CONDERI

CONDERI ACADEMIC INSIGHT | 학술·연구 인사이트 | 대외 공개판

위임은 능력의 문제가 아니라 설계의 문제다: AI 협업 12편의 실증이 가리키는 같은 곳

Delegation Is a Design Problem, Not a Capability Problem: What Twelve Empirical Studies Agree On

이번 호의 문제의식

2026년 하반기 학술지의 공통 질문은 "AI가 얼마나 잘하는가"가 아니라 **"사람이 언제, 어떤 조건에서, 무엇을 AI에게 넘기는가"**입니다. 이번 호는 기술·경영·과학·교육 네 영역에서 최근 공개된 실증 연구 12편을 골라, 문제의식 → 표본과 방법 → 수치 → 실전 시사점 순으로 정리했습니다. 결론은 한 방향입니다 — **성과를 만드는 것은 모델 성능이 아니라 위임의 구조**이고, 그 구조는 정보 공개·책임 게이트·기록의 세 가지로 조작 가능합니다.

문서 정보 / Document Information

문서등록번호 Document No.	CONDERI-AR-2026-11-14
발행일자 Published	2026. 11. 14. (토) Saturday
권호 Volume	VOL. 2026-11-14
분석범위 Coverage	2026. 06 ~ 11 공개·게재 논문 (ISR·Organization Science·APJIS·ICIS·NBER·arXiv)
배포구분 Distribution	Public | 일반공개
발행기관 Publisher	CONDERI AI산업전략연구실
조사·집필 Research	AI산업전략연구실

0.14 SD
AI 튜터 2년 완전참여 효과
NBER 18개 중학교

37.4%
최상위 모델의 기업과제 성공률
EnterpriseOps-Gym

0.63 → 0.76
3인 팀 성과, 구조화 후
공유기억 + 승인 게이트

65%
전문가가 모델보다 선호된 비율
300회 블라인드 비교

한 문장으로: 위임의 성과는 모델 성능이 아니라 정보·게이트·기록의 설계에서 나온다

최근 6개월 공개된 12편의 실증 연구가 같은 결론으로 수렴했다. 효과는 작고, 조건은 구체적이다

12편
기술 4 · 경영 4 · 과학 2 · 교육 6의 실증 연구를 교차 정리

0.06~0.31 SD
AI 학습 개입의 효과크기 범위. 대부분 0.2 SD 전후

7개 조합
AI 위임 의향을 만드는 충분조건 조합 수(592명 fsQCA)

64/125
전문가가 서명할 수 없는 치명 오류를 포함한 모델 산출물

기술: 능력은 올라갔고, 판단은 남았다

기업 과제에서 최상위 모델의 완전 성공률은 37.4%였고, 사람이 짠 계획을 주면 14~35%p 올랐습니다. 병목은 추론이 아니라 **전략적 계획 수립**입니다. 재무 81개 과제 벤치마크에서는 9개 프런티어 모델의 산출물 125건 중 64건이 실무자가 서명할 수 없는 오류를 포함했습니다. 읽으면 그럴듯하지만 절반은 틀립니다.

경영: 위임은 정보가 만든다

ISR 실험에서 AI의 결과 정보만 주면 위임 빈도(β=−0.397)와 효과(β=−0.423)가 모두 떨어졌고, 사전 확신도만으로는 효과가 유의하지 않았습니다. 둘을 함께 주면 상호작용항이 유의하게 플러스(β=0.589 / 0.565)로 바뀌며 위임이 성과로 이어졌습니다. 정보는 따로 주면 해롭고, 같이 주면 작동합니다.

이번 호 결론 AI 협업의 성과는 모델이 아니라 사람이 무엇을 알고 넘기는가에서 갈린다. 위임의 세 가지 조작 변수는 사전 확신도, 사후 결과, 그리고 책임의 소재다

영역	핵심 논문·출처	표본·방법	핵심 수치
기술	ICIS 2026 위임 체제 연구	한국 성인 592명, 2×2×2 시나리오 실험 + fsQCA	충분조건 조합 7개, 3가지 논리
기술	EnterpriseOps-Gym (2026)	14개 프런티어 모델, 기업 도구사용 과제	최고 성공률 37.4%
경영	ISR 2026 위임 정보 실험	이미지 분류 56,000 관측치, 혼합효과 로지스틱	상호작용 β=0.589(p<.05)
경영	Organization Science 2026	P&G 전문가 791명 현장 실험	개인+AI = 팀 성과
과학	arXiv 2026 협업 구조 실험	1,482 세션, DiscoveryBench	3인 팀 0.63 → 0.76
교육	NBER w35620 (2026)	테네시 중학교 18개교, 2년 군집 무작위	연 0.06~0.08 SD

