, 문서의 법적 형식은 열려 있으며 협약에 대한 정식 검토도 완료되지 못해 구속력 있는 규율로 이어질지는 2028년 당사국회의의 결정에 달려 있음
- 한편 다자 규범 논의의 속도와 실제 군사적 도입 속도 간 괴리가 지속적으로 지적되고 있으며, 해당 격차가 확대될 수 있다는 우려가 제기됨

주요 관찰사항

■ 아래는 위 분석 결과에서 공통적으로 확인되는 사항을 정리한 것임

- (제도 측면) 인간 통제를 운용 단계의 승인뿐 아니라 개발·조달·시험평가·법적 검토 단계에서도 다루는 경향이 확인됨
- (기술 측면) 설명가능성·추적가능성·불확실성 정보 제공이 인간 통제의 기술적 전제조건으로 다뤄지고 있으나, 이를 검증할 평가체계는 초기 단계에 있음
- (운용 측면) 자동화 편향 완화를 위한 운용자 교육, 인간–AI 팀 훈련, 이의제기 절차의 제도화가 통제 실효성의 조건으로 부각되고 있음
- (국제 측면) 규범 논의와 실제 도입 속도의 괴리, 동맹·다자 협력에서 상호 검증 가능한 보증 요건의 부재가 지적되고 있음

동향 AI 안전 동향 분석 리포트 26-02

군사 AI의 인간 판단·통제: 국제 정책·연구 동향

Human Judgment and Control over Military AI:
International Policy and Research Trends



ETRI 한국전자통신연구원 www.etri.re.kr
인공지능안전연구소 AI Safety Institute
95560

AI 안전 동향 분석 시리즈 26-02(2026년 11월)
『군사 AI의 인간 판단·통제: 국제 정책·연구 동향』

발 행 일 2026년 11월 27일
발 행 처 한국전자통신연구원 인공지능안전연구소 AI안전정책 및 대외협력실
발 행 인 최 민 석 (AI안전정책 및 대외협력실 실장)
연 구 진 유 고 운 (AI안전정책 및 대외협력실 석사후연수연구원)
전 화 번 호 031-739-7651
주 소 (13488) 경기도 성남시 분당구 대왕판교로712번길 22, 글로벌 R&D센터 A동 4층
편집·디자인 ※ 본 자료의 표지 디자인에는 생성형 AI 기반 이미지 제작 도구가 활용됨.

본 자료의 저작권은 한국전자통신연구원 인공지능안전연구소에 있으며, 무단 복제·배포·상업적 이용을 금합니다.