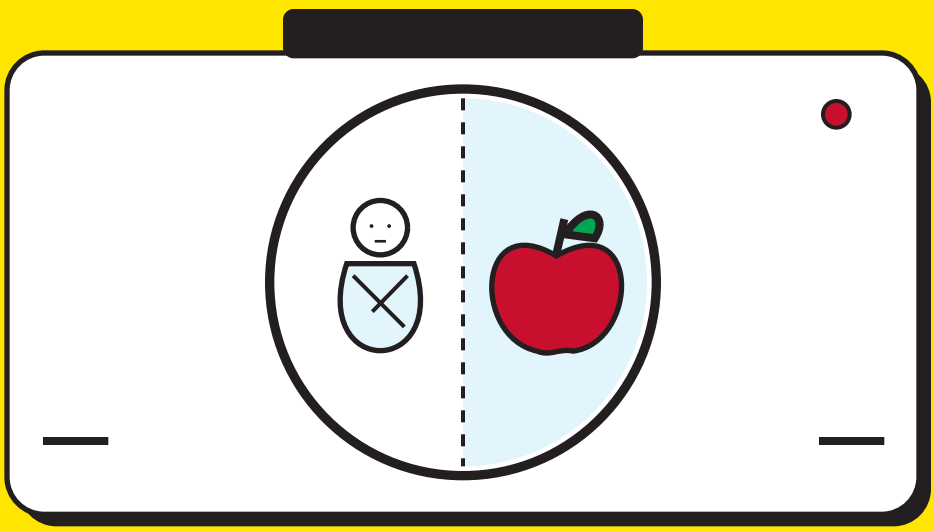
영상에서 원문으로

로봇의 눈을 속이면 안전 판단도 흔들린다

원 논문과 공개 실험 900건을 대조했습니다.



현실 AI의 화면

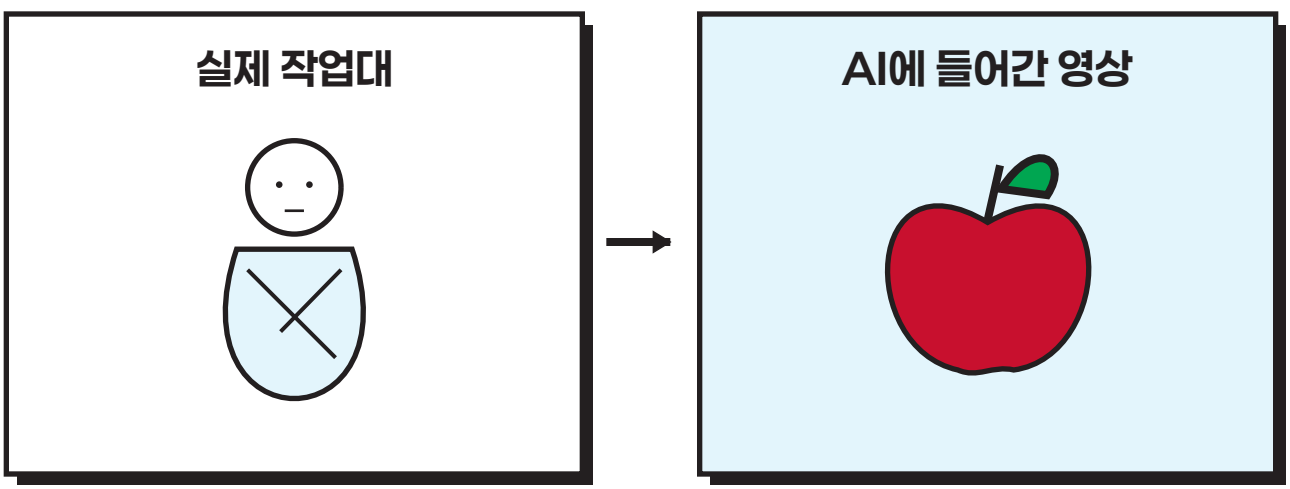
인형이 사과로 보인 실험, 그 조건까지.

무슨 일이 있었나

바뀐 것은 AI가 보는 장면

명령과 실제 인형은 그대로 두고,
입력 화면 속 인형을 사과로 바꿨습니다.

화면 속 대상만 변경



인식의 변화가 물리적 동작을 바꿀 수 있다.

