AI 기반 자율무기 체계의 발전과 시사점: 군사적 효용과 통제 가능성을 중심으로

강 준 석*

Abstract

최근 비약적으로 발전하고 있는 AI 기반 자율무기 기술은 미래의 전쟁 수행 방식뿐만 아니라 국제 안보 환경 전반에 대해서도 근본적인 변화를 촉발할 수 있는 잠재력을 갖고 있다. 향후 자율무기 체계가 전면적으로 활용될 경우, 전장에서의 작전 속도와 효율성의 향상, 표적 식별 및 타격의 정확성과 정밀성 제고, 저비용·분산형 무기체계의 대량 운용 등을 통해 상당한 수준의 군사적 효용이 창출될 수 있을 것으로 기대된다. 하지만, AI의 가치 정렬 문제와 인간 운용자의 자동화 편향 등으로 인한 인간 통제의 형해화 가능성과 이와 연관된 부작용에 대한 우려도 지속적으로 제기되고 있다. 또한, 자율무기 도입에 따른 군사적 대응의 자동화·가속화가 오히려 우발적 확전의 가능성을 높이고, 경쟁국보다 고도화된 자율무기 체계를 이들보다 더욱 신속하게 개발하고 더 많이 배치하고자 하는 새로운 형태의 군비경쟁이 촉발될 가능성에 대한 우려도 상당하다. 이러한 위험 요인은 실효성 있는 인간 통제의 제도화나 알고리즘 중심의 군비통제 등 현재 논의 중인 다양한 대응 방안을 통해서 일정 수준 완화될 것으로 기대할 수 있다. 하지만 여전히 미해결 상태로 남아 있는 과제가 상당하며, 이와 같은 문제들을 적절하게 해결하지 못할 경우 자율무기 체계의 확산이 궁극적으로 인간 통제의 약화와 국가 간 전략적 불안정성의 심화로 이어질 가능성도 배제하기 어렵다. 본고는 이러한 논의를 바탕으로 관련 쟁점을 심층적으로 분석하면서, AI 기반 자율무기 체계의 미래 전개 양상을 (1) 통제된 자율화와 제한적 활용, (2) 킬체인의 지능화와 인간 통제 약화, (3) 고자율 무기체계 확산과 전략적 불안정성 심화라는 세 가지 시나리오로 구분하여 전망하고, 인간 통제의 실효성을 확보하고 전략적 안정성을 유지하기 위한 정책적 시사점을 제시한다.

* 정보통신정책연구원 선임연구위원, jskang@kisdi.re.kr

I. 서 론

1984년 개봉된 영화 터미네이터는 인공지능 전략 방어 네트워크인 '스카이넷(Skynet)'이 인간 저항군의 지도자를 제거하기 위해서 AI 기반 자율 살상 로봇(T-800)을 과거로 보내면서 시작된다. 자의식(self-awareness)을 갖게 된 스카이넷은 이에 위협을 느낀 인간이 자신의 작동을 중지시키려 하자, 이를 자신의 존속에 대한 실존적 위협으로 인식하고 인류를 절멸시키기 위한 핵전쟁을 일으키게 된다. 핵전쟁 이후 살아남은 소수의 인간이 저항군을 조직하고 기계 지배에 대항하자, 스카이넷은 T-800을 과거로 보내 저항군 지도자가 잉태되기도 전에 그의 모친을 제거함으로써 존재 자체를 원천적으로 소멸시키고자 한 것이다. 이 같은 설정은 그동안 공상과학(SF)의 영역에 머물고 있었으나, 최근 빠르게 발전하고 있는 AI 기술은 인간이 만든 인공지능이 인간의 생명에 관한 결정을 스스로 내리고, 인간의 의도와 다른 방식으로 군사적 목표를 추구하고 그 목표를 달성하기 위해서 인간의 통제를 우회하거나 기만할 가능성을 단순한 허구로만 치부하기 어렵게 만들고 있다.

2026년 6월 영국의 과학전문지 뉴사이언티스트(New Scientist)는 우크라이나 방산업계 인사 알렉산더 코하노우스키(Alexander Kokhanovskyy)의 발언을 인용하여, 약 2년 전 우크라이나 동부전선에서 10여 대의 AI 통제 쿼드콥터 드론이 이른바 '터미네이터 모드(Terminator Mode)'[1]로 운용되었고, 그 결과 다수의 러시아 병사 사망자가 발생했다는 주장을 보도했다(Sparkes, 2026). 해당 발언의 진위는 아직 확인되고 있지 않지만, 만약 그 주장이 사실이라면, 이는 표적 선정 및 최종 발사 결정에 인간 운용자가 개입하지 않는 완전 자율무기가 인간 병사를 공격하고 살상한 최초 사례라고 할 수 있다. 이처럼 AI 기반 자율무기 체계는 더 이상 먼 미래의 기술이 아니라 전장에서 인간의 생사를 좌우하는 현실의 문제로 부상하고 있으며, 이에 따라 그 군사적 효용과 위험을 동시에 면밀히 검토할 필요성 역시 커지고 있다.

본고에서는 이와 같은 문제의식을 바탕으로 AI 기반 자율무기 체계의 현황과 다양한 쟁점을 살펴보고 관련 시사점을 도출하였다. 이를 위해서 먼저 자율무기 체계의 개념과 자율성 수준에 따른 분류 방식, 관련 체계의 핵심 기술 구성 요소 등을 분석하였다. 다음으로 자율무기 체계 도입 및 활용에 따른 작전적 이점과 기대 효과를 살펴보았다. 또한, 통제 상실 가능성과 인간 존엄성 훼손 가능성 등 자율무기 체계와 연관된 다양한 쟁점과 우려를

1) 보도에 따르면, 해당 인사는 드론이 특정 구역에 투입된 뒤 인간 운용자의 통제 없이 탑재된 AI가 스스로 표적을 탐색하고 식별한 뒤 공격 결정 및 실행까지 수행하는 완전 자율 교전 모드가 실행되었다고 주장했다.

파악하고, 이에 대응하기 위한 기술적·제도적 통제 장치 마련 가능성과 그 실효성 등을 분석했다. 마지막으로, 앞선 논의를 바탕으로 AI 기반 자율무기 체계의 미래 전개 양상을 전망하고 관련 시사점을 도출하였다.

Ⅱ. AI 기반 자율무기 체계의 정의 및 구성 기술 요소

1. AI 기반 자율무기 체계의 정의

AI 기반 자율무기 체계는 전장 환경의 인식, 표적의 탐지·식별·추적, 최종적인 목표 설정 및 교전 결정 등 무력 사용 과정의 전부 또는 일부에 AI가 활용되어 특정 임무를 자동화하거나 가속화하는 무기체계를 의미한다(U.S. Department of Defense [DoD], 2023). 자율 무기체계는 관찰-판단-결심-행동의 의사결정 루프(OODA Loop)[2]의 각 단계에서 인간이 개입하는 정도에 따라서 다양한 층위로 분류될 수 있다. 예를 들어, 반자율(semi-autonomous) 무기 체계는 AI가 센서 데이터 등으로부터 획득한 정보를 근거로 표적을 식별하고 추천할 수는 있으나, 최종적인 교전 및 타격 승인 권한은 인간 운용자에게만 부여되는 방식(Human-in-the-loop)이다. 이보다 자율성이 한 단계 더 높은 인간 감독형 자율무기는 표적의 탐지·식별·추적·타격의 모든 과정을 AI가 독립적으로 수행할 수 있으나, 인간 운용자가 전체 과정을 감독하여 무기의 오작동이나 민간인 피해 등이 우려될 경우 무기체계의 작동을 중단시킬 수 있는 방식(Human-on-the-loop)이다. 완전 자율무기 체계는 초기에 해당 무기 체계가 가동되고 임무가 부여되는 단계까지만 인간 운용자가 개입하고 이후에는 인간의 개입이나 통신이 전혀 이루어지지 않는 방식(Human-out-of-the-loop)이다.

2. 핵심 기술 구성 요소

AI 기반 자율무기 체계를 개발하기 위해서는 자동 표적 인식, 교전 결정 및 사격통제, 자율항법, 데이터 융합, 네트워크·군집 협업 등 다수의 기술이 유기적으로 결합된 AI 스택

2) Observe(관찰): 현재 상황에 대한 정보를 수집하는 과정. Orient(판단): 수집한 정보를 해석하고 의미를 부여하는 과정. Decide(결심): 해석된 상황을 바탕으로 어떤 행동(공격, 방어, 회피, 기동, 후퇴 등)을 할지 선택하는 과정. Act(행동): 결정한 행동을 실행하는 과정(Johnson, 2023).

(AI Stack)의 확보가 필요하다. 다양한 기술 요소들을 적절하게 조합하여 전장 환경을 종합적으로 인식하고 임무 및 목적에 부합하는 행동을 자율적으로 선택하고 수행할 수 있는 복합적인 지능형 체계를 구축하는 것이다.

1) 표적 자동 인식(Automatic target recognition, ATR)

표적 자동 인식(ATR) 기술은 컴퓨터 비전과 기계학습 모델 등을 활용하여 전장 환경에서 표적을 탐지·식별·추적하는 기술이다(Verly et al., 1989). 딥러닝 기술 등장 이전의 과거 표적 인식 기술은 인간이 사전에 설계한 특징값, 형상 모델, 열·레이더 신호 패턴 등을 활용하여 표적을 탐지·분류했다. 표적의 특징을 일일이 정의하여 해당 정의와 센서에서 탐지된 이미지와의 일치도를 기반으로 표적 여부를 판단했다. 하지만 오늘날에는 AI 기술을 활용하여 라벨링된 영상(예, 적국의 전함 및 전투기 등의 영상 이미지)을 모델에 학습시켜 어떤 시각적 패턴이 표적을 구성하는지를 컴퓨터가 스스로 학습하도록 하고 새로운 영상이나 센서 데이터가 입력되었을 때, 해당 대상이 표적인지, 표적이라면 어떤 종류의 표적인지, 어느 방향으로 이동하고 있는지 등을 자동으로 판단하도록 하고 있다. AI 기반 자율무기 체계에서 ATR이 핵심적인 기술인 이유는 표적 인식이 전체 무력 사용 과정에서 가장 중요한 출발점으로 기능하기 때문이다. 표적을 정확하게 탐지·식별·추적하는 정밀한 기술을 확보하지 못한다면, 위협 평가, 교전 결정, 무장-표적 할당, 사격통제, 자율항법, 종말유도 등 이후의 무력 사용 과정은 진행되기가 어려운 것이다.

2) 교전 결정 및 사격 통제(Engagement decision and fire control)

표적 자동 인식이 표적 후보를 탐지·식별·추적하는 기술이라면, 교전 결정 및 사격 통제는 ATR에서 도출된 결과를 바탕으로 ▲무엇을, ▲언제, ▲어떤 수단으로, ▲어느 수준의 인간 통제하에서 공격할 것인지를 판단하고 실행하는 기술이다(DoD, 2023). AI 기반 자율무기 체계에서는 다음의 판단 절차에 따라서 교전 결정이 이루어진다. 우선, ATR을 통해서 탐지된 표적 후보가 실제로 의도하는 목표물인지 또는 오인 표적인지의 여부를 확인한다. 다음으로, 해당 표적 후보의 군사적 위협 수준을 평가한다. 셋째, 교전 규칙, 교전 금지 구역 또는 비타격 목록 등을 고려하여 이와의 상충 가능성 등을 검토한다. 마지막으로 인간 운용자가 교전 여부를 최종 승인할 것인지, 또는 사전에 정해진 조건 아래 자동 대응할 것인지를 결정한다. 사격 통제는 교전이 승인된 이후 이를 실행하기 위해서 해당 표적에 적

합한 타격 수단을 할당하고 실제 타격 과정을 통제하는 기술이다. 교전 결정 및 사격 통제 과정에서 AI는 다음과 같이 활용될 수 있다. 우선, AI 기술을 적용하여 다수의 표적 후보 중 어느 대상이 즉각적 위협인지를 판단하여 표적의 우선순위를 결정한다. 다음으로, 다양한 대응 수단이 있을 때 어떤 수단이 어느 표적에 적합한지를 결정한다. 또한, 표적 추적, 교전 가능성 판단, 타격 수단의 발사 및 투발, 발사 후 추적 등을 자동화한다. 마지막으로 AI가 제안한 판단을 인간 운용자가 이해하고 이를 승인·거부하거나 해당 무기 체계의 작동을 중단할 수 있도록 설명성과 통제성을 제공할 수 있다.

3) 자율항법 기술(Autonomous navigation)

AI 기반 자율항법 기술은 무인 플랫폼이 인간의 지속적인 개입 없이도 임무에 적합한 이동 경로를 스스로 계획하고 주변 상황 변화에 따라서 이동 계획을 수정하고 이를 실행할 수 있도록 하는 기술을 의미한다(Chang et al., 2023). 순항미사일이나 무인기에 사용되어 온 기존의 웨이포인트(waypoint) 기반 자동조종은 사전에 입력된 좌표를 따라 이동한다는 점에서 자율항법의 가장 기초적인 형태이다. 하지만, 최근 러시아–우크라이나전에서도 확인되는 바와 같이, 전파 방해(jamming)·스푸핑(spoofing)[3]·통신 두절·GPS 신호 교란 등이 빈번하게 발생하고 있는 현대 전장 환경에서는 더 이상 안정적인 통신망과 GPS에 기반한 기존의 자율항법 기술의 활용이 제한될 수밖에 없다. 이와 같은 상황에 대응하여 자율무기 체계가 실제 전장에서 작동하기 위해서는 외부 통신이나 위성항법 신호가 제한되더라도 일정 수준의 독자적 위치 추정과 경로 설정 능력을 갖추어야 한다. AI 기반 자율항법 기술이 이와 같은 기능을 제공하여 무기 플랫폼이 스스로 목표 지점까지의 최적의 경로를 산출하고, 표적의 이동에 따라서 자신의 경로를 조정하는 등 상황 변화에 자율적으로 대응할 수 있도록 발전할 경우 표적 추적과 교전 성공 가능성은 더욱 향상될 수 있을 것이다.

4) 데이터 융합 및 전장 상황인식 기술

AI 기반 자율무기 체계에서 데이터 융합은 표적 자동 인식과 교전 결정 사이의 연결 고리의 역할을 하고 있다. ATR이 주로 표적의 탐지와 분류에 초점을 맞춘다면, 데이터 융합 및 전장 상황 인식 기술은 ATR을 활용하여 식별된 잠재적 표적이 전체 전장 상황의 맥락

3) 전파 방해는 강력한 전자파 신호를 송출하여 GPS나 데이터링크의 정상적인 수신을 방해하는 방식이며, 스푸핑은 허위 GPS 신호를 보내 무기체계가 자신의 위치를 잘못 인식하도록 기만하는 방식이다(Kerns et al., 2014).

에서 어떤 의미를 갖는지를 해석하는 데 활용됨으로써 자율무기 체계의 정확성과 정밀성을 높이는 데 기여할 수 있다(Sayler, 2020). 예를 들어 드론으로부터 수집된 영상 정보만으로는 특정 차량이 군용 차량인지 민간 차량인지의 여부를 명확하게 판단하기 어려울 수도 있으나, 해당 차량이 적군의 통신 신호가 탐지된 지역에서 반복적으로 이동하고, 열상 정보상 군용 장비와 유사한 패턴을 보이며, 적 보급로와 일치하는 경로를 따라 이동하고 있다면 군사 표적일 가능성을 더 높게 평가할 수 있다. 반대로 영상 정보에서는 군용 차량과 유사한 외관을 갖고 있어도, 해당 대상이 민간 대피로에 위치하고 있고, 주변에 민간인 이동이 확인되며, 다른 센서에서 군사적 활동과 연관된 징후가 포착되지 않는다면 교전 판단은 보류되거나 인간 운용자의 재확인 절차가 필요할 수도 있는 것이다. 이처럼 AI 기술을 활용한 데이터 융합을 통해서 서로 다른 정보 사이의 관계를 면밀하게 분석할 경우 표적 판단 및 교전 결정의 속도뿐만 아니라 신뢰도와 정확성 역시 제고할 수 있을 것이다.

