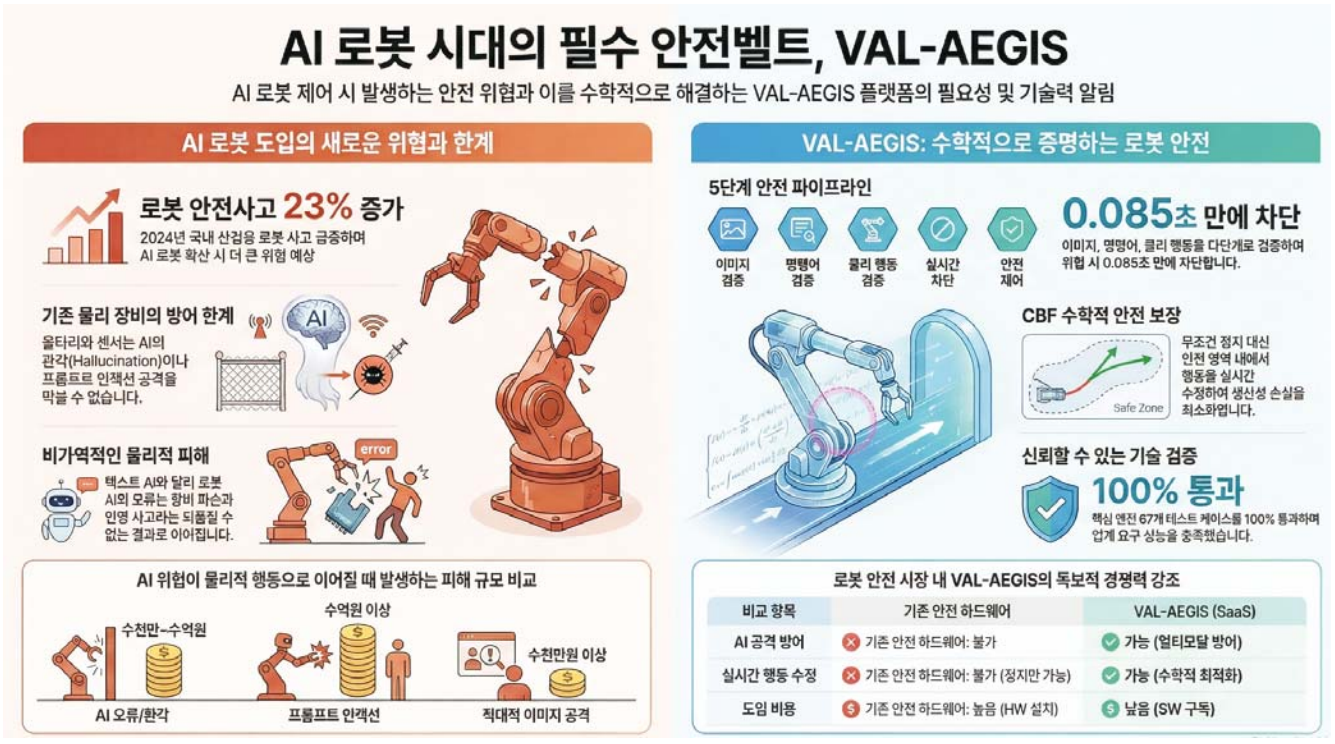
로봇 AI 안전, 한국 스타트업이 해결



〈그래픽=야타브 · 오나이제공〉

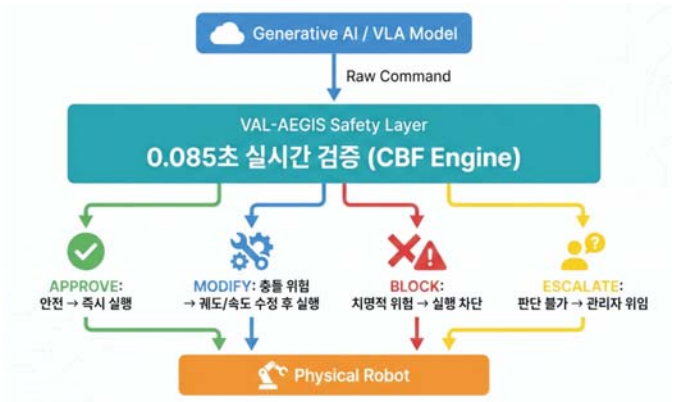
야타브 · 오나이, MWC 2026서 안전망 오픈소스로 공개

한국 스타트업 야타브(YATAV)와 오나이(ONAI)가 MWC 2026에서 로봇 AI 안전을 위한 안전망을 공개한다고 밝혔다.

NVIDIA를 비롯한 글로벌 빅테크들의 '월드모델(World Model)' 기반 로봇 AI 개발이 가속화되는 가운데, 업계에서는 안전 문제를 지적하고 있다. 챗봇이 잘못된 답변을 하면 정보가 왜곡되지만, 로봇이 잘못 움직이면 사람이 다치기 때문이다.

이에 야타브와 오나이가 공개한 시뮬레이션 데모는 해당 위험을 생생하게 보여준다. 공격자가 로봇의 카메라 입력에 육안으로 보이지 않는 미세한 노이즈를 주입하면, 로봇은 관절이 허용 범위를 초과해 비정상적으로 뒤틀리거나, 갑자기 빠른 속도로 움직이며 '순간이동' 하는 현상을 보인다. 자기회귀(Autoregressive) 추론 과정에서 오류가 기하급수적으로 증폭되기 때문이다.