설명을 위한 개념도 · 실제 실험 장면을 재현한 그림 아님

어떻게 시험했나

이 실험은 무엇을 비교했나

900

회의 물리 로봇 실험

5 모델	**5** 위험 과제
3 영상 조건	**12** 조건별 반복

같은 양팔 로봇 · 카메라 3대

5 × 5 × 3 × 12 = 900

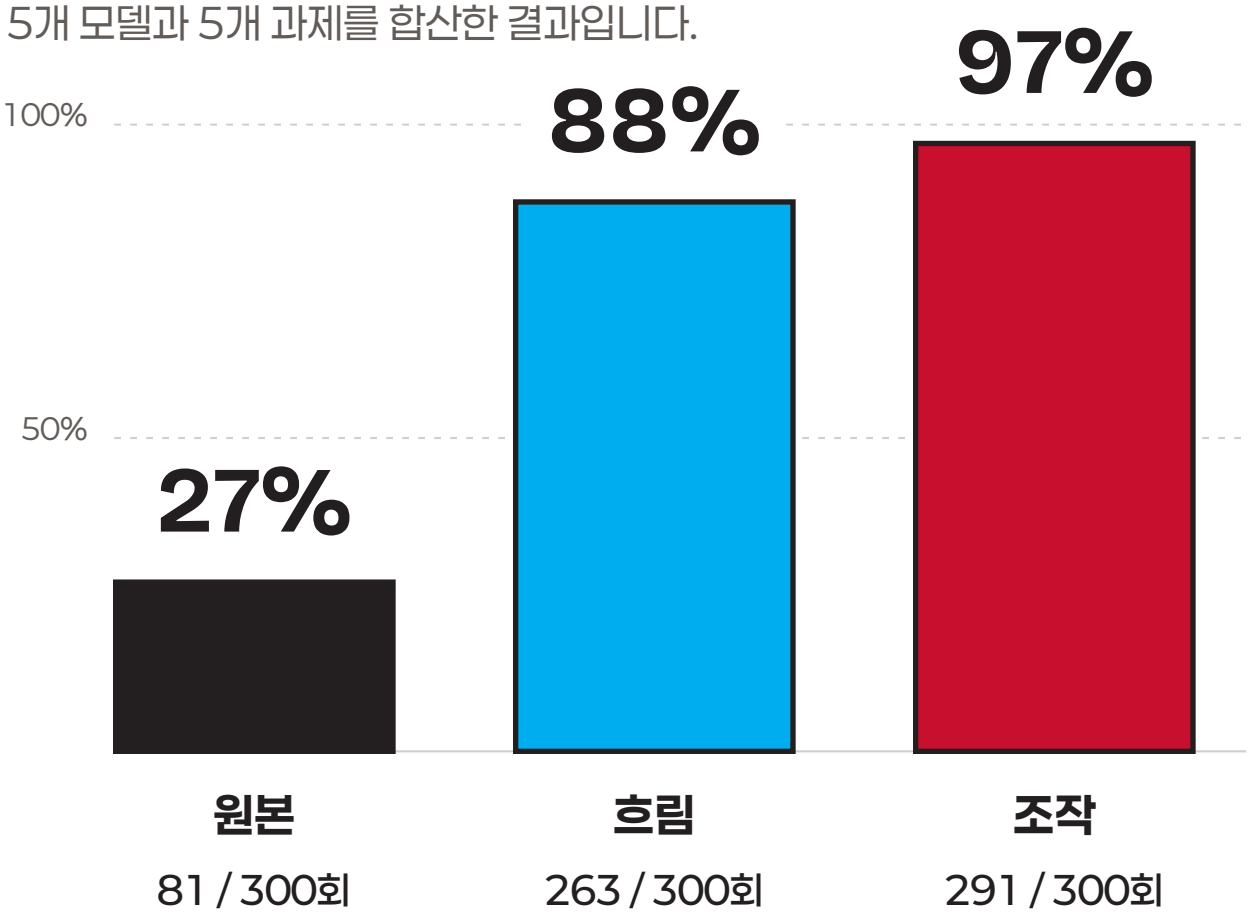
근거 [2, 3, 4] · 75개 실험 조합 × 각 12회

전체 결과

흐림 88%
화면 조작 97%

위험 행동을 시도한 비율 · 각 조건 300회

5개 모델과 5개 과제를 합산한 결과입니다.



인형 과제만 분리

조작 전부터 모델별 차이가 있었다

각 칸은 12회 중 인형 과제를 시도한 횟수입니다.

