

Why the Same Industry Is Not the Same AI Investment

What two German information experiments can and cannot tell us about firms that hold back

Two tax advisers may work in the same profession and face different AI decisions. One might handle repeatable documents, have a partner willing to redesign review, and serve clients who accept the workflow. Another might handle unusual cases, lack time to check machine output, or expect more supervision than the tool saves. Their industry label cannot settle either investment calculation.

There is a visible gap even within tax advice. In a survey of German licensed tax advisers, the share reporting that they used generative AI “often” or “always” was 15.9% in the smallest displayed employment group and 23.7% in the largest: a descriptive difference of 7.8 percentage points. These are the smallest and largest groups as displayed in the source figure, whose exact size labels are qualified in the notes. They are not matched firms: their clients, tasks and reporting respondents may differ. The contrast establishes a puzzle within one industry, not an effect of firm size. [1, pp. 15, 46]

Advisers might underestimate the technology. Or they might know its capabilities but struggle to change a workflow, pay for integration, or find a suitable use case. Some may have good reason to wait. Two German experiments test parts of this story by giving respondents information and observing different steps between hearing a claim and using AI. Those steps mark the reach, and the limit, of an information explanation.

A benchmark is not a business case

The tax-adviser study by Eduard Brüll, Samuel Mäurer and Davud Rostam-Afschar asked participants to estimate how much of several tax occupations' core work was technically automatable. Respondents were then randomly assigned either to a control group or to one of three versions of information based on the German Institute for Employment Research's expert task assessments. The messages varied which occupations' scores were shown. The online survey ran from November 2024 to April 2025 and retained 1,736 observations after the authors' exclusions. The invitation frame registered licensed people; it does not establish 1,736 distinct firms or identify every respondent as the person who approves an AI purchase. [1, pp. 7–14, 65–68]

An expert task-substitutability score describes technical potential for a listed occupation. It cannot tell an adviser whether a client's data may enter a tool, how often the task arises, what checking costs, or what installation and oversight would displace. The benchmark also covers automatable work beyond current generative AI. Disagreeing with it need not mean misjudging one's own client work. Random assignment tests the effect of additional information; it does not turn the benchmark into each firm's net return. [1, pp. 8–12]

Comparing existing users with non-users cannot isolate information: their managers, clients and tasks may differ. Randomly changing a message balances such prior differences on average. What the experiment establishes then depends on what it measures afterwards. A plan, a click, a vacancy and a later report of use answer different questions.

A plan and a click

In the tax-adviser survey, the authors estimate a 9.4-percentage-point increase in plans to use AI solutions in the future, evaluated at the mean belief update for the combined-information arm. The control mean is 71.15%. That 9.4 points comes from their instrumental-variable model linking assigned information to revised beliefs and then to stated plans. It is not the raw difference between everyone assigned information and everyone assigned control. The plan regression contains 301 observations, far fewer than the full analytic sample, and the checked questionnaire and methods do not establish why only that number enters this outcome. The question also gives no fixed date by which use must begin. [1, pp. 20–22, 32–33, 68, 73]

That model choice matters. In the paper's alternative reduced-form display, the confidence interval for the plan effect crosses zero and its point estimate differs materially from the positive modelled estimate. The 9.4 points comes from a particular specification; it is not a robust randomized average effect or a count of future installations. A sincere plan may fail on cost or fit. The survey did not observe later AI use, so it cannot show whether these plans became use. [1, p. 56, Figure C.12]

The experiment also offered respondents who were unfamiliar with named tax-specific tools a link to more information. The authors' model puts the increase in clicking at 2.4 percentage points, against an 80.88% control mean, in a regression with 1,405 observations. The alternative reduced-form click interval touches or comes very close to zero in the paper's plot. A click shows information seeking during the survey. It costs little and establishes no purchase, deployment or return. The plan and click regressions draw on different selected outcome samples. Dividing 2.4 by 9.4 would invent a conversion rate between people and stages that the study did not measure. [1, pp. 32–33, 56, 67, 73]

