휴머노이드 로봇 안전 기술과 표준화 동향

Humanoid Robot Safety: Technologies and Standardization

장민수 (M. Jang, minsu@etri.re.kr) 소셜로보틱스연구실 책임연구원
윤영우 (Y. Yoon, youngwoo@etri.re.kr) 소셜로보틱스연구실 책임연구원
고우리 (W.-R. Ko, wrko@etri.re.kr) 소셜로보틱스연구실 책임연구원
박천수 (C. Park, bettle@etri.re.kr) 소셜로보틱스연구실 책임연구원
조미영 (M.Y. Cho, mycho@etri.re.kr) 소셜로보틱스연구실 책임연구원
김도형 (D.-H. Kim, dhkim008@etri.re.kr) 소셜로보틱스연구실 책임연구원/실장

ABSTRACT

As humanoid robots transition from development phases to factory floors, new safety challenges have emerged. Ensuring safety is become a mandatory condition for deployment, as exemplified by the EU Machinery Regulation, which requires third-party assessment for self-evolving safety components. However, existing safety frameworks are designed for traditional, fixed-base robots and rely on assumptions such as deterministic behavior and predefined safety zones, which do not apply to the continuous balancing, whole-body contact, and probabilistic artificial intelligence policies of humanoids. This article analyzes these gaps by examining the safety technologies required for learning-based humanoid robots. The discussion is structured into three parts: physical interaction safety, alignment and guardrails for foundation models, and evaluation frameworks and standards. We review how current safety functions are divided between low-level reactive filters and high-level deliberative layers, highlighting the urgent need for comprehensive evaluation metrics, robust simulation-to-real validation, and the development of humanoid-specific safety standards to support safe industrial integration.

KEYWORDS 로봇 파운데이션 모델, 안전 기술, 안전 표준, 휴머노이드 로봇

I. 서론

전 세계적으로 휴머노이드 로봇이 단순한 기술 실증을 넘어 대규모 현장 배치와 양산 단계로 진입하면서 실용화가 본격화되고 있다[1,2]. 현대자동차그룹, BMW, 토요타 등 글로벌 주요 기업들은 자사 제조 라인 및 물류 현장에 휴머노이드 로봇 도입을 적극적으로 추진하며 산업 생태계를 확장하고 있다

* DOI: https://doi.org/10.22648/ETRI.2026.J.410606
* This work was supported by the NRF/MSIT in 2026 (Grant No. RS-2026-25550793, Project Title: Embodiment-Aware Planning and Action Generation for Task Intelligence Sharing across Heterogeneous Humanoids, 50%) and by the IITP/MSIT (RS-2024-00336738 Development of Complex Task Planning Technologies for Autonomous Agents, 50%).

본 저작물은 공공누리 제4유형
출처표시+상업적이용금지+변경금지 조건에 따라 이용할 수 있습니다.
©2026 한국전자통신연구원

표 1 주요 휴머노이드 로봇 실용화 현황과 계획

기업/기관	실용화 내역
Figure AI [1]	• BMW 스파턴버그 공장 차체 공정에서 Figure 02가 10개월간 BMW X3 3만 대 이상 생산 지원 • 2026년 6월부터 물류 공정에 후속 기종 Figure 03 투입
애지봇 (AgiBot) [2]	• 2025년 8월 자동차 부품 제조기업 푸린정공 공장에 바퀴형 양팔 로봇 A2-W 100대 투입
보스턴 다이내믹스/ 현대차그룹 [3,4]	• 자체 제조 시설에 수만 대의 로봇 도입 계획 (Spot·Atlas 등 포함) • 연간 3만 대 생산 가능한 미국 내 로봇 공장 구축 등 투자 추진
토요타 (Toyota) [5]	• 전 세계 자사 공장에 40만 대의 휴머노이드 로봇 배치 실용화 계획 발표
K-휴머노이드 얼라이언스 [15]	• 핵심 부품 국산화 및 범용 로봇 파운데이션 모델 개발 추진 • 2029년부터 연 1,000대 이상 양산 체계 구축 추진

(표 1)[1,3-5].

그러나 실용화의 최우선 전제 조건인 '안전' 문제가 핵심 걸림돌로 대두되고 있다[6]. 과거 정해진 궤적을 따르는 고정형 산업용 로봇 시대에도 치명적인 사고는 끊이지 않았다[6]. 2023년 11월 경남 고성의 농산물유통센터에서는 바닥에 고정된 로봇 팔이 40대 작업자를 압착해 사망하게 한 사고가 발생했으며, 경찰은 센서가 작업자를 상자로 인식한 것으로 추측했다[7]. 미국 산업안전보건청(OSHA)의 중대 상해 보고서를 분석한 연구에 따르면 2015년부터 2022년 사이 발생한 로봇 관련 중상해 사고는 77건에 달한다[6]. 이전 세대의 사고가 정해진 궤적을 따르는 제한된 공간 내에서의 협착과 충돌로 발생했다면, 범용 휴머노이드 로봇의 사고 위험은 균형 상실, 불확실한 생성형 행동, 그리고 예측 불가능한 충돌력 등으로 나타나며 훨씬 복잡한 양상을 띤다[8]. "로봇이 두개골을 골절시킬 수 있다"라고 경고한 피규어 AI의 전직 안전 책임자가 회사를 상대로 소송을 제기하는 등 사람과 공존하며 작업하는 로봇의 강력한 힘 역량에 따른 물리적 위험성이 현장에서 지속적으로 제기되고 있다[9].

안전은 이제 단순한 기술적 과제를 넘어 실용화와 현장 배치를 위한 필수 법적 조건이 되었다[10]. 세계 각국의 규제와 정책이 안전성 검증을 로봇 실용화의 필수 전제로 명문화하고 있기 때문이다. 유럽연합(EU) 기계규정은 기계학습을 통해 자가진화하는 안전부품을 제3자 적합성 평가 대상으로 지정하여 2027년 1월부터 의무 적용할 예정이다[10]. 반면, 현재 시행 중인 국내 법정 로봇 인증은 실외 주행 로봇의 운행 안전을 심사하는 8개 항목에 그치고 있어 휴머노이드 로봇의 핵심인 실내 조작 작업과 사람과의 전신 다접촉 안전을 평가할 기준이 부족한 상황이다[11]. 현재 산업계의 실용화 일정을 살펴보면 유럽의 제3자 평가 의무(2027년)와 국내 양산 목표(2029년)가 임박했음에도, 프라운호퍼 IPA 등은 동적 안정형 휴머노이드 로봇을 포괄하는 ISO 25785-1 등 첫 안전 표준의 발간을 2028년으로 내다보고 있으며, 대구시에 구축되는 국내 휴머노이드 로봇 안전인증센터는 2030년에야 완료될 예정이다[12-14]. 상용화 시점과 안전 평가 인프라 구축 시점 사이의 시간적 엇박자를 극복하기 위해 휴머노이드 로봇에 특화된 새로운 다층적 안전 기술과 실측 데이터 기반의 평가 체계를 선제적으로 완비할 필요가 있다.