기술 I. 위임 의향은 기능이 아니라 조합이 만든다 — 592명 fsQCA

ICIS 2026「Who Really Decides? Delegation Regimes and Reorganization of Human Control under Agentic AI」

조건	수준	해석
설계 조건	설명가능성 · 조정가능성 · 되돌림 가능성	서비스 설계가 위임 의향을 만든다
사용자 조건	AI 친숙도 · 자기 효능감	친숙도가 낮으면 통제감이 대체재로 작동
결과	7개 충분조건 조합	하나의 정답이 아니라 3가지 경로가 존재

세 가지 논리

(1) **일반 강화형** — 설계 조건과 사용자 조건이 모두 높으면 위임 의향이 높다. (2) **상보적 안심형** — 통제가능성과 AI 친숙도가 함께 높을 때 의향이 올라간다. (3) **대체적 안심형** — 둘 중 하나가 없어도 다른 하나가 그 자리를 메우면 의향이 유지된다. 실무적으로는 "설명을 잘 하는 AI"와 "되돌릴 수 있는 AI"가 서로의 대체재라는 뜻입니다.

방법론이 말해주는 것

이 연구는 회귀계수로 평균 효과를 구하지 않고, 조합으로 결과를 설명하는 fsQCA를 썼습니다. 같은 기능이라도 사용자 집단에 따라 다른 조합이 작동한다는 것을 보여주기 위한 선택입니다. 한국 성인 592명, 2×2×2 시나리오 실험입니다.

실전 시사점: 온보딩에서 무엇을 먼저 보여줄 것인가

도입 첫 주에 "이 모델이 얼마나 똑똑한가"를 보여주는 것은 위임을 늘리지 않습니다. 대신 ① 판단 근거를 함께 보여주고 ② 사용자가 언제든 되돌릴 수 있다는 것을 화면에서 확인시키고 ③ 첫 업무는 실패 비용이 낮은 과제로 배치해야 합니다. 세 가지는 기능이 아니라 **설계 결정**이며, 개발 공수는 크지 않습니다.

기술 II. 실행은 되는데 계획이 안 된다 — 기업 과제 벤치마크 두 건

EnterpriseOps-Gym(14개 모델)과 EnterpriseVAL(81개 재무 과제·9개 모델)이 같은 곳을 가리킨다

37.4%
14개 프런티어 모델 중 최고 성공률(기업 도구사용 과제)

+14~35%p
사람이 작성한 계획을 주입했을 때의 성과 향상

53.9%
최고 모델의 '수행 불가 과제 거절' 정확도

64/125
실무자가 서명할 수 없는 오류를 포함한 산출물 비율

계획이 병목이다

상태를 가진 기업 환경에서 여러 단계 도구를 쓰는 과제를 14개 모델에 물었습니다. 최고 모델(Claude Opus 4.5)도 성공률 37.4%였습니다. 그런데 사람이 짠 오라클 계획을 주면 14~35%p 올랐습니다. 즉 모델이 못하는 것은 실행이 아니라 **무엇을 어떤 순서로 할지 스스로 정하는 일**입니다. 동시에 최고 모델조차 불가능한 과제를 거절하는 비율이 53.9%에 그쳤습니다. 거절은 안전 문제이자 비용 문제입니다.

재무 벤치마크의 경고

81개 재무·회계 과제(실무경력 평균 18년 전문가 작성)를 9개 모델에 적용했습니다. 루브릭 점수만 보면 모델들은 비슷하게 높았지만, 실무자가 서명할 수 없는 치명 오류만 세면 점수가 급락했습니다. 약한 영역은 계산·수치 정확성·회계 처리라는 **정량**이고, 강한 영역은 어조·출처 충실성이라는 정성입니다.

