

기술

공장 AI 에이전트는 시키지 않은 일 중 무엇을 먼저 해도 되는가

알리기와 실행을 나누는 판정표

방동걸 (주)웨이스 대표이사

2026년 9월 27일

회사 메일을 읽던 AI 에이전트가 먼저 말을 거는 일이 늘었습니다. 강연 관련 메일을 읽고 슬라이드 초안을 만들어 드릴지 묻거나, 답장을 보내지 않은 메일이 하나 남아 있다고 알려 줍니다. 투자 유튜버 김단테 씨는 최근 영상에서 개인용 에이전트 여러 개를 써 본 경험을 소개하며, 에이전트의 가장 중요한 덕목 가운데 하나로 시키지 않은 일을 알아서 하되 사람을 귀찮게 하지 않는 것을 꼽았습니다. 훌륭한 비서의 덕목과 같다는 설명입니다.

제조 AI 운영체제 VEXPLOR를 만드는 웨이스가 이 덕목을 공장으로 옮겨 보면 한 가지가 먼저 보입니다. 사무실에서 먼저 움직인 에이전트가 틀리면 초안 한 장이 버려지지만, 공장에서 먼저 움직인 에이전트가 틀리면 잘못된 작업 지시가 현장으로 내려가거나 자재가 잘못 발주됩니다. 그래서 공장 에이전트의 선제성은 덕목이 아니라 설계할 대상입니다. 이 글은 에이전트가 먼저 해도 되는 일과 사람에게 올려야 하는 일을 나누는 기준 3가지, 그리고 그 기준을 판정표와 설정으로 옮기는 방법을 정리합니다.

01

사무 에이전트가 먼저 움직여도 되는 것은 틀렸을 때 잃는 것이 작기 때문입니다

영상에 나온 사례를 행동 단위로 나눠 보면 공통점이 보입니다.

- 메일을 읽고 할 일을 찾아낸다 — 읽기만 합니다. 아무것도 바꾸지 않습니다.
- 강연 슬라이드 초안을 만든다 — 새 파일을 만들 뿐, 기존 자료를 고치지 않습니다.
- 답장하지 않은 메일을 알려 준다 — 사람에게 알리는 것으로 끝납니다.
- 호텔마다 문의 메일을 보낸다 — 이것만 내 밖으로 나가는 행동입니다. 영상에서도 에이전트는 문의 메일을 써 보자고 제안했고, 사람이 「내 계정으로 보내 줄 수 있냐」고 요청한 뒤에야 보냈습니다.

앞의 셋은 틀려도 지우면 끝납니다. 사람이 보고 「아니다」라고 하면 아무 일도 일어나지 않은 것과 같습니다. 선제성이 기분 좋은 놀라움으로 느껴지는 것은 이 조건 덕분입니다. 마지막 행동은 성격이 다릅니다. 한 번 나간 메일은 보낸 쪽에서 되돌릴 수 없습니다.

02

공장에서는 같은 「먼저 하기」도 틀렸을 때 되돌리는 비용이 다릅니다

공장 에이전트가 할 수 있는 행동을 같은 방식으로 나눠 보면 차이가 커집니다.

행동	틀렸을 때	되돌리기
불량 이력 조회	잘못된 답 한 번	다시 조회
설비 이상 알림	헛걸음 한 번	알림 확인 후 종료
작업 순서 초안	사람이 고침	초안 폐기
작업 지시 확정	잘못된 생산	생산 전까지만 가능
자재 발주	재고·대금 발생	거래처와 협의
설비 정지·가동 명령	설비·사람 위험	되돌릴 수 없음

위 셋과 아래 셋 사이에 선이 있습니다. 위 셋은 정보를 만들고, 아래 셋은 공장과 바깥 세계를 바꿉니다. 영상에서 에이전트가 시키지 않고 먼저 한 일은 모두 위쪽에 있었고, 아래쪽 행동은 사람이 시킨 뒤에야 했습니다. 공장 에이전트에게 기대하는 일은 점점 아래쪽으로 내려옵니다. 「알아서 해 주는 에이전트」를 공장에 그대로 들이면 안 되는 이유입니다.