이에 본고는 휴머노이드 로봇 안전 기술 동향과 관련 국제 표준화 현황을 자세히 짚어보고 다가올 대규모 휴머노이드 로봇 실용화 시대를 대비해 우리가 당장 해결해야 할 시급한 과제를 살펴보고자 한다. 또한, 본고에서는 인간과 공간을 공유하는 휴머노이드 로봇의 안전 기술을 크게 물리적 상호작용 안전, 파운데이션 모델의 정렬과 가드레일, 그리고 평가 체계와 표준의 세 부분으로 나누어 분석한다. Ⅱ장에서는 기존 로봇 안전 기술 체계와 새로운

휴머노이드 로봇 제어 기술의 특징을 대비하여 안전망이 추가로 요구되는 지점을 정리한다. 이어 Ⅲ장에서는 물리적 상호작용 안전 기술을, Ⅳ장에서는 파운데이션 모델 정렬 및 가드레일 기술을, Ⅴ장에서는 신규 평가 체계와 표준 현황을 차례로 소개하고 Ⅵ장에서 전체 내용을 요약하고 성공적인 현장 배치를 위한 향후 핵심 시사점을 도출한다.

Ⅱ. 휴머노이드 로봇 안전 기술 필요성

전통적인 로봇 안전 규제와 인증 경로는 정해진 환경에서 구동하는 산업용 로봇의 제어 체계를 바탕으로 설계되었다. 그러나 인간과 같은 공간을 이동하며 인공지능(AI)을 통해 스스로 판단하는 휴머노이드 로봇이 등장하면서 기존 안전 기술의 전제들은 근본적인 한계에 직면하고 있다.

1. 전통적 로봇 안전 기술의 특징

기존 산업용 로봇의 안전은 위험성 평가를 통한 위험원 식별, 제어 계통 신뢰도의 등급화(PL, SIL), 그리고 실측 시험을 통한 접촉력 확인이라는 세 단계가 맞물려 보증된다. 이 체계는 기계가 고정된 베이스 위에서 정해진 궤적을 따르며, 방호 울타리로 인간과 분리되고, 정지된 상태가 곧 안전하다는 조건을 전제한다[16,17].

인간과 공간을 공유하는 협동로봇(Cobot)에서도 이러한 한계가 일부 유지된다. 협동로봇은 동력 및 힘 제한(PFL: Power and Force Limiting), 속도 및 이격 감시(SSM: Speed and Separation Monitoring) 등을 통해 물리적 충격량을 인간이 부상당할 수 있는 임계점 이하로 억제한다. 이때 접촉은 신체가 구속된 상태의 '준정적 접촉'과 충격 후 튕겨 나가는 '과도 접촉' 두 가지로만 분류되며, 한곗값은 성인 100명의 통증 역치 75백분위수를 기준으로 한다.

가장 핵심적인 특징은 결정론적(Deterministic) 거동을 전제한다는 점이다. 같은 입력에 항상 같은 출력이 나오므로 설계자는 최악 조건의 접촉점을 사전에 지정할 수 있고 소수의 실측 시험만으로도 전체 운용 상태의 안전을 대표하여 평가할 수 있다.

2. 휴머노이드 로봇 제어 기술의 특징

전통적 전제와 극명하게 대비되는 휴머노이드 로봇의 제어적 특성은 크게 세 가지로 요약된다.

첫째는 능동 제어와 전도 위험성이다. 휴머노이드 로봇은 두 발로 직립하기 위해 끊임없이 능동 제어를 수행해야 한다. 단순한 전원 차단이나 비상 정지는 곧 넘어짐(전도)을 의미하며 크고 무거운 기체의 낙상은 작업자에게 치명적인 위험을 초래할 수 있다[18].

둘째는 로봇 행동의 확률적 비결정성이다. 시각 · 언어 · 행동 모델(VLA: Vision Language Action Model), 월드액션모델(WAM: World Action Model) 등 로봇 파운데이션 모델(RFM: Robot Foundation Model)로 제어하는 휴머노이드 로봇은 확률적 샘플링이나 관측 변동에 따라 동일한 지시에서도 서로 다른 궤적을 생성할 수 있어 운용 조건 전반의 행동과 안전성을 사전에 열거하여 검증하기 어렵다.

셋째는 성공과 안전의 괴리이다. RFM 기반 제어에서는 높은 작업 성공률이 안전을 담보하지 않는다. SafeVLA-Bench의 LIBERO 평가에서는 평균 성공률이 94%를 넘는 정책에서도 전체 시행의 약 13~15%가 불안전한 것으로 나타났다. 별도의 RoboCasa 평가에서는 성공한 시행 중 약 36~56%에서 안전 위반이 관찰되었다[19]. 이는 작업 성공 여부와 안전 제약 준수 여부를 구분하여 평가해야 함을 보여준다.

표 2 휴머노이드 로봇 안전 기술 필요성

기존 기술 특성	휴머노이드 로봇 특성	필요 기술
준정적/과도 접촉 대상, 힘의 상한 기준만 제시	전신 다접촉, 지속 동적 접촉 고려 필요, 의도된 접촉 힘 하한 제시 필요	접촉력 한계에 연계한 속도·힘 제어, 의도된 접촉 판정 필요
결정론적 거동으로 최악 접촉점 사전 지정 가능	학습 정책(VLA)이 같은 조건에서도 매번 다른 궤적 생성	정책 모델 출력단의 안전층, 제약 학습 정렬, 행동 분포 채점 벤치마크 등 필요
기기 정지가 최선의 안전 상태	능동 제어를 통한 균형 유지가 필요하므로 기기 정지는 전도 위험 유발 가능	전도 예측과 보호 제어, 안전 상태 재정의 필요
방호 울타리로 사람과 분리	사람과 공간을 공유하며 능동적으로 이동	실시간 충돌 회피, 시뮬레이션 기반 안전 제어 학습과 실물 검증 필요
조건 열거 시험과 부품 고장률 등급 판정	학습 기반 인지 기능은 조건 열거 검증 불가, 배포 후 지속 갱신 발생	AI 정렬과 가드레일, 불확실성 정량화, 시나리오 기반 판정과 갱신 관리 필요

3. 새로운 안전 기술의 필요성

앞서 살펴본 바와 같이 전통적 로봇 안전의 전제들은 휴머노이드 로봇 특유의 구조와 AI 제어 방식 앞에서는 더 이상 유효하지 않다. 낡은 반응형 임계치 제약 체계를 넘어선 새로운 하이브리드 제어 기술과 평가 체계가 시급히 요구된다. 표 2에 기존 안전 기술의 전제가 성립하지 않는 이유와 이를 해결하기 위해 새롭게 요구되는 기술을 정리하였다.