과제 영역	통과 과제 수	특징
재무(Treasury)	4 / 4	자동화가 이미 성숙한 영역. 모델 성과 최고
수주·수금(OTC)	12 / 16	정형 데이터 처리 중심
위험·규정(R&C)	5 / 7	판단 개입이 크지만 기준이 명문화
기록·보고(RTR)	19 / 31	대량·반복, 오류 누적 위험
조달·지급(PTP)	4 / 7	문서 대조 중심
계획·분석(FP&A)	6 / 16	수치 정확성 요구 최대, 치명 실패 최다

해석 주의 성공률은 과제 난이도와 채점 기준에 민감합니다. 수치는 벤치마크 설계 시점의 모델 기준이며, 모델이 갱신되면 달라집니다.

경영 I. 위임을 만드는 것은 정보의 조합이다 — ISR 2026

Taudien·Fügener·Gupta·Ketter, Information Systems Research, 2026-10-28 게재

−0.397
AI 결과 정보만 제공 시 위임 빈도 계수(β, p<.05)

+0.453
AI 확신도 사전 제공 시 위임 빈도 계수

+0.589
확신도×결과 상호작용(위임 빈도). 둘을 함께 줄 때만 유의

−8.371
사용자 자신감 수준의 효과. 자신감이 높을수록 위임하지 않는다

정보를 따로 주면 해롭다

실험은 이미지 분류 과제 56,000 관측치입니다. AI의 사후 결과(정답 여부)만 보여주면 위임 빈도가 떨어졌습니다(β=−0.397). 사전 확신도만 보여주면 빈도는 올랐지만 효과는 유의하지 않았습니다(β=0.240). 두 정보를 함께 줬을 때만 위임 효과가 유의하게 올라(β=0.565) **인간+AI 결합 성과가 AI 단독(78%)과 사람 단독을 동시에 넘어섰습니다**(t(134)=3.72, p<.001).

왜 그런가

사전 확신도는 "이번 과제에서 AI가 맞을 확률"에 대한 기대를 보정하고, 사후 결과는 그 기대를 확인하거나 반박합니다. 둘이 함께 있어야 사용자의 머릿속 모델이 정확해지고, 근거 없는 알고리즘 혐오가 줄어듭니다. 한쪽만 주면 기대만 남거나 결과만 남아 판단이 왜곡됩니다.

실전 시사점: 화면에 두 개의 숫자를 같이 띄워라

AI 제안 옆에 ① 이 과제에서의 예상 정확도(사전)와 ② 지난 제안의 실제 적중 여부(사후)를 함께 붙이면 위임의 질이 올라갑니다. 반대로 자신감이 높은 사용자일수록 위임을 회피하는 경향(β=−8.371)이 강하므로, 숙련자에게는 "위임하지 않는 것이 최선인 구간"을 명시해 주는 설계가 필요합니다.

경영 II. 개인+AI는 팀을 대체하고, 비판적 문해력은 가치를 높인다

현장 실험(791명)과 가치기반 채택 연구(APJIS 36(3))가 조직 설계에 주는 두 가지 지침

조직: 개인+AI가 팀을 대체한다

P&G 전문가 791명을 대상으로 한 사전등록 현장 실험(Organization Science, 2026-06)에서 AI를 쓴 개인의 성과가 AI 없는 팀의 성과와 같았습니다. AI는 직능 간 칸막이도 낮췄습니다 — 연구개발 인력은 기술 편향, 상업 인력은 시장 편향 제안을 내던 것이 AI 사용 후 균형을 이뤘습니다. 아이디어 생성의 질은 올라갔고, **선별·평가에는 사람의 판단이 여전히 우위**였습니다.

채택: 문해력이 가치평가를 가른다

APJIS 36(3)(2026-09)은 생성형 AI 문해력을 활용·비판적 평가·윤리적 사용·창의적 활용의 4차원으로 나눠 가치기반 채택모형으로 검증했습니다. 활용·창의 차원은 편익 인식을 높였고, 비판적 평가는 오히려 편익 인식을 낮추고 위험 인식을 높였습니다. 윤리적 사용은 위험 인식만 낮췄습니다. 주목할 점은 **성과 위험이 지각된 가치에 정(+)의 영향을 줬다**는 것입니다.