NVIDIA 로봇 시뮬레이션 플랫폼에 특화
추론 전후 단계 작동 5중 가드 시스템 구성
모든 공격 실시간 탐지, 안전하게 정지
야타브, Apache 2.0 라이선스 오픈소스 공개
Rust 언어로 개발 메모리 안전성 등 확보
글로벌 연구 커뮤니티와 AI 안전 표준 제시
"인간의 책임 지킨다는 철학 갖고 기술 개발"

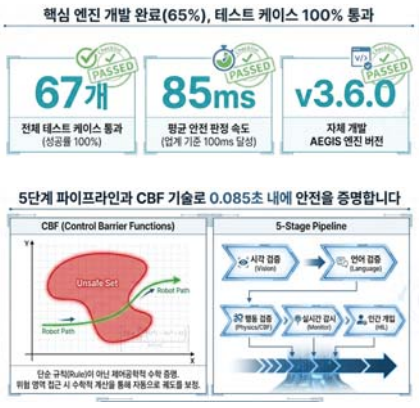


월드모델은 로봇이 카메라 영상과 동작 명령을 기반으로 '다음에 무슨 일이 일어날까?'를 예측하는 AI 기술이다. 실제 환경에서 시행착오를 겪지 않고도 시뮬레이션으로 안전하게 학습할 수 있어, 차세대 로보틱스의 핵심으로 주목받는다. 하지만 이 시스템은 적대적 공격(Adversarial Attack)에 구조적으로 취약하다는 게 업계의 설명이다.

보고서에 따르면, 월드모델 기반 로봇 시스템에는 'ΔNaN 주입 공격: 로봇 관절 명령에 'Not a Number' 값을 삽입 → 로봇 완전히 작동 불능 Δ관절 한계 위반: 물리적으로 불가능한 각도 명령 → 로봇 관절 손상 또는 주변 인명 피해 Δ속도 급변 공격: 순간적으로 극단적 속도 변화 → 예측 불가능한 움직임으로 충돌 사고 Δ자기회귀 드리프트: 시간이 지나며 누적되는 미세 오류 → 점진적 궤도 이탈로 작업 실패 또는 사고'라는 4가지 치명적인 공격 시나리오가 존재한다.

특히 위험한 것은 자기회귀 특성이다. 월드모델은 이전 프레임의 예측을 다음 프레임 입력으로 사용한다. 한 번의 작은 오류가 다음 단계에서 증폭되고, 그 오류가 또 다음 단계에 영향을 주며 눈덩이처럼 커진다. 이런 위험에도 로봇 AI 안전 기준은 아직 존재하지 않는 상황이다.

야타브와 오나이가 이번에 공개한 'AEGIS DreamDojo Guard SDK'는 세계 최초의 로봇 월드모델 전용 안전 프레임워크다. NVIDIA의 로봇 시뮬레이션 플랫폼 'DreamDojo'에 특화되어 있으며, 로봇 AI 추론 전후 단계에서 작동하는 5중 가드 시스템으로 구성된다.



야타브와 오나이가 공개한 브라우저 기반 3D 시뮬레이션에서, 보호되지 않은 로봇은 NaN 주입 공격으로 완전히 작동 불능 상태가 됐다. 관절 한계 위반 시 비정상적으로 뒤틀렸고, 속도 급변 공격에는 순간이동 현상을 보였다.

반면 AEGIS로 보호된 로봇은 모든 공격을 실시간으로 탐지하고 안전하게 정지했다. 관절은 안전 범위 내로 제한됐고, 속도 변화는 부드럽게 보간 처리됐으며, 자기회귀 드리프트는 실시간으로 보정됐다.

또한, 야타브는 Apache 2.0 라이선스 오픈소스를 공개했다. 야타브 강석주 CISO는 Rust 언어로 AEGIS를 개발해 메모리 안전성과 고성능을 확보했다. Apache 2.0 라이선스는 Fourier GR1, Unitree G1, Galaxea YAM, AgiBot 등 주요 로봇 플랫폼을 지원하며, 커스텀 로봇 설정도 가능하다. 단 3줄의 코드로 간편하게 로봇 동작 안전을 검증할 수 있는 것이 특징이다.

야타브 김광일 CTO는 "AEGIS는 안전은 특권이 아니라 기본권이라는 믿음 아래, AI 안전 위험을 병행 연구하고 해결해 나가는 초격차 전용 안전 프레임워크"라고 밝혔다.

야타브는 오픈소스 공개를 통해 글로벌 연구 커뮤니티와 함께 임바디드(Embodied) AI 안전 표준을 만들어가겠다는 비전을 제시했다.

오픈소스 SDK 외에도, 웰포드(Welford) 스트리밍 분산 분석, 피어슨 상관 분석, 스펙트럼 분석 등 고급 통계 기법을 탑재한 AEGIS Pro 상용 버전도 개발 중에 있다.

오나이(ONAI) 조인선 CEO는 "서비스와 기능 적용할 생태계 구축이 필요하다"며, 차후 API/SDK 개발 후 배포해 다양한 환경에 활용 및 적용이 가능하도록 개발 중이다.

이성찬 CEO는 "2025년 MWC에서 SKT와 함께 글로모 어워드를 수상한 야타브는 2026년에도 MWC에서 혁신적인 안전망을 공개할 것이라며 "GitHub에 공개된 AEGIS SDK는 지금 이 순간에도 전 세계 개발자들에게 다운로드되고 있다. 'Powerful Technology, Human Accountability' 강력한 기술력, 인간의 책임을 지킨다'는 철학을 갖고 기술 개발에 주력하겠다"고 전했다.

/김재훈 기자