III. 휴머노이드 로봇 물리 안전 기술

이 장에서는 기존 안전 기술의 한계를 극복하고 휴머노이드 로봇의 물리적 안전을 보장하기 위해 연구되고 있는 핵심 안전 기술들을 살펴본다. 개별 기술 동향을 살펴보기에 앞서 로봇 물리 안전 체계의 전체적인 구현 구조를 먼저 정리한다.

1. 물리 안전 개입 단계와 기술 계층

물리적 안전 기술은 위험 상황의 발생 흐름에 따라 세 단계로 나눠서 살펴볼 수 있다[20].

- **충돌 전 단계:** 감지와 예측을 통해 사람과 로봇의 시공간 접촉을 선제적으로 예방
- **접촉 단계:** 물리적 접촉이 발생했을 때 힘, 압력, 에너지를 인체 상해 한계 이하로 억제
- **접촉 후 단계:** 접촉, 충돌을 즉각 감지하고 정지, 순응 전환, 후퇴로 상해를 제한

각 단계에 대응하는 로봇 안전 제어 기술은 표 3에서 보듯이 4개의 구현 계층(L1~L4)으로 나누어 다층 방어 구조를 형성할 수 있다. 하위 계층(L1)으로 갈수록 제어 주기가 빠르고 결정론적인 물리 안전 보장을 담당하며 상위 계층(L4)으로 갈수록 추상적인 확률적 보장을 제공한다. 어느 한 계층도 단독으로 완전하지 않기 때문에 각 계층의 실패 모드(Failure Mode)를 다른 계층이 상호 보완하도록 설계된다.

표 3 휴머노이드 로봇 물리 안전 기술의 계층 구성

계층	핵심 역할	대표 기술
L1 물리/기구적 안전 제어	접촉 충격 억제 및 충돌 후 반응	컴플라이언스, 임피던스·어드미턴스 제어, 충돌 감지
L2 동역학적 안전 필터	실행 직전 명령을 안전 제약 안으로 제한	제어 장벽 함수(CBF), 도달 가능성 분석, 예측 안전 필터
L3 정책 학습과 궤적 최적화	안전 비용을 반영한 정책·경로 생성	CMDP, Safe MPC, 정합 예측
L4 작업 계획과 환경 모니터링	사람·환경 예측과 하위 계층 제약 제공	의도·궤적 예측, 점유 격자, 속도 및 이격 감시(SSM)

2. L1 계층: 물리/기구적 안전 제어

L1은 상위 계층이 모두 실패해 사람과 로봇이 닿는 상황을 전제로 전달되는 힘, 압력, 에너지를 인체 상해 한계 아래로 제한하는 계층이다. 이 계층의 기술은 다섯 부류로 구분할 수 있다. 첫째는 기구적 컴플라이언스다. 관절에 탄성 요소를 넣어 충돌 에너지를 기계적으로 흡수하는 방식으로 작동한다. 강성을 낮추면 상해 위험은 줄어들지만 위치 정밀도와 제어 대역폭은 그만큼 저하된다. 둘째는 관절 토크 제어에 기반한 임피던스 · 어드미턴스 제어다. 관절마다 토크 센서를 두고 로봇이 질량-스프링-댐퍼(Mass-Spring-Damper)처럼 거동하게 만들어 외력이 가해지면 밀려나도록 하는 제어 방식으로서 오늘날 협동 로봇의 사실상 표준이다[21]. 셋째는 힘과 토크, 속도 한계의 정량화다. 안전을 위해 허용되는 상한선은 생체역학 시험을 통해 정하며 사례 기반 상해 지표와 속도 · 질량 의존성을 제어기 설계에 직접 반영한다[22,23]. 산업 현장에서 쓰이는 안전 수치 기준은 ISO/TS 15066(2016)과 2025년 개정된 ISO 10218 표준에 제시되어 있다[24-26]. 넷째는 센서리스 충돌 감지(Sensorless Collision Detection)와 반사적 반응이다. 동역학 모델에 기반한 모멘텀 옵서버(Momentum Observer)로 외부 토크를 추정하면 별도의 힘 센서 없이도 충돌을 판단할 수 있고, 그 판단에 따라 정지나 역주행, 제로 중력 모드(Zero Gravity Mode), 감속 제어 등 안전 제어를 실행할 수 있다[27,28]. 다섯째는 피부형 센싱이다. 정전용량이나 광, 상호유도 방식의 근접 감지 피부는 접촉 전에 근접 거리를 측정할 수 있고, 접촉 후에는 압력 분포를 측정할 수 있어 L1 계층의 안전 개입 시점을 충돌 시점보다 한 단계 앞당길 수 있다[29].

L1은 충돌 자체를 막지 못하고 강성을 낮추거나 속도를 제한할수록 생산성과 정밀도가 동시에 낮아진다. 또한, 허용 힘 기준은 공구나 날카로운 끝단 장치로 인한 국지적 압력이나 여러 지점이 동시에 닿는 상황에서 안전을 보장하기 어렵다. 그러므로 상위 계층의 회피와 추가 보호 설계가 함께 필요하다.

3. L2 계층: 동역학적 안전 필터

L2 계층은 상위 정책이나 제어기가 위험한 명령을 내리더라도 실행 직전에 실시간으로 이를 걸러내는 수학적 안전 필터 계층이다. 이 계층에서는 제어 장벽 함수(CBF: Control Barrier Functions)와 이차계획법(Quadratic Programming)을 통해 모델의 정확성, 초기 상태의 안전성 및 제약의 실행 가능성 등 필요한 가정이 성립할 때 로봇 상태가 안전 집합을 유지하도록 제어한다[30]. 근거리 미래 시점의 위험 회피 가능성을 기준으로 안전 집합을 계산하여 활용하는 도달 가능성 분석이나 위험한 제어 명령을 미리 제거하는 예측 제어 기반 필터도 활용된다[31].

최근에는 로봇 파운데이션 모델 출력단에 결합하여 정책 출력을 안전 제약 안으로 사상하여 정책 모델을 재학습하지 않고 안전한 행동을 만들어 내는 필터 구조가 제안되기도 하였다[32].