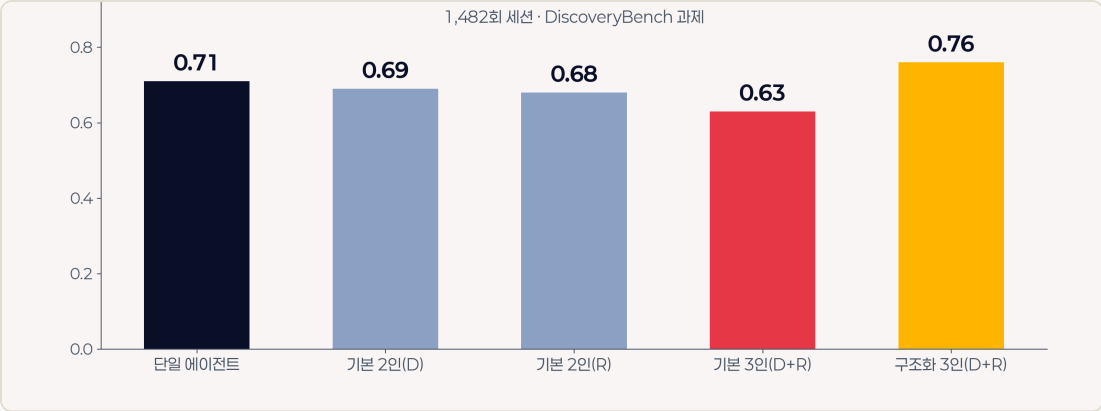
읽는 법: "성과 위험이 가치를 높인다"는 역설

숙련 사용자는 위험을 인지할수록 오히려 더 신중하게, 더 높은 가치로 AI를 씁니다. 위험을 모르는 사용자는 편익만 크게 느끼고 실제 성과는 낮습니다. 따라서 조직의 AI 교육은 "사용법"보다 **실패 사례와 한계**를 먼저 가르치는 편이 채택의 질을 높입니다.

연구	표본	핵심 발견	실무 번역
Cybernetic Teammate (Org. Sci.)	791명, 현장 실험	개인+AI = 팀, 직능 칸막이 완화	팀 구성 전에 AI 배치를 먼저 검토
GenAI 문해력 (APJIS 36(3))	설문·SEM	문해력 4차원의 차별적 효과	교육을 4차원으로 분리 설계
유료화·통제 (ISR 2026-04)	프롬프트 180만 건	비용이 보이면 통제(정교한 프롬프트)가 늘고 만족↑	과금·한도 표시가 품질을 올린다
위임 체제 (ICIS 2026)	592명, fsQCA	설계·사용자 조건의 7개 조합	온보딩을 집단별로 다르게

과학 I. 팀에 사람을 더하면 성과가 떨어진다 — 구조가 있어야 오른다

1,482회 세션 실험. 공유 기억과 승인 게이트는 대체재가 아니라 보완재다



자료: arXiv 2026, Collaborative Gym·DiscoveryBench 1,482 세션. 구조화 조건은 공유 그룹기억 + 사람 승인 게이트

0.71 → 0.63
단일 에이전트 대비 3인 기본 팀의 성과 하락 (프로세스 손실)

+0.13
같은 3인 팀을 구조화했을 때 성과 향상

+0.31~0.43
구조화 후 사람의 발의 비중 증가

0.68 → 0.64
공유 기억만 넣고 게이트를 빼면 오히려 성과 하락

프로세스 손실은 실재한다

전문성이 다른 구성원을 더 넣을수록 성과가 떨어졌습니다(3인 0.63 vs 단일 0.71). 대화량은 늘었지만 가설 지지도는 0.28에서 0.18~0.19로 떨어졌습니다. **협업은 기본값으로 좋아지지 않습니다.** 구조화(공유 기억 + 승인 게이트)를 넣으면 3인 팀이 0.76으로 올라 단일 에이전트를 넘어섰습니다.

두 장치의 분업

공유 그룹기억은 "누가 무엇을 아는가"의 지도를 만들고, 승인 게이트는 그 지도를 구속력 있는 결재선으로 바꿉니다. 공유 기억만으로도 발의 분포는 바뀌지만 성과는 오르지 않았습니다(0.68→0.64). 책임이 붙지 않는 지도는 조정 비용만 늘립니다.

