[AI and Workforce] AI Adoption and the Changing Value of Human Work

SIAI Research Editorial

Swiss Institute of Artificial Intelligence, Chaltenbodenstrasse 26, 8834 Schindellegi, Schwyz, Switzerland

Abstract

Indicators of exposure to artificial intelligence, hereinafter referred to as AI, rank occupations according to how many of their tasks a model can technically affect and are often read as projections of productivity, employment and wages. The paper examines why jobs with similar exposure result in different outcomes and argues that the value of human labor is judged in three sequential filters: the technical capability of the model, the economics of its deployment within the workflow and the demand for the final product. A review of ten studies, including field experiments, randomized trials, employee-based competency assessments, Danish administrative data and U.S. payroll data, shows that verification and integration costs explain part of the gap between technical capability and useful workplace outcomes, that individual productivity gains coexist with zero changes in earnings and that improving the performance of beginners does not demonstrate skill development. A sensitivity scenario shows that a few minutes of control and a small chance of reworking are enough to nullify a measured 40% reduction in completion time. The analysis does not document a causal effect of AI on youth employment nor long-term skill loss, for which no follow-up data are yet available.

1 Introduction - The Outcome Puzzle

Three datasets for the same period produce results that are difficult to reconcile. In a U.S. customer support firm, access to a generative AI assistant increased the number of requests resolved by each employee per hour by an average of 15% in a sample of 5,172 employees, with the greatest gains among the less experienced.[1] In Denmark, linking two surveys from the end of 2023 and 2024, with 25,000 workers in eleven exposed occupations, to administrative employment registers did not find a significant effect of conversational models on earnings or recorded hours in any occupation; confidence intervals exclude average effects of more than 2% and users themselves reported savings of just 2.8% of their working hours.[2] In the U.S., payroll data up to June 2026 show that the employment of 22- to 25-year-olds in the most exposed occupations is about 19% below the level it would have had if it had moved like that of their peers in less exposed occupations, while for experienced workers there is no corresponding gap.[3] The same technology seems to help beginners more at the task level, to coincide with fewer hires of beginners, and, in a country with high use, to leave no corresponding trace in earnings. This discrepancy motivates the examination of the changing value of human work.

The usual explanation begins with exposure. If a profession contains many tasks that a model can technically perform, it is considered to be threatened or to benefit, depending on the reading. This explanation omits two steps that mediate between the technical feasibility and the result in the labor market. The first is economic and organizational, since a task that the model can perform is not assigned to it when the control of the result, its integration into an existing process and the expected cost of an error that will go unnoticed outweigh the benefit. The second concerns demand. A productivity gain reduces the demand for labor only when the demand for the product does not increase enough to absorb the time saved. The Brookings analysis of workforce policy gets to the same point on the policy side and argues that exposure is a misguided organizing principle, because the question that counts is how AI changes the value of human expertise and whether it broadens or limits access to productive labor.[4]

The period from 2023 to 2026 made the issue pressing, because technical capability and adoption moved quickly while measured outcomes did not. In evaluations by experienced workers, models in 2024-Q2 completed text-based tasks that take humans about 1.5 hours with roughly 60% success, rising above 70% by 2025-Q3.[5] In the same period, the share of American workers using generative AI at work rose from 33% in August 2024 to 45% in May 2026.[6] Eurostat reported that among European companies that considered the use of AI without adopting it, 70.3% cited a lack of relevant expertise in 2025.[7] This paper examines the gap at the level of the job, where the assignment decision is actually made.

The research question is why tasks and occupations with similar exposure can experience different productivity, employment and earnings outcomes. The contribution is a task-level explanation that separates technical capability from deployment economics and then from labor-market outcomes and that identifies when expertise becomes more valuable, less scarce, or more difficult to acquire. The first hypothesis is that verification and integration costs explain part of the gap between task-level technical potential and useful workplace deployment; it would be challenged by cases in which those costs are small yet adoption remains weak. The second is that individual productivity gains can coexist with lower demand for labor when output demand grows too

little to absorb released capacity; it would be challenged by cases in which demand expansion, new tasks, or quality improvements increase labor demand. The third is that AI can improve novice performance in some tasks while reducing the opportunities through which novices learn, so immediate convergence of performance is not evidence of long-term skill development or deterioration; it would be weakened by longitudinal evidence showing that AI-assisted learners acquire equal or greater competence when working without it. The paper does not predict whether AI will increase or decrease jobs overall. It examines the mechanisms that determine which outcome occurs in which task.