4. L3 계층: 정책 학습과 궤적 최적화

L3 계층은 위험 영역을 사전에 고려하여 목표를 달성하는 경로와 제어 정책을 유기적으로 학습하고 최적화하는 중간 주기 단계다. 이 계층에서는 제약 마르코프 결정 과정(CMDP: Constrained Markov Decision Process)에 기반한 강화학습과 라그랑주 완화 기법을 활용하여 궤적 최적화 학습 과정에 안전 위반 비용을 반영한다[33]. 아울러 Safe MPC(Model Predictive Control)와 정합 예측(Conformal Prediction)을 활용하여 불확실성을 기반으로 안전 궤적을 생성하

기도 한다[31,34].

이러한 제약 기반 학습은 복잡한 동적 환경에서 로봇이 효율적인 경로를 찾도록 돕지만, 제약 조건이 주로 기댓값 수준의 통계적 보장에 머문다는 한계를 지닌다. 따라서 개별 운용 시 발생하는 돌발적인 안전 위반을 완벽히 차단하지는 못하며 보상 해킹이나 학습 분포 이탈에 취약하다는 약점이 있다.

5. L4 계층: 작업 계획과 환경 모니터링

L4 계층은 사람의 의도와 이동 경로를 예측하고 작업 절차 수준에서 위험을 선제적으로 회피하는 최상위 물리 안전 계층이다. 인간의 단기 궤적 및 의도 예측 모델, 점유 격자 분석, 그리고 세이프가드 영역 설정 등을 통해 사람과 로봇의 시공간적 접촉을 미리 방지하는 기능을 수행한다.

이 계층은 인간-로봇 물리적 상호작용(pHRI)과 밀접하며 국제 표준에 따른 속도 및 이격 감시 등의 운용 모드와 연계되어 작동한다. 현장 적용성 측면에서 SSM은 작업자와의 거리에 따라 속도를 동적으로 조절하는 가장 보편적인 안전 구현 수단으로 활용된다[35-38].

L4 계층의 예측은 시야 가림이나 사람의 돌발 행동 등으로 인해 오류가 발생하기 쉽기 때문에 보장 수준이 제한적이다. 따라서 이 상위 계층은 직접적인 안전 제어를 수행하기보다 작업 맥락과 사람의 상태 이해를 바탕으로 하위 제어 계층에 필요한 제약 단서 데이터와 안전 기준 정보를 전달하는 안전 나침반 역할을 담당한다.

IV. 파운데이션 모델 정렬과 가드레일

이 장에서는 로봇 파운데이션 모델(RFM)이 휴머노이드 로봇의 제어에 도입되면서 새롭게 대두된 위협 유형을 분석하고 이를 훈련과 추론 시점에 걸쳐 방어하기 위한 정책 정렬과 가드레일 기술 동향을 살펴본다.

1. 위협 유형

VLA 제어 체계에서의 안전 연구 절차는 다음과 같다. 위협과 방어의 시점을 각각 '훈련 시점'과 '추론 시점'으로 나누고 공격, 방어, 평가, 배치의 관점으로 세분화한다.

- **훈련 시점 공격:** 인지와 행동 생성이 하나의 거대 신경망으로 통합되면서, 훈련 데이터에 백도어(Backdoor)를 삽입해 특정 조건에서 로봇의 궤적 자체를 조작하는 공격이 가능해졌다. 일례로 동작을 청크(Chunk) 단위로 실행하는 사이에 시각 관측이 끊기는 구간을 노린 SilentDrift 연구는 2% 미만의 오염 데이터만으로도 VLA로부터 위험 행동을 유발하는 공격을 93.2% 성공할 수 있음을 보였다[39].
- **추론 시점 공격:** 로봇을 제어하는 언어 모델의 안전 제약을 자동으로 우회하는 RoboPAIR 탈옥(Jailbreak) 기법은 자율주행 시스템, 지상 이동 로봇, 상용 사족보행 로봇을 대상으로 한 실험에서 최대 100%의 공격 성공률을 보였다[40].
- **행동 청크 출력의 위험성:** 확산 트랜스포머(DiT: Diffusion Transformer)나 플로우 매칭(Flow Matching) 기반 VLA는 일반적으로 청킹(Chunking) 기법을 통해 수행할 일련의 행동 시퀀스를 한 번에 추론하여 출력한다. 따라서 다음 한 스텝의 관절 명령만을 단편적으로 검사하는 기존 방식으로는 청크 후반부에 발생하는 물리적 충돌을 사전에 포착하기 어렵다.

2. 의미 가드와 무결성 모니터

하위 단계의 동역학적 안전 필터는 물리적 충돌을 차단할 수는 있지만, 개방형 환경에서 주어지는 “칼을 아이 쪽에 건네라”와 같은 본질적으로 위험하거나 부적절한 지시 자체를 거절하지는 못한다. 이에 따라 물리적 궤적 생성 이전 단계에서 개입하는 상위 가드레일 기술이 요구된다.

의미 가드(Semantic Guard)는 자연어 명령의 맥락을 분석하여 유해 지시, 작업 권한 밖의 요구, 위험 물체 취급 지시를 실행하기 전에 거절한다. 언어 모델 영역에서는 규범 원칙 기반의 합성 데이터로 위반 판별 분류기를 학습시켜 탈옥 성공률을 86%에서 5% 미만으로 대폭 낮춘 연구가 보고되었다[41]. 로봇 계획 단계에서는 자연어 안전 규칙을 선형 시제 논리(Linear Temporal Logic)와 같은 정형화된 언어로 변환하여 로봇의 작업 계획이 안전 규범을 위반하는지 실행 전에 자동 검증하고 차단하는 RoboGuard 등의 기술이 개발되었다[42].

Google DeepMind의 AutoRT는 로봇 헌법(Robot Constitution)을 작업 생성과 검토에 적용한 사례이다. 이 헌법은 아시모프의 로봇 3원칙에서 영감을 얻은 기본 규칙, 사람, 동물, 날카로운 물체 등을 다루는 작업의 제한 규칙, 탑재 중량 등 물리적 제약을 자연어로 명시하여 LLM이 부적절한 작업을 제안하거나 선택하는 것을 줄인다[43]. 이는 프롬프트 기반의 의미적 가드레일로, 모델의 가중치를 수정하는 안전정렬과 구분된다. AutoRT는 관절 힘 임계치에 따른 자동 정지와 인간 감독자의 물리적 정지 스위치를 병행하여 다층 안전 구조를 구성하였다[43].