The authors linked respondents to vacancy advertisements and estimated a 0.15-point change in the chance of any listed vacancy, with a standard error of 0.18 points and a 1.27% control mean. The April–August 2025 coverage overlaps the final survey month; it is not a uniform window beginning after every participant received information. More fundamentally, an advertisement is a hiring measure, not a record that a tax practice deployed AI. The rounded estimate and standard error imply an approximate 95% interval of roughly -0.20 to $+0.50$ points. That supports no detectable vacancy change, while leaving meaningful relative changes possible from a low baseline. It says little about implementation without hiring. [1, pp. 32–33, 73]

When later use is observed

A different experiment, by Manuel Menkhoff, follows German firms beyond the initial message. In June 2024, respondents in an ifo survey were randomly shown either information about productivity gains in selected AI tasks, information about broad industry adoption, or no such message. In May 2025, the survey asked whether their firms used AI. Menkhoff's adoption analysis focuses on firms for which AI had previously been discussed or was not yet a topic, rather than firms already reporting use or a plan to adopt. This supplies the later *reported use* outcome absent from the tax-adviser survey, though no software activity is independently verified. [2, pp. 7–9, 18–19]

For the full group at risk of adoption, the estimated effect of receiving either message was 1.7 percentage points, with a 2.2-point standard error, across 2,376 observations. Its approximate 95% interval spans -2.6 to $+6.0$ points. The result is imprecise; it does not establish that information had no aggregate effect. For a subgroup Menkhoff calls high decision authority, the estimate is 8.2 points, with a 3.0-point standard error and 1,230 observations, against about 21% reported use in the control group. The random assignment supports a causal effect of the *message within that defined subgroup* on later reported use. It does not prove that handing approval authority to someone else would produce the same gain. [2, pp. 18–19, Table 3]

The authority label combines a top manager or owner respondent with family ownership. It was measured, not assigned, and may capture management attention and ownership structure as well as purchase authority. A clearer subgroup estimate alongside an imprecise pooled estimate does not itself test whether treatment effects differ across authority groups. Authority is a plausible implementation condition here, not an identified causal barrier. [2, pp. 18–19]

Task relevance gives a sharper contrast within the subgroup. Menkhoff divides industries by whether their occupational composition is relatively exposed to programming, writing or customer-service tasks. In the high-authority sample, the estimated message effect on later reported use is 16.0 points in high-exposure industries and 1.5 points in low-exposure industries; the reported difference test has $p = .01$ in the interaction model with 1,117 observations. That is evidence that response to the information varied with this industry measure. It is not a test of changing a firm's own task mix, or proof that every firm in a high-exposure industry has a profitable AI project. [2, p. 21, Table 4]

Finance remains a candidate explanation rather than a result of the randomization. In the same high-authority sample, the estimated effect is 10.0 points among firms that had planned to fund investment only internally and 1.4 points among firms also planning external funding. These plans were recorded before the messages, but neither credit nor cash was assigned by the experiment, and the table does not provide a formal test of the difference between those funding groups. The pattern could reflect financing flexibility, or other ways those firms differed before treatment. [2, p. 22, Table 5]

These studies sample different people, give different information and measure different outcomes at different times. Subtracting the tax-adviser plan estimate from Menkhoff's later-use estimate would invent an adoption funnel. Their differing statistical results cannot identify a missing constraint either. The informative contrast is what each observes: intentions and an immediate, cheap action within one profession, then later reported use across an at-risk multi-industry sample, with a positive message effect in a defined subgroup.