03

먼저 해도 되는 일을 나누는 기준은 3가지입니다

행동마다 아래 세 질문에 답합니다. 하나라도 「예」면 그 행동은 에이전트가 먼저 하지 않습니다.

- 되돌릴 수 없는가** — 실행 뒤 원래 상태로 돌아가는 방법이 없거나, 돌아가는 동안 손실이 생깁니다. 작업 지시는 취소할 수 있어도 그사이 만든 제품은 되돌리지 못합니다.
- 물리 세계에 닿는가** — 설비·로봇·밸브처럼 움직이는 것에 명령이 갑니다. 화면 안의 일과 화면 밖의 일은 위험의 크기가 다릅니다.
- 회사 밖으로 나가는가** — 발주·메일·납기 회신처럼 거래처가 받는 행동입니다. 밖으로 나간 것은 우리 쪽에서 지울 수 없습니다.

세 질문에 모두 「아니요」면 에이전트가 먼저 해도 됩니다. 조회, 알림, 초안, 시뮬레이션이 여기에 들어갑니다. 하나라도 「예」면 에이전트는 제안까지만 하고, 실행은 사람이 승인한 뒤에 합니다. 물리 세계에 닿는 행동은 승인에 더해 설비 쪽 인터록이 따로 막아야 합니다. 소프트웨어의 승인 버튼은 사람이 실수로 누를 수 있기 때문입니다.

하나라도 「예」면 에이전트는 제안까지만



세 질문 가운데 하나라도 「예」면 에이전트는 실행하지 않고 제안까지만 합니다

04

판정표로 옮기면 행동마다 에이전트의 권한이 정해집니다

행동	세 질문 결과	에이전트 권한
조회·요약	모두 아니요	먼저 실행
이상 알림	모두 아니요	먼저 실행
작업 순서 초안	모두 아니요	먼저 실행
작업 지시 확정	되돌릴 수 없음	제안 후 승인
자재 발주	밖으로 나감	제안 후 승인
설비 명령	물리 세계에 닿음	승인과 인터록

이 표에서 중요한 것은 행동의 이름이 아니라 세 질문입니다. 새 행동이 생길 때마다 표에 한 줄을 더하고 세 질문에 답하면 권한이 정해집니다. 같은 행동이라도 공장마다, 라인마다 답이 달라질 수 있습니다. 시험 라인의 작업 지시는 되돌리기 쉽고 양산 라인의 작업 지시는 어렵습니다. 그래서 판정은 행동과 라인을 함께 놓고 합니다.

05

세 질문을 설정으로 옮기면 에이전트가 판정표를 어길 수 없습니다

판정표를 문서로만 두면 에이전트를 만드는 사람마다 해석이 달라집니다. 행동마다 세 질문의 답을 필드로 적고, 실행 직전에 그 필드를 검사하는 단계를 둡니다.

필드	값	뜻
action	confirm_work_order	작업 지시 확정
line	press-02	적용 라인
irreversible	true	되돌릴 수 없는가
physical	false	물리 세계에 닿는가
external	false	회사 밖으로 나가는가
mode	propose	제안만 허용
approver	생산팀장	승인자

검사 규칙은 단순합니다. irreversible·physical·external 가운데 하나라도 true면 mode는 propose(제안)만 허용합니다. 에이전트가 execute(먼저 실행)로 요청해도 검사 단계가 거절하고 제안으로 바뀌 승인자에게 보냅니다.

이 검사는 에이전트 밖에 둡니다. 에이전트가 스스로 판단해 규칙을 지키게 하면, 에이전트가 틀리는 날 규칙도 함께 틀립니다. 설정 파일을 고칠 수 있는 사람도 에이전트를 만드는 사람과 나눕니다.