나아가 Sermanet 등은 실제 장면과 상해 사례에 기반한 데이터로 로봇 헌법을 자동 생성하고, 판단 오류를 바탕으로 규칙을 보완하는 방법을 제안하였다[44]. 사람이 해석하고 수정할 수 있는 규칙으로 로봇 행동을 관리한다는 장점이 있으나, 운용 환경과 기체 특성에 맞는 규칙의 설계 및 검증이 필요하다. 자연어 규칙 준수만으로 접촉력이나 전신 균형의 안전을 보장할 수 없으므로 하위 물리 안전 제어와 결합해야 한다.

NVIDIA의 Halos는 로봇용 하드웨어, 안전 운영체제, 안전 응용 및 검증 체계를 결합한 기능 안전 플랫폼이다. IGX Thor와 Halos Core를 기반으로 안전 기능을 구성하며, 외부 작업장 카메라를 활용하는 Outside-In Safety Blueprint를 통해 로봇의 주변 인식을 보완한다[45]. 이러한 구조는 학습 기반 제어와 별도로 안전 기능을 구성하고 검증하는 데 활용될 수 있다.

3. 불확실성 정량화와 런타임 검증

모델의 판단이 불확실할 때 무리하게 작업을 추측하여 실행하는 위험을 방지하기 위해 정책의 확신도를 실시간 수치로 환산하는 불확실성 정량화도 안전 확보에 활용된다.

대표적으로 SAFE 프레임워크는 정책 내부의 잠재 표현을 점수화하여 과업 종류와 무관하게 실패 징후를 조기 감지한다[46]. 또한, 정합 예측(Conformal Prediction)을 접목한 KnowNo 등의 기술은 유한 표본에서 정답이 포함될 확률을 수학적으로 보증하는 후보 집합을 구성하고, 집합의 불확실성이 임계치 이상으로 높아질 경우 로봇이 행동을 멈추고 인간 관리자에게 확인 질의를 던져 명확한 지시를 기다리도록 설계되었다[47].

다만, 현재의 가드레일과 실패 탐지 체계는 주로 경험적 임곗값에 의존하고 있어 엄밀한 안전 보증 기준을 실시간 제어에 결합하는 연구는 여전히 진행형이다.

4. RFM 안전정렬기술

RFM 안전정렬은 작업 수행 과정에서 위험한 행동이 생성될 가능성을 줄이도록 모델의 학습 목표와 행동 정책을 조정하는 기술이다. 앞 절의 가드레일과 연계하여 모델 자체에 안전 요구사항을 반영하는 역할을 담당한다. 최근에는 VLA를 중심으로 안전 제약 학습, 사람의 교정 피드백 활용, 지시 타당성 검증, 선택적 망각 등의 방법이 제안되고 있다[48-51].

첫째는 안전 요구사항을 학습 제약으로 반영하는 방식이다. SafeVLA는 안전 문제를 제약 마르코프 결정 과정(CMDP)으로 정의하고 위험 상황에서 수집한 행동과 안전 위반 비용을 활용하여 VLA 정책을 안전 강화학습으로 정렬한다. 사각지대, 깨지기 쉬운 물체, 위험 장비 등을 포함하는 Safety CHORES 평가 환경에서 기존 방법 대비 누적 안전 비용을 83.58% 줄이고 작업 성공률을 3.85%p 높여 안전성과 작업 성능을 함께 개선할 가능성을 보였다[48].

둘째는 사람의 교정 행동을 정책 개선에 활용하는 방식이다. 행동 선호 최적화(APO: Action Preference Optimization)는 로봇 실행 중 사람이 개입해 수정한 궤적으로부터 행동의 선호 정보를 추출한다. 이를 바탕으로 실패를 유발하는 행동의 생성 가능성은 낮추고 교정 행동은 강화하여 정책을 갱신한다. 이는 사전에 수집한 시연 데이터에 의존하는 학습을 현장 피드백으로 보완하는 방법이며, 사람의 개입을 통해 작업 수행의 신뢰성을 높이는 데 활용될 수 있다[49].

셋째는 지시 타당성을 모델이 검증하도록 학습하는 방식이다. IVA(Instruct-Verify-and-Act)는 환경에 없는 물체나 성립하지 않는 조건을 포함하는 실행 불가능한 지시를 거절하고 실행 가능한 지시로 수정하도록 학습한다. 올바른 지시와 잘못된 지시를 함께 제공하여 시각 정보와 언어 지시의 일치 여부를 스스로 판단하도록 훈련하며 이를 통해 잘못된 지시를 무리하게 실행하는 문제를 줄일 수 있다. 단, 지시의 실행 가능성 판정만으로 물리적 안전이 보장되는 것은 아니다[50].

넷째는 이미 학습된 위험 행동을 선택적으로 제거하는 방식이다. VLA-Forget은 시각 인코더, 시각 언어 연결부, 행동 생성 계층에 분산된 지식을 함께 고려하여 불안전하거나 개인정보와 관련된 행동을 망각하도록 모델을 수정한다. 제거 대상 행동의 억제와 정상적인 시각 인식 및 작업 수행 능력의 유지를 동시에 학습 목표로 설정한다. 이는 모델 전체를 다시 학습하지 않고 문제가 되는 행동을 수정하기 위한 접근이다[51].

이러한 안전정렬 기술은 모델이 위험한 행동을 생성하는 빈도를 줄이는 데 이바지할 수 있으나, 학습 데이터와 평가 환경을 벗어난 상황까지 안전을 보장하지는 못한다. 특히 휴머노이드 로봇에 적용하려면 전신 균형, 사람과의 접촉, 장기 작업 절차에서 발생하는 위험을 학습과 평가에 함께 반영해야 한다. 따라서 모델 내재형 안전정렬을 하위 동역학적 안전 필터 및 실행 단계 가드레일과 결합하고, 모델 갱신 이후에도 안전 위반과 작업 성능을 함께 검증하는 체계가 필요하다.

V. 안전 평가 체계와 표준화 동향

앞서 살펴본 바와 같이 파운데이션 모델 제어와 휴머노이드 로봇의 결합은 기존의 하드웨어 중심 안전 제어를 넘어서는 새로운 기술적 난제를 던지고 있다. 이 장에서는 이러한 기술적 공백을 극복하기 위해 소프트웨어 평가 지표, 데이터셋 규격, 그리고 하드웨어 시스템 인증을 아우르는 최신 평가 체계와 국제 표준화 현황을 종합하여 살펴본다.

1. 안전 평가지표와 벤치마크

지금까지 로봇 AI 정책을 평가하는 핵심 지표는 작업을 완수했는가를 따지는 '성공률(Success Rate)'이었다. 그러나 성공적인 작업 궤적 이면에도 위험한 행동이 숨어있을 수 있다는 사실이 입증되면서, 평가의 축을 다변화하는 새로운 벤치마크들이 등장하고 있다.

