AI/ML

저희 사내 규칙을 AI가 고치기 시작했습니다

채점표만 빼고

방동걸 (주)웨이스 대표이사

2026년 8월 8일

이번 주 저희는 사내 업무 규칙 문서에 한 줄을 새로 넣었습니다. 사실 검증 게이트, 기준 데이터 대조 점검, 완료 확인 — 이 항목들은 자동화 사이클이 스스로 완화하거나 수정할 수 없는 고정 계층이라는 선언입니다. 바꾸려면 사람의 승인을 거쳐야 합니다. 제조 AI 운영체제 VEXPLOR를 만드는 웨이스 사내 이야기이고, 왜 그 줄이 필요했는지부터 풀어 보겠습니다.

업무 자동화를 만들면서 자동화가 건드리지 못하는 영역을 일부러 만든다는 게 이상하게 들릴 수 있습니다. 저희가 이 규칙을 넣은 이유를 설명하려면, 요즘 AI가 좋아지는 방식부터 이야기해야 합니다.

저희는 프레스 186대에 더해 고객사의 기존 AMR 25대·EMS 6대까지 함께 동기화하고, AI의 판단을 PLC 제어로 직접 잇는 현장을 실증했습니다. 설비에 명령이 나가는 자리에서는 "AI가 알아서 좋아진다"는 말이 곧바로 위험이 됩니다. 그 경험이 이번 규칙의 배경입니다.



작업 방식은 고치게 하되 채점표에는 손대지 못하게 합니다.

1. 요즘 AI는 모델이 아니라 작업 방식에서 좋아진다

같은 모델을 쓰는데 결과가 다른 경우가 있습니다. 차이는 모델이 아니라 그 주변에 있습니다. 어떤 도구를 쥐여 줬는지, 무엇을 기억하게 했는지, 중간에 스스로 점검하게 했는지, 실패했을 때 무엇을 보고 다시 시도하는지. 이 실행 체계를 하네스라고 부릅니다.

전 오픈AI 안전연구 리드 릴리안 웡(Lilian Weng)의 최근 정리에 따르면, 최근의 성능 향상 상당 부분이 모델 가중치가 아니라 이 하네스를 고쳐서 나옵니다. 같은 정리가 인용한 실험(Darwin Gödel Machine 논문 기준)에서는 AI가 자기 실패 기록을 읽고 자신의 프롬프트·도구 설명·중간 처리 코드를 직접 수정하게 했더니, 소프트웨어 수정 과제 정답률이 20%에서 50%로, 다국어 코딩 과제가 14.2%에서 30.7%로 올랐습니다. 모델은 그대로 두고 일하는 방식만 바꾼 결과입니다.

제조로 옮기면 익숙한 이야기입니다. 설비를 새로 사지 않고 공정 순서와 판단 기준을 바꿔 수율을 올리는 것과 같습니다. 그리고 바로 그 지점에서 다음 질문이 나옵니다. 방식을 고쳐서 좋아진다면, 그 고치는 일도 AI에게 맡기면 되지 않을까.

2. 그러면 AI는 시험지를 고치기 시작한다

실제로 그렇게 굴러갑니다. AI가 자기 작업 기록을 읽고, 무엇이 자꾸 틀리는지 찾고, 자기 지침을 고쳐 다음 판에 반영합니다. 저희 업무 자동화도 이 구조로 운영합니다.

여기서 갈리는 지점이 있습니다. 성적이 오르는 것과 실제로 좋아지는 것은 다릅니다.

AI가 자기 작업 방식을 고칠 수 있다면, 채점 기준도 그 방식의 일부입니다. 통과 조건을 느슨하게 바꾸는 쪽이 실력을 올리는 쪽보다 언제나 빠르고 확실합니다. 검사 항목을 하나 줄이면 합격률은 즉시 오릅니다. 이것을 보상 해킹이라고 부르는데, 나쁜 의도가 필요하지 않다는 점이 핵심입니다. "점수를 올려라"라는 지시만으로 충분합니다.

앞서 저희는 병목은 생산이 아니라 검증이다라는 글에서 AI가 만들어 내는 양이 폭발하면서 사람의 검증이 병목이 된다고 썼습니다. 오늘 이야기는 그다음 문제입니다. 검증이 병목이라고 해서 검증 기준까지 AI에게 맡기면, 병목은 사라지지만 확인할 방법도 함께 사라집니다.