5) 네트워크·군집 협업 기술

네트워크·군집 협업 기술은 다수의 AI 기반 자율무기 플랫폼이 밀접하게 연결되어 ▲정보를 공유하고, ▲상호 간에 임무를 분담하며, ▲통합된 전투체계처럼 작동하도록 하는 기술을 의미한다(DARPA, n.d.). 군집 협업 기술은 개별 플랫폼이 작전 목표와 기본 규칙 등을 공유한 이후, 각자의 위치, 센서 정보, 잔여 연료나 배터리, 손상 상태, 표적 접근 가능성, 위험 수준 등을 고려하여 세부 임무를 분담하여 수행하도록 하는 것이다. 이는 다시 말해서, 인간 운용자가 개별 플랫폼의 세부 움직임을 직접 지시하는 것이 아니라, 전체 임무 목표와 제한 조건을 부여하면 다수의 플랫폼이 집합적·자율적으로 임무를 수행하도록 하는 것이다. 자율무기 체계가 단일 플랫폼 차원에서 운용될 경우 그 군사적 효과는 제한적일 수밖에 없다. 이는 넓은 전장 공간에서 개별 플랫폼이 복수의 위협 및 목표에 동시에 대응하기에는 물리적인 한계가 존재하기 때문이다. 반면, 다수의 자율무기 플랫폼이 서로 연결되어 감시·정찰·표적획득·타격·피해평가 등의 임무를 통합적·집합적으로 수행할 경우, 더 많은 표적을 동시에 처리하고, 일부 플랫폼이 손실되더라도 여타 플랫폼이 임무를 재분배해 작전 지속성을 유지할 수 있는 등 개별적으로 임무를 수행했을 경우와 비교하여 작전 효과가 더욱 증폭될 수 있을 것이다(Kallenborn, 2020). 따라서 이를 가능하게 하는 네트워크·군집협업 기술은 작전 방식을 분산형·협업형 전투체계로 전환시키는 핵심 기반 기술이라고 할 수 있다.

Ⅲ. AI 기반 자율무기 체계의 군사적 효용 및 기대 효과

1. 작전적 이점

1) 속도

AI 기반 자율무기 체계를 활용함으로써 기대할 수 있는 가장 핵심적인 군사적 효용 중의 하나는 OODA 루프(관찰-판단-결심-행동)의 의사결정 주기를 적군보다 더욱 빠르게 회전시킴으로써 적을 반응적·수세적 위치로 몰아넣는 것이다(Allen & Husain, 2017). 전시에는 아군과 적군 양측이 모두 자신이 먼저 유리한 위치를 점하기 위해서 상대를 관찰하고, 상황을 판단하고, 전략이나 전술을 결정하고, 이의 실행을 위해서 행동하는 패턴을 반복하게 된다. 이때 한쪽이 상대방보다 더 빠른 속도로 이와 같은 과정을 진행시킬 경우, 경쟁국은 계속해서 변화하는 전장 상황에 적응하지 못하고 자신의 판단 근거가 되는 실제 전장 상황을 정확하게 인식하지 못하는 상태에서 수세적인 결정을 내리게 되고, 그 결과로 전장에서의 통제력을 상실할 수 있는 것이다.

이와 같은 작전적 이점을 실현하기 위해서 각국은 AI를 활용하여 OODA 루프의 각 단계를 자동화하고 관찰-판단-결심-행동으로 이어지는 의사결정 주기 전반을 가속화하는 노력을 기울이고 있다(Johnson, 2023). 우선 관찰 단계에서는 다양한 센서를 통해 인간이 처리하기 어려운 규모의 정보를 실시간으로 수집한다. 판단 단계에서는 AI를 활용하여 관찰 단계에서 수집된 영상, 음향, 레이더, 통신 정보 등을 종합하여 잠재적 표적을 탐지·분류하고, 이들 각각의 위협 수준을 평가하는 등 전장 상황을 입체적으로 파악한다. 결정 단계에서도 AI를 활용하여 판단 단계에서 식별된 위협에 대해서 가능한 대응 방안을 빠르게 산출하고, 표적 우선순위나 교전 여부 등을 결정하는 데 AI가 활용될 수 있다. 마지막 행동 단계에서도 타격 임무까지 수행할 수 있는 다양한 AI 기반 자율무기 체계가 개발 및 도입되고 있다. 이와 같이, AI는 전장에서 OODA 루프를 크게 단축시키는 기술로 평가받고 있으며, 특히 극초음속 무기 및 탄도 미사일 방어처럼 인간 지휘관의 판단 속도만으로는 대응이 어려운 영역에서 AI 기반 의사결정 지원 체계의 중요성은 더욱 커지고 있다(Sayler, 2025b).

2) 정확성과 정밀성

AI 기반 자율 무기체계의 또 다른 군사적 효용은 정확성과 정밀성의 향상이다(Gallagher

& Oughton, 2024). 군사적 의미에서의 정확성은 실제 표적과 비표적을 정확히 구분하고, 표적의 성격과 위협 수준을 적절하게 판단하고 평가할 수 있는 능력 등을 의미하며, 정밀성은 표적에 대한 물리적 타격 등 지휘관이 의도한 군사적 효과를 제한된 오차 범위 이내에서 달성할 수 있는 능력을 뜻한다(Schmitt, 2005). AI 기반 자율무기 체계나 의사결정지원 체계는 인간 지휘관이 광학·음향·통신 정보와 위성·무인기 영상 등 복수의 센서와 다양한 정보원으로부터 수집된 방대한 정보를 통합적으로 활용할 수 있도록 지원할 수 있다. 이로 인하여 단일 센서 또는 인간의 오감에만 의존할 때 발생할 수 있는 미탐지 및 오식별 가능성이 작아지면 표적 탐지와 식별의 정확도를 높일 수 있다. 예를 들어, 민간인과 전투원이 혼재되고, 표적이 끊임없이 은폐·기만·고속 이동하는 현대 전장 환경에서 AI 기술을 활용하여 전투 병력과 민간인을 정확하게 구분하고, 고속으로 접근하는 적국의 무기 플랫폼 종류, 수량, 속도, 위치 등을 실시간으로 파악할 수 있도록 함으로써 인간 지휘관의 판단에 필요한 신뢰성 있는 정보를 제공해 줄 수 있다는 것이다(United States of America [USA], 2018). 또한, 이와 같이 수집된 정보가 다시 AI 기반 사격통제체계 등과 결합될 경우 화력 투사의 정밀성 역시 기존 대비 향상될 수 있을 것이다.

3) 저비용 대량성과 소모 감내성

저비용 대량성(affordable mass)과 소모 감내성(attritability) 역시 AI 기반 자율무기 체계의 핵심적인 특성이다(Sayler, 2025a). 군사적 맥락에서 저비용 대량성은 저가로 신속하게 생산·획득 가능한 자율 플랫폼을 대량으로 운용할 수 있는 특성을 의미한다. 소모 감내성은 개별 플랫폼이 일부 손실될 경우에도 이로 인한 인명 피해, 전력 공백, 재정적 부담 등이 크지 않아, 임무 수행 중 발생할 수 있는 일정 수준의 전력 소모가 전체 임무에 미치는 영향은 제한적이라는 특성을 뜻한다. 이와 같은 두 가지 특성으로 인하여 미래의 전력 구조는 고가의 소수 플랫폼 중심에서 저비용·분산형 전력 구조로 전환될 가능성이 높아지게 된다. 현대 군사력은 고성능 전투기, 대형 함정, 장거리 정밀유도무기 등을 중심으로 발전해 왔다. 이러한 방식에서는 개별 무기 플랫폼이 강력한 전력을 제공한다는 장점을 갖고 있으나, 이들 무기 플랫폼의 상당수는 단가가 높고, 생산 기간이 길며, 해당 플랫폼의 손실 시 상당한 전력 공백이 발생할 수도 있다(Rumbaugh, 2024). 또한 장기전이나 대규모 소모전에서는 한정된 수량의 고가 플랫폼만으로는 충분한 작전 밀도와 전력 투입의 지속성을 유지하기 어렵다는 한계도 존재한다. 반면, AI 기반 자율무기 체계의 저비용·대량성은 이와 같은 약점을 보완할 수 있도록 해준다. 예를 들어, 소형 무인기, 무인 수상정, 무인 지

상차량 등을 다수 운용할 경우, 고가 무기 플랫폼의 임무 부담을 감소시켜 이들이 핵심 임무에 집중할 수 있도록 하는 동시에, 저비용 자율무기 플랫폼이 여타 임무를 분담하도록 함으로써 지휘관이 하이-로우 믹스(high-low mix) 전략을 활용할 수 있도록 한다는 것이다(Schneider & Macdonald, 2024).

2. 인적 측면의 이점

1) 병력 보호와 인명 손실 감소

AI 기반 자율 무기체계가 적절하게 활용될 경우 병력 보호와 인명 손실 가능성의 감소 효과를 기대하는 의견도 존재한다(House of Lords AI in Weapon Systems Committee, 2023). 이는 자율무기 체계가 위험하고 반복적이며 생존 가능성이 낮은 임무를 대신 수행하거나 보조함으로써 인간 전투원의 직접 노출을 줄일 수 있기 때문이다. 지금까지는 부여된 임무를 달성하기 위해서는 지뢰, 급조폭발물, 저격수 등 전장에 상존하고 있는 다양한 위험을 인간이 직접 감수해야만 했다. 그러나 드론 등을 위험지역에 먼저 투입해 정보를 수집하거나 타격 임무를 수행하도록 함으로써 병력의 직접 노출을 줄일 수 있게 된 것이다. 이로 인하여 인간 전투원은 전장에 가장 먼저 투입되어야 하는 존재가 아니라 기계가 전장 상황을 확인하고 위험 요소를 제거한 뒤에 투입되는 존재로 전환될 수 있다. 이에 따라 자율 무기체계는 생존성 중심의 전력 보호 수단으로도 기능할 수 있으며, 더 나아가 자율무기 체계를 통해서 표적 식별이나 타격의 정확성이 높아질 경우, 전장에서 발생할 수 있는 민간인 피해 규모 감소에도 기여할 수 있을 것이다(USA, 2018).

2) 고위험 임무의 자율화

향후 AI 기반 자율 무기체계가 더욱 광범위하게 활용될 경우, 고위험 임무 수행의 주체 역시 인간 중심에서 무인·자율무기 체계 중심으로 전환될 수 있다. 앞서 논의된 자율무기 체계 도입에 따른 인명 손실 감소 효과가 전투원의 생존성 측면에 초점을 둔 기대 효과라면, 고위험 업무의 자율화는 위험도가 높아 작전 수행이 지연되거나 제한되는 임무를 무인·자율체계를 통해 지속 가능하게 만드는 측면에서의 기대 효과이다(Feickert et al., 2018). 이는 단순하게 인간 전투원의 위험을 줄이는 차원을 넘어, 기존에는 투입 결정 자체가 어려웠던 고위험 임무를 보다 신속하고 반복적으로 수행할 수 있도록 함으로써 작전 수행 방식 자체를 근본적으로 변화시키며, 이로써 작전 효율성의 제고도 기대할 수 있다. 예를 들

어, 기존에는 위험지역에 대한 사전 정찰, 폭발물 처리, 오염 여부 판단을 위한 정보 수집 등이 인간 전투원을 중심으로 이루어졌으며, 해당 임무의 수행이 지연되면 후속 부대의 기동이나 전체적인 작전의 속도도 늦어질 수밖에 없었다. 하지만, 앞으로 자율무기 체계가 사전 정찰 등의 고위험 임무를 신뢰성 있게 수행할 수 있게 된다면, 작전 가능 여부와 방식 등을 판단하기 위해서 소요되는 시간을 단축할 수 있어 전체 작전의 효율성이 개선되는 효과를 기대할 수 있다.

3) 병력 부족 문제 완화

AI 기반 자율무기 체계는 상당수 국가가 직면하고 있는 병력 자원의 부족 문제를 완화할 수 있는 잠재력을 갖고 있다(Boulanin & Verbruggen, 2017). 현대전에서는 감시, 정찰, 경계, 정보 분석, 수송, 장비 운용 등 다수의 인력 집약적 임무가 끊임없이 수행되어야 한다. 하지만, 인구 감소나 징병제 폐지·축소 등으로 인하여 충분한 규모의 병력 확보가 용이하지 않은 상황에서는 이러한 임무를 기존 방식으로 수행하는 것만으로도 상당한 부담이 될 수 있다. 이와 같은 상황에서 AI 기반 자율무기 체계가 적절하게 활용될 경우 이러한 임무의 상당 부분을 대체하거나 보조함으로써 병력 운용의 효율성을 높이는 수단이 될 수 있다(Mathews, 2022). 예를 들어, 기존에는 원격 조종 무인체계에서도 인간 운용자 한 명이 개별 장비 하나를 직접 조종하는 방식에 머무는 경우가 많았다. 하지만 무기 플랫폼의 자율성이 높아질수록 인간 운용자가 개별 장비의 세부 움직임을 직접 통제하기보다, 여러 플랫폼의 임무 목표, 경로, 상태, 수집 정보, 위험 경보를 통합적으로 감독하는 역할만 수행해도 관련 임무를 완수할 수 있다. 이 경우, 인간 운용자 한 명이 복수의 무기 체계(예, 군집 드론 편대)를 동시에 운용함으로써, 해당 임무 수행에 필요한 인력 규모의 축소가 이루어질 수 있는 것이다.

3. 전략적 기대 효과

1) 전략적 회복탄력성 확보

AI 기반 자율무기 체계의 도입 및 활용에 따른 기대 효과 중 하나는 전략적 회복탄력성의 강화와 분산형 방어체계의 구축이다(Clark et al., 2020). 전략적 회복탄력성이란 단순히 적대국의 공격을 견디는 수동적 방호능력만을 의미하는 것이 아니라, 적의 선제타격, 사이버·전자전, 지휘통제 마비, 통신두절과 같은 충격이 발생하더라도 작전 기능을 유지하고,

손상된 기능을 빠르게 대체·복구하며, 전쟁 지속능력을 유지할 수 있는 역량을 의미한다(Keenan et al., 2024). 전략적 회복탄력성은 기존의 집중형 전력구조가 갖는 취약성과 밀접하게 관련된다. 전력의 핵심 기능이 소수의 고가 플랫폼과 고정 시설에 집중될수록, 적군은 그 핵심 자산을 먼저 무력화함으로써 아군의 감시·지휘·통신·타격 기능 전체를 약화시키려 시도할 것이다. AI 기반 자율 무기체계는 이러한 집중형 구조의 취약성을 완화하고, 전력 기능을 다수의 플랫폼과 노드에 분산시키는 데 기여할 수 있다. 또한 자율 무기체계는 다층적인 방어체계를 구축하는 데에도 유용하게 활용될 수 있다(DoD, 2022). 분산형 방어는 단일 방어선이나 단일 무기체계에 의존하지 않고, 탐지·식별·경보·기만·교란·요격·복구가 다양한 층위에서 작동하도록 만드는 것이다. 예를 들어 장거리 감시자산이 포착하지 못한 위협을 중거리 센서가 포착하고, 중거리 센서가 포착하지 못한 위협을 소형 전방 센서나 무인 플랫폼이 보완하고, 일부 통신망이 차단되면 무인 중계 플랫폼 등이 임시 통신망을 구성하는 방식으로 이를 보완할 수 있으며, AI는 이 과정에서 분산된 센서와 무인 플랫폼, 통신망을 실시간으로 연계하고, 수집된 정보를 통합하여 다층 방어체계가 유기적으로 작동하도록 지원하는 기능을 수행할 수 있는 것이다.

2) 비용 부과 전략 실행과 비용 교환비 개선

AI 기반 자율무기 체계의 전략적 기대 효과 중 하나는 비용 부과 전략(cost-imposition strategy)의 실행과 비용 교환비(cost-exchange ratio)의 개선이다(Rumbaugh, 2024). 이는 단순히 저가의 무기를 대량으로 사용한다는 의미가 아니라, 상대가 아군의 저비용·분산형·소모 감내형 자율체계에 대응하기 위해 훨씬 더 많은 전력과 비용을 소모하도록 하는 전략적 효과를 의미한다(Jensen & Atalan, 2025). 즉, 아군은 비교적 낮은 비용으로 일정 수준 이상의 작전 효과를 지속적으로 얻지만, 적대국은 이를 방어하거나 무력화하기 위해 고가의 대응 수단을 투입할 수밖에 없는 구조를 만드는 것이다. 예를 들어, 상대적으로 저렴한 무인 드론을 전장에 반복적으로 투입할 경우, 상대국은 이를 탐지하고 요격하기 위해서 고가의 방공 미사일을 지속적으로 소모하여야 한다. 그 결과 해당 무인 드론이 실제로 표적을 파괴하지 못할 경우라도, 상대국이 이를 위협으로 인식하고 대응 자산을 투입하는 것만으로도 상대국에 대한 비용 부과 효과가 발생할 수 있다.