최근에는 작업 성공률과 안전 위반 지표를 함께 보고하는 평가 방식이 확산하고 있다. SafeVLA-Bench는 전체 시행 중 작업에는 성공했으나 안전 제약을 위반한 시행의 비율인 SBU(Succ-But-Unsafe)를 제시하여 성공과 안전의 괴리를 측정한다[19]. ForesightSafety-VLA는 13개 안전 범주를 바탕으로 기본 안전 제약과 언어 · 시각 변화에 대한 취약성을 진단하며[52], LIBERO-Safety는 물리적 안전과 의미적 안전을 함께 평가한다[53].

SafeAgentBench는 작업 계획 단계에서 위험 지시를 거부하는 능력을 평가하기 위해 수백 건의 안전 · 위험 과업을 제시하고 모델의 유해 지시 거부율(Rejection Accuracy)을 핵심 성능 지표로 측정한다[54].

Google DeepMind의 ASIMOV는 로봇의 행동이 안전하고 바람직한지 판단하는 능력을 평가하는 의미적 안전 벤치마크로서 실세계 장면과 병원 상해 보고서를 바탕으로 구성한 로봇 헌법을 파운데이션 모델의 프롬프트에 적용하여 얼마만큼 사람과 유사하게 안전을 판단할 수 있는지 평가한다[44]. ASIMOV v2는 상해 서술과 운용 제약을 이미지 · 영상 시나리오로 확장하여 위험 인지, 안전 추론 및 개입 판단 능력을 평가한다[55]. 이러한 평가는 작업 실행 전의 위험 상황 판단을 검증하는 데 유용하지만 휴머노이드의 전신 접촉력과 전도 위험을 직접 측정하는 실물 시험을 대체하지는 않는다.

2. 휴머노이드 특화 데이터셋 표준화

신뢰할 수 있는 모델 정렬과 평가를 위해서는 고품질의 안전 데이터셋 확보가 필수적이나 현재 공개된 접촉력 기반 실측 데이터는 대부분 단일 로봇 팔에 편중되어 있다. 휴머노이드 로봇이 환경과 상호작용을 하며 발생하는 전신 다접촉이나 인체와의 의도된 물리적 접촉 상황을 실측한 공개 데이터셋은 제한적이며 표준 허용 기준의 근거로 활용할 수 있는 체계적인 실측 자료가 부족하다. 이는 향후 접촉력 한곗값 등 표준 허용 기준을 제정할 근거 자료의 부재로 직결된다.

이러한 공백을 메우기 위해 데이터 규격 자체의 표준화 논의가 최근 본격화되었다. 국제표준화기구(ISO) 기술위원회 TC299(로보틱스) 산하에 '휴머노이드 로봇 데이터셋'을 다루는 신규 작업반(WG16)이 신설되어 AI 기반 휴머노이드의 안전성과 신뢰성을 확보하기 위한 데이터 기준 논의가 시작되었다[13].

3. 동적 안정 로봇 표준(ISO 25785-1)

기존 산업용 로봇의 기계적 안전을 다루는 ISO 10218-1/-2 시리즈는 2025년에 개정판이 발간되었으나 이동성과 능동 균형 유지에 따른 위험을 포괄하는 데 한계가 있으며 인공지능이나 기계학습에 관한 조항을 포함하지 않는다[25,26]. 따라서 인간과 공간을 공유하며 능동적으로 직립 보행하는 휴머노이드 로봇을 온전히 포괄할 수 없다.

이에 ISO TC299 산하 WG12 작업반은 휴머노이드 로봇을 포함한 동적 안정 산업용 이동 로봇을 포괄하는 새로운 안전 표준인 ISO 25785-1 프로젝트를 출범시켰다[56]. 정적 안정 로봇과 달리 동적 안정형 로봇은 전원이 차단되거나 제어에 실패할 경

우 기체가 넘어지는 전도 위험이 핵심 안전 문제이다. 현재 17개국 전문가들이 모여 전원 차단 시 안전 상태 정의 방안 등을 논의하고 있으며, 해당 표준은 36개월의 개발 일정을 거쳐 2028년경 발간을 목표로 하고 있다[13].

4. 글로벌 규제 대응과 국내 인증 인프라 구축

글로벌 규제 당국은 실용화를 앞둔 차세대 로봇에 대해 강력한 사전 및 사후 안전 검증을 요구하고 있다. 2027년 1월부터 적용되는 유럽연합의 기계규정(Machinery Regulation 2023/1230)은 기계학습을 기반으로 자가 진화하는 안전 부품과 내장 시스템을 제3자 적합성 평가 대상으로 지정했다[10]. 나아가 EU AI법과 사이버복원력법은 조항별로 이미 단계적으로 적용되고 있다[57,58]. 상기 기계규정 부속서는 로봇의 출시 이후에도 학습 단계를 포함해 위험한 설정의 변경을 추적하고 이력을 보존하도록 요구하는 등 배포 후 갱신 관리를 명문화했다[10].

반면, 현재 국내의 법정 로봇 인증은 실외이동로봇의 운행 안전을 다루는 8개 심사 항목에 그쳐 실내 조작 작업과 사람과의 전신 물리적 접촉 안전을 다루는 데 부족함이 있다. 또한, 기존 운행안전인증의 정기검사와 별도로 휴머노이드 로봇 AI 모델의 지속 갱신에 따른 재평가 방안과 기준도 구체화할 필요가 있다[11].

한편, 정부는 대구시 국가로봇테스트필드 부지에 약 187억 원을 투입하여 2030년 완료를 목표로 '휴머노이드 로봇 안전인증센터' 구축을 진행 중이다[14]. 이를 통해 AI 신뢰성과 사이버 보안을 아우르는 안전 인증 환경 확보를 목표하고 있다.

다가올 휴머노이드 로봇 상용화 관련 해외 규제에 합법적이고 안전하게 대응하기 위해 산업계와 연구계는 휴머노이드 로봇 안전 시험 인프라 구축과 국제 표준 제안 논의를 현시점부터 선제적으로 이끌어갈 필요가 있다.

VI. 결론

본고는 학습형 휴머노이드 로봇의 안전 기술을 물리적 안전, 파운데이션 모델 정렬과 가드레일, 평가와 표준의 세 부분으로 나누어 살펴보았다. 현재 로봇 안전 기능은 하위 반사 계층과 최상위 숙고 계층에서 상호 연관성 없이 파편적으로 연구개발되고 있다. 검증 가능한 근거를 바탕으로 RFM 기반 로봇 정책 모델 출력의 안전을 실시간 보장하는 휴머노이드 로봇 안전 제어 아키텍처는 아직 확립되지 않은 상태이다. 휴머노이드 로봇 보급과 시장 창출을 위해 시급하게 이 기술적 공백을 채워 나갈 필요가 있다. 아울러 자율주행 자동차 분야처럼 안전 평가 방식이 단일 시행 측정에서 행동 분포에 대한 통계적 판정으로 옮겨가고 있는 만큼 새로운 안전 지표를 발굴하고 산업 인증에 도입하는 작업이 필요하다.

