REVENUE OPERATIONS & AI

# What AI actually changes in revenue operations

Two field experiments and a staggered rollout. Two of the harms an average cannot see; the third it records as an improvement.

**Sinan Isoglu**, MBA (Quantic)

Commercial growth leader, lecturer and doctoral researcher

| PUBLISHED | UPDATED | READING TIME | WORDS |
| --- | --- | --- | --- |
| 27 July 2026 | 27 August 2026 | 14 min | 3,079 |

https://isoglu.com/insights/journal/what-ai-changes-in-revenue-operations/

CONTENTS




MANAGEMENT SUMMARY

Three studies measured what AI does to commercial work, and two of them found a harm that an average cannot see while the third found a harm that the average records as an improvement: the most experienced workers losing quality while the mean rises, a task boundary nobody can feel themselves crossing, and customers disengaging when a bot is disclosed. None of the three is about revenue operations and none of their numbers should be carried into one. What transfers is the instrumentation problem: two of those harms are invisible to a throughput metric read alone, and the third is booked as an improvement. The essay sets out four cuts, one of them a re-cut of data most commercial stacks already hold and three of them requiring data they do not, and is explicit about where the evidence stops applying.



Keywords: Revenue operations · Artificial intelligence in sales · Productivity measurement · Field experiments · Evidence quality

The closest peer-reviewed match I found for generative AI in B2B sales measures sales performance by asking salespeople about themselves. It finds a positive effect. It is honest, competent adoption research. What it establishes is that people think AI improved their commercial output: a different question from whether it did.

Meanwhile the three studies that actually measured what AI does to commercial work were published outside the sales literature, go uncited in the two sales papers this essay examines, and two of them found a harm an average cannot see while the third found one an average records as an improvement: in the tail of a skill distribution, at a task boundary, or on the customer's side of the call. That is the argument here: an instrument a rollout may use to decide whether AI helped cannot resolve the three ways it hurts.

## What does empirical evidence show about AI in revenue operations?

The study is Rodriguez, Deeter-Schmelz and Krush, published last year in the Journal of Business & Industrial Marketing. It is the closest peer-reviewed match to the question, and closeness to the question is what a retrieval layer ranks on. Generative AI, B2B sales, an empirical model, partial least squares. How often it is the paper that actually comes back is not something I have measured. Its finding is that GenAI technology "has a positive impact on the effectiveness of the sales process, administrative efficiency and sales performance."

Now look at what those three outcomes are made of. The abstract describes the whole apparatus: in-depth interviews and feedback from leadership to build a questionnaire, a pilot survey, then "an online survey distributed to a larger sample," analysed with partial least squares. Every stage of it is an instrument for collecting what people say. The limitations paragraph adds the two things that follow: the study "focuses on a single health-care company" and "relies on self-reported data from sales professionals."

Read that as a commercial leader rather than as a methodologist. What the study establishes is an association, inside one firm, between how much salespeople say they use GenAI and how well they say their selling is going. That is a real finding about a real thing. It is not a measurement of output.

The authors put both limits in their own abstract. The paper is answering its own question well. Its stated purpose is a moderation model: whether support from upper management moderates the path from technology self-efficacy to GenAI use, and perceptual measurement is the right instrument for that question. The problem is what happens when it is asked a different one: whether AI worked, answered by a study designed to find out whether people think it did. The retrieval test described in a separate piece does not record that distinction, and this one takes the question as settled and moves on.

## The field's working assumption runs the other way

There is a second recent paper worth reading, and it frames the problem differently from everything that follows. Pia Hautamäki and Minna Heikinheimo built a framework from interviews with thirty-two top-level managers in B2B sales organisations, using a grounded theory approach, in the May 2025 issue of the Journal of Business Research. Their starting point is that "studies have demonstrated that artificial intelligence (AI) can enhance sales efficiency" in business-to-business contexts, and that the live problem is the gap that follows it: "despite the wide accessibility of AI, its adoption in B2B sales remains limited."

So the question they ask is what lets an organisation exploit AI at all. Their answer is managerial capability: "data-based human capital," the social capital of a knowledge-sharing culture, and "a transformative AI-positive mindset."

Hold that while reading the rest, because it is a fair description of where the field's attention sits: on adoption, and on the capability that makes adoption work. Adoption is a different question from effect, and three specific harms sit outside anything an adoption frame is built to see.