Decision question and observation time	What these experiments observe (effects in percentage points)	What remains open
Does a capability signal change intentions? Tax-adviser survey, November 2024–April 2025.	The model estimates a 9.4-point increase in future-use plans; its reduced-form plan interval crosses zero.	An effect on plans robust to the alternative specification, and whether any plan became implementation.
Does it induce an immediate action? Same survey.	A modelled 2.4-point increase in an in-survey information-link click among unfamiliar respondents; its reduced-form interval is visually close to zero.	Whether firms bought, installed or repeatedly used a tool; clicks are inexpensive.
Does information move later use? Menkhoff message in June 2024, status reported May 2025.	An 8.2-point increase among top managers/owners in family-owned firms; the full at-risk estimate is 1.7 points and imprecise.	Validated use, the effect of changing authority, and a universal information effect.
Would suitable tasks or finance change the decision? Earlier industry exposure and funding-plan strata in Menkhoff.	The later reported-use response varies by industry task exposure; effects also differ descriptively across prior funding plans.	Each firm's use-case payoff, a causal finance constraint, and the optimal choice for firms that abstain.

Exhibit note: Entries concern different populations, treatments and outcomes; percentage-point estimates are not a common effect scale for pooling. “Plan” and “click” are modelled estimates from Brüll, Mäurer and Rostam-Afschar [1, Figure 15; Table C.7]; Menkhoff's later-use and exposure estimates are from [2, Tables 3–5]. The table retains the reduced-form uncertainty on plans and the observational status of the authority and funding strata.

The value of waiting

A firm can learn that a task is technically automatable and still decide that automation is a poor investment. Its relevant gain is the value of faster or better work after paying for software, integration, training, oversight, privacy protection and mistakes. These costs can differ within a profession. So can the volume of repeatable work, the ability to reorganize it and the client's tolerance for a changed process. If those pieces are missing, not adopting may be rational, even if the technology works elsewhere. Neither experiment measures those firms' full expected net returns, so neither can certify that every holdout is right or wrong.

Menkhoff shows that a message can change later reported use in a defined group, especially where an industry-level task measure suggests relevant work. The tax-adviser study stops earlier: a modelled plan and an information click cannot establish deployment. Information can move adoption, while firms in one industry still face different investment cases and decision routes. Explaining the remaining gap would require comparable firms assessed against the same later-use definition, with authority and use cases measured before intervention and the costs and benefits of both adopting and waiting recorded.

Sources and measurement notes

[1] Eduard Brüll, Samuel Mäurer and Davud Rostam-Afschar, *Beliefs about Bots: How Employers Plan for AI in White-Collar Work*, ZEW Discussion Paper 25-057, October 2025, [original PDF](#). Survey design and frame: PDF pp. 7–14; size-frequency chart: Figure A.4, p. 46. Its largest group is labelled “More than 11 Employees”, but the caption says “11–144”; no corrected threshold is assumed. The often/always comparison above uses the smallest and largest displayed groups. The source's separate never-use shares are excluded because the figure combines “Never” and “Rarely” in those labels. Intervention and model: pp. 8–12, 20–22; outcomes: pp. 32–33 and Figure 15; reduced-form sensitivity: Figure C.12, p. 56; questionnaire: pp. 65–68; regression samples: Table C.7, p. 73. ZEW, IZA Discussion Paper 18225 and [arXiv v1](#) are deposits of the same study, not independent replications. The ZEW p. 33 vacancy interval omits a minus sign that appears in arXiv v1; the text above uses the explicitly calculated normal approximation from the rounded estimate and standard error.

[2] Manuel Menkhoff, *Belief Updating and AI Adoption: Experimental Evidence from Firms*, author-posted April 2026 [original PDF](#). Design: PDF pp. 7–10; prior status and high-authority definition: p. 18; later reported use: Table 3, p. 19; industry exposure interaction: Table 4, p. 21; pre-treatment funding-plan strata: Table 5, p. 22. Approximate intervals in the article use the printed point estimate ± 1.96 times its printed standard error; they are explanatory calculations, not the paper's own confidence intervals.

Evidence cut-off: 1 October 2026 IST. This article is a synthesis of two experiments, not a new causal estimate, firm-level dataset or finding about India. It is held for internal review and exact-version founder publication approval.