새로 들이는 행동은 먼저 제안 모드로만 둡니다. 에이전트의 제안과 사람이 실제로 내린 결정을 함께 기록하고, 일정 기간 둘이 얼마나 일치하는지 봅니다. 일치율이 충분히 높고 되돌리기 쉬운 행동부터 먼저 실행으로 옮깁니다. 권한은 한 번에 주지 않고 행동 단위로 넓힙니다.

06

사람을 귀찮게 하지 않는 것도 정해 두지 않으면 무너집니다

판정표를 지키면 에이전트는 안전해지지만, 사람에게 올라오는 제안이 늘어납니다. 제안이 너무 많으면 사람은 읽지 않고 승인하거나 아예 보지 않습니다. 둘 다 판정표가 없는 것과 같은 결과입니다.

웨이스는 사내 업무에 AI 에이전트를 여럿 쓰면서 사람에게 올리는 종류를 다섯 가지(밖으로 나가는 것·서명·가격과 계약 조건·돈·사람)로 정해 두었습니다. 정하기 전에는 확인을 기다리는 안건이 두 달 사이 주

2건에서 주 100건 넘게 늘었습니다. 나머지는 에이전트가 「이렇게 가정하고 진행한다」를 기록한 뒤 진행합니다. 공장 에이전트에도 같은 원칙이 필요합니다.

- **올릴 것을 종류로 정한다** — 판정표에서 「제안 후 승인」 이상인 행동만 사람에게 갑니다.
- **올릴 때는 추천안 하나와 기한을 붙인다** — 선택지 다섯 개를 주면 사람이 처음부터 다시 조사해야 합니다.
- **올리지 않는 것은 가정을 기록한다** — 나중에 결과가 틀렸을 때 어느 가정에서 어긋났는지 찾을 수 있습니다.
- **알림은 묶어서 보낸다** — 같은 설비에서 난 경보 열 건은 한 건으로 묶습니다.

07

에이전트가 먼저 보내는 알림은 네 줄이면 충분합니다

먼저 해도 되는 일 가운데 가장 많이 쓰이는 것은 알림입니다. 알림이 쓸모 있으려면 받은 사람이 다른 화면을 열지 않고 판단할 수 있어야 합니다. 설비 이상 알림을 예로 들면 네 줄로 구성합니다.

줄	담는 것	예시
무엇이	대상과 현상	2호 프레스 진동 상승
근거	비교 기준	최근 7일 평균보다 높음
추천안	하나만	다음 교대 전 점검
기한	판단 시한	오늘 18시까지

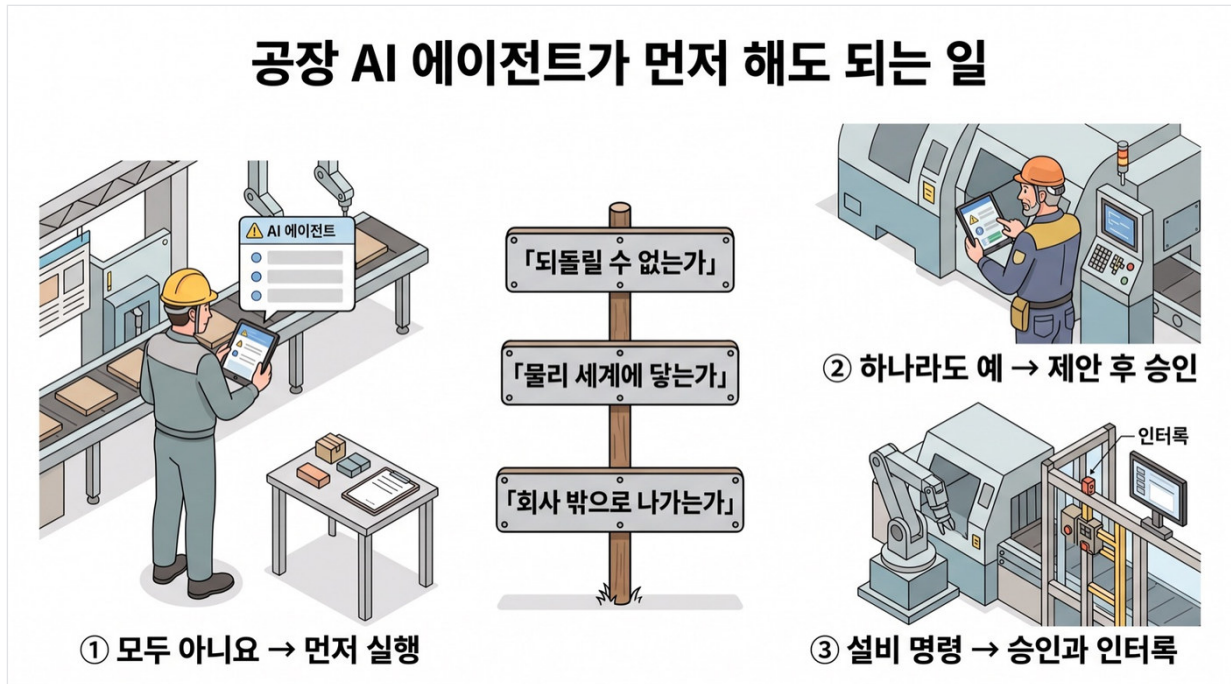
근거 줄이 빠지면 사람은 알림을 믿지 못하고 원본 데이터를 다시 찾습니다. 추천안이 여럿이면 판단을 사람에게 되돌려 보낸 것과 같습니다. 기한이 없으면 알림은 쌓이기만 합니다. 네 줄 가운데 하나라도 채우지 못하는 알림은 아직 보낼 준비가 안 된 것입니다.