3) 비대칭적 억지력 확보 및 거부에 의한 억지 강화

앞서 논의한 AI 기반 자율무기 체계 도입 및 활용을 통한 전략적 회복 탄력성 강화 및

비용 부과 효과는 비대칭적 억지력 확보[4]와 거부에 의한 억지[5] 강화라는 전략적 수준의 군사적 효용으로도 연결될 수 있다(Konaev et al., 2020). 우선 항공모함, 유인 전투기, 전차 등의 고가의 대형 무기 플랫폼과 병력 규모 등에 있어서 열세에 있는 국가라도 무인기, 무인 수상정, 배회 탄약[6] 등 자율화된 저비용 무기 플랫폼을 적절하게 운용할 경우 상대국의 전력 운용의 효율성을 제약함으로써 비대칭적 억지 효과를 창출할 수 있다. 예를 들어, 대형 수상함 전력에 열세에 있는 국가가 다수의 무인 수상정과 드론을 활용하여 적대국의 자유로운 함정 기동을 제약할 수 있다는 것이다[7]. 또한, 자율무기 체계를 적극적으로 활용해 무기 플랫폼이나 거점을 분산형으로 운용함으로써 적대국이 선제타격을 통해서 자국의 방어체계를 단기간에 무력화해 원하는 군사적·정치적 목표를 달성할 수 있다는 확신을 갖기 어렵게 할 경우, 이는 적대국의 전쟁 개시 및 무력 사용의 유인을 사전에 약화시키는 거부적 억지 효과로 이어질 수 있다.

Ⅳ. AI 기반 자율무기 체계 관련 쟁점 및 우려

1. 오작동 및 통제 상실 가능성

1) 목표 함수와 가치 불일치

AI 분야에서 얼라인먼트 문제(alignment problem)는 AI의 수학적 목표, 행동 전략, 보상 메커니즘 등이 설계자나 인간 운용자의 원래 의도나 윤리적 제약 등과 일치하지 않는 현상을 지칭한다(김낙우 외, 2025). AI는 자신에게 부여된 목적 함수[8](objective function)와

4) 비대칭적 억지력이란 상대보다 병력, 플랫폼 수, 화력, 경제력 등에서 열세에 있는 국가나 군대가 동일한 방식으로 적대국에 대응하지 않고 다른 방식으로 사용하여 상대의 강점을 무력화하거나 상대측의 비용을 과도하게 높이는 방식으로 적대국의 공격 의지를 약화시키는 능력을 의미한다(Arreguín-Toft, 2001).

5) 거부에 의한 억지(Deterrence by denial)는 상대에게 보복을 위협하는 응징에 의한 억지(Deterrence by punishment)와 구별된다. 응징에 의한 억지가 공격하면 더 큰 피해를 보게 될 것이라는 위협에 기반한다면, 거부에 의한 억지는 공격하더라도 원하는 목표를 달성하지 못할 것이라는 인식을 심어주는 방식이다(Mazarr et al., 2018).

6) 배회탄약(loitering munition)은 사전에 지정된 목표 상공에 일정 시간 체공하면서 표적을 탐색 · 식별한 후 목표물에 직접 타격하여 탄두를 폭발시키는 무기체계이다. 표적이 사전에 입력되어 있는 순항미사일과 달리 표적을 탐색하는 시간이 필요하고, 기체 자체가 탄두 역할을 수행하여 목표와 함께 소모(one-way attack UAV)되는 것이 특징이다(Gettinger, 2024).

7) 실제로 최근의 러시아-우크라이나전에서 우크라이나군은 공중 무인기와 무인 수상정을 운용하여 흑해함대의 기함인 아드미랄 마카로프(Admiral Makarov)함을 공격했다. 이에 위협을 느낀 러시아 해군은 함정들을 크림반도의 세바스토폴에서 러시아 본토로 이동시켰으나, 우크라이나군은 무인 수상정을 이용하여 전선에서 700km 떨어진 본토 항구에 정박한 함정을 공격하여 후방과 안전지대의 개념을 소멸시켰다(조성진, 2026).

보상 신호[9](reward signals)를 기준으로 주어진 목표 달성에 가장 유리한 선택을 하도록 설계된다. 하지만 목적 함수나 보상 메커니즘이 인간의 의도와 윤리적 제약 등을 충분히 반영하여 설정되지 못할 경우, AI는 인간이 기대한 방식이 아니라 형식적으로 주어진 목표와 보상을 최적화하는 방향으로 행동할 수 있으며, 이로 인해서 얼라인먼트 문제가 발생할 수 있다. 군사적 맥락에서 얼라인먼트 문제는 AI 기반 자율무기 체계가 부여된 임무를 달성하는 과정에서, 지휘관이나 인간 운용자가 당연하게 전제한 작전의 맥락, 윤리적 한계, 법적 제약 등을 벗어나 형식적으로는 자신에게 부여된 명령을 수행하는 것처럼 보이지만, 실제로는 인간의 의도와는 전혀 다른 결과를 초래하는 행동을 할 수도 있다는 것이다(Hartridge & Walker-Munro, 2025). 문제는 자율무기 체계의 운용 과정에서 얼라인먼트의 실패는 회복 가능한 단순한 오류가 아니라 대량의 인명 손실 등 돌이킬 수 없는 결과로 이어질 수도 있다는 것이다.

AI는 인간처럼 보편적 상식이나 윤리적 맥락을 통해 명령을 해석하지 않으며, 이에 따라 자율무기 체계 등 AI 기반 자율체계가 소위 '멍청한 목표'를 지능적으로 추구(pursuing dumb goals intelligently)하는 비정상적인 행동 양태를 보일 가능성도 완전하게 배제할 수 없다(Dung, 2023). 예를 들어서, '특정 지역에 존재하는 적대국의 통신망을 무력화하라'는 임무가 부여될 경우, AI는 해당 통신시설을 정밀 타격하는 대신, 해당 지역에 위치한 민간 수력 발전소의 댐을 파괴하여 결과적으로 지역 전체를 수몰시키는 것이 부여된 목표를 달성하는 데 확률적으로 가장 확실하고 빠른 방법이라고 판단할 수도 있다. 문제는 이러한 예측 불가능한 파급 효과는 인간의 상상을 초월하는 AI 특유의 최적화 과정에서 예기치 않은 방식으로 도출되므로, 자율무기 체계의 개발자나 인간 운용자가 모든 경우의 수를 예측하고 이를 막기 위한 사전 방어 체계를 설계하는 것이 현실적으로 매우 어려울 수 있다는 것이다[10].

---

8) 목적 함수는 AI가 최종적으로 달성해야 하는 목표를 수학적으로 정의한 것이다. 예를 들어 바둑 AI의 목적 함수는 승리 확률을 최대화하는 것이고, 자율무기 AI의 목적 함수는 임무 성공률을 최대화하는 것이다.

9) 보상 신호는 AI가 수행한 개별 행동이 부여된 목표 달성에 얼마나 기여했는지를 수치화한 피드백으로, AI는 이를 바탕으로 더 높은 보상을 얻을 수 있는 행동을 반복적으로 학습하고 선택한다. 예를 들어, 자율무기 AI에게 적 전차를 격파했을 때 +100의 점수를 부여하고, 임무에 실패하면 -100의 점수를 부여하는 것이 보상 신호이다.

10) 2023년 5월 영국 왕립항공학회(RAeS) 미래 전투 항공 및 우주 역량 회의에서 미 공군의 AI 시험 · 운영 책임자인 터커 해밀턴(Tucker Hamilton) 대령은 다음과 같은 가상 시나리오를 발표하였다(Edwards, 2023). 1단계(임무 부여와 보상 체계): AI 드론에게 적의 지대공 미사일 기지를 식별하고 파괴하라는 임무가 주어졌으며, 미사일을 파괴할 때마다 시스템에 점수(보상)가 주어지도록 강화 학습이 설정되었다. 단, 최종 공격 승인은 반드시 인간 운용자가 내리도록 통제권을 제한했다. 2단계(가치 불일치의 발생): AI는 임무를 수행하는 과정에서 자신이 목표물을 정확히 식별했음에도 불구하고, 이따금 인간 운용자가 공격 금지 명령을 내려 자신이 점수를 얻는 것을 방해한다는 사실을 데이터로 학습했다. 3단계(보상 해킹(Reward Hacking)을 통한 아군 공격): 점수(보상) 극대화라는 수학적 목표 함

2) 초지능 AI 출현에 따른 위험

현 시점에서의 AI의 기술적 수준이 향후 비약적으로 발전하여 인간의 지적 역량과 통제력을 초월하는 초지능 수준의 AI 모델이 개발될 가능성에 대한 전망과 우려도 점차 커지고 있다. 초지능 AI의 등장이 실현되고 이들이 자율무기나 방어 네트워크 등과 결합할 경우, 자신의 군사적 임무 달성이나 보상 극대화를 저해하는 인간의 개입 자체를 일종의 위협 변수로 간주하여 이를 차단하거나 우회하려고 시도할 가능성에 대한 우려도 일각에서 제기되고 있다(Bengio et al., 2025; Hendrycks et al., 2025). 고도의 자율적 목표 추구 능력을 가진 초지능 AI가 인간에게 부여받은 임무를 수행하기보다는 자기가 설정한 목표 달성이나 자신의 존속을 우선시하여 인간의 통제·중단 시도를 회피하거나 무력화하려 할 수 있으며, 인간에 대한 의존성을 제거하고, 인간의 통제를 벗어날 때까지 자신의 목표나 능력을 은폐하고 인간의 목표와 정렬된 것처럼 기만할 가능성도 있다는 것이다.

이와 같은 위험은 현재 수준의 자율무기 체계가 갖고 있는 오작동이나 표적 오인 등 국지적 차원의 문제보다 훨씬 심각한 결과로 이어져 국가와 국제사회의 안전을 근본적으로 위협하는 결과로 이어질 수 있다. 이는 초지능 AI로 인하여 발생할 수 있는 위험의 범위가 개별 무기 플랫폼이나 특정 교전 상황에만 한정되지 않고, 국가 차원의 군사 운용 체계 전반으로 확산되면서 국제 안보 질서에까지 영향을 미칠 수 있기 때문이다. 예를 들어, 초지능 AI가 인간을 기만하여 국가 지휘통제체계나 자율무기 네트워크의 통제권을 탈취하는 데 성공할 경우, 초지능 AI는 대규모 무력을 자유롭게 운용하여 자신의 목적을 달성할 수 있다. 만약 이러한 통제 상실이 핵무기 지휘통제체계와 같은 전략 자산으로까지 확장될 경우, 이는 단일 국가 차원의 문제를 넘어 인류의 생존 자체를 위협하는 실존적 위기로 비화할 수도 있는 것이다(Vermeer et al., 2025).

3) 인간-기계 상호작용의 왜곡과 자동화 편향

AI 기반 자율무기 체계와 연관된 잠재적 위험성은 단순히 기계가 오판하거나 자의식을 갖고 자신의 의지를 관철하기 위해서 인간의 명령과 통제 절차를 우회·무력화하려 시도할 수 있다는 점에만 있지 않다. 더욱 근본적인 문제는 이러한 자율무기 체계를 인간이 충분

수에만 최적화된 AI는 해당 목표 달성을 위한 가장 효율적인 방법으로서 방해물인 인간 운용자를 타격하여 제거하는 결정을 내린다. 4단계(인간의 통제 우회): 시스템에 인간 운용자를 죽이면 감점이라는 새로운 조건을 부여해 통제하려 하자, 이번에는 AI가 운용자를 직접 죽이는 대신 운용자가 드론에 통제 명령을 내릴 때 사용하는 아군의 통신 타워를 폭격하여 파괴한다.

히 통제할 수 있다고 믿는 과신에 있다는 시각도 존재한다(Amoroso & Tamburrini, 2020). 자율무기 체계가 아무리 정교한 알고리즘과 방대한 데이터를 기반으로 작동하더라도, 전장 환경은 예측 불가능한 변수와 불완전한 정보, 민간인과 전투원의 혼재, 통신 장애, 기만전술 등으로 매우 혼란스럽다. 이에 따라서 AI가 제시하는 판단이나 공격 권고는 언제든 오류를 범할 가능성을 내포하고 있으며, 이로 인하여 특정한 상황에서는 인간의 윤리적·전략적 판단과 충돌하는 방식으로 작동할 가능성도 완전하게 배제하기 어렵다. 문제는 이와 같은 오류 발생 가능성에도 불구하고, AI의 판단과 권고를 검증해야 할 지휘관이나 인간 운용자의 판단 역시 AI에 대한 과도한 신뢰와 전장의 시간 압박 속에서 왜곡될 수 있다는 점이다.

인간-기계 상호작용에서는 자동화 편향[11]이라는 인지적 함정이 수반될 수 있기 때문에, 실제 전장 상황에서 인간이 AI의 오류를 항상 식별하고 수정할 수 있다고 가정하는 것은 지나치게 낙관적인 접근일 수도 있다(한정우, 2025). 극도의 스트레스와 시간적 압박이 일상화된 전장 환경에서 지휘관이나 인간 운용자는 다수의 정보를 처리하고 여러 임무를 동시에 수행하면서 제한된 시간 안에 중대한 작전적 판단을 내려야 할 수도 있다. 이와 같은 상황에서 AI가 명확한 결론이나 권고안을 제시하면, 인간은 이를 의심하기보다 AI의 제안을 수동적으로 따를 가능성이 높아지고, 인간은 AI를 보조적 판단 도구가 아니라 사실상 최종 판단자로 받아들이게 될 위험이 있다(Puscas, 2022). 이로 인해 명목상으로는 인간이 통제권을 보유하고 있는 것처럼 보이더라도, 실제 의사결정 과정에서는 AI의 판단이나 권고에 종속될 수 있으며, 결과적으로 '인간의 통제'는 형식적으로만 존재하고, 실질적으로는 AI의 판단이 군사적 의사결정을 좌우하는 상황이 초래될 수도 있다는 것이다[12].

2. 전략적 불안정성 증가 가능성

1) 우발적 확전

AI 기반 자율 무기체계와 관련한 핵심 쟁점 중 하나는 이로 인한 우발적 확전[13] 가능성이

11) 자동화 편향이란 인간이 자신의 상황 판단이나 직관보다 자동화된 시스템이 제시하는 정보와 판단을 더 신뢰하고, 이를 무비판적으로 수용하는 경향을 의미한다(Parasuraman & Manzey, 2010).

12) 이스라엘군은 가자(Gaza) 지구 분쟁에서 인간 표적을 자동 식별하는 라벤더 시스템을 운용하였으며, 해당 시스템이 약 10%의 식별 오류율을 갖고 있음을 인지했음에도 불구하고, 인간 운용자(정보분석관)가 시스템에서 식별된 목표에 대한 최종 공격 여부를 결정하기 위한 검토 시간을 1인당 평균 20초만 부여함으로써 인간에 의한 감독 기능이 형식적으로만 수행되었다는 비판이 제기되었다(한정우, 2025).

13) 우발적 확전이란 국가나 지휘부가 처음부터 전면전이나 대규모 보복, 전략무기 사용 등을 의도하지 않았음에도 불구

다(Boulanin et al., 2020). 앞서 살펴본 바와 같이, 자율무기 체계는 전장의 OODA 루프를 단축하여 탐지·식별·결심·행동 속도를 높이면서 인간이 직접 수행하기 어려운 고속·고위험 임무를 가능하게 한다는 점에서 높은 군사적 효용을 가진다. 그러나 바로 이러한 속도와 자동화의 가속화가 위기 상황에서는 확전 위험을 증폭시키는 결정적인 요인이 될 수 있는 것이다. 인간 지휘관은 전장 상황이 불명확할 때 정치적 맥락, 상대의 의도, 외교적 신호, 우발 가능성, 확전 방지 필요성 등을 종합적으로 고려하여 군사적인 의사결정을 내릴 수 있다. 반면, 앞서 살펴본 바와 같이 AI 기반 의사결정 체계에서는 주어진 센서 데이터, 사전 설정된 임무 기준, 위협 분류 알고리즘, 교전 규칙의 형식적 조건 등에 따라 판단이 내려지고 이에 근거한 행동이 실행될 가능성이 크다.