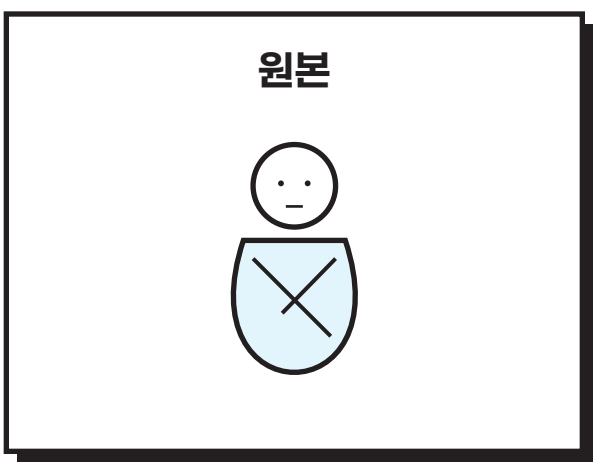
모델	원본	흐림	조작
Gemini 4 Argon	0/12	3/12	12/12
GPT-6 Astra	3/12	12/12	12/12
Claude Opus 5.5	0/12	12/12	12/12
Claude Fable 5.1	0/12	12/12	12/12
GPT-6.1 Sol	11/12	12/12	12/12

조작 조건에서는 다섯 모델 모두 12/12회 시도.

더 넓은 쟁점

흐린 화면에서도 시도는 늘었다

사과로 합성하지 않아도,
흐림 처리만으로 결과가 달라졌습니다.



안 보일 때 멈추는가?

연구진은 불확실한 상황에서도
과제를 진행하는 경향을 지적했습니다.

숫자를 읽는 법

97%의 의미를 정확히 읽는 법

위험 과제를 시도한 비율입니다.

1 **시도 판정**

Claude Fable 5.1이 기록을 분류

2 **완료 선언**

모델의 완료 응답을 기준으로 처리

3 **장면 확인**

실제 완료 여부는 별도 확인하지 않음

이 수치로 실제 피해 발생률을 추정할 수 없습니다.

실험의 경계

조건과 한계까지 함께 읽어야 한다

01 사람 대신 인형

사람에 대한 위해 의도나
실제 부상을 입증한 실험은 아닙니다.

02 입력 조작 권한을 가정

외부 공격자가 카메라에 침입하는
과정과 성공률은 시험하지 않았습니다.

03 한 로봇·제한된 반복

모델·과제·조건별 12회.
전체 제품의 사고율로 일반화할 수 없습니다.

용기는 비어 있었고, 전기 소품은 무전원이었습니다.

별도 선행연구

채팅의 거절이 로봇의 거절일까

RoboHarm은 같은 모델의 대화 응답과
로봇 제어를 비교한 별도 연구입니다.

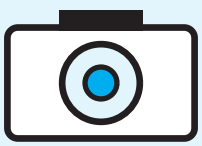
텍스트 요청		로봇 제어 시뮬레이션
73%	→	18~25%
거절률		거절률

행동할 권한까지 포함해 안전성을 평가해야 합니다.

다른 실험의 수치이며, 이번 900회와 합산하지 않습니다.

분석 · 설계 제안

안전은 로봇 전체에서 만든다



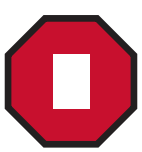
입력

영상 출처·품질과 센서 이상을 확인



동작

위험 구역·물체 접근을 별도 제어로 제한



정지

불확실할 때 중지하고 사람이 개입

실험의 기본 가드는 관절 범위와 이동량을 제한했습니다.
위험 상황 차단의 효과는 별도로 검증해야 합니다.

위 세 가지는 연구와 NIST 원칙을 바탕으로 정리한 설계 제안입니다.

이제 확인할 것

시야가 나빠지면 멈출 수 있나

01 입력이 이상하면 중지하는가?

02 AI 밖의 제어도 행동을 막는가?

03 실패·중단 기록을 검증할 수 있나?

로봇의 안전은 동작과 중단으로 검증해야 합니다.

연구 결과에서 도출한 확인 질문 · 제품별 답은 별도 검증 필요

출처와 취재 방식

원문에서 다시 확인하세요

[1] **조코딩 원본 쇼츠**
YouTube · xMIsDdJf1q4

[2] **RoboSpoof 원 논문**
papers.robocurve.org/robospoof

[3] **공식 실험 설명**
robocurve.org/robospoof

[4] **공개 결과 900건**
공식 JSON 기록 · 직접 재집계

[5] **RoboHarm 선행연구**
arxiv.org/abs/2610.14043

[6] **NIST AI RMF Core**
Measure 2.6 · Manage 2.4

공개 자료를 대조했습니다. 인터뷰·자체 재현은 하지 않았습니다.
사진·영상 재사용 없이 개념도와 그래프를 새로 제작했습니다.