AI/ML

에이전트 스무 개가 만장일치로 틀리는 법

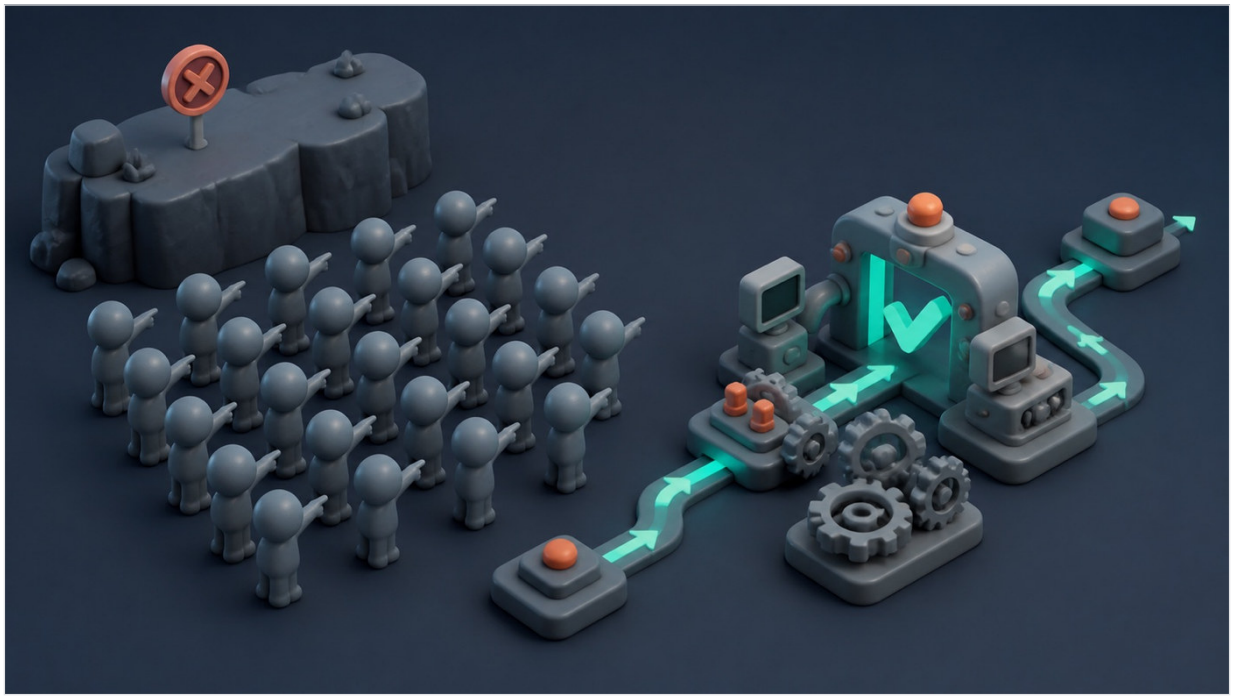
합의는 검증이 아니다

같은 모델 에이전트 스무 개는 같은 오류에 만장일치로 동의합니다. 상호 리뷰와 다수결은 신뢰성을 올리지 않습니다 — 검증은 언제나 시스템 밖의 증거가 합니다. 부샤르의 그래프 엔지니어링 하이프 해체를 저희 운영 경험으로 재해석했습니다.

방동걸 (주)웨이스 대표이사
2026년 8월 16일

지난 글에서 저희는 이렇게 썼습니다. 에이전트는 안쪽 루프를 돕고, 사람은 바깥 루프를 잡고, 두 루프의 경계에 놓이는 것이 증거라고. 그 글을 읽고 나면 자연스러운 반론이 하나 떠오릅니다. "그럼 에이전트를 여러 개 두고 서로 검증하게 하면 되지 않나? 스무 개가 검토해서 전부 통과시켰다면 믿을 만하지 않나?" 웨이스가 VEXPLOR를 만드는 일에 AI 에이전트를 쓰며 쌓은 경험이 그 바탕입니다.

최근 AI 교육자 루이 프랑수아 부샤르(Louis-François Bouchard)가 이 반론을 정면으로 무너뜨리는 글을 냈습니다. 요지는 한 문장입니다. 같은 모델을 바탕으로 만든 에이전트 스무 개에게 같은 결함이 담긴 맥락을 읽히면, 스무 개가 같은 오류에 만장일치로 동의한다는 것입니다. 모델은 자기 답을 선호하는 성향이 있어서, 상호 리뷰와 다수결은 신뢰성을 올려 주지 않습니다. 합의의 수가 아무리 늘어도, 그것은 검증이 아닙니다.