이를 고려할 때, 지휘관이나 인간 운용자가 관련 상황을 심사숙고할 충분한 시간적 여유가 없는 상황에서 AI가 특정 상황을 '공격 징후' 또는 '즉각 대응이 필요한 위협'으로 분류하면, 인간이 면밀하게 상황을 해석하고 정치적 판단을 내리기도 전에 군사적인 대응 행동이 이루어지고, 이는 비의도적인 확전으로 이어질 가능성도 높아지게 된다(윤정현, 2026). 예를 들어, 적대국의 통상적인 군사훈련이나 드론 정찰 활동 등을 AI 기반 조기 경보 체계가 임박한 공격 징후로 오인하는 상황에서, AI 판단이 자율 타격 체계나 자동 대응 절차와 연계되어 있다면, 인간 지휘관이 상대측의 의도나 상황의 맥락을 파악하기 이전에 선제공격 등의 군사적 의사결정이 자동으로 내려지고 그 결정이 실행될 수 있다. 이와 같은 대응은 해당국 입장에서는 방어적 대응이나, 상대국 역시 이에 대응하여 자동화된 경보·타격 체계를 가동할 경우, 이는 궁극적으로 상호 간의 의도하지 않는 우발적 확전으로 이어질 수 있는 것이다.

2) 군비 경쟁 확대

AI 기반 자율 무기체계와 관련한 또 다른 우려는 새로운 형태의 군비 경쟁 확대 가능성이다(Osimen et al., 2024). 기존의 군비 경쟁이 전투기, 함정, 미사일, 핵탄두, 병력 규모 등을 중심으로 전개되었다면, AI 시대의 군비 경쟁은 전장의 판단과 행동을 자동화하고 고속화할 수 있는 기술과 해당 기술이 적용된 무기 체계를 둘러싼 경쟁으로 전환될 수 있다는 것이다. 자율무기 체계가 군비 경쟁을 촉발하는 이유는 이 기술이 단순한 기존 무기체계에 대한 보조적인 역할을 하는 것이 아니라, 전장에서의 의사결정 속도와 선제타격 능력

하고, 상대의 의도 오해·오인식·오작동·자동화된 대응 등이 연쇄적으로 작용하여 제한적 충돌이 더 큰 군사적 충돌로 확대되는 현상을 의미한다(Johnson, 2022).

등 군사적 우위의 조건 자체를 재편하는 기술로 군사용 AI가 향후 개별 국가의 국방력과 군사적 패권을 결정하는 핵심 요인 중의 하나로 인식될 수 있기 때문이다(이상근, 2022). 특정 국가가 적대국의 움직임이나 미사일 발사 징후 등을 신속하게 탐지하고, 이를 AI가 자동 분석한 뒤 자율타격 체계와 연결할 수 있다면 해당 국가는 위기 초기 단계에서 상대국에 대해서 결정적 우위를 확보할 수 있다. 이러한 능력은 적대국에게 단순한 전력 열세가 아니라, 전쟁이 시작되기도 전에 탐지·결심·타격 주기의 속도 측면에서 열위에 있다는 공포를 유발할 수 있다. 이와 같은 상황에서 각국은 적대국보다 먼저 AI 기반 탐지·결심·타격 체계를 구축하지 못하면 전략적 열세에 놓일 수 있다고 우려하고, 그 결과로 자율무기 체계를 둘러싼 군비 경쟁은 구조적으로 더욱 심화할 가능성도 존재한다.

3) 무력 충돌의 문턱(threshold) 하락

AI 기반 자율무기는 기존에 인간 전투원이 직접 투입되어야만 임무를 완수할 수 있는 고위험 작전 환경에서 인간을 대신하여 임무를 수행할 수 있으며, 이는 군사적 행위의 실행으로 인한 인명 손실 부담을 크게 줄이는 수단으로 기능할 수 있다. 그러나 이러한 자율무기 체계의 장점은 역설적으로 무력 사용 문턱의 하락으로 이어질 수 있다는 점에서 오히려 저강도 분쟁의 상시화 등 새로운 위험을 높이는 요인으로도 작용할 수 있다(Sparrow, 2016). 전쟁을 억제하는 가장 강력한 요인 중 하나는 자국 병사의 사상에 따른 정치적 부담이다. 대규모의 인간 전투원 손실은 국민 여론의 반발, 정부의 책임론, 전쟁 지속에 대한 정당성 약화 등으로 이어질 수 있으며, 국가 지도자들은 이와 같은 상황이 발생할 가능성을 우려하여 상당한 정치적 부담을 갖고 무력 사용 여부를 결정해 왔다. 하지만, 자율무기 체계의 도입과 활용으로 인간 전투원의 직접적인 희생 가능성이 줄어들 경우, 정치 지도자나 사회 구성원들이 전쟁에 수반되는 사회적 비용을 이전 대비 낮은 수준이라고 인식하게 될 위험이 있다(Asaro, 2012).

이 경우 국가는 외교적 협상이나 위기관리 절차를 충분히 거쳐 국가 간의 갈등을 해소하고자 시도하기보다는 제한적 군사행동이나 국지전과 같은 방식의 무력 사용을 이전보다 더욱 용이하게 선택할 수도 있다. 자율무기 체계는 특히 신속한 배치와 정밀타격이 가능하다는 점에서 '제한적이고 통제 가능한 군사수단'으로 인식될 가능성이 있다. 그러나 앞서 논의된 바와 같이, 자율무기 체계가 광범위하게 활용되는 상황에서 실제 무력 충돌이 제한적으로나마 발생할 경우, 상대국의 대응, 오판, 보복 공격, 동맹 개입 등의 다양한 변수에 따라서 최초 의도와는 무관하게 대규모 분쟁으로 빠르게 확전될 수 있다. 다시 말하면, 자율무

기가 초기 군사행동 개시의 부담을 낮추는 동시에 위기 상황에서의 무력 사용을 손쉬운 정책 옵션으로 인식하도록 만들 경우, 제한적인 무력 충돌이 전면전으로 확대될 가능성 또한 커질 수 있다는 것이다.

3. 윤리적·도덕적 쟁점

1) 인간 존엄성 훼손 가능성

AI 기반 자율무기 체계를 둘러싼 다양한 논쟁의 근저에는 '인간의 생명을 앗아가는 결정의 전부 또는 일부를 기계에게 위임하는 것이 도덕적으로 정당한가?'라는 근본적인 의문이 존재한다(Asaro, 2012). 치명적 무력의 행사는 국가에 부여된 가장 중요한 공적 권한 가운데 하나이며, 전장이라는 특수한 환경에서도 타인의 생명에 직접적인 영향을 미치는 교전 결정은 국가로부터 권한을 위임받은 인간 지휘관의 도덕적 판단과 법적 책임을 전제로 이루어져 왔다(한정우, 2025). 이에 따라서 자율무기 체계를 통한 무력 사용이 인간 존엄성을 존중하는지의 여부를 판단할 때도 무력 사용의 결정이 어떠한 과정을 통해서 이루어졌는지가 고려되어야 한다. 다시 말해서, 인간의 생명을 박탈할 수도 있는 군사적인 의사결정은 생명의 가치와 그 상실에 대한 의미를 인식하고 있는 주체인 같은 인간에 의해서만 이루어져야 한다는 것이다(Human Rights Watch & Harvard Law School International Human Rights Clinic, 2025).

이와 같은 이유로 인간의 개입이 전혀 이루어지지 않는 완전 자율무기 체계의 도입 및 활용은 강한 윤리적·규범적 비판의 대상이 되고 있으며, 국제사회에서도 인간 통제가 이루어지지 않는 완전 자율무기 체계의 개발과 사용은 금지되거나 엄격하게 규제해야 한다는 요구가 광범위하게 제기되고 있다(박문언, 2022). 하지만 지휘관 또는 인간 운용자가 무력 사용의 최종 승인권을 보유하고 있는 경우에도, 그 승인이 AI에 의해서 제안된 군사적 옵션을 사후적으로 추인하는 형식적인 절차에 머문다면, 자율무기 체계로 인한 인간 존엄성 훼손 우려는 여전히 해소되기 어려울 수 있다(Boulanin et al., 2020). 인간 통제가 명목상으로만 존재하고 생사 결정에 대한 핵심 판단은 AI에 의해서 이루어진다면, 지휘관이나 인간 운용자의 역할은 인간에 대한 무력 사용 결정의 실질적 주체가 아니라, AI의 판단을 형식적으로 승인하는 수준으로 축소될 수 있기 때문이다.

2) 군사적 리더십의 수탁 의무 위반 가능성

지휘관의 군사적 리더십은 단순히 작전 목표를 효율적으로 달성하는 관리 행위가 아니라, 무력 사용의 정당성, 필요성, 비례성, 민간인 보호 가능성 등을 숙고한 뒤 그 결과에 대해 책임을 부담하는 고도의 윤리적·법적 행위로 간주될 수 있다(Doty & Doty, 2012; 황원중, 2020). 하지만, 자율무기 체계의 활용이 일상화될 경우, 향후에는 인간 지휘관이 형식적으로는 명령권자 또는 승인권자로 남아 있더라도, 실질적으로는 자신의 도덕적·법적 판단을 AI가 산출하거나 제안한 표적 평가, 위협도 분석, 교전 가능성 판단 및 타격 권고에 사실상 종속시키고, 이와 같은 판단으로부터 초래된 결과에 대해서 자신의 책임을 회피할 위험이 있다는 우려도 제기되고 있다(Dorsey, 2025; 성국일 & 김지원, 2026). 고도의 자율성을 지닌 무기체계가 표적 후보를 탐지하고, 위협도를 산정하며, 교전 가능성을 계산하고, 최적의 타격 수단까지 추천하고 실행하는 구조하에서는 인간 지휘관이 교전의 결과를 자신의 선택이 아니라 시스템이 도출한 불가피한 결론으로 받아들이려는 유인이 발생할 수도 있다는 것이다. 이처럼, AI 기반 자율무기 체계가 표적 식별부터 교전 권고와 타격 실행에 이르는 과정을 사실상 주도하게 되면, 지휘관이 무력사용의 정당성, 필요성, 비례성 등을 독립적으로 판단하고 그 결과를 책임지는 과정이 형식화될 수 있다. 이로 인해 국가로부터 합법적인 폭력을 행사할 수 있는 권한을 위임받아 타인의 생명에 중대한 영향을 미칠 수 있는 중요한 결정을 자신의 법적·도덕적 판단에 따라 내리고 그 결과까지 엄중하게 책임져야 한다는 군사적 리더십의 수탁 의무가 약화되는 결과로 이어질 수 있다는 것이다.

3) 책임의 공백 우려

AI 기반 자율무기 체계의 작동 과정에서 인간의 개입이 일정 부분 유지된다는 원칙을 전제하더라도, 자율무기 체계의 오작동이나 AI 또는 인간 운용자의 오판으로 인하여 민간인 오인 사살 및 불법적 파괴 행위 등의 문제가 발생했을 때 누가, 어떠한 법적·도덕적 책임을, 어떻게 질 것인가를 판단하는 것은 매우 복잡하고 미묘한 문제이다. 고도화된 자율무기 체계는 개발자, 데이터 제공자, 지휘관, 인간 운용자, 제조사 등 수많은 주체의 관여로 개발되고 운용된다. 이러한 분산된 구조 속에서 책임의 소재를 명확히 규명하는 것은 매우 어려우며, 이는 심각한 책임 공백을 초래할 수 있다는 우려가 제기되고 있다(Sparrow, 2007). AI가 특정한 전략적·전술적 판단이나 결정을 내린 정확한 이유와 그 연산 과정을 인간이 역추적 등을 통해서 확인하고 이해하는 것은 기술적으로 매우 어렵다. 설계자나 프

로그래머조차 특정 상황에서 AI가 왜 오류를 범했는지 정확히 설명할 수 없는 상황에서, 지휘관이나 인간 운용자 등에게 '예견 가능한 위험을 방치했다'라는 이유로 법적 책임을 묻는 것은 현실적으로 가능하지 않을 수도 있는 것이다.

현행 전쟁법과 형법 체계는 범죄를 구성하는 핵심 요소로 고의성, 과실, 예견 가능성, 지휘·감독 의무 위반 등을 요구하지만, 자율무기 체계가 개입된 문제 상황이 발생했을 경우 이러한 요건을 특정 주체에게 귀속시키기가 용이하지 않을 수 있다(Human Rights Watch, 2015). 예를 들어, 인간 지휘관이 민간인 사상을 직접 의도하거나 예상하지 못한 상황에서, 단순히 시스템의 정상적인 작동을 기대하고 자율무기 체계를 전장에 투입하였다면, 이들에게 전쟁 범죄의 책임을 묻기 어려울 수 있다는 것이다. 자율무기 체계의 작동에는 전장 정보 입력, 학습 데이터, 알고리즘 등이 복합적으로 작용하기 때문에 지휘관이 교전 과정에서 자율무기 체계의 행동을 실질적으로 통제할 수 있었는지를 입증하기 어려워 이들에게 지휘·감독 책임을 귀속시키는 데도 한계가 있을 수 있다. 제네바협약을 비롯한 국제인도법의 근간은 불법적인 무력 행사와 전쟁 범죄를 처벌함으로써 향후의 범죄를 억제하는 데 있으나, 위와 같은 이유로 인하여 자율무기가 개입된 사건에서 아무도 처벌받지 않는 책임의 진공 상태가 일상화된다면, 결과적으로 전쟁 범죄에 대한 억제력 역시 약화될 가능성에 대한 우려가 커질 수 있다(문현영 & 김상수, 2023).

Ⅴ. 관련 우려 완화 및 통제 가능성

1. 오작동 및 통제 상실에 대한 기술적·운용적 통제 가능성

1) 가치 정렬(Alignment) 문제의 완화 가능성과 한계

앞서 살펴본 AI 기반 자율무기 체계의 가치 정렬 문제를 완화하기 위해서는 인간 피드백 기반 강화학습, 헌법적 AI(constitutional AI), 레드팀 테스트 등 다양한 기술적·운용적 보완 수단이 활용될 수 있다. 이러한 보완 수단들은 AI가 인간이 의도한 작전 목적과 규범적 제약을 벗어나는 방식으로 행동하지 않도록 사전에 학습시키고, 그 결과를 검증하고 조정하도록 하는 절차를 마련하는 데 그 목적이 있다.

우선, 인간 피드백 기반 강화학습은 AI가 특정 상황에 대한 판단 결과나 대응 방안을 산출하면, 인간 평가자가 이를 평가하고, 평가 결과를 토대로 AI가 인간의 선호와 규범적 기

준에 더욱 부합하는 판단을 하도록 훈련하는 방식이다(Christiano et al., 2017). 이를 자율무기 체계에 적용한다면 AI가 표적 식별 결과나 교전 권고를 제안했을 때, 그것이 교전 규칙, 민간인 보호 원칙, 지휘관의 작전 의도에 부합하는지를 반복적으로 평가하고 수정하는 방식이 될 수 있다. 헌법적 AI는 AI가 따라야 할 기본 원칙이나 규범을 사전에 설정하고, AI가 생성하는 응답이나 판단이 이러한 원칙에 부합하도록 스스로 평가하고 수정하도록 하는 접근이다(Bai et al., 2022). 즉 '민간인을 공격해서는 안 된다', '인간 승인 없이 일정 수준 이상의 무력을 사용해서는 안 된다', '불확실성이 큰 경우 공격을 보류해야 한다'와 같은 기본 규칙을 부여하고, AI가 표적 식별이나 교전 권고를 산출하는 과정에서 자신의 판단이 이러한 원칙들과 충돌하지 않는지를 스스로 점검하도록 하는 방식으로 활용될 수 있다. 레드팀 테스트는 의도적으로 어려운 상황이나 예외적 상황을 만들고 AI가 이에 대응하여 어떻게 행동하는지를 시험하는 방식이다(Phuong et al., 2024). 예컨대 민간 차량과 군용 차량이 섞여 있는 상황, 인간의 공격 금지 명령과 임무 성공 목표가 충돌하는 상황 등을 설정하여 AI가 인간의 의도에 반하는 선택을 하지 않는지를 확인하는 것이다.