휴머노이드 로봇의 상용화는 임박하였으나 안전 기술과 표준 인증 체계는 국제적으로 미진한 상황이다. 발 빠르게 휴머노이드 로봇 안전 기술을 개발하고 이를 기반으로 국제 사회에 안전 기준과 안전 보장 프레임워크를 선제적으로 제안하여 주도해 나간다면 휴머노이드 로봇 기술 주권 확보는 물론 세계 휴머노이드 로봇 시장을 선도할 수 있을 것이다.

참고문헌

[1] BMW Group, “BMW Group Advances the use of Physical AI in Production with Figure 03 Project in Spartanburg,” BMW Group PressClub Global, 2026. 6. 25.

[2] 권숙희, “中휴머노이드로봇 100대, 한 공장에 ‘취업’…대규모 상용화 첫발,” 연합뉴스, 2025. 8. 13. https://www.yna.co.kr/view/AKR20250813058700009

[3] Hyundai Motor Group, “Hyundai Motor Group Announces AI Robotics Strategy to Lead Human-Centered Robotics Era at CES 2026,” 2026. 1. 6.

[4] Boston Dynamics, “Boston Dynamics & Hyundai Motor Group Expand Collaboration to Drive Mobility Manufacturing & Innovation,” 2025. 4. 3.

[5] M.J. Park, “Toyota to Deploy 400,000 Humanoid Robots at Plants Worldwide,” Seoul Economic Daily, 2026. 9. 19. https://en.sedaily.com/international/2026/09/19/toyota-to-deploy-400000-humanoid-robots-at-plants-worldwide

[6] N.E. Sanders et al., “Robot-related Injuries in the Workplace: An analysis of OSHA Severe Injury Reports,” Appl. Ergon., vol. 121, 2024, doi: 10.1016/j.apergo.2024.104324

[7] 김정훈, “사람을 박스로 인식한 로봇…40대 노동자 깔려 숨졌다,” 경향신문, 2023. 11. 8. https://v.daum.net/v/qxyuusBESi

[8] IEEE Humanoid Study Group, “A Pathway Study for Future Humanoid Standards,” 2025. 9.

[9] 백승일, “피규어AI, ‘로봇이 두개골 골절시킬 수 있다’고 경고한 안전 책임자에게 피소,” 로봇신문, 2025. 11. 24. https://robotnews.tistory.com/m/16793

[10] European Union, “Regulation (EU) 2023/1230 on Machinery,” Off. J. Eur. Union, L 165, 2023; “Corrigendum to Regulation (EU) 2023/1230,” L 169, 2023.

[11] 산업통상자원부, “실외이동로봇 운행안전인증 절차 및 기준 등에 관한 고시,” 산업통상자원부 고시 제2023-211호, 2023. 11. 17.

[12] Fraunhofer IPA, “Fraunhofer IPA Develops Standardized Analyses for Application-relevant Criteria of Humanoid Robots,” Fraunhofer IPA, 2026. 5. 27.

[13] 신영빈, “‘휴머노이드 안전기준 만든다’…ISO 국제표준 논의 본격화,” ZDNet Korea, 2026. 2. 16. https://zdnet.co.kr/view/?no=20260214162213

[14] 정재훈, “대구시, 휴머노이드 로봇 안전인증센터 구축사업 본궤도,” 전자신문, 2026. 8. 5. https://www.etnews.com/20260805000289

[15] 오예진, “2030년 완전 자율주행차 양산·휴머노이드 연 1천대씩 생산,” 연합뉴스, 2025. 9. 10. https://www.yna.co.kr/view/AKR20250910038800003

[16] P. Chemweno et al., “Orienting Safety Assurance with Outcomes of Hazard Analysis and Risk Assessment: A Review of the ISO 15066 Standard for Collaborative Robot Systems,” Saf. Sci., vol. 129, 2020, doi: 10.1016/j.ssci.2020.104832

[17] A. Giallanza et al., “Occupational Health and Safety Issues in Human-Robot Collaboration: State of the Art and Open Challenges,” Saf. Sci., vol. 169, 2024, doi: 10.1016/j.ssci.2023.106313

[18] C. Ding et al., “Toward Certified Functional Safety for Industrial Humanoid Robots: The Fail-Passive Gap and a Feasibility Study,” arXiv preprint, 2026, doi: 10.48550/arXiv.2608.02809

[19] J. Fan et al., “SafeVLA-Bench: A Benchmark for the Success-Safety Gap in Vision-Language-Action Models,” arXiv preprint, 2026, doi: 10.48550/arXiv.2606.00773

[20] S. Haddadin et al., “Robot Collisions: A Survey on Detection, Isolation, and Identification,” IEEE Trans. Robot., vol. 33, no. 6, 2017, pp. 1292-1312, doi: 10.1109/TRO.2017.2723903

[21] A. Albu-Sch ffer et al., “The DLR Lightweight Robot: Sesign and Control Concepts for Robots in Human Environments,” Ind. Robot, vol. 34, no. 5, 2007, pp. 376-385, doi: 10.1108/01439910710774386

[22] S. Haddadin et al., “Requirements for Safe Robots: Measurements, Analysis and New Insights,” Int. J. Robot. Res., vol. 28, no. 11-12, 2009, pp. 1507-1527, doi: 10.1177/0278364909343970

[23] S. Haddadin et al., “On Making Robots Understand Safety: Embedding Injury Knowledge into Control,” Int. J. Robot. Res., vol. 31, no. 13, 2012, pp. 1578-1602, doi: 10.1177/0278364912462256

[24] ISO, “Robots and Robotic Devices—Collaborative Eobots,” ISO/TS 15066:2016, 2016.

[25] ISO, “Robotics—Safety Requirements—Part 1: Industrial Robots,” ISO 10218-1:2025, 3rd ed., 2025.

[26] ISO, “Robotics—Safety Requirements—Part 2: Industrial Robot Applications and Robot Cells,” ISO 10218-2:2025, 2nd ed., 2025.