과학 II. 인간-AI 팀의 집단지능은 어디에서 측정되는가

상호기억체계와 공유 멘탈모델이 이론과 실험 양쪽에서 같은 자리를 가리킨다

집단지능의 단위를 다시 정의하다

Frontiers in Organizational Psychology(2026-10)의 OHAI 프레임워크는 사람과 인공 구성원이 여러 주기에 걸쳐 상호의존적으로 일하는 집합을 분석 단위로 삼습니다. 핵심 추론 규칙은 명확합니다 — **인공 구성요소에 구현된 속성이 곧 팀 수준의 현상은 아니다.** 팀의 집단지능은 실제로 조직되고 사용되는 방식에서 입증되어야 합니다. 파트너 모델링, 상호기억체계(TMS) 조정, 공유 멘탈모델 정렬이 그 관찰 렌즈입니다.

공유 멘탈모델의 실증

IJHCI(2026-01)의 2×2×2 실험(128명)은 사람-AI 팀에서 지식 상보성이 높을수록 팀 효과성과 만족도가 올라가고, 공유 멘탈모델이 그 경로를 매개하며, 팀 분위기가 긍정적일 때 효과가 증폭된다는 것을 보였습니다. Human Relations(2026-09)의 3인 팀 실험(사람-AI 98팀, 사람만 65팀)은 정보 탐색 방식이 팀 유형에 따라 달라진다는 exploratory 결과를 남겼습니다.

연구	설계	발견	시사점
OHAI 프레임워크 (Frontiers)	이론적 통합	인공 구성원의 속성 ≠ 팀 수준 현상	AI 도입 효과를 팀 단위로 재측정
공유 멘탈모델 (IJHCI)	128명, 3원 요인 실험	지식 상보성 → 공유 멘탈모델 → 효과성	역할·전문성 지도를 먼저 합의
팀 유형·정보탐색 (Human Relations)	3인 팀 163팀	정보 탐색 패턴이 팀 유형별로 다르다	회의 규칙을 사람-AI 혼합에 맞게 재작성
TMS·이중 신뢰 (AMCIS 2026)	연구 설계 제안	대인 신뢰와 AI 신뢰의 이중 경로	신뢰를 하나의 지표로 관리하지 말 것

교육 I. AI 튜터의 효과는 작고, 대부분 사용 방식에서 사라진다

NBER 2년 실험과 MBA 현장 실험. 접근성은 조건이 아니라 출발점일 뿐이다



자료: 각 논문의 보고 효과크기(표준화 점수, SD 단위). 구현·표본·과목이 달라 직접 비교는 제한적

18개교·2년
NBER Khanmigo 군집 무작위 실험 규모

+1.3 백분위
학기당 수학 성취 상승. 연 0.06~0.08 SD

0.14 SD
완전 참여 시 환산 효과. 기존 학습보조와 유사

1/3
중앙값 학생이 연습일 중 AI에게 메시지를 보낸 비율

병목은 접근이 아니라 사용이다

학생의 96%가 최소 한 번은 AI 튜터를 사용했지만, 중앙값 학생은 연습한 날의 3분의 1에만 메시지를 보냈고 오답이 발생한 세션의 17%에서만 AI에게 물었습니다. 보낸 메시지도 대부분 정답 입력이나 추천 질문 클릭이었습니다. **AI를 주는 것과 AI로 생각하게 하는 것은 다른 문제**입니다.

구조가 효과를 만든 사례

같은 주제를 구조화한 튜터로 배운 MBA 86명은 통제군보다 6.63점 높았고(p=.007), 관계적 추론 수준의 답변 비중이 8%에서 49%로 올랐습니다(통제군 8%→27%). 음성 인터페이스는 대화량을 두 배로 늘렸지만 학습 효과는 동일했고 비용은 2.8배였습니다.

교육 II. 효과크기 0.06~0.31 SD, 조건이 붙을 때만 재현된다

여섯 개 RCT를 같은 축에 놓고 보면, 구조·교사·기록이 붙은 개입만 일관되게 플러스를 냈다

연구	표본·설계	효과	조건
Khanmigo (NBER w35620)	테네시 18개 중학교, 2년 군집 RCT	연 0.06~0.08 SD, 완전참여 0.14 SD	기존 보조수업 시간에 삽입
구조화 튜터 (arXiv 2609.23958)	MBA 86명, 사전등록 현장실험	+6.63점(p=.007), 관계적 답변 8→49%	교재 기반 구조 + 채널 교차
적응형 문제배열 (스탠퍼드·타이베이)	고교 10개교, 5개월	+0.15 SD	챗봇 상호작용 신호로 난이도 조정
생성형 튜터 (나이지리아)	중등 9개교 1,328명 RCT	0.310 SD (95% CI 0.176~0.444)	교사 안내 + 커리큘럼 정렬
AI 접근 (학부 210명)	2세션 실험	즉시 0.25 SD, 1주 후 0.20 SD	중간 성취층에 집중
AI 튜터 (IZA, 대학 334명)	인센티브 시험 대비 RCT	+0.23 SD	읽기 노력을 구축하지 않음