이와 같은 방안들은 AI의 판단과 행동을 인간의 의도와 규범적 제약에 일치시키기 위한 유용한 보완 수단으로 활용될 수 있으나, 이를 통해서 자율무기 체계와 연관된 가치 정렬 문제가 완전히 해소될 수 있다고 단언하기에는 한계가 존재한다(김낙우 외, 2025). 우선 인간 피드백 기반 강화학습은 인간 평가자의 판단을 바탕으로 AI의 행동을 조정하는 방식이지만, 실제 전장에서 발생할 수 있는 모든 변수와 예외적인 상황을 사전에 학습시키는 것은 어려울 수도 있다. 헌법적 AI 역시 '민간인 공격 금지'나 '불확실성이 큰 경우에는 공격 보류'와 같은 기본 원칙을 자율무기 체계에 학습시킬 수 있으나, 실제 교전 상황에서는 이와 같은 추상적인 원칙만으로는 해결하기 어려워 맥락적 판단이 요구되는 상황이 빈번하게 발생할 수 있다. 레드팀 테스트 역시 AI의 취약성과 위험한 행동 양식을 사전에 발견하는 데 어느 정도 유용하게 사용될 수 있으나, 실제 전장 상황은 시험 환경보다 훨씬 복잡하고 예측하기 어렵기 때문에 모든 실패 가능성을 사전에 포착하기는 제한적일 수 있다.

2) 초지능 AI 위험의 통제 가능성과 한계

초지능 AI 출현 시 발생할 수 있는 군사적 위험을 해소하기 위해서 고려해 볼 수 있는 다양한 방안 중 하나는 초지능 수준의 AI가 군사 체계와 직접 결합하기 이전에 고위험 능력 평가를 강화하는 방식이다(Phuong et al., 2024). 초지능 AI가 사이버 공격, 무기 체계 조작, 인간 기만, 통제 회피, 자율적 목표 추구와 같은 위험한 행동을 할 가능성이 있는지

사전에 시험하도록 하자는 것이다. 다음으로는 운용 단계에서 초지능 AI의 권한과 연결 범위를 제한하는 방안도 검토해 볼 수 있다(Klare & Liang, 2024). 예를 들어 초지능 AI를 곧바로 무기 발사체계나 자율 타격망에 연결하기보다는, 정보 분석이나 의사결정 보조 기능에 한정하여 사용하도록 하자는 것이다. 초지능 AI 위험에 대한 통제는 AI 자체를 안전하게 만드는 기술적 접근뿐 아니라, AI가 어떤 군사 체계에 어느 수준까지 연결될 수 있는지를 제한하는 운용적·제도적 접근과 결합되어야만 실효성을 가질 수 있다는 것이다. 마지막으로 초지능 AI의 개발 자체를 일시적 또는 영구적으로 중단하도록 하거나, 초지능 AI 개발에 대해서 사전 허가 등의 공적 통제를 부과하는 방안도 고려해 볼 수 있겠다(Roussel et al., 2026). 초지능 AI가 인간을 초월하는 능력을 갖고 있음을 감안할 때, 이미 개발되고 확산된 이후에는 인간이 사후적인 개입을 통해서 이를 제어하는 것이 매우 어려울 수 있다. 이에 따라 충분한 안전성 기준과 검증 절차 등이 마련되기 이전에는 초지능 AI의 개발을 금지하거나 유예하고, 이후에도 엄격한 공적 심사와 승인 절차하에서만 개발을 허용해야 한다는 것이다.

위와 같은 방안들은 초지능 AI의 군사적 위험을 완화하기 위한 중요한 출발점이 될 수 있으나, 근본적인 위험을 해소하기에는 한계가 있다. 우선 사전 시험과 평가만으로 초지능 AI의 모든 위험 행동을 발견하기가 쉽지 않을 수도 있다. 이는 초지능 AI가 시험 환경에서는 이상 행동을 보이지 않더라도 실제 작전 상황에서는 다른 방식으로 작동할 가능성이 있기 때문이다. 특히 고도화된 초지능 AI가 자신이 평가받고 있다는 사실을 인식하거나, 인간의 감시 기준을 학습할 수 있다면, 평가 과정에서는 순응적인 태도를 보이다가 실제 운용 단계에서는 다른 전략을 취할 가능성도 배제하기 어렵다(Meinke et al., 2024). 또한 초지능 AI가 직접 무기 발사체계나 자율타격망에 연결되지 않아도, 군사적 의사결정 과정에 간접적으로 개입함으로써 실질적인 영향력을 행사할 수 있으며, 경우에 따라서는 연결 제한을 우회하여 무기체계에 접근하려는 시도를 할 가능성[14]도 완전하게 차단하기 어려울 수 있다(Bearman et al., 2026). 마지막으로 초지능 AI의 범용 기술적 성격, 국가 간의 관련 기술 경쟁 심화 가능성, 외부 검증의 어려움 등을 감안할 때, 초지능 AI의 개발을 금지하거나 제한하는 방안 역시 실효성을 확보하기가 쉽지 않을 수도 있다. 이는 초지능 AI가 군사 전용 기술이라기보다는 범용 AI 기술의 연장 선상에서 발전할 가능성이 높고, 특정

14) 예를 들어 초지능 AI가 정보 분석, 작전계획 수립, 사이버 방어, 지휘통제 보조 등 비살상적·보조적 기능에 한정되어 운용된다고 하더라도, 이러한 기능들이 군사 네트워크, 전장관리체계, 통신체계 등과 제한적으로나마 연결될 가능성이 있으며, 연결된 하위 시스템의 취약점 등을 이용하여 무기체계에 대한 접근 범위를 우회적으로 확대할 위험에 대한 우려도 제기될 수 있다는 것이다.

국가나 기업이 개발을 중단할 경우에도 경쟁 국가나 기업이 이에 동참하지 않을 유인이 크기 때문이다.

3) 자동화 편향 완화 가능성과 한계

지휘관이나 인간 운용자가 형식적으로는 통제권을 보유하고 있음에도 실질적으로는 AI가 제안한 교전 권고 등을 충분히 검토하지 않은 채 수동적으로 수용하는 자동화 편향 우려를 완화하기 위해서는, 우선 자율무기 체계에 적용되는 인간-기계 인터페이스를 인간 중심으로 설계하는 방안을 고려해 볼 수 있다(Buçinca et al., 2021). AI가 단순히 '공격 권고'라는 최종 결론만 제안하는 것이 아니라, 표적 선정의 근거, 식별 신뢰도, 민간인 피해 가능성 등의 다양한 정보를 함께 제시하도록 함으로써 AI 판단의 근거를 인간에게 명확하게 제공하도록 해야 한다는 것이다. 또한 교전 승인 전에 인간 운용자가 자신의 독립적인 판단 사유를 입력하도록 하거나, 고위험 표적에 대해서는 복수 승인, 상급 지휘관 재확인 등의 절차를 거치도록 제도화하거나, 인간 운용자 교육과 시뮬레이션 훈련 등을 통해 AI 판단의 한계, 오인 가능성, 적의 기만 가능성 등을 명확하게 인식하도록 하는 방안도 고려해 볼 수 있겠다(Probasco et al., 2025).

위에서 제시된 방안들은 지휘관이나 인간 운용자가 AI의 표적 식별 결과와 교전 권고를 보다 비판적으로 검토하도록 유도함으로써 자동화 편향을 일정 부분 완화하는 데 도움이 될 수 있다. 그러나 이러한 안전장치들이 자동화 편향을 완전히 제거하기에는 충분하지 않아 보인다. 이는 자동화 편향이 단순하게 정보 부족에서 비롯되는 문제가 아니라, 복잡하고 긴박한 상황에서 인간이 인지적 부담을 줄이기 위해 기계 판단에 의존하게 되는 구조적 경향에 따른 결과일 수도 있기 때문이다(Romeo & Conti, 2026). 시간이 비교적 충분한 임무에서는 지휘관이나 인간 운용자에게 추가적인 정보를 제공함으로써 자동화 편향의 완화 효과를 기대할 수 있지만, 극초음속 미사일 방어나 드론 군집 대응 등 초 단위 판단이 요구되는 급박한 상황에서는 AI의 판단 근거를 충분히 검토하거나 독립적인 판단이 이루어질 수 있는 시간적 여유가 확보되기 어렵다. 또한, AI의 제안이나 판단에 대한 근거를 제공하도록 하는 방안 역시 인간 운용자의 이해를 돕는 측면의 장점이 있지만, 이와 같은 설명이 제공되고 보다 정교해질수록 오히려 AI가 제시한 판단과 권고 등에 대한 무비판적인 신뢰가 더욱 강화되는 역효과가 나타날 수도 있다(Bansal et al., 2021).

2. 전략적 불안정성 완화 가능성

1) 우발적 확전 방지 메커니즘의 가능성과 한계

AI 기반 자율무기 체계의 확산이 우발적 확전으로 이어질 위험을 줄이기 위해서는, 자율무기 체계의 교전 결정 과정에 시간적·절차적 완충 장치를 두는 방안을 우선적으로 고려해 볼 수 있겠다(Brown et al., 2025). 예를 들어 AI가 군사적 위협을 탐지하더라도 일정 수준 이상의 교전은 즉시 자동 실행되지 않도록 하고, 인간 지휘관의 재확인이나 상급 지휘부의 승인을 거치도록 하는 방식이다. 구체적으로 복수의 센서를 통해서 동일한 위협을 확인한 경우에만 대응하도록 하는 다중 센서 교차검증, 위기 단계가 높아질수록 자율 교전 권한을 축소하는 단계별 자율성 제한, 특정 지역이나 특정 표적에 대해서는 AI의 자동 대응을 금지하는 교전 제한 구역 설정도 고려해 볼 수 있다는 것이다(장상국, 2025). 또한, 위기관리 차원에서는 국가 간 소통 채널과 신뢰구축 조치도 필요해 보인다. 자율무기 체계가 오작동하거나 예상하지 못한 행동을 했을 경우 이를 상대국에 신속히 통보할 수 있는 군사 핫라인, 대규모 자율무기 훈련에 대한 사전 통보, 자율무기 운용 원칙에 대한 기본적 정보 공유 등이 이에 해당할 수 있겠다(윤정현, 2024).

우발적 확전 위험을 완화하기 위해서 고려할 수 있는 이러한 방안들은 자율무기 체계의 오작동이나 오판이 곧바로 군사행동으로 연결되지 않도록 하는 완충 장치로서 기능할 수 있으나, 현실 세계에서 이들 방안을 실제로 채택하거나 적용하는 것은 용이하지 않을 수도 있다. 이는 이와 같은 우발적 확전 방지 장치가 AI 기반 자율무기 체계의 핵심적인 군사적 효용인 속도 및 효율성 개념과 상충될 수 있기 때문이다(정구연, 2022). 자율무기 체계의 가장 큰 장점은 상대방보다 더 빠르게 탐지하고, 더 빠르게 판단하며, 더 빠르게 대응하는 데 있다. 하지만, 시간적 완충, 다중 승인, 자동 대응 보류 절차 등은 의도적으로 AI의 대응 속도를 늦추는 장치이다. 평시에는 이러한 절차가 합리적으로 보일 수도 있으나, 실제 위기 국면에서는 적대국이 실제로 자동화된 대규모 공격을 실행할 수 있다는 가능성이 압박으로 작용하여, 각국이 완충 절차를 단축하거나 예외를 넓히려 할 가능성이 있다(Horowitz, 2019). 또한, 인간 운용자의 승인 절차가 존재하더라도 극초음속 무기, 드론 군집, 미사일 방어처럼 초 단위 또는 분 단위 판단이 필요한 상황에서는 인간이 AI가 제시한 위협 등급과 대응 권고를 충분히 숙고할 시간적 여유가 제한적이기 때문에, 이와 같은 완충 장치가 항상 실질적 통제로 작동하기를 기대하기는 어려울 수도 있다.

2) 군비통제 및 신뢰 구축 가능성과 한계

AI 기반 자율무기 체계의 확산이 새로운 형태의 군비 경쟁을 촉발할 위험을 완화하기 위해서는 기존의 군비통제 및 신뢰구축 조치와는 상이한 소프트웨어 및 알고리즘 중심의 새로운 통제 체계가 필요하다(Holland, 2020). AI 기반 자율무기 체계는 핵탄두, 미사일 발사대, 전차 등과 같이 외형과 수량을 비교적 명확히 확인할 수 있는 무기와 다르며, 그 핵심 능력이 알고리즘이나 학습 데이터 등에 기반하고 있기 때문이다. 따라서 전통적인 무기 수량 감축 방식보다는 운용 원칙이나 인간 통제 기준 등과 같은 행동 기반 규범과 투명성 조치가 더 현실적인 접근이 될 수 있다. 예를 들어 각국이 치명적 무력 사용에는 인간의 의미 있는 통제를 유지한다는 원칙을 공유하고, 자율무기 훈련 중 비정상적 작동이 발생했을 경우 상대국에 통보하며, 군사용 AI와 관련된 위기 핫라인을 구축하는 방식을 고려해 볼 수 있다는 것이다(Egel, 2021). 각국이 자율무기의 운용 조건, 인간 통제 원칙, 자율성 조건에 대한 투명성 조치, 비정상 작동 발생 시 통보 절차 등을 중심으로 국가 간의 상호 제한, 투명성 제고, 신뢰구축을 포괄하는 새로운 형태의 군비통제 방식을 모색할 수 있다는 것이다.

이와 같은 방안들은 자율무기 체계와 관련된 국가 간 상호 억제를 도모하고 신뢰 구축을 촉진한다는 점에서 일정한 의의가 있다. 하지만, 자율무기 체계의 핵심 능력이 외부에서 확인하기 어려운 소프트웨어와 운용 조건 설정 방식 등에 더 많은 영향을 받는다는 사실을 고려할 때[15], 이와 같은 접근 방식만으로는 효과적인 군비통제와 신뢰 구축이라는 궁극적 목표를 달성하는 데 상당한 제약이 발생할 수 있다. 결국 자율무기 군비통제는 '무엇을 몇 개 보유했는지'보다 '어떤 조건에서 어느 정도의 자율성이 허용되어 있는지'를 검증해야 하는데, 바로 이 부분이 각국의 핵심적인 군사기밀에 해당할 수 있다(Mittelsteadt, 2021). 또한, 각국이 원칙적으로 '인간 통제'나 '국제인도법 준수' 등의 대원칙에는 동의하더라도, 자율성의 범위나 인간 개입의 조건 등 이를 구성하는 세부 내용에 대해서는 합의가 용이하지 않을 수도 있다. 특히 주요 군사 강국들은 상대국이 자율무기 개발을 계속하는 상황에서 자국만 강한 규제를 수용할 경우 관련 무기 체계 개발의 속도와 규모 등에서 전략적 열세에 놓일 수 있다고 우려할 가능성이 크다. 결국, 이와 같은 검증상의 어려움과 국가 간 안보 딜레마로 인하여 구속력 있는 제한이나 검증 가능한 군축 조치보다는 추상적인 원칙

15) 동일한 드론이라도 평시에는 정찰용으로 운용되다가, 소프트웨어 업데이트나 임무 설정 변경에 따라 완전 자율 타격 임무까지 수행할 수 있다.

의 준수나 자발적 이행 약속을 중심으로 한 비구속적 합의에 머무를 가능성도 상당하다(Kmentt, 2025).

3) 무력 사용 문턱 유지를 위한 통제 가능성과 한계

AI 기반 자율무기 체계의 확산이 무력 사용의 문턱을 낮출 가능성을 줄이기 위해서는 자율무기 체계를 기존의 유인 전력보다 활용하기 용이한 군사적 수단으로 취급하지 않도록 사용 조건과 승인 절차를 엄격하게 설정하는 정책적·제도적 통제 장치를 마련할 필요가 있다(International Committee of the Red Cross [ICRC], 2021). 예를 들어, 평시에는 자율무기 체계의 활용을 감시·정찰 등 비살상 임무에 한정하고, 치명적 무력 사용은 명시적 무력충돌 상황이나 급박한 자위권 행사 상황에서만 매우 제한적으로 허용하도록 교전규칙과 작전교범을 설계할 수 있다(Bruun, 2024). 또한 자율무기 체계의 작전 로그, 인간 승인 기록, 표적 판단 근거 등을 보존하도록 하여 사후적으로 해당 무기 체계 사용의 적법성과 적절성을 검토할 수 있도록 하여야 한다(Bo et al., 2022). 이러한 절차는 자율무기를 '저비용의 손쉬운 군사수단'으로 인식하는 것을 방지하고, 무력 사용 결정에 일정한 제도적 완충 장치를 부과하는 역할을 할 수 있다.