## The strongest evidence lives somewhere else entirely

The three studies that actually measured what AI does to commercial work sit outside the sales literature, and neither of the two sales papers above cites any of them. Rodriguez et al. carry 113 references and Hautamäki and Heikinheimo 81; the only Luo in either list is a different paper of his, and Brynjolfsson and Dell'Acqua are in neither.

Best performers can get worse. Erik Brynjolfsson, Danielle Li and Lindsey Raymond studied the staggered introduction of a generative AI assistant across 5,172 customer-support agents. Their results appeared last year in the Quarterly Journal of Economics. The paper highlights a 15% average lift in issues resolved per hour. That number arrives with its own qualifier attached: the same sentence ends "with substantial heterogeneity across workers," and the abstract then says what the heterogeneity is: "Less experienced and lower-skilled workers improve both the speed and quality of their output, while the most experienced and highest-skilled workers see small gains in speed and small declines in quality." The authors state it plainly; it gets lost downstream of them.

Read that as a commercial leader rather than as an economist. What happened at the top of the distribution was a trade: small gains in speed against small declines in quality. And these are support agents at one firm, not people carrying accounts: what transposes is the position in the distribution, not the job. A 15% mean is entirely compatible with your best sellers getting slightly worse. A mean may be the only output a rollout dashboard exposes.

The same study locates where the gain came from. It was largest "for moderately rare problems, where human agents have less baseline experience but the system still has adequate training data." The common cases people already handle well, and the genuinely novel ones may offer little relevant training data to draw on. The gain sat in the band between them, while a volume-oriented deployment can concentrate on common cases.

One more finding sits in the same abstract, and it runs against the customer-side harm this essay comes to below: "customers are more polite and less likely to ask to speak to a manager." Where the assistant sat behind a human agent, the customer's side of this rollout got better.

There is a boundary, and the person crossing it cannot feel it. Fabrizio Dell'Acqua and colleagues ran a preregistered experiment with 758 knowledge workers at Boston Consulting Group across eighteen realistic tasks, all of them inside what they call the jagged technological frontier, plus a nineteenth: "a complex managerial task selected to be outside the frontier." It was published this year in Organization Science. On the eighteen, subjects using GPT-4 completed 12.2% more tasks, 25.1% faster, at higher quality. On the nineteenth, they were 19% less likely to reach the correct answer.

The word doing the work is jagged. Their description: AI assistance "improves performance for some tasks but worsens it for others, even within the same knowledge workflow and with a seemingly similar level of difficulty." Two tasks that look equally hard to the person doing them can sit on opposite sides. Which side a task of yours is on shows up only in the output, which is cut 3 below.

Some of the loss has nothing to do with capability. Xueming Luo, Siliang Tong, Zheng Fang and Zhe Qu randomised more than 6,200 customers between chatbots and human agents on highly structured outbound sales calls. The study is in Marketing Science. Undisclosed, the bots were "as effective as proficient workers." Disclosed before the conversation, purchase rates fell by more than 79.7%. The mechanism the authors reach for is not performance: the effect "seems to be driven by a subjective human perception against machines." Customers "are curt and purchase less because they perceive the disclosed bot as less knowledgeable and less empathetic," and the authors are explicit that this happens "despite the objective competence of AI chatbots." They also record the limit of the finding in the next sentence: "such negative impact can be mitigated by a late disclosure timing strategy and customer prior AI experience."

A throughput metric cannot see a customer who disengaged. It records a shorter call, which the study also found, reporting that disclosure "substantially decreases call length."

Table 1 What each of these studies actually measured

| Study | Design | Subjects | Setting | Where a harm would show |
|---|---|---|---|---|
| Rodriguez et al. (2025) | Survey, partial least squares | Sales reps, one health-care firm | B2B selling | Not measured: the outcome is a report |
| Hautamäki and Heikinheimo (2025) | Grounded theory, interviews | 32 top-level managers | B2B sales organisations | Not measured: the outcome is capability |
| Brynjolfsson et al. (2025) | Staggered rollout, not randomised | 5,172 support agents | Post-sale support, one firm | Top of the skill distribution |
| Dell'Acqua et al. (2026) | Preregistered randomised experiment | 758 knowledge workers | 18 constructed tasks inside the frontier, 1 outside | One side of a task boundary |
| Luo et al. (2019) | Randomised field experiment | More than 6,200 customers | Structured outbound calls | The customer's side of the call |

Author's summary of the five papers cited, from the published abstract of each. Designs are described as the authors describe them. No result figures are reproduced here; the argument of the essay is that those numbers should not travel.