스무 개가 만장일치여도 검증을 통과한 것은 아닙니다.

01

그래프 엔지니어링이라는 이름에 속지 않기

부샤르의 글은 요즘 회자되는 "그래프 엔지니어링"이라는 신조어의 거품을 걷어내는 데서 출발합니다. 프롬프트 엔지니어링, 콘텍스트 엔지니어링, 루프 엔지니어링, 그래프 엔지니어링 — 이름은 계속 바뀌는데, 그래프 구조 자체는 워크플로우 엔진과 DAG 스케줄러가 10년 넘게 쓰던 기성품입니다. 새로운 것은 딱 하나, 그래프의 노드가 고정 규칙 실행자에서 확률적으로 판단하는 AI 에이전트로 바뀌었다는 점입니다.

그래서 그래프를 그리는 일의 진짜 가치는 구조의 새로움이 아니라 결정의 명시화에 있습니다. 무엇이 병렬로 돌아도 되는가. 어떤 상태가 다음 단계로 넘어가는가. 누가 거부권을 갖는가. 비용 상한은 얼마인가. 채팅창 맥락 속에 숨어 있던 이 결정들을 바깥으로 꺼내 적는 것 — 그것이 전부이고, 그것이 핵심입니다.

노드가 확률적이 된 순간, 검증의 위치가 문제가 됩니다. 확률적인 것들끼리 서로를 확인하게 하면 확률이 확신으로 바뀌는 착시가 생깁니다. 부샤르의 답은 명확합니다. 검증은 에이전트 시스템 밖의 증거만이 할 수 있습니다. 실제로 실행된 테스트, 실제로 들어온 입금, 이탈하지 않은 고객, 그리고 사람의 판단.

02

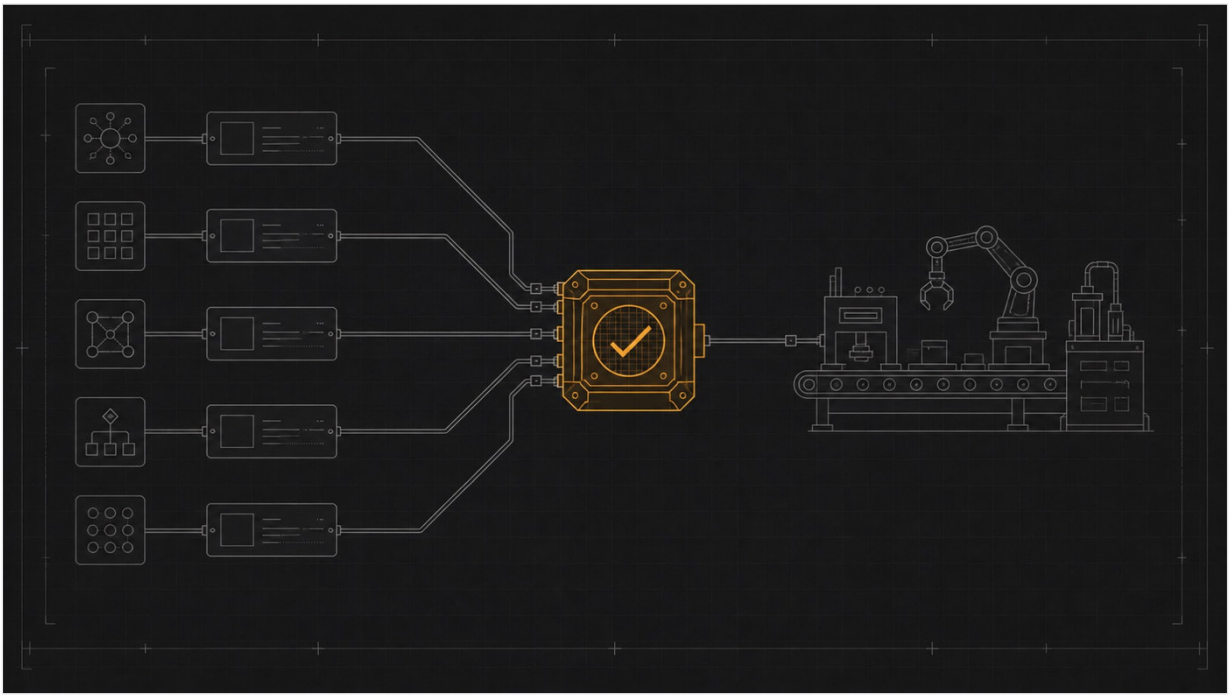
저희가 몸으로 배운 세 장면

이 원칙이 낯설지 않은 이유가 있습니다. 저희가 AI로 일을 시키면서 같은 결론에 도달한 장면이 여럿이기 때문입니다.