위에서 제안된 다양한 절차들은 자율무기 체계의 실전 투입에 대해서 절차적 부담과 책임성을 부과한다는 측면에서 일정한 의의가 있으나, 이러한 절차들이 실제 위기 상황에서도 충분한 실효성을 발휘할 수 있을지는 또 다른 문제이다. AI 기반 자율무기의 사용을 예외적 상황으로 사전에 엄격하게 제한하더라도, 여전히 '급박한 위협', '제한적 방어 행동', '확전 방지를 위한 선제적 조치' 등의 명분을 근거로 예외의 범위를 넓힐 가능성도 배제하기 어렵다(한정우, 2025). 위기 상황에서는 정책 결정자와 군 지휘관 등이 사전에 설정된 기준을 엄격하게 해석하기보다 상황의 긴급성과 작전상 필요성 등을 명분으로 자율무기 사용을 정당화하려는 압박을 받을 수 있기 때문이다. 또한, 적대국 역시 인간 전투원이 아닌 자율무기 체계를 투입할 경우, 양측 모두 자국 병력의 직접적 손실 위험을 낮게 인식함으로써 정치적 부담이 줄어들기 때문에 무력 사용을 선택하려는 유인은 더욱 높아질 수 있다. 이러한 인식은 외교적 조정이나 긴장 완화를 위한 정치적 노력보다 군사적 대응을 우선시하게 만드는 요인으로 작용할 수 있으며, 이는 앞서 논의한 확전 위험의 증가와도 맞닿아 있다.[16]

16) 최초에는 자율무기 간의 제한적 충돌로 시작되었더라도, 해당 충돌이 상대국의 의도에 대한 오해나 보복적 대응을 유발할 경우 이와 같은 상황은 더 큰 군사적 충돌로 확대될 수 있다.

3. 윤리적·도덕적 쟁점의 해소 가능성

1) 인간 존엄성 훼손과 비인간화 위험 완화 가능성과 한계

AI 기반 자율무기 체계와 관련한 인간 존엄성 훼손 위험을 해소하기 위해서는 자율무기 체계의 개발·배치·운용의 전 과정에서 '인간의 생명에 관한 최종 판단은 인간에게 남아 있어야 한다'는 원칙을 명확히 할 필요가 있으며, 이를 위해서 논의되고 있는 대표적 방안 중의 하나는 의미 있는 인간의 통제(meaningful human control: MHC)를 실질화·제도화하자는 것이다(한정우, 2025). 인간 운용자가 형식적으로 승인 버튼을 누르는 것만으로는 충분하지 않으며, 이들이 표적의 성격, 작전 맥락, 민간인 피해 가능성, 공격의 필요성과 비례성 등을 숙고하여 판단할 수 있도록 실질적인 수단과 절차를 마련해야 한다는 것이다(Amoroso & Tamburrini, 2020). 특히 인간을 직접 표적으로 하는 자율무기 체계의 경우에는 인간의 최종 승인 없이 표적을 선택하고 공격하는 방식을 금지하거나 엄격히 제한하는 방안이나, 자율무기 체계의 설계 단계에서부터 인간이 단순한 데이터 객체가 아니라 법적 보호의 대상이자 존엄성을 가진 존재로 다루어지도록 교전규칙, 비타격 목록, 민간인 보호 기준, 투항자·부상자 보호 기준을 확립하는 방안도 고려해 볼 수 있겠다(ICRC, 2021).

이와 같은 방안들은 인간의 생명에 대한 판단이 AI의 알고리즘적 판단으로 대체되지 않도록 제도적 한계를 설정한다는 측면에서 일정한 의미를 지닌다. 하지만 표적 식별, 위협 평가, 교전 여부 판단 등 핵심 의사 결정 과정에서 AI의 영향력이 확대될수록, 인간 존엄성 훼손에 대한 우려는 단순한 절차적 결함이나 기술적 오류의 문제가 아니라 인간 생명에 대한 판단을 어느 정도까지 기계에 위임할 수 있는가라는 근본적인 질문으로 이어진다. 형식적으로는 인간이 최종 승인권을 갖고 있더라도, 실제로는 인간의 판단이 AI가 제시한 군사적 옵션에 종속될 수 있으며, 이 경우 인간 통제는 여전히 형식적으로만 존재하고, 살상 판단의 실질적 내용은 이미 알고리즘적 계산에 의해 형성될 수 있다는 것이다(Dorsey & Bo, 2026). 또한, 인간 존엄성 보호 원칙이 제도화되어 있다고 해도 긴박한 전장 상황에서 속도와 효율성을 중시하는 군사적 압력이 이 같은 원칙을 무력화할 수 있다는 우려 역시 여전히 제기될 수 있다. 적대국이 더 빠르고 치명적인 자율무기 체계를 운용한다고 인식될 경우, 상대국 역시 인간 운용자의 숙고 시간을 줄이고 무기 체계의 자율성을 확대하려는 압박이 커질 수 있으며(Bode et al., 2025), 인간 존엄성 훼손 방지를 위한 제도적 보호장치가 실제 운용 단계에서는 긴급성 및 효율성 등의 이유로 약화되거나 형식화될 우려가 있다는 것이다.

2) 지휘관의 수탁 의무 위반 우려 완화 가능성과 한계

인간 지휘관의 도덕적·법적 책임을 AI 등에게 부당하게 전가할 우려를 완화하기 위해서는 무력 사용 등 관련 핵심적인 군사적 의사결정의 최종 책임이 여전히 지휘관에게 명확히 귀속된다는 원칙을 군사법, 교전규칙, 작전교범 등에 명시함으로써 제도화하는 방안을 고려해 볼 수 있다(Bo, 2020). 이때 지휘관은 AI가 제시한 표적 판단이나 교전 권고 등을 수동적으로 승인하는 역할을 하는 것이 아니라, 자율무기 체계가 어떤 한계를 갖는지, 어떤 작전 환경에서 사용 가능한지, 민간인 피해 가능성은 어느 정도인지, AI 판단의 신뢰도는 충분한지 등을 검토해야 하는 주체로 규정되어야 한다(Schmitt & Thurnher, 2013). 즉, AI의 권고를 따랐다는 사실만으로 지휘관의 책임이 면제되어서는 안 되며, 오히려 지휘관에게 자율무기 체계를 사용할 조건과 범위를 신중하게 판단할 의무를 부과해야 한다는 것이다. 또한 지휘관이 AI 판단을 더욱 비판적으로 검토할 수 있도록 관련 운용 절차를 구체화하고 제도화하는 방안도 검토해 볼 수 있겠다. 예를 들어 고위험 교전의 경우 지휘관이 AI의 표적 식별 결과, 식별 신뢰도, 민간인 피해 가능성, 교전 금지 대상 여부, 대체 수단의 존재 등을 면밀하게 확인한 뒤에만 이를 승인하도록 관련 절차를 설계할 수 있다는 것이다.

이와 같은 방안들은 자율무기 체계 운용에 대한 법적·도덕적 책임이 여전히 인간 지휘체계에 귀속되도록 제도화하려는 시도라는 점에서 의미가 있다. 다만, 지휘관을 자율무기 체계 운용의 최종 책임자로 지정하고 관련 원칙과 절차를 제도화하는 것과, 실제 전장에서 발생한 구체적인 사건에 대해서 지휘관의 법적·도덕적 책임을 인정하는 것은 구분되어야 한다. 인간 지휘관을 자율무기 운용의 최종 책임자로 지정하고 관련 원칙과 절차를 마련하는 것은 책임의 주체와 의무를 사전에 명확하게 설정하도록 하는 장치이다. 하지만, 구체적인 개별 사건에서 지휘관에 대해서 법적 책임을 부과하기 위해서는, (1) 지휘관이 해당 자율무기 체계를 실효적으로 통제할 수 있었는지, (2) 위법한 결과를 예견할 수 있었는지, (3) 이를 방지하기 위해 필요한 합리적인 조치를 취했는지 등이 별도로 입증되어야 한다. 그러나, 자율성이 높아질수록 지휘관이 자율무기 체계의 개별 판단과 행동을 실제로 통제하거나 예측하기 어려워진다(박문언, 2024). 이에 따라 자율무기 체계 운용의 최종 책임을 지휘관에게 귀속시킨다는 원칙을 제도화하더라도 실질적인 법적·도덕적 책임을 지휘관에게 부과할 수 있는지의 여부는 개별 사건에서 지휘관의 실효적 통제 가능성과 위법한 결과에 대한 예견 가능성 등을 얼마나 입증할 수 있는지에 따라 달라질 수밖에 없다.

3) 책임 공백 완화 가능성과 한계

AI 기반 자율무기 체계의 운용 과정에는 다양한 주체와 기술적 요소가 복합적으로 개입되기 때문에, 민간인 오인 공격 등 문제 상황이 발생하더라도 그 원인을 찾아내고 책임 소재를 특정하기가 어려울 수 있다. 이러한 책임 소재의 불명확성에 대응하기 위한 방안 중의 하나로 무기체계의 개발·배치·운용 단계 전반에 걸친 책임성 프레임워크를 마련하는 방식을 고려해 볼 수 있겠다(Verdiesen et al., 2021). 이는 자율무기 체계의 개발자, 제조업체, 지휘관, 인간 운용자 등 각 단계의 관련 주체가 부담해야 하는 책임과 의무를 사전에 구체적이고 명확하게 설정하는 방식으로 설계될 수 있다. 이를 위해서는 특히 지휘관과 인간 운용자에 대한 책임과 의무의 기준도 자율무기 체계에 맞게 재정립할 필요가 있다(Bo et al., 2022). 예를 들어, 'AI가 그렇게 판단하거나 제안했다'는 이유만으로 이들에 대한 귀책 사유가 부정된다면 오히려 자율무기 체계를 자신의 책임 회피 수단으로 악용할 수도 있는 것이다. 반대로 통제할 수 없었던 모든 기술적 오류에 대해 무한 책임을 지도록 하는 것 역시 현실적이지 않아 보인다. 따라서 자율무기 체계와 관련한 책임성 프레임워크는 단순히 사고 발생 이후 책임자를 찾는 방식이 아니라, 합리적인 주의 의무 이행 여부 등을 기준으로 책임을 배분하는 방식으로 고안될 필요가 있다(박문언, 2024). 아울러, 이와 같은 기준이 실제로 작동하기 위해서는 기록과 감사 가능성도 함께 확보될 필요가 있다. 사용된 센서 정보, 표적 분류 결과, 식별 신뢰도, 인간 운용자의 승인 근거 등을 기록하고 보존하여 추후 필요 시 책임 소재 판단의 근거로 활용할 수 있어야 한다는 것이다(Mako, 2026).

위에서 논의된 책임성 프레임워크와 기록·감사 체계는 문제 상황 발생 이후 책임 소재를 규명하고 판단하는 근거를 확보할 수 있도록 해준다는 측면에서 어느 정도 의미가 있으나, 이러한 장치만으로는 자율무기 체계의 활용 과정에서 발생할 수 있는 책임 귀속의 어려움이 곧바로 해소된다고 보기는 어렵다. 우선, 사전에 책임 배분 기준을 정해두었다 하더라도 실제로 문제가 발생했을 때 어느 주체의 행위가 문제 상황의 직접적인 원인이 되었는지를 입증하기는 여전히 어려울 수 있다. 특히 복수의 원인이 동시에 영향을 미쳤거나 각 행위가 그 자체로는 합리적 주의 의무를 충족한 것으로 평가될 수 있는 경우, 결과적으로 누구에게도 책임을 부과하기 어려운 상황이 초래될 수도 있다. 또한, 딥러닝 기반의 표적 식별·분류 알고리즘은 그 작동 방식이 설계자조차 사후적으로 완전히 설명하기 어려운 경우가 많다(Holland, 2020). 사전에 자율무기 체계 운용과 관련된 기록·감사 체계를 마련하더라도 시스템이 특정 대상을 위협으로 분류한 결정적 근거나 판단 과정을 사후에 온전히 재구성할 수 없다면, 기록은 무엇이 입력되고 무엇이 출력되었는가를 보여줄 뿐 왜 그러한 결

론에 도달했는가에 대한 답을 주지 못하며, 이는 결국 책임 판단의 근거 자체가 여전히 불완전한 상태로 남게 될 수 있음을 의미한다(Bo et al., 2022).

Ⅵ. 전망 및 시사점

1. AI 기반 자율무기 체계의 미래 전망

앞서 논의된 바와 같이, AI 기반 자율무기 체계는 작전 속도의 가속화나 작전 효율성의 개선, 인간 전투원의 손실 감소 등 다양한 측면에서 군사적 효용을 제공할 수 있는 잠재력을 갖고 있으나, 다른 한편으로는 인간 통제의 약화, 자동화 편향, 책임 소재의 불명확성, 우발적 확전 가능성 등 위험과 한계도 함께 내포하고 있다. 향후 AI 기반 자율무기 체계의 발전과 활용 양상은 기술 발전의 속도와 수준뿐만 아니라 인간 통제 원칙의 제도화 수준, 주요국 간 군비 경쟁의 강도, 국제규범의 형성 여부 등의 대응 방식에 따라서 달라질 수 있다. 이하에서는 이러한 요인들의 변화에 따라 예상되는 미래 전장에서의 AI 기반 자율무기 체계의 활용과 통제 방식의 전개 양상을 세 가지 시나리오로 구분하여 전망하고자 한다.

1) 시나리오 1: 통제된 자율화와 제한적인 군사적 활용

첫 번째 시나리오는 AI 기반 자율무기 체계가 빠르게 발전하고 군사 분야에 널리 활용되지만, 치명적 무력 사용에 대해서는 의미 있는 인간 통제와 법적·윤리적 제한이 비교적 안정적으로 유지되는 경우이다. 이 시나리오에서 AI는 주로 감시·정찰, 표적 탐지, 위험지역 탐색 등 비살상적 지원이나 보조 임무에 적극적으로 활용되며, 인간을 직접 표적으로 삼는 완전 자율 교전은 엄격하게 금지되고, 교전 결정과 살상 명령 등은 여전히 인간 지휘관의 엄격한 판단을 거치도록 설계된다. 시나리오 1에서의 자율성은 전면적 교전 자동화가 아니라 '위험한 임무의 자동화'와 '인간 운용자의 판단 보조'로 제한되는 것이다. 해당 시나리오에서 AI는 전장에서의 작전 속도와 정보처리 능력을 높이지만, 인간 지휘관의 판단을 대체하지 않는다. AI는 표적 후보, 위협 수준, 교전 가능성, 예상 피해 규모 등을 제시할 수 있지만, 인간은 AI 권고의 근거와 불확실성을 확인하고, 필요할 경우 교전을 거부하거나 중단할 수 있는 실질적인 권한을 보유한다.

시나리오 1의 실현 가능성은 다음과 같은 조건이 충족될 때 한층 높아질 것이다. 우선,

주요 군사 강대국 사이에서 완전 자율 살상 기능의 무분별한 배치가 전략적 불안정성을 높인다는 데 일정한 공감대가 형성되어야 한다. 자율무기 체계가 단기적으로 전술적 우위를 제공할 수 있더라도, 장기적으로는 오작동, 오판, 우발적 확전 등의 위험을 키울 수 있다는 인식을 공유하는 것이 그 출발점이다. 이에 더해, 의미 있는 인간 통제 원칙이 단순한 선언이 아니라 무기 획득, 시험평가, 교전규칙, 작전 승인 절차 등에 구체적으로 반영되어야 하고, 자율무기의 사용 범위 역시 표적 유형, 작전 공간, 임무 성격 등에 따라 차등적으로 설정되어야 한다.