## What does not transfer

None of these three studies is about revenue operations. Brynjolfsson is post-sale support at one firm, and a staggered rollout rather than a randomised one. Dell'Acqua is knowledge workers on constructed tasks. Luo is pre-LLM fieldwork, published in 2019, on highly structured outbound calls, which tells you nothing about a GPT-class agent. Nobody should carry 15%, 19% or 79.7% into a pipeline conversation. I am not going to.

What transfers is the shape. Three methods, three settings, three different mechanisms. In two of them the harm sits where a mean cannot reach: in the tail of a skill distribution, and at a task boundary. In the third it sits in plain sight, on the customer's side of the call, and the mean reports it as an improvement. That is a claim about instrumentation, and instrumentation does transfer.

It is worth saying where this sits relative to Daron Acemoglu and Pascual Restrepo. Their argument is that recent technological change has been biased towards automation rather than towards the creation of new tasks, and that the consequences have been "stagnating labour demand, declining labour share in national income, rising inequality and lowering productivity growth." That is the same family of concern at a different altitude. Their frame is task-level: AI "can automate tasks previously performed by labour or create new tasks", and their outcome is national income. This one is a single commercial organisation, and the outcome is what you put on a dashboard on Monday. Both can be true. Only one of them tells you what to instrument.

## How should commercial leaders sequence AI deployment across sales operations?

If your rollout produced a modest average lift, the average is the least interesting thing you now know. Four splits. The first is a re-cut of data you already hold, provided the rollout left some sellers untreated. The other three need something you do not have yet: a rarity distribution over deal types, and a graded quality measure.

1. Split the effect by experience, against a comparison group. Brynjolfsson, Li and Raymond condition on experience and skill, and they can identify the effect at all because agents not yet switched on are measured over the same window. Both halves have to travel. Group sellers by something fixed before the rollout that is not itself an output measure: tenure, ramp status, a skill rating that predates it, and compare the change in each group against sellers who do not have the tool yet. If the most experienced group gains less than the least experienced in the treated arm, and that gap does not appear in the untreated one, you have the Brynjolfsson pattern. The comparison arm is what makes that sentence mean anything. Rank people on a noisy baseline and then measure their change by that rank. If the top group declines even without treatment, that is regression to the mean. At small seller counts, calculate precision and regression-to-the-mean risk for the design rather than assuming a stable decile. With forty sellers, split them in half only as an exploratory contrast unless a power or precision analysis supports finer groups.
2. Split by how rare a case type is. Bucket cases, deals or tickets by how often that type occurs. The prediction is a middle band: little effect on the routine, little on the genuinely novel, most in between. It rides on a precondition Brynjolfsson's assistant met: "the system still has adequate training data." A precondition your model may not meet, so rare in the firm meant thin in the model. An off-the-shelf model has no such gradient: a public-sector RFP that is rare in your book can be common in its training data, and your product's edge cases are rare in both. Check that the two rarities line up before reading a middle band as the same finding. If your gain is concentrated in the common cases, you are measuring time saved on work that was already cheap.
3. Find your boundary by grading a sample of the output. Take a stratified sample of AI-assisted output across task types and have it graded against a written rubric by someone who was not involved. The frontier is jagged. The only way to know which side a task is on is to check the output, not the confidence of the person who produced it. Grade a sample of work done without the tool alongside it: without that second sample you have a quality level, not a treatment effect. And what you are grading for is output that is wrong and looks right, so the grader has to be able to redo the work rather than read it.
4. Measure quality and throughput together or not at all. Two of the three findings above are invisible to a throughput metric read alone. The third is worse than invisible: disclosure shortened the calls, so throughput booked the harm as an improvement. Resolution rate against quality score. Meetings booked against a graded sample of how well they were qualified. Proposals sent against the rate at which they come back for rework. A single-axis measure cannot represent a trade, and every result above is a trade.

None of this needs new software. Three of the four need data you do not have yet, and between them it comes to two things: a rarity taxonomy over deal types, and a quality measure with a rubric and a grader behind it. Neither is missing by oversight: what gets recorded is what a system owns, which is also why the interval between a lead arriving and a human first acting on it has no field anywhere. The rest is the same numbers cut a different way, a comparison group chosen before the rollout rather than after: one of the three questions any number entering a decision should face, and a willingness to look at the tail rather than the middle. There is also a timing floor under all four: at a three-to-nine-month cycle, a rollout reviewed at ninety days holds no converted-deal data for anyone who started after it. Cut on stages that resolve inside the window, or wait for the cohort.