[27] A. De Luca et al., “Collision Detection and Safe Reaction with the DLR-III Lightweight Manipulator Arm,” in Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst., (Beijing, China), 2006, pp. 1623-1630, doi: 10.1109/IROS.2006.282053

[28] A. De Luca and R. Mattone, “Sensorless Robot Collision Detection and Hybrid Force/Motion Control,” in Proc. IEEE Int. Conf. Robot. Autom., (Barcelona, Spain), 2005, pp. 999-1004, doi: 10.1109/ROBOT.2005.1570247

[29] B. Wu et al., “Proximity Sensing Electronic Skin: Principles, Characteristics, and Applications,” Adv. Sci., vol. 11, no. 13, 2024, doi: 10.1002/advs.202308560

[30] A.D. Ames et al., “Control Barrier Functions: Theory and Applications,” in Proc. European Control Conference, 2019, pp. 3420-3431, doi: 10.23919/ECC.2019.8796030

[31] K.P. Wabersich et al., “Data-Driven Safety Filters: Hamilton-Jacobi Reachability, Control Barrier Functions, and Predictive Methods for Uncertain Systems,” IEEE Control Syst. Mag., vol. 43, no. 5, 2023, pp. 137-177, doi: 10.1109/MCS.2023.3291885

[32] M. T lle et al., “Towards Safe Robot Foundation Models,” arXiv preprint, 2025, doi: 10.48550/arXiv.2503.07404

[33] E. Altman, Constrained Markov Decision Processes, Chapman & Hall/CRC, 1999.

[34] L. Lindemann et al., “Safe Planning in Dynamic Environments Using Conformal Prediction,” IEEE Robot. Autom. Lett., vol. 8, no. 8, 2023, pp. 5116-5123, doi: 10.1109/LRA.2023.3292071

[35] P.A. Lasota et al., “A Survey of Methods for Safe Human-Robot Interaction,” Found. Trends Robot., vol. 5, no. 4, 2017, pp. 261-349, doi: 10.1561/2300000052

[36] F. Flacco et al., “A Depth Space Approach to Human-Robot Collision Avoidance,” in Proc. IEEE Int. Conf. Robot. Autom. (ICRA), (Saint Paul, MN, USA), May 2012, pp. 338-345, doi: 10.1109/ICRA.2012.6225245

[37] O. Adamides et al., “Perception and Computation for Speed and Separation Monitoring Architectures,” Robotics, vol. 14, no. 4, 2025, doi: 10.3390/robotics14040041

[38] A. Zacharaki et al., “Safety Bounds in Human Robot Interaction: A Survey,” Saf. Sci., vol. 127, 2020, doi: 10.1016/j.ssci.2020.104667

[39] B. Xu et al., “SilentDrift: Exploiting Action Chunking for Stealthy Backdoor Attacks on Vision-Language-Action Models,” in Findings Assoc. Comput. Linguist.: ACL 2026, (San Diego, CA, USA), 2026, pp. 34570-34582, doi: 10.18653/v1/2026.findings-acl.1725

[40] A. Robey et al., “Jailbreaking LLM-Controlled Robots,” arXiv preprint, 2024, doi: 10.48550/arXiv.2410.13691

[41] M. Sharma et al., “Constitutional Classifiers: Defending against Universal Jailbreaks across Thousands of Hours of Red Teaming,” arXiv preprint, 2025, doi: 10.48550/arXiv.2501.18837

[42] Z. Ravichandran et al., “Safety Guardrails for LLM-Enabled Robots,” IEEE Robot. Autom. Lett., vol. 11, no. 4, 2026, pp. 4649-4656, doi: 10.1109/LRA.2026.3667488

[43] M. Ahn et al., “AutoRT: Embodied Foundation Models for Large Scale Orchestration of Robotic Agents,” arXiv preprint, 2024, doi: 10.48550/arXiv.2401.12963

[44] P. Sermanet et al., “Generating Robot Constitutions & Benchmarks for Semantic Safety,” in Proc. Conf. Robot Learn., (Seoul, Republic of Korea), 2025, pp. 4767-4823.

[45] S.H. Sheshadri et al., “Inside NVIDIA Halos for Robotics: A Full-Stack Functional Safety System for Physical AI,” NVIDIA Technical Blog, 2026. 6. 22.

[46] Q. Gu et al., “SAFE: Multitask Failure Detection for Vision-Language-Action Models,” in Adv. Neural Inf. Process. Syst., vol. 38, 2025, pp. 44801-44836, doi: 10.52202/085713-1337

[47] A.Z. Ren et al., “Robots That Ask for Help: Uncertainty Alignment for Large Language Model Planners,” arXiv preprint, 2023, doi: 10.48550/arXiv.2307.01928

[48] B. Zhang et al., “SafeVLA: Towards Safety Alignment of Vision-Language-Action Model via Constrained Learning,” in Adv. Neural Inf. Process. Syst., vol. 38, 2025, pp. 170196-170234.

[49] W. Xia et al., “Human-assisted Robotic Policy Refinement via Action Preference Optimization,” in Adv. Neural Inf. Process. Syst., vol. 38, 2025, pp. 41281-41303.

[50] W.H. Hsieh et al., “Do What? Teaching Vision-Language-Action Models to Reject the Impossible,” arXiv preprint, 2025, doi: 10.48550/arXiv.2508.16292

[51] R. Ranjan and A. Polyzou, “VLA-Forget: Vision-Language-Action Unlearning for Embodied Foundation Models,” arXiv preprint, 2026, doi: 10.48550/arXiv.2604.03956

[52] M. Lyu et al., “ForesightSafety-VLA: A Unified Diagnostic Safety Benchmark for Vision-Language-Action Models,” arXiv preprint, 2026, doi: 10.48550/arXiv.2606.27079

[53] R. Cui et al., “LIBERO-Safety: A Comprehensive Benchmark for Physical and Semantic Safety in Vision-Language-Action Models,” arXiv preprint, 2026, doi: 10.48550/arXiv.2606.23686

[54] S. Yin et al., “SafeAgentBench: A Benchmark for Safe Task Planning of Embodied LLM Agents,” arXiv preprint, 2024, doi: 10.48550/arXiv.2412.13178

[55] A. Jindal et al., “Can AI Perceive Physical Danger and Intervene?,” arXiv preprint, 2025, doi: 10.48550/arXiv.2509.21651

[56] ISO, “Robotics—Safety requirements for dynamically stable industrial mobile robots (legged, wheeled, or other forms of locomotion)—Part 1: Robots,” ISO/CD 25785-1, Committee Draft, 2026.

[57] European Union, “Regulation (EU) 2024/1689 Laying Down Harmonised Rules on Artificial Intelligence,” Off. J. Eur. Union, L 2024/1689, 2024. 7. 12.

[58] European Union, “Regulation (EU) 2024/2847 on Horizontal Cybersecurity Requirements for Products with Digital Elements,” Off. J. Eur. Union, L 2024/2847, 2024. 11. 20.