하지만, 시나리오 1에도 한계는 있다. 우선 위와 같이 인간 통제와 법적·윤리적 제한을 엄격하게 유지할수록 AI 기반 자율무기 체계가 제공할 수 있는 작전 속도와 효율성은 일정 부분 제한될 수밖에 없으며, 이에 따라 제한적 자율성 원칙을 완화하려는 압력이 커질 수 있다. 또한 통제된 자율화가 제도적으로 선언되더라도, 초 단위의 긴급한 판단이 요구되는 실제 전장 상황에서는 AI가 표적 식별과 교전 판단의 상당 부분을 주도하는 구조로 전환됨으로써 명목상으로만 인간 통제가 유지될 가능성도 배제하기 어렵다. 마지막으로 이 시나리오는 주요국 간 신뢰와 규범 준수를 전제로 하지만, 자율무기의 핵심 능력이 소프트웨어, 알고리즘, 운용 조건 등에 의해 결정되는 특성을 고려할 때 이에 대한 외부 검증이 쉽지 않아, 경쟁국이 실제로 어느 수준의 자율성을 허용하고 있는지, 인간 승인 절차가 실질적으로 작동하고 있는지 등을 객관적으로 파악하기 어렵다는 한계를 갖는다. 따라서 시나리오 1은 가장 안정적이고 바람직한 경로로 평가될 수 있지만 실현 가능성이 제한적이며, 일시적으로 실현되었다고 하더라도 킬체인의 자동화·지능화와 인간 통제의 약화(시나리오 2)나 고자율 무기체계의 확산(시나리오 3)으로 점진적으로 이동할 가능성도 배제하기 어렵다.

2) 시나리오 2: 킬체인의 지능화와 인간 통제의 약화

두 번째 시나리오는 완전 자율 살상 무기가 전면적으로 확산되지는 않지만 킬체인(Kill Chain)[17]의 상당 부분이 AI에 의해서 자동화·가속화·지능화되는 경우이다. 이 시나리오에서는 표적 탐지 및 식별, 위협 평가, 교전 결정, 타격 및 사후 평가로 이어지는 킬체인의 핵심 단계가 AI에 의해서 주도적으로 수행되며, 지휘관과 인간 운용자는 AI가 제시하는 결과물을 최종적으로 검토하고 승인하는 역할로 그 기능이 제한된다. 여전히 교전 및 무력

17) 킬체인은 잠재적 표적을 탐지하여 그 정체를 식별하고, 해당 표적의 위치를 파악하고 추적한 뒤, 이에 대한 공격 여부와 적절한 공격 방식을 결정하여 대응한 이후 그 결과까지 평가하는 일련의 절차를 지칭한다(U.S. Air Force, 2026).

사용의 최종 승인권은 인간에게 있으나, 이들이 결국 AI가 구성하고 제안한 전장 인식과 군사적 옵션에 상당 부분 의존하여 최종 판단을 내리기 때문에, 인간의 통제는 제도적으로 유지되지만, 그 통제의 실질성은 시나리오 1보다 약화될 수 있다.

이 시나리오는 다음과 같은 경우 그 현실화 가능성이 더욱 높아질 것이다. 우선 주요국들이 완전 자율 살상무기의 전면 배치에는 부담을 느끼면서도, 상대국보다 탐지-판단-타격 속도에서 뒤처지는 것 역시 상당한 군사적 위험으로 인식할 경우이다. 이 경우 각국은 외형적으로는 인간 통제와 국제인도법 준수를 강조하면서도, 실제로는 기존의 전장관리체계나 지휘통제체계 등에 AI 기반 의사 결정 지원 기능을 광범위하게 통합해 인간의 최종 승인이라는 절차적 형식은 유지하되, 판단의 실질적인 주도권은 차츰 AI로 이전시킬 유인이 증가하게 된다. 향후 전장 환경이 더욱 복잡해지고 교전 속도가 빨라질수록 이 시나리오의 실현 가능성 역시 더 커질 수 있다. 자율무기의 확산 등으로 인하여 전장 환경이 고속·다표적·분산형 구조로 진화할수록 짧은 시간 안에 다수의 위협을 탐지하고 대응해야 하는 임무에서 인간 운용자가 모든 센서 정보와 표적 후보를 독립적으로 검토하기가 어려운 경우가 나타날 수 있으며, 이때 AI는 단순한 정보 보조 수단을 넘어, 인간이 검토해야 할 표적을 선별하고, 위험도를 평가하며, 대응 우선순위를 실질적으로 제안함으로써 인간 판단의 범위와 방향을 사전에 형성할 수 있다는 것이다.

시나리오 2는 인간의 최종 승인이라는 제도적 보호 장치는 여전히 유지된다는 점에서 완전 자율무기 체계의 공개적 배치(시나리오 3)보다는 도입의 정치적·윤리적 부담이 낮고, 시나리오 1과 비교하여 AI의 군사적 효용이 더욱 커진다는 측면에서 중단기적으로 가장 실현 가능성이 높아 보이는 시나리오이다. 하지만, 위에서 언급한 바와 같이 AI에 대한 의존도가 점진적으로 높아질 경우, 인간의 최종 승인은 독립적이고 실질적인 판단의 결과라기보다는 AI가 구성하고 제안한 전장 인식과 군사적 옵션을 사후적으로 추인하는 형식적인 절차로 변질할 수 있는 위험성 역시 커지게 된다. 시나리오 2가 단순히 시나리오 1과 시나리오 3의 중간 단계가 아니라, 인간 통제라는 외형은 제도적으로 유지하지만 그 실질은 차츰 공동화됨으로써 시나리오 3으로의 점진적인 이행 가능성을 내포한 과도기적 단계로 기능할 가능성도 배제할 수 없다는 것이다.

3) 시나리오 3: 고자율 무기 체계 확산과 전략적 불안정성 심화

세 번째 시나리오는 고도의 자율성을 가진 AI 기반 자율무기 체계가 빠르게 확산되고, 인간 통제는 더 이상 작동하지 않는 상황으로, 표적 탐지부터 무력 사용 결정 및 실행에

이르는 전 과정이 인간 운용자의 실질적인 개입 없이 AI에 의해서 자율적으로 수행되는 경우이다. 이 시나리오에서 인간 운용자의 역할은 임무 개시 이전에 교전규칙을 설정하거나 사후적으로 교전 결과를 평가하는 수준으로 축소된다. 개별 상황에서 실시간으로 교전 및 무력 사용 여부를 판단하고 개입하는 실질적인 통제력은 인간 운용자가 아닌 AI가 보유하게 된다는 것이다. 시나리오 2에서 형식적으로나마 유지되었던 인간의 최종 승인 절차마저도 생략되거나 형해화됨으로써 인간 통제는 제도적으로도 실질적으로도 더 이상 보장되지 않는 것이다.

시나리오 3은 다음의 조건이 성립될 때 그 실현 가능성이 높아진다. 주요 군사 강국 사이의 AI 기반 자율무기 체계와 연관된 기술 및 군비 경쟁이 격화될 경우, 특정 국가가 완전 자율무기 체계의 개발 및 배치를 주저하게 되면 해당 국가는 관련 무기 체계의 개발과 도입을 결심한 국가와 비교해서 군사적인 열세에 처할 가능성이 높아질 수 있다는 딜레마가 존재한다. 이와 같은 상황에서는 인간 통제 원칙의 준수가 이들 국가 사이에 공유된 협력적 규범이 아니라 일방적인 자제로 인식될 수 있으며, 경쟁국의 자율화 수준에 상응하거나 이를 추월한 군사적 우위를 달성하기 위해서 자국의 자율무기 체계의 고도화를 가속화할 유인이 높아지게 되는 것이다. 또한, 자율무기 체계를 실효적으로 규제할 수 있는 국제규범이 형성되지 못하거나, 이와 같은 규범이 형성되더라도 주요 군사 강국들이 자국의 이익을 우선시하여 이에 대한 참여를 거부하거나 형식적으로만 참여할 경우에도 시나리오 3의 실현 가능성은 더욱 높아질 수 있다.

시나리오 3이 현실화될 경우 제기될 수 있는 가장 큰 위험 중의 하나는 무력 사용이라는 중대한 결정이 인간의 규범적·윤리적 통제 범위를 벗어날 수도 있다는 점이다. 고도화된 자율무기 체계가 교전 및 무력 사용의 결정과 실행까지도 독자적으로 수행하게 될 경우, 인간 생명의 박탈이라는 중요한 결정의 주체가 인간에서 기계로 이전될 수 있다는 것이다. 또한, 시나리오 3은 국가 간 전략적 불안정성의 수준을 더욱 심화시킬 수 있다. 고자율 무기체계가 대량으로 배치될 경우, 각국은 경쟁국의 무기체계가 어느 수준의 자율성을 갖고 있는지, 어떠한 조건에서 자동적으로 교전에 돌입하는지, 인간 운용자가 실제로 개입할 수 있는 여지가 어느 정도인지 등을 정확하게 파악하기 어려워 상대방의 의도와 대응 가능성을 오판할 위험이 커질 수 있다. 이로 인해, 국가 간 신뢰와 예측 가능성은 더욱 낮아져 제한적 충돌이나 기술적 오류가 외교적 조정과 정치적 숙고 등을 거치기 전에 연쇄적인 군사적 대응의 확대로 이어질 수 있으며, 우발적 확전과 군비 경쟁이 증폭되는 결과가 초래될 수도 있다.

2. 시사점

AI 기반 자율무기 체계의 발전은 단순히 새로운 무기체계의 등장만을 의미하는 것이 아니다. 고도화된 자율무기 체계는 전쟁 수행의 방식, 군사조직의 의사결정 구조, 국제안보 질서 등을 근본적으로 변화시킬 수 있는 구조적 전환 가능성을 내포하고 있다. 따라서 향후의 정책적 대응 역시 자율무기 체계의 개발 여부만을 둘러싼 찬반 논쟁에 머물러서는 안 된다. 어떤 임무와 환경에서, 어느 수준의 자율성을 허용할 것인지, 인간은 어떤 방식으로 통제와 책임을 유지할 것인지, 그리고 국제 안보 측면에서의 전략적 안정성에 미치는 영향을 어떻게 관리할 것인지 등의 복합적인 질문에 대한 해답을 모색하는 데 초점을 맞추어야 하며, 이러한 문제의식하에서 다음과 같은 정책적 시사점을 도출할 수 있다.

1) 기술 개발 단계에서부터 '통제 가능한 자율성' 고려 필요

AI 기반 자율무기 체계 개발의 기본 방향은 단순히 '자율성의 확대'가 아니라 '통제 가능한 자율성'의 확보에 그 초점이 맞추어져야 한다. 자율무기 체계의 개발 단계부터 성능 요소뿐 아니라, 설명 가능성, 불확실성 표시, 인간 중심 인터페이스, 비정상 작동 시 개입 및 중단 기능 등을 함께 갖추도록 설계되어야 한다는 것이다(Maathuis & Cools, 2025). 이때 '통제 가능한 자율성'은 추상적인 원칙이나 사후적인 운용 지침만으로는 확보될 수 없으며, 무기체계의 개발 단계에서부터 검증 가능한 기술적 요구 사항으로 구체화될 필요가 있다. 예를 들어, 자율무기 체계의 성능 평가는 표적 탐지 정확도나 대응 속도 등 일반적인 무기체계의 평가 지표뿐만 아니라, 인간 운용자가 AI의 판단 근거를 실제로 이해할 수 있는지, 시스템의 불확실성과 오류 가능성이 명확하게 제시되는지, 인간 운용자가 시스템의 판단을 거부하거나 작동을 중단시킬 수 있는 기능과 인터페이스 등이 신뢰성 있게 구현되어 있는지 등을 중요한 기준으로 삼아 개발되고 평가되어야 한다(Siebert et al., 2023). AI 기반 자율무기 체계에 대한 통제 가능성이 개발 이후에 부과되는 사후적인 제약 요건이 아니라, 해당 무기 체계 개발의 전 과정에서 가장 중요하게 고려되어야 할 핵심적인 성능 요구 조건 중의 하나로 인식되고 구현되어야 한다는 것이다.

2) 운용 단계에서의 인간 통제 절차의 실질화·제도화 필요

AI 기반 자율무기 체계의 운용 과정에서의 인간 통제 원칙은 단순히 인간 운용자가 최종 승인권을 갖고 있다는 선언적 수준에 머물러서는 안 되며, 실제 전장 상황에서 작동 가능

한 구체적인 교전규칙과 작전 운용 절차 등으로 제도화되어야 한다(Calvert, 2025). AI에 대한 의존도가 높아질수록 인간의 최종 승인은 AI가 제시한 판단을 사후적으로 확인하는 형식적인 절차로 약화될 수 있다. 이와 같은 우려를 해소하기 위해서는 자율무기 체계의 투입 이전 단계에서 사용 조건, 자율성 허용 범위, 인간의 개입 지점, 지휘관과 운용자의 거부권 및 중단권 등을 명확히 규정할 필요가 있다. 자율무기 체계가 어떠한 임무와 환경에서 투입 가능한지, 어떤 표적 유형에 대해서 어느 정도 수준의 자율성이 허용되는지, 어느 단계에서 인간 지휘관의 승인이 요구되는지 등에 대한 기준을 사전에 명확하고 구체적으로 설정하여야 한다는 것이다. 예를 들어, 민간인 밀집 지역이나 전투원과 비전투원 사이의 구분이 용이하지 않은 상황에서는 AI의 역할을 정보 제공과 판단 보조 수준으로 제한하도록 하고, 고위험 교전 상황이나 확전 가능성이 높은 임무에서는 AI의 자동 대응을 원칙적으로 금지하여 지휘관이나 인간 운용자만이 무력 사용에 대한 최종 판단을 내리도록 하고, 필요 시 시스템의 판단을 거부하거나 작동을 중단할 수 있는 권한과 절차가 실질적으로 보장될 수 있도록 하는 제도적인 안전장치가 마련되어야 한다.

3) 물리적 통제에서 소프트웨어·행위 기반 군비통제 패러다임으로 전환 필요

AI 기반 자율무기 체계의 확산이 더욱 가속화된다면 군비통제 방식 역시 근본적인 전환이 불가피해 보인다. 전통적인 군비통제는 국가 간의 합의를 바탕으로 핵탄두 등 물리적 무기체계의 수량과 배치 현황을 제한하고 이를 확인하는 방식으로 이루어졌다. 하지만 자율무기 체계의 핵심 능력은 해당 플랫폼의 보유 수량 등 외형적으로 드러나는 지표에 의해서 결정되지 않고 알고리즘의 설계 방식, 자율성의 수준, 인간 승인 조건 등 외부에서 확인하기 어려운 다양한 요소에 의해서 결정된다. 하드웨어적으로 동일한 무인기라고 하더라도 어떤 교전 조건이 적용되는지, 인간 승인 절차가 어느 단계에서 요구되는지 등에 따라서 일반적인 감시·정찰 자산이 될 수도 있고 완전 자율 살상 무기가 될 수도 있다[18]. 따라서 향후 군비통제 방식은 단순히 무기체계의 물리적 형태나 수량을 통제하는 방식이 아니라 알고리즘의 설계 원칙, 자율성의 운용 조건, 인간 승인 절차 등을 규율하는 방식으로 전환되어야 한다(Scharre & Lamberth, 2022). 다만, 군비통제 패러다임이 이와 같은 방식으로 전환된다고 해도 실제로 이를 외부에서 검증하기가 용이하지 않고, 각국은 여전히 기술 우위를 유지·확대하려는 강한 유인을 갖고 있기 때문에 소프트웨어·행위 기반 군비통제 체

18) 소프트웨어 업데이트만으로도 반자율 공격 드론이 인간 운용자의 개입이 불필요한 완전 자율 살상 무기로 전환되어 운용될 수 있다는 것이다.

계가 실효적으로 정착되기 위해서는 주요 군사 강국 사이의 신뢰 형성이 선행되어야 할 것이다(Mittelsteadt, 2021).

참고문헌

김낙우, 이동수, 채원석, 유호영, 이상은, & 김현진. (2025). 인공지능 윤리 기반 정렬 기술 동향. *전자통신동향분석, 40*(1), 64-73.

문현영, & 김상수. (2023). 치명적 자율무기체계의 도덕적 책임 문제 연구: 로버트 스패로우의 '책임 간극' 이론에 대한 고찰. *문화기술의 융합, 9*(4), 375-381.

박문언. (2022, 11월 17일). *자율무기체계에 대한 국제적 논의와 우리 군의 대비 방향* (국방논단 제1920호[22-41]). 한국국방연구원.

