When should a support assistant reveal its diagnosis?

A firm has one paid hour to train a new support worker. Should it show the AI's diagnosis immediately or ask the worker to attempt the case first? Studying a worked solution gives the worker an example to follow. Attempting a diagnosis may help the worker recognize the same principle when a different customer problem arrives. I would test which sequence produces more correct, unaided resolutions of unfamiliar tickets two weeks later.

What remains open

Support-agent studies report productivity gains and evidence of learning during outages, randomize access to diagnostic assistance, and evaluate AI role-play through employees' follow-up self-reports. The order of instruction has closer precedents. Van Gog reversed identical example–problem pairs for novices and found that an early example-first advantage disappeared by the test. Fässler and colleagues found no sequencing advantage on next-day near or far diagnostic transfer among medical students. Wong and Qiu found that students who generated ideas before collaborating with ChatGPT outperformed unrestricted users on an unaided invention task. That comparison also changed how participants used ChatGPT. Attempt-first experiments also report gains on programming transfer and no learning benefit on food choices. The narrow contribution here is to measure delayed transfer across support issues under identical guidance and paid practice time. [S1] · [E2] · [E6] · [R1] · [R2] · [R3] · [E5] · [E4]

Training and test

Within baseline-score groups, I would randomly assign equal numbers of workers to write a diagnosis and then read the AI-generated solution, or to read the same solution, close it and then write. Both groups work individually with the same reading and writing intervals. Two support practitioners would verify every solution against the fictional product manual.

For each pair of issue families, a separate draw determines which family each worker practices. Consider refunds and account access. The manual says the latest event in the audit log governs when the summary screen is stale. In a refund ticket the screen says "paid" but the log says "queued." Full credit requires citing the queued event and keeping the case open. In an access ticket the screen says "pending" but the log says "activated." Full credit requires citing activation and closing the case. Across variants, both families have open and closed cases. Workers practice one family and receive new tickets from both at follow-up.

Estimand

The primary effect is the percentage-point difference between assigned groups in the fraction of held-out tickets resolved correctly within the time limit during a supervised, AI-disabled session. I would report the practiced-family effect and the difference between effects, with worker-level uncertainty. An equally large effect on tickets governed by an unrelated, untrained rule would weaken the rule-specific learning explanation. I would report return rates by arm and bound the effect by assigning missing workers opposite extreme scores, then reversing them.

Feasibility

I would seek platform approval to recruit 24 consenting workers through the User Interviews professional panel. A screener would require weekly ticket handling and under six months' total support experience. Each visit pays separately. All cases are synthetic. Graders receive shuffled responses and case IDs. An independent check must find zero assignment or participant identifiers in their export. I would proceed to sizing the trial only if 24 workers enroll within four weeks, at least 10 of 12 per arm return, and two graders independently agree on at least 90% of ticket scores. Pilot score variation and recruitment cost would indicate whether the budget can detect one additional correct resolution per twenty unfamiliar tickets. If a gate fails I would revise recruitment, follow-up or scoring and repeat the pilot. Live-queue value would depend on how often unfamiliar issues arise. [Recruitment panel]

Selected references

[S1] Generative AI at Work

[E2] Generative AI in Action: Field Experimental Evidence from Alibaba's Customer Service Operations

[E4] Do People Engage Cognitively with AI? Impact of AI Assistance on Incidental Learning

[E5] Answer First or Google First? Using the Internet in ways that Enhance, not Impair, One's Subsequent Retention of Needed Information

[E6] Preventing escalation in customer service: immediate motivation and 4-week outcomes in a workplace deployment of LLM role-play training

[R1] Effects of identical example–problem and problem–example pairs on learning

[R2] Problem-solving in virtual environment simulations prior to direct instruction for differential diagnosis in medical education: An experimental study

[R3] Think First, ChatGPT Later: Guiding Human–AI Collaboration for Learning Gains in Independent Human Creativity