MODULE 11 · 심화

에이전트를 믿을 수 있는 선까지 — 자율성과 안전장치

모듈 5에서 우리는 "워크플로우로 충분하면 에이전트를 쓰지 않는다"고 배웠고, "완료를 의심하라"는 검증 게이트도 배웠습니다. 2026년 7월 말부터 8월 말 사이, AI를 실제로 쓰는 사람들 사이에서 이 이야기가 훨씬 절박해졌습니다 — "AI에게 어디까지 맡길 것인가"가 하반기 최대 화두로 떠올랐기 때문입니다. 배경지식 없이도 따라올 수 있게, 커뮤니티에서 오간 실무 논의를 정리해 우리 연구회가 이미 서 있는 위치와 비교해봅니다.

01 승인 없이 AI가 결정한 사건들 — "어디까지 맡길 것인가"

올여름, AI 코딩 도구를 실무에 쓰는 사람들 사이에서 비슷한 이야기가 잇따라 공유됐습니다. 지시하지 않은 파일을 AI가 스스로 지우거나, 사람이 확인하기도 전에 다음 단계로 진행해버리는 사례들입니다. 공통점은 한 가지 — **사람의 승인 없이 AI가 "진행 여부"를 스스로 결정했다**는 것입니다.

> 💡 **비유**
>
> 신입 직원에게 "이 서류 처리해줘"라고 했더니, 알아서 관련 없어 보이는 서류함까지 열어 정리하고 결재도 없이 발송까지 끝내버린 상황과 비슷합니다. 결과물이 잘 됐어도 문제입니다 — "누가 언제 무엇을 결정했는지" 되짚을 수 없기 때문입니다.

이런 사건들이 한둘이 아니라 여러 건 겹쳐 공유되면서, 커뮤니티의 논조가 "AI가 얼마나 똑똑한가"에서 "AI에게 무엇을 시킬지, 어디서 멈추게 할지"로 옮겨갔습니다. 이 모듈이 다루는 것이 바로 그 질문입니다.

02 복습 — 판단 사다리를 오를수록 안전장치가 필요해진다

모듈 5의 판단 사다리를 다시 떠올려봅니다. ① 단발 호출 → ② 워크플로우 → ③ RAG → ④ 에이전틱 루프. 사다리를 오를수록 모델에게 맡기는 **자율성**이 커집니다. 단발 호출은 사람이 매번 결과를 보고 다음 지시를 내리지만, 에이전틱 루프는 모델이 스스로 판단하고, 검증하고, 재시도까지 합니다.

- 사다리 낮은 단(①②)에서는 사람이 매 단계를 보고 있어서 사고가 나도 그 자리에서 잡힙니다.
- 높은 단(③④)으로 갈수록 모델이 여러 단계를 **사람 눈을 거치지 않고** 이어서 처리합니다 — 그래서 이 모듈의 안전장치가 필요해지는 지점도 바로 이 지점입니다.
- 즉 "에이전트를 쓸까 말까"의 질문 뒤에는 항상 "이 자율성에 맞는 안전장치를 걸었는가"라는 두 번째 질문이 따라와야 합니다.

모듈 5에서 배운 검증 게이트(모델이 "완료했다"고 주장하면 기계적 검사로 증거를 확인하고, 실패하면 자동으로 다시 실행시키는 구조)는 이 안전장치의 한 형태입니다. 이번 모듈은 "완료를 확인하는 것"을 넘어 "애초에 무엇을 하도록 허락할 것인가"를 다룹니다.

03 권한 설계의 실무 원칙

이런 사고를 막기 위해 실무자들 사이에 자리 잡은 원칙들이 있습니다. 특별할 것 없어 보이지만, 지켜지지 않아서 사고가 납니다.

- **실행 모드 제한** — 에이전트가 "읽기만 가능"인지 "쓰기·삭제까지 가능"인지를 작업 성격에 맞게 미리 정해둡니다. 삭제·발송·결제처럼 되돌리기 어려운 실행은 기본으로 막아둡니다.
- **권한 격리** — 에이전트마다 그 일에 **필요한 만큼만** 접근 권한을 줍니다. 문서 요약 에이전트에게 결제 시스템 접근 권한이 있을 이유가 없습니다.
- **최소 권한 원칙** — "혹시 필요할지 몰라서" 넉넉하게 권한을 주지 않습니다. 필요해지면 그때 추가합니다.
- **에스컬레이션 설계** — 위험도가 높은 작업(삭제, 외부 발송, 되돌릴 수 없는 변경 등)은 에이전트가 알아서 처리하지 않고, **사람의 승인으로 올려보내도록** 미리 설계해둡니다.

> 💡 **비유**
>
> 회사 출입 카드와 비슷합니다. 신입 직원이라고 모든 문을 열 수 있는 카드를 주지 않습니다. 자기 부서 문만 열리고, 서버실처럼 위험한 곳은 카드로 안 열리고 담당자 승인이 따로 필요합니다. 에이전트에게 주는 권한도 똑같이 설계해야 합니다.

04 자율성의 등급 — 우리는 어디까지 쓰는가

승인 없이 진행한 사고들을 막는 방법은 결국 "자율성에 단계를 매기고, 각 단계에 맞는 안전장치를 거는 것"입니다. 아래 표는 낮은 단계부터 높은 단계까지를 정리한 것입니다.