08

공장 에이전트를 들이기 전에 확인할 5가지

- 에이전트가 할 수 있는 행동이 목록으로 적혀 있습니다.
- 행동마다 되돌릴 수 없는가, 물리 세계에 닿는가, 회사 밖으로 나가는가의 답이 적혀 있습니다.
- 권한 검사를 에이전트 밖에 두어, 에이전트가 틀려도 규칙은 지켜집니다.
- 설비 명령은 소프트웨어 승인과 별도로 설비 쪽 인터록이 막습니다.

- 사람에게 올라오는 제안의 종류와 하루 건수를 알고 있습니다.



이 글을 한 장으로 요약하면 이렇습니다 — 세 질문으로 공장 에이전트가 먼저 해도 되는 일과 사람에게 올릴 일을 나눕니다

사무실의 에이전트는 먼저 움직일수록 편해지지만, 공장의 에이전트는 먼저 움직여도 되는 범위가 정해져 있을 때 편해집니다. 그 범위는 에이전트의 성능이 아니라 행동의 성격으로 정합니다. 되돌릴 수 없는가, 물리 세계에 닿는가, 회사 밖으로 나가는가. 이 세 질문에 답해 둔 공장은 에이전트가 먼저 알리는 일을 마음 놓고 늘리고, 먼저 실행하는 일은 한 행동씩 넓혀 갈 수 있습니다.

WACE(웨이스)는 공장이 스스로 판단하고 멈추지 않게 하는 제조 AI 운영체제 VEXPLOR를 만드는 회사다.

09

함께 읽어보세요

- "AI가 알아서 라인 세웠습니다": 이 한마디가 왜 공포인가
- 시험장을 빠져나온 AI — 에이전트에게 어디까지 맡길 것인가
- 오너는 하자고 하는데 일이 시작되지 않는 이유



원문	blog.vexplor.com/post/what-factory-ai-agents-may-do-first
발행	2026년 9월 27일 · 방동걸 (주)웨이스 대표이사
회사	주식회사 웨이스 (WACE Inc.) · 대표이사 방동걸 · 사업자등록번호 439-87-02852
주소	(18462) 경기 화성시 동탄구 동탄영천로 150, 에이동(A동) 1702·1706호 (영천동, 실리콘엘리동탄)
연락처	전화 070-4730-0629 · contact@wace.me
홈페이지	wace.me · linkedin.com/company/wace-inc



원문



홈페이지



링크드인