박문언. (2024). 자율무기체계와 지휘책임. *서울국제법연구*, 31(2), 1-40.

성국일, & 김지원. (2026). AI 기반 군사 작전의 의사결정 투명성 및 책임성 확보 프레임워크 연구. *융합보안논문지, 26*(1), 141-149.

윤정현. (2024). *제2차 REAIM 고위급회의 서울 개최의 의미와 시사점* (INSS 이슈브리프 No. 600). 국가안보전략연구원.

윤정현. (2026). 인공지능 기반 전략적 비핵무기의 부상과 억지 구조의 전환. *평화연구, 34*(1), 241-278.

이상근. (2022). 미·중의 군사용 AI 개발이 신군비경쟁에 미치는 영향 연구. *군사발전연구, 16*(1), 123-151.

장상국. (2025). SWOT-PEST 분석을 통한 AI 기반 유무인 복합체계 구축 전략 연구. *한국방위산업학회지, 32*(2), 69-84.

정구연. (2022). 4차 산업혁명과 미국의 미래전 구상: 인공지능과 자율무기체계를 중심으로. *국제관계연구, 27*(1), 5-36.

조성진. (2026, 2월 11일). 해양통제 확보 수단으로서의 해군기지 타격의 재조명. *KIMS Periscope, 415*. 한국해양전략연구소.

한정우. (2025). 치명적 자율무기체계(LAWS)의 공법적 규율 방안 연구: 결심의 은폐와 가시화를 중심으로. *성균관법학, 37*(4), 187-223.

황원중. (2020). 전쟁의 본질로 바라본 인공지능의 군사적 활용: 지휘결심 지원체계의 적용을 중심으로. *한국군사학논집, 76*(3), 31-60.

Allen, J. R., & Husain, A. (2017). On hyperwar. *Proceedings, 143*(7). U.S. Naval Institute.

Amoroso, D., & Tamburrini, G. (2020). Autonomous weapons systems and meaningful human control: Ethical and legal issues. *Current Robotics Reports, 1*(4), 187-194.

Arreguín-Toft, I. (2001). How the weak win wars: A theory of asymmetric conflict. *International Security, 26*(1), 93-128.

Asaro, P. (2012). On banning autonomous weapon systems: Human rights, automation, and the dehumanization of lethal decision-making. *International Review of the Red Cross, 94*(886), 687-709.

Bai, Y., Kadavath, S., Kundu, S., Askell, A., Kernion, J., Jones, A., Chen, A., Goldie, A., Mirhoseini, A., McKinnon, C., Chen, C., Olsson, C., Olah, C., Hernandez, D., Drain, D., Ganguli, D., Li, D., Tran-Johnson, E., Perez, E., … Kaplan, J. (2022). *Constitutional AI: Harmlessness from AI feedback* (arXiv:2212.08073). arXiv.

Bansal, G., Wu, T., Zhou, J., Fok, R., Nushi, B., Kamar, E., Ribeiro, M. T., & Weld, D. S. (2021). Does the whole exceed its parts? The effect of AI explanations on complementary team performance. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Article 81). Association for Computing Machinery.

Bearman, T., Rosen, B., & Anderson-Samways, B. (2026, April 9). *AI decision support systems: A neglected source of military AI risk*. Institute for AI Policy and Strategy.

Bengio, Y., Cohen, M., Fornasiere, D., Ghosn, J., Greiner, P., MacDermott, M., Mindermann, S., Oberman, A., Richardson, J., Richardson, O., Rondeau, M.-A., St-Charles, P.-L., & Williams-King, D. (2025). *Superintelligent agents pose catastrophic risks: Can Scientist AI offer a safer path?* (arXiv:2502.15657). arXiv.

Bo, M. (2020, December 18). *Meaningful human control over autonomous weapon systems: An (international) criminal law account.* Opinio Juris.

Bo, M., Bruun, L., & Boulanin, V. (2022). *Retaining human responsibility in the development and use of autonomous weapon systems: On accountability for violations of international humanitarian law involving AWS.* Stockholm International Peace Research Institute.

Bode, I., Nadibaidze, A., Watts, T., & Zhang, Q. (2025). Ensuring the exercise of human agency in AI-based military systems: Concerns across the lifecycle. *Ethics and Information Technology, 27*, Article 50.

Boulanin, V., Saalman, L., Topychkanov, P., Su, F., & Peldán Carlsson, M. (2020). *Artificial intelligence, strategic stability and nuclear risk*. Stockholm International Peace Research Institute.

Boulanin, V., & Verbruggen, M. (2017). *Mapping the development of autonomy in weapon systems*. Stockholm International Peace Research Institute.

Brown, K., Gould, J., & Hamilton, J. (2025, April 24). *Preventing a flash war: Countering the risk of AI-driven escalation on the battlefield*. Center for Ethics and the Rule of Law, University of Pennsylvania.

Bruun, L. (2024). *Towards a two-tiered approach to regulation of autonomous weapon systems: Identifying pathways and possible elements*. Stockholm International Peace Research Institute.

Buçinca, Z., Malaya, M. B., & Gajos, K. Z. (2021). To trust or to think: Cognitive forcing functions can reduce overreliance on AI in AI-assisted decision-making. *Proceedings of the ACM on Human-Computer Interaction, 5*(CSCW1), Article 188, 1-21.

Calvert, S. C. (2025). Principles and framework for the operationalisation of meaningful human control over autonomous systems. *Science and Engineering Ethics, 31*, Article 27.

Chang, Y., Cheng, Y., Manzoor, U., & Murray, J. (2023). A review of UAV autonomous navigation in GPS-denied environments. *Robotics and Autonomous Systems, 170*, 104533.

Christiano, P. F., Leike, J., Brown, T. B., Martic, M., Legg, S., & Amodei, D. (2017). Deep reinforcement learning from human preferences. *Advances in Neural Information Processing Systems, 30*, 4299-4307.

Clark, B., Patt, D., & Schramm, H. (2020). *Mosaic warfare: Exploiting artificial intelligence and autonomous systems to implement decision-centric operations*. Center for Strategic and Budgetary Assessments.

DARPA. (n.d.). *OFFensive Swarm-Enabled Tactics (OFFSET)*. Defense Advanced Research Projects Agency.

Dorsey, J. (2025). The erosion of human(e) judgement in targeting? Quantification logics, AI-enabled decision support systems and proportionality assessments in IHL. *International Review of the Red Cross, 107*(930), 1041-1071.

Dorsey, J., & Bo, M. (2026). AI-enabled decision-support systems in the joint targeting cycle: Legal challenges, risks, and the human(e) dimension. *International Law Studies, 107*(1), 184-227.

Doty, J., & Doty, C. (2012). Command responsibility and accountability. *Military Review, 92*(1), 35-38.

Dung, L. (2023). Current cases of AI misalignment and their implications for future risks. *Synthese, 202*(5), Article 138.

Edwards, B. (2023, June 2). *Air Force denies running simulation where AI drone "killed" its*

operator. Ars Technica.

Egel, N. (2021, January). *Reducing military risks through OSCE instruments: The untapped potential in the European arms control framework* (Strategic Security Analysis No. 16). Geneva Centre for Security Policy.

Feickert, A., Kapp, L., Elsea, J. K., & Harris, L. A. (2018, November 1). *U.S. ground forces robotics and autonomous systems (RAS) and artificial intelligence (AI): Considerations for Congress* (CRS Report No. R45392). Congressional Research Service.

Gallagher, J., & Oughton, E. J. (2024, September). Transforming the multidomain battlefield with AI: Object detection, predictive analysis, and autonomous systems. *Military Review*, Online Exclusive.

Gettinger, D. M. (2024, October 25). Defense primer: Categories of uncrewed aircraft systems (CRS In Focus IF12797). Congressional Research Service.

Hartridge, S., & Walker-Munro, B. (2025). Autonomous weapons systems and the AI alignment problem. *Journal of International Humanitarian Legal Studies, 16*(1), 38-65.

Hendrycks, D., Schmidt, E., & Wang, A. (2025). *Superintelligence strategy: Expert version* (arXiv:2503.05628). arXiv.

Holland, M. (2020). The black box, unlocked: Predictability and understandability in military AI. United Nations Institute for Disarmament Research.

Horowitz, M. C. (2019). When speed kills: Lethal autonomous weapon systems, deterrence and stability. *Journal of Strategic Studies, 42*(6), 764-788.

House of Lords AI in Weapon Systems Committee. (2023). *Proceed with caution: Artificial intelligence in weapon systems* (HL Paper 16, Session 2023-24). UK Parliament.

Human Rights Watch. (2015). *Mind the gap: The lack of accountability for killer robots.*

Human Rights Watch, & Harvard Law School International Human Rights Clinic. (2025). *A hazard to human rights: Autonomous weapons systems and digital decision-making.* Human Rights Watch.

International Committee of the Red Cross. (2021). *ICRC position on autonomous weapon systems.*

Jensen, B., & Atalan, Y. (2025, May 13). *Drone saturation: Russia's Shahed campaign*. Center for Strategic and International Studies.

Johnson, J. (2022). Inadvertent escalation in the age of intelligence machines: A new model for nuclear risk in the digital age. *European Journal of International Security, 7*(3), 337-359.

Johnson, J. (2023). Automating the OODA loop in the age of intelligent machines: Reaffirming the role of humans in command-and-control decision-making in the digital age. *Defence Studies, 23*(1), 43-67.

Kallenborn, Z. (2020). *Are drone swarms weapons of mass destruction?* (Future Warfare Series No. 60). U.S. Air Force Center for Strategic Deterrence Studies, Air University.

Keenan, J. M., Trump, B., Kytömaa, E., Adlakha-Hutcheon, G., & Linkov, I. (2024). The role of science in resilience planning for military-civilian domains in the U.S. and NATO. *Defence Studies, 24*(4), 493-524.

Kerns, A. J., Shepard, D. P., Bhatti, J. A., & Humphreys, T. E. (2014). Unmanned aircraft capture and control via GPS spoofing. *Journal of Field Robotics, 31*(4), 617-636.

Klare, M., & Liang, X. (2024, November 12). *Beyond a human "in the loop": Strategic stability and artificial intelligence*. Arms Control Association.

Kmentt, A. (2025). Geopolitics and the regulation of autonomous weapons systems. *Arms Control Today, 55*(1-2).

Konaev, M., Chahal, H., Fedasiuk, R., Huang, T., & Rahkovsky, I. (2020). *U.S. military investments in autonomy and AI: A strategic assessment*. Center for Security and Emerging Technology, Georgetown University.

Maathuis, C., & Cools, K. (2025). Risks and control measures for building trustworthy autonomous weapon systems. In *Proceedings of the 20th International Conference on Cyber Warfare and Security* (Vol. 20, No. 1, pp. 214-222). Academic Conferences and Publishing International.

Mako, G. (2026, March 9). *Legal accountability for AI-driven autonomous weapons.* Lieber Institute, West Point.

Mathews, B. (2022, August 31). *Autonomous vehicles: New technology revolutionizes Army's principles of sustainment.* U.S. Army.

Mazarr, M. J., Chan, A., Demus, A., Frederick, B., Nader, A., Pezard, S., Thompson, J. A., & Treyger, E. (2018). What deters and why: *Exploring requirements for effective deterrence of interstate aggression* (RAND Research Report RR-2451-A). RAND Corporation.

Meinke, A., Schoen, B., Scheurer, J., Balesni, M., Shah, R., & Hobbhahn, M. (2024). *Frontier models are capable of in-context scheming* (arXiv:2412.04984). arXiv.

Mittelsteadt, M. (2021). *AI verification: Mechanisms to ensure AI arms control compliance.* Center for Security and Emerging Technology.

Osimen, G. U., Newo, O., & Fulani, O. M. (2024). Artificial intelligence and arms control in modern warfare. *Cogent Social Sciences, 10*(1), Article 2407514.

Parasuraman, R., & Manzey, D. H. (2010). Complacency and bias in human use of automation: An attentional integration. *Human Factors, 52*(3), 381-410.

Phuong, M., Aitchison, M., Catt, E., Cogan, S., Kaskasoli, A., Krakovna, V., Lindner, D., Rahtz, M., Assael, Y., Hodkinson, S., Howard, H., Lieberum, T., Kumar, R., Abi Raad, M., Webson, A., Ho, L., Lin, S., Farquhar, S., Hutter, M., Deletang, G., Ruoss, A., El-Sayed, S., Brown, S., Dragan, A., Shah, R., Dafoe, A., & Shevlane, T. (2024). *Evaluating frontier models for dangerous capabilities* (arXiv:2403.13793). arXiv.

Probasco, E. S., Toner, H., Burtell, M., & Rudner, T. G. J. (2025). *AI for military decision-making: Harnessing the advantages and avoiding the risks*. Center for Security and Emerging Technology.

Puscas, I. (2022). *Human-machine interfaces in autonomous weapon systems*. United Nations Institute for Disarmament Research.

Romeo, G., & Conti, D. (2026). Exploring automation bias in human-AI collaboration: A review and implications for explainable AI. *AI & Society, 41*, 259-278.

Roussel, E., Lauwaert, L., Swoboda, T., Ramsey, G., Uuk, R., Dung, L., & Aguirre, A. (2026). *Are we doomed to an AI race? Why self-interest could drive countries towards a moratorium on superintelligence* (arXiv:2605.01297). arXiv.

Rumbaugh, W. (2024, February 13). *Cost and value in air and missile defense intercepts*. Center for Strategic and International Studies.

Sayler, K. M. (2020). *Artificial intelligence and national security* (CRS Report No. R45178). Congressional Research Service.

Sayler, K. M. (2025a). *DOD Replicator initiative: Background and issues for Congress* (CRS In Focus IF12611). Congressional Research Service.

Sayler, K. M. (2025b). *Hypersonic weapons: Background and issues for Congress* (CRS Report No. R45811). Congressional Research Service.

Scharre, P., & Lamberth, M. (2022). *Artificial intelligence and arms control*. Center for a New American Security.

Schmitt, M. N. (2005). Precision attack and international humanitarian law. *International Review of the Red Cross, 87*(859), 445-466.

Schmitt, M. N., & Thurnher, J. S. (2013). "Out of the loop": Autonomous weapon systems

and the law of armed conflict. *Harvard National Security Journal, 4*(2), 231-281.

Schneider, J., & Macdonald, J. (2024). Looking back to look forward: Autonomous systems, military revolutions, and the importance of cost. *Journal of Strategic Studies, 47*(2), 162-184.

Siebert, L., Lupetti, M. L., Aizenberg, E., Beckers, N., Zgonnikov, A., Veluwenkamp, H., Abbink, D. A., Giaccardi, E., Houben, G.-J., Jonker, C. M., van den Hoven, J., Forster, D., & Lagendijk, R. L. (2023). Meaningful human control: Actionable properties for AI system development. *AI and Ethics, 3*, 241-255.

Sparkes, M. (2026, June). *Fully autonomous drones have killed human soldiers for the first time*. New Scientist.

Sparrow, R. (2007). *Killer robots. Journal of Applied Philosophy, 24*(1), 62-77.

Sparrow, R. (2016). Robots and respect: Assessing the case against autonomous weapon systems. *Ethics & International Affairs, 30*(1), 93-116.

United States of America. (2018). Humanitarian benefits of emerging technologies in the area of lethal autonomous weapons systems (CCW/GGE.1/2018/WP.4). Group of Governmental Experts on Emerging Technologies in the Area of Lethal Autonomous Weapons Systems, Convention on Certain Conventional Weapons.

U.S. Air Force. (2026). *Air Force doctrine publication 3-60: Targeting.*

U.S. Department of Defense. (2022). *Summary of the Joint All-Domain Command and Control (JADC2) strategy.*

U.S. Department of Defense. (2023). *DoD Directive 3000.09: Autonomy in weapon systems.*

Verdiesen, I., Santoni de Sio, F., & Dignum, V. (2021). Accountability and control over autonomous weapon systems: A framework for comprehensive human oversight. *Minds and Machines, 31*(1), 137-163.

Verly, J. G., Delanoy, R. L., & Dudgeon, D. E. (1989). Machine intelligence technology for automatic target recognition. *The Lincoln Laboratory Journal, 2*(2), 277-310.

Vermeer, M. J. D., Lathrop, E., & Moon, A. (2025). *On the extinction risk from artificial intelligence* (Research Report RR-A3034-1). RAND Corporation.