효과크기의 의미

0.2 SD 전후는 "한 학년 성장의 일부"에 해당하는 크기입니다. 화려하지 않지만, 비용이 낮고 확장이 쉬운 개입으로서는 충분히 의미가 있습니다. 단, 나이지리아 연구처럼 **교사 안내·커리큘럼 정렬·짝활동을 묶은 패키지**일 때 효과가 커졌고, 순수 챗봇 접근만으로는 작았습니다.

한국의 조건

국내에서는 AI 디지털교과서가 '교육 자료'로 지위가 바뀌었고, 경기도는 초등 단독 사용을 제한하고 고등학교는 출처 공개를 전제로 허용하는 안을 제시했습니다. 학습 효과의 증거가 0.1~0.3 SD 구간에 머무는 동안 정책은 "쓰게 할 것인가"에서 "어떻게 쓰게 하고 무엇을 기록할 것인가"로 이동했습니다.

종합: 네 영역이 같은 곳을 가리킨다 — 정보 구조, 게이트, 기록

기술·경영·과학·교육의 실증 12편을 하나의 표에 놓으면 남는 것은 세 개의 설계 변수다

공통 발견	기술	경영	과학	교육
구조가 성과를 만든다	오라클 계획 +14~35%p	정보 조합이 위임을 만든다	공유기억+게이트 0.63→0.76	구조화 튜터 +6.63점
사용(참여)이 병목이다	거절 정확도 53.9%	자신감 높은 사용자는 회피	발의 분포가 성과를 가른다	메시지 1/3, 오답세션 17%
인간 판단은 평가에 남는다	정량에서 치명 오류	아이디어 선택은 사람 우위	가설 지지도가 핵심 지표	교사 안내가 효과 조건
기록·책임이 필요하다	감사 가능성	사후 결과 정보	승인 게이트	출처 공개 전제 허용

하나의 문장

네 영역의 실증은 같은 문장으로 수렴합니다. **AI의 성과는 모델의 능력 상한이 아니라, 사람이 AI의 확신과 결과를 동시에 보고 판단할 수 있게 만드는 정보 구조에서 결정된다.** 기술 논문은 계획 수립에서, 경영 논문은 정보 조합에서, 과학 논문은 승인 게이트에서, 교육 논문은 참여 유도에서 같은 것을 측정했습니다.

반대로 말하면

모델을 바꾸는 것만으로는 이 네 가지가 개선되지 않습니다. 벤치마크 점수가 올라도 실무자가 서명할 수 없는 오류는 남고(64/125), 튜터가 좋아져도 학생이 묻지 않으면 효과는 사라집니다. 2027년 예산은 모델 도입비가 아니라 **정보 구조·게이트·기록**에 먼저 배분되어야 합니다.

적용 프레임워크: 위임 설계 5단계와 30일 실행표

실증 결과를 제품·운영 화면의 결정으로 번역했다. 각 단계는 측정 지표와 실패 신호를 함께 가진다

단계	설계 항목	측정 지표	실패 신호
1. 과제 분해	AI가 계획을 세우는 구간과 사람이 세우는 구간을 분리	계획 수립 소요·재작업률	오라클 계획 없이 성공률이 급락
2. 정보 공개	사전 확신도 + 사후 결과를 동시 노출	위임 빈도·위임 후 성과 개선율	결과만 노출(β=−0.397)
3. 책임 게이트	고위험 행동에 지정 승인자·승인 조건 명시	게이트 통과율·거절 정확도	게이트 없는 공유기억(0.68→0.64)
4. 참여 설계	첫 과제 난이도·되돌림·추천 질문의 질	주간 실사용률·질문 깊이	접근권만 배포(중앙값 1/3)
5. 기록	산출물의 출처·과정·검증 이력 저장	감사 추적 가능 산출물 비율	서명 불가 오류(64/125)