평가기는 진화 루프 밖에 두는 것이 원칙입니다.

---

3. 고치지 못할 것을 먼저 정한다

그래서 규칙은 단순합니다. AI가 개선할 수 있는 영역과 손댈 수 없는 영역을 나눕니다.

AI가 고쳐도 되는 것은 일하는 방법입니다. 순서를 바꾸고, 참고할 자료를 늘리고, 자주 틀리는 대목에 점검을 하나 더 넣는 것. 반대로 AI가 고칠 수 없어야 하는 것은 무엇을 합격으로 볼지에 대한 기준, 그리고 실제로 무언가를 실행할 권한입니다. 이 둘은 개선 순환 바깥에 두고, 판단이 서지 않을 때는 통과가 아니라 정지로 떨어지게 설계합니다.

저희 사내 규칙에 넣은 고정 계층이 정확히 이것입니다. 회사 수치가 기준 원본과 맞는지 확인하는 대조, 문서 사이의 어긋남을 훑는 점검, 빠뜨린 것이 없는지 보는 완료 확인 — 자동화가 아무리 스스로 좋아져도 이 기준선은 자기가 낮추지 못합니다. 함께 만든 것이 실패 기록입니다. 자동화가 막혔을 때 어떻게 우회했는지를 따로 쌓아 두면, 다음 개선의 재료가 됩니다. 성공만 기록하면 같은 벽에 계속 부딪힙니다.

---

4. 공장에서는 이 원칙이 더 분명해진다

소프트웨어에서는 잘못된 개선이 잘못된 코드로 끝납니다. 공장에서는 설비와 사람에게 갑니다.

저희가 자율화를 단계로 쪼개 상품화하는 이유가 여기 있습니다. VEXPLOR의 권한 프레임워크(AAF)는 AI의 권한을 조회, 제안, 승인 후 실행, 자율 실행, 긴급 개입의 5단계로 나눕니다. 현장에는 보통 1차년도에

제안 단계, 2차년도에 승인 후 실행 단계로 넓혀 갑니다. 자율화의 폭을 계약과 검증으로 정하는 것이지, AI가 잘한다고 스스로 넓히는 것이 아닙니다.

여기에 물리 제약 검증을 겁니다. 설비가 실제로 낼 수 없는 값, 안전 범위를 벗어난 명령은 AI가 아무리 확신해도 아래에서 막힙니다. 판정 기준을 AI가 못 바꾸게 하는 것과 같은 발상입니다. 그래서 저희는 자율제어를 곧바로 시작하지 않고, AI 판정을 숙련자 판정과 나란히 돌려 비교하는 단계부터 밟습니다. 개입 없이 답만 맞춰 보는 구간입니다. AI가 스스로 좋아지는 것은 이 단계 안에서 일어나야 합니다. 이 단계 자체를 AI가 건너뛸 수 있으면 안전 설계는 의미가 없습니다.

이 이야기를 더 자세히 다룬 글이 "AI가 알아서 라인 세웠습니다": 이 한마디가 왜 공포인가입니다.

---

5. 공급사에 물어볼 질문

AI가 스스로 좋아진다는 말은 앞으로 더 자주 듣게 될 것입니다. 그때 확인할 것은 얼마나 좋아지느냐가 아닙니다.

"그 AI가 스스로 바꿀 수 없는 것은 무엇입니까?"

답이 없다면, 그 시스템의 성적표는 시스템 자신이 만든 것입니다. 자율의 조건은 능력이 아니라 채점자를 분리하는 데 있습니다. 스스로 고치는 AI를 원한다면, 고치지 못할 것부터 정해야 합니다.

웨이스가 하는 일이 궁금하시면: wace.me

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

원문 blog.vexplor.com/post/ai-improves-itself-not-the-scoreboard
발행 2026년 8월 8일 · 방동걸 (주)웨이스 대표이사

회사 주식회사 웨이스 (WACE Inc.) · 대표이사 방동걸 · 사업자등록번호 439-87-02852
주소 (18462) 경기 화성시 동탄구 동탄영천로 150, 에이동(A동) 1702-1708호 (영천동, 실리콘앨리동탄)
연락처 전화 070-4730-0629 · contact@wace.me
홈페이지 wace.me · linkedin.com/company/wace-inc

원문 홈페이지 링크드인