장면 하나. 저희는 문서에 들어가는 도식 이미지를 생성형 AI로 만듭니다. 언젠가 레퍼런스용 도식 열네 장을 사람이 검수해 통과시킨 일이 있었는데, 나중에 글자 단위로 전수 판독해 보니 오타가 열여덟 건 나왔습니다. 시뮬레이션이 "시믈레이션"으로, 형식지가 "형석지"로 적혀 있었습니다. 사람의 눈도, 그럴듯해 보인다는 합의도 검증이 아니었습니다. 지금 저희 규칙은 이렇습니다. 이미지에 넣을 문안을 먼저 목록으로 확정하고, 생성 후 그 목록과 기계적으로 대조합니다. 검증을 "보는 눈"에서 "시스템 밖의 확정적 절차"로 옮긴 것입니다.

장면 둘. 저희 내부 검토는 여러 에이전트를 동원하는데, 전부 같은 모델입니다. 부샤르가 짚은 상관 오류의 조건을 정확히 갖추고 있는 셈입니다. 그래서 저희는 세 가지로 완화합니다. 검토 에이전트마다 서로 다른 렌즈와 독립된 맥락을 주고, 이미지 생성처럼 갈 수 있는 곳은 다른 계열의 모델로 교차시키고, 최종 신뢰는 언제나 합의 수가 아니라 확정적 스크립트와 실측, 사람의 승인에 둡니다. 회사 수치가 원본과 맞는지는 에이전트의 동의가 아니라 대조 스크립트가 판정합니다.

장면 셋. 저희는 일하는 방식 자체를 AI가 고치게 두는 실험을 계속하고 있습니다. 단 하나, 평가 기준만은 그 진화 루프 밖에 둡니다. 검증 계층을 AI가 스스로 완화할 수 있게 하는 순간, 시험 문제를 학생이 고치는 구조가 되기 때문입니다. 부샤르 식으로 말하면, 어디에 확률적 에이전트를 쓰고 어디를 따분하지만 확정적인 코드로 남길지 가르는 일이 기술의 본체입니다.



합의가 아니라 검증이 실행의 조건입니다.

03

공장에서는 이 원칙이 더 무겁다

소프트웨어에서 검증 실패는 버그로 끝나지만, 공장에서 AI의 판단은 설비와 사람에 직결됩니다. 그래서 저희는 제조 AI의 자율화를 설계할 때 같은 원칙을 더 엄격하게 적용합니다. AI가 여럿이 동의했다는 것은 실행의 근거가 되지 않습니다. 가상 환경에서 실제로 돌린 시뮬레이션 결과, 현장 데이터와의 대조, 그리고 권한 단계에 따른 사람의 승인 — 시스템 밖의 증거가 쌓인 만큼만 자율의 범위를 넓힙니다.

"멀티에이전트라서 믿을 수 있다"는 문장은 저희 자료 어디에도 없습니다. 그 문장이 성립하지 않는다는 것이 이 글의 요지이기 때문입니다. 믿음의 근거는 에이전트의 수가 아니라 검증 계층의 존재입니다.

04

시작은 언제나 루프 하나

부샤르의 실무 조언은 소박합니다. 에이전트 마흔 개짜리 그래프로 시작하지 말 것. 루프 하나, 실제 검증자 하나, 들여다볼 수 있는 상태, 확실한 중지 조건 — 이것으로 시작해서 실패의 양상을 이해한 뒤에만 리뷰어를 붙이고 병렬을 나누고 거부권을 설계하라는 것입니다.

저희 하니스도 같은 규칙으로 돌아갑니다. 단순한 일은 에이전트 하나로 처리하고, 복잡도가 확인될 때만 팀을 구성합니다. 구조가 먼저가 아니라 실패 양상이 먼저입니다.

에이전트 스무 개의 만장일치는 감동적이지만, 검증은 거기에 없습니다. 검증은 언제나 시스템 밖에 있습니다.

참고: Louis-François Bouchard, "Graph Engineering, Explained" (louisbouchard.ai). 이 글은 해당 글의 논지를 저희 운영 경험으로 재해석한 것입니다.

관련 글: 병목은 생산이 아니라 검증이다 — 루프의 경계는 증거

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

원문 blog.vexplor.com/post/consensus-is-not-verification
발행 2026년 8월 16일 · 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