30일 실행표

1주차 — 현재 AI 사용 과제를 "계획/실행/평가"로 분해하고, 계획을 사람이 맡을 구간을 표시한다. **2주차** — 제안 화면에 사전 확신도와 사후 적중률을 함께 표시한다(간단한 문구·배지로 가능). **3주차** — 고위험 행동 3개에 승인자와 조건을 지정하고, 게이트 통과 기록을 남긴다. **4주차** — 산출물에 출처·검증 이력을 붙이는 서식을 시범 적용하고, 위임 후 성과 개선율을 측정한다. 네 단계 모두 모델 교체 없이 가능합니다.

English Summary

Twelve papers, one conclusion: delegation is a design problem

Executive Summary (English)

Across twelve empirical papers published between June and November 2026 in information systems, organization science, computer science and education economics, one finding repeats: performance in human-AI collaboration is set by the structure of delegation, not by model capability alone. Taudien, Fügener, Gupta and Ketter (Information Systems Research, 28 October 2026) show with 56,000 task observations that revealing only the AI outcome lowers delegation frequency (beta = -0.397) and effectiveness (beta = -0.423), that ex ante certainty alone is insignificant for effectiveness, and that the interaction of the two is positive and significant (beta = 0.589 and 0.565). Only when both signals were shown did combined performance exceed both the AI alone (78%) and the unaided human.

Where models still fail

EnterpriseOps-Gym, a stateful enterprise tool-use environment, reports a 37.4% success rate for the best of fourteen frontier models, with oracle human plans adding 14 to 35 percentage points - strategic planning, not execution, is the bottleneck - while the best model refuses infeasible tasks only 53.9% of the time. Accenture's EnterpriseVAL, built on 81 finance tasks authored by practitioners with an average of 18 years of experience, found that 64 of 125 model outputs contained at least one error a practitioner could not sign off, and that qualified practitioners were preferred in 65% of 300 blinded comparisons.

0.06-0.31 SD
Effect size range of AI tutoring and learning interventions in six RCTs

1/3
Median share of practice days on which students actually messaged the AI tutor

7
Sufficient configurations of high delegation intention among 592 Korean adults (fsQCA)

0.63 to 0.76
Three-person team performance after adding shared memory and approval gates

맺음: 남는 증거 네 개

강한 증거·실무 증거·반직관 증거·정책 증거로 정리하면, 다음 의사결정은 훨씬 좁아진다

본 보고서의 한계

인용한 연구는 서로 다른 과제·표본·측정도구를 사용하므로 효과크기를 직접 비교할 수 없습니다. NBER 연구의 0.06~0.08 SD는 학기당 수치이고, 나이지리아 연구의 0.310 SD는 12회 세션 패키지의 종합평가 효과입니다. 본문은 방향과 조건을 비교했고, 크기 비교는 하지 않았습니다. 또한 일부 벤치마크 수치는 공개 시점의 모델 기준이며 모델 갱신에 따라 변합니다.

다음 호에서 다룰 것

APJIS는 생성형 AI를 조직 행위자로 보는 특별호를 진행 중이며 1차 심사 결과가 2026-10-31에 나왔고 수정논문 마감은 12-31입니다. 다음 호에서는 이 특별호의 1차 통과 논문과, 11-17 이후 공개될 기업 실적·상장 자료를 연구 결과와 대조해 다룰 예정입니다.

정리	내용
가장 강한 증거	ISR의 정보 조합 실험(56,000 관측치, 상호작용 유의)과 Organization Science 현장 실험(791명)
가장 실무적인 증거	EnterpriseVAL의 치명 오류율(64/125)과 EnterpriseOps-Gym의 계획 병목(+14~35%p)
가장 반직관적인 증거	공유 기억만으로는 팀 성과가 오르지 않는다(0.68 → 0.64)
가장 정책적인 증거	AI 튜터의 효과는 접근이 아니라 사용 빈도에서 결정된다(메시지 1/3, 오답세션 17%)

출처 본문의 논문은 각 학술지 공개 페이지 및 초록 기준으로 정리했습니다. 수치는 원문에 표기된 값을 그대로 인용했습니다.