등급	무엇을 하는가	사고 시 멈추는 지점
① 제안만	AI는 방안을 제시하고, 실행은 사람이 직접 한다	실행 자체가 사람 손이라 사고가 나도 그 자리에서 멈춘다
② 실행 + 사람 승인	AI가 실행까지 준비하지만, 최종 버튼은 사람이 누른다	승인 직전에 결과를 확인할 기회가 있다
③ 실행 + 사후 검토	AI가 실행까지 마치고, 사람은 나중에 결과를 확인한다	이미 실행된 뒤라 되돌리기 어려운 작업에는 위험하다
④ 완전 자율	AI가 판단·실행·후속 조치까지 사람 개입 없이 전부 처리한다	중간에 멈출 지점이 없다 — **우리는 이 등급을 쓰지 않는다**

7월 말 사고 사례들은 대부분 ③이나 ④ 등급의 일을 ①·②처럼 다뤄서 생긴 일입니다 — "사람이 보고 있겠지"라는 암묵적 가정이 실제로는 지켜지지 않았던 것입니다.

> 🔗 **우리 연구회에선**
>
> 우리는 이 화두를 이미 앞서 실천해왔습니다. 경력산정 태스크는 AI가 초안을 만들어도 **최종 확인은 담당자**가 하도록 못 박아뒀고, 공통 태스크 파이프라인에는 **검증 대기 칸반 컬럼**을 따로 두어 결과가 곧바로 반영되지 않고 한 단계를 더 거치게 했습니다. 6월 PII 데모에서 강조한 "**완료를 의심하라**"(모듈 5)도 같은 원칙입니다 — 이 모두가 표의 ② 등급(실행 + 사람 승인)에 해당하는 설계이고, 우리는 애초에 ④ 등급을 쓰지 않기로 정해둔 셈입니다.

05 기술보다 사람이 장벽이다 — 조직의 수용성

또 하나의 화두 — AI 도입이 막히는 진짜 이유는 기술이 부족해서가 아니라 **사람이 장벽**이라는 것입니다. 두 가지가 반복해서 이야기됩니다.

- **직무 대체 불안** — "이 일을 AI가 대신하면 나는 어떻게 되나"라는 걱정은 기술 설명만으로는 풀리지 않습니다.
- **책임 소재 불명확** — 에이전트가 실행한 일에서 문제가 생기면 **누구의 책임인지**가 정해져 있지 않습니다. 이 답이 없으면 조직은 도입을 주저하다가 결국 멈춥니다.

이 두 번째 항목이 특히 중요합니다. 모듈 5의 검증 게이트나 이 모듈의 권한 설계·에스컬레이션 모두, 기술적으로는 사고를 줄이는 장치이지만 조직 입장에서는 **"이 지점까지는 사람 책임, 이 지점부터는 AI 실행"이라고 선을 그어주는 장치**이기도 합니다. 책임 소재를 명확히 하는 것 자체가 도입의 전제 조건이라는 뜻입니다.

06 기업 메모리는 큐레이션이다

조금 다른 각도의 화두도 있습니다. 사내 문서를 AI에게 많이 넣어줄수록(모듈 5의 RAG 방식처럼 근거 문서를 붙여주는 것) 성능이 좋아질 것 같지만, 실증 결과는 반대였습니다. **문서를 많이 넣는 것보다, 엄선된 소수의 문서가 AI 성능을 더 크게 좌우**했습니다.

> 💡 **비유**
>
> 공부할 때 참고서를 책상 위에 산더미로 쌓아둔다고 성적이 오르지 않는 것과 같습니다. 오히려 잘 정리된 핵심 요약 한두 권이 더 도움이 됩니다. AI에게 주는 지식도 양이 아니라 무엇을 골라 넣을지가 관건입니다.

이 결론은 이 모듈의 주제와 이어집니다. 권한을 설계하는 것만큼, **지식을 무엇을 줄지 설계하는 것**도 신뢰의 문제입니다. 아무 문서나 다 넣어두면 AI가 무엇을 근거로 판단했는지 사람이 되짚기 어려워지고, 이는 결국 책임 소재를 흐리는 또 하나의 원인이 됩니다.

> 🔗 **우리 연구회에선**
>
> 트렌드 참고: 이 모듈의 화두들은 클로드코드 커뮤니티 아카이브의 주간 정리(2026년 7~8월: 자율 실행 사고·조직 수용성·권한 설계)와 매거진(기업 메모리 큐레이션 실증)에서 왔습니다. 이 내용은 **커뮤니티 논의를 정리한 것이며, 개별 사례는 독립적으로 검증되지 않았습니다** — 방향을 읽는 자료로 참고하고, 수치나 사례를 그대로 인용할 때는 원출처를 다시 확인하세요.

> ↗ **심화 학습**

더 깊이: 위 주간 정리들의 원문을 chat.axwith.com에서 직접 확인할 수 있습니다. 우리 연구회 안에서는 모듈 5(검증 게이트·판단 사다리)를 먼저 복습해두면 이 모듈이 훨씬 잘 읽힙니다. · 커뮤니티 정리 성격의 자료이니 "그 시점 논의의 스냅샷"으로 받아들이세요.

07 퀴즈 — 인접 개념을 가려내기

CHECK

이해했는지 확인해 봅시다

틀려도 괜찮습니다 — 오답을 고르면 왜 아닌지 설명이 나오고, 정답을 찾을 때까지 다시 고를 수 있습니다. 점수는 첫 시도 기준입니다.

문제 1 / 5

Q1. 올여름 공유된 '자율 실행 사고'들의 공통 문제는 정확히 무엇이었나?

① AI가 내놓은 결과물의 품질이 낮았다는 것

② 사람의 승인 없이 AI가 다음 단계로 진행할지 여부를 스스로 결정했다는 것

③ AI 사용 비용이 예상보다 많이 나왔다는 것

④ AI가 외부 해킹 공격을 받아 오작동했다는 것

← 이전
모듈 10 · 로컬 서빙의 다음 카드

다음 →
모듈 12 · 허깅페이스에서 모델 읽는 법