Figure 1 Four cuts of data you already have

| THE SPLIT YOU RAN | WHAT THE MEAN SAID | WHAT THE TAIL SAID | DO THEY AGREE? |
|---|---|---|---|
| Experience, rarity, task type, or quality. | The number the rollout reported. | Most experienced group, rare band, outside the frontier. | If not, the mean was the wrong statistic. |
| | | | |
| | | | |
| | | | |
| | | | |

Four rows because there are four cuts, and they are not alternatives: a rollout can pass three of them and fail the fourth. Cut 1 needs a comparison group; without one the third column fills itself. A blank in the third column is not a

Author's own worksheet.

## Where this stops holding

Suppose your effect is genuinely uniform: the same lift for your strongest and weakest sellers, the same across problem types, no machine anywhere a customer can see. Then the mean is a perfectly good statistic and none of the above applies. That case exists. How often it is the real one is not known: every study in this piece was selected because it found heterogeneity, so this set cannot give you a base rate. The four cuts are how you find out which case you are in.

The honest summary is that “did it work” got answered before it was measured. The two sales papers discussed here are both built from what people say about AI rather than from what changed when they used it. The studies that measured something sit in economics, marketing science and organisational behaviour. None of them is about you. Each of them measured an outcome instead of asking about one, and that is what made the damage findable at all.

## Boundary

Boundary. The evidence separates self-reported improvement from measured task outcomes. Use the diagnostic by role and task, then recheck it longitudinally before treating it as an operating rule.

Evidence base. The analytical frame also draws on these additional sources: Acemoglu and Restrepo 2020. The links identify the exact works; they support the mechanisms and boundary conditions discussed here, not every claim in isolation.

Acemoglu, D., & Restrepo, P. (2020). The wrong kind of AI? Artificial intelligence and the future of labour demand. Cambridge Journal of Regions, Economy and Society, 13(1), 25–35. https://doi.org/10.1093/cjres/rsz022

Brynjolfsson, E., Li, D., & Raymond, L. (2025). Generative AI at work. The Quarterly Journal of Economics, 140(2), 889–942. https://doi.org/10.1093/qje/qjae044

Dell'Acqua, F., McFowland, E., III, Mollick, E. R., Lifshitz, H., Kellogg, K. C., Rajendran, S., Krayer, L., Candelon, F., & Lakhani, K. R. (2026). Navigating the jagged technological frontier: Field experimental evidence of the effects of artificial intelligence on knowledge worker productivity and quality. Organization Science, 37(2), 403–423. https://doi.org/10.1287/orsc.2025.21838

Hautamäki, P., & Heikinheimo, M. (2025). Fully leveraging AI in B2B sales: Exploring sales managers' capabilities and organizational knowledge processes. Journal of Business Research, 194, 115396. https://doi.org/10.1016/j.jbusres.2025.115396

Luo, X., Tong, S., Fang, Z., & Qu, Z. (2019). Frontiers: Machines vs. humans: The impact of artificial intelligence chatbot disclosure on customer purchases. Marketing Science, 38(6), 937–947. https://doi.org/10.1287/mksc.2019.1192

Rodriguez, M., Deeter-Schmelz, D. R., & Krush, M. T. (2025). The impact of generative AI technology on B2B sales process and performance: An empirical study. Journal of Business & Industrial Marketing, 40(10), 2013–2027. https://doi.org/10.1108/JBIM-02-2025-0097

END OF ESSAY

## HOW TO CITE THIS ESSAY

Isoglu, S. (2026, July 27). What AI actually changes in revenue operations. Sinan Isoglu. https://isoglu.com/insights/journal/what-ai-changes-in-revenue-operations/

## KEEP READING

Revenue operations & AI

### The function without a German name

https://isoglu.com/insights/journal/the-function-without-a-german-name/

Revenue operations & AI

### The number you call

https://isoglu.com/insights/journal/the-number-you-call/

Revenue operations & AI

### What is RevOps?

https://isoglu.com/insights/journal/what-is-revops/

**Sinan Isoglu**, MBA (Quantic)
Commercial growth leader, lecturer and doctoral researcher


Online version: https://isoglu.com/insights/journal/what-ai-changes-in-revenue-operations/