2 Analytical Framework

Exposure means that a technology can affect a task, feasibility means that it can meet a specific functional requirement and adoption means that it is actually used in a defined population and over a defined period. Automation exists when task execution is transferred to the system; augmentation exists when humans continue to perform the task with AI assistance. Productivity refers to output relative to labor input at a specified quality standard, so time saved alone is not automatically a productivity gain if quality, review, or rework changes. Employment, hours, earnings, job quality and output are separate outcomes and none replaces another as a measure. The analysis primarily concerns domain professionals using AI and the broader workforce affected by it. Applied AI engineers appear only where the evidence concerns their own work, such as experiments with programmers; frontier AI researchers are outside the paper's scope

The first filter is technical competence. The revised Global Occupational Exposure Index of the International Labour Office, published in May 2025, uses a representative sample from the 29,753 tasks in the Polish occupational classification system, a survey of 1,640 workers and expert review and ranks occupations in four exposure grades. According to it, one in four workers worldwide is employed in a profession with some exposure to generative AI, but only 3.3% of global employment belongs to the highest classification. The revision of the index itself is instructive. The 2023 scores for tasks such as keeping minutes of meetings or scheduling appointments, which in some cases reached 0.9, were deemed overly optimistic about full automation, since two years of experimentation showed that many such tasks still require significant human effort and the agency emphasizes that its estimates describe a maximum limit of exposure rather than a real impact on employment.[8] An exposure index, no matter how carefully constructed, measures the first filter and is silent about the next.

The second filter is the economics of deployment. For a business or employee, the decision to assign a task to a model depends on the total cost per completed task, i.e. license and computing power, workflow adjustment, control time, privacy restrictions and the expected cost of errors, compared to the cost and performance of the person already performing it. With representative data of 9,835 German workers from a 2024 survey, Lindenlaub, Oh, Rodríguez and Veldkamp found that an exposure-only model explains about 14% of the variation in adoption between occupations, while an index that also accounts for AI user costs and worker productivity relative to pay explains almost 60%. For occupations that account for about 30% of employment, the two approaches make opposite predictions and for accountants, despite the high exposure, audit costs and privacy restrictions limit the advantage of AI.[9] The finding that comparative advantage better predicts adoption than absolute

capability agrees with Svanberg and colleagues' assessment of computer vision, according to which, with the integration costs of the time, only 23% of worker wages currently paid for vision tasks would be economically attractive to automate at the modeled costs.[10] American data from 2026 show the same discrepancy on the part of employees. Medical secretaries use AI at a rate of 16.8%, compared with 61% predicted by their exposure measure and the Federal Reserve Bank of St. Louis notes that adoption falls short of exposure in professions and jobs where privacy rules and the cost of error are high.[11]

Verification costs need special attention, because they are often treated as a temporary defect of the models. A task is advantageous to entrust to a model when the production time saved exceeds the sum of the control time, the correction time and the expected cost of an error that will escape. This sum increases when errors are plausible and difficult to detect, when the correct answer depends on knowledge not contained in the command and when a mistake has major or irreversible consequences. The Brookings analysis compares model success rates to human reliability frameworks used in commercial environments, in which the odds of human error have historically ranged from 2% in routine tasks to 16% in complex tasks that require understanding and notes that the measures are not directly comparable but appear noticeably lower than the average error rates of models.[12] This point leads to the question of who can audit. Auditing requires the same expertise as production and often greater, because the auditor must recognize a mistake that looks right, without having gone through the steps that produced it himself.

The third filter turns a productivity gain into an outcome for workers. When an employee completes a task faster, the business can produce more with the same staff, produce the same with less, or convert time to quality. Which option prevails depends on how much demand increases when costs fall and whether profits are passed on to wages or stay with the business. The context of Autor and Thompson's expertise adds a second dimension: when automation removes the less demanding tasks of a profession, the remaining work becomes more specialized, wages tend to go up and employment decreases, while when it removes the more demanding ones, the profession opens up to less skilled workers, with more jobs and lower wages.[13] The same technical progress can therefore increase or decrease wages, depending on the position that the subtracted work has within the bundle of the profession and an exposure index does not distinguish the two cases. At the same task-level gain, a fixed-demand case can translate released capacity into fewer labor hours, whereas an expansion case can translate it into additional output, shorter waiting times, or higher quality. These are partial-equilibrium firm responses; economy-wide employment also depends on adjustment across firms, sectors and demand, so time savings do not determine aggregate employment.

Figure 1 connects the three filters and adds two feedback loops that the linear readings of exposure omit. The first is a demand loop, through which reducing the cost per job can expand the market and increase the demand for work in the rest of the work. The second is a learning loop and for this paper it is the most important. If the tasks through which new workers were learning were assigned to models, fewer people would, in the future acquire the expertise needed to test the results of the models and the cost of verification, instead of decreasing with the improvement of technology, may increase with a delay of years. The diagram is conceptual; it organizes the items and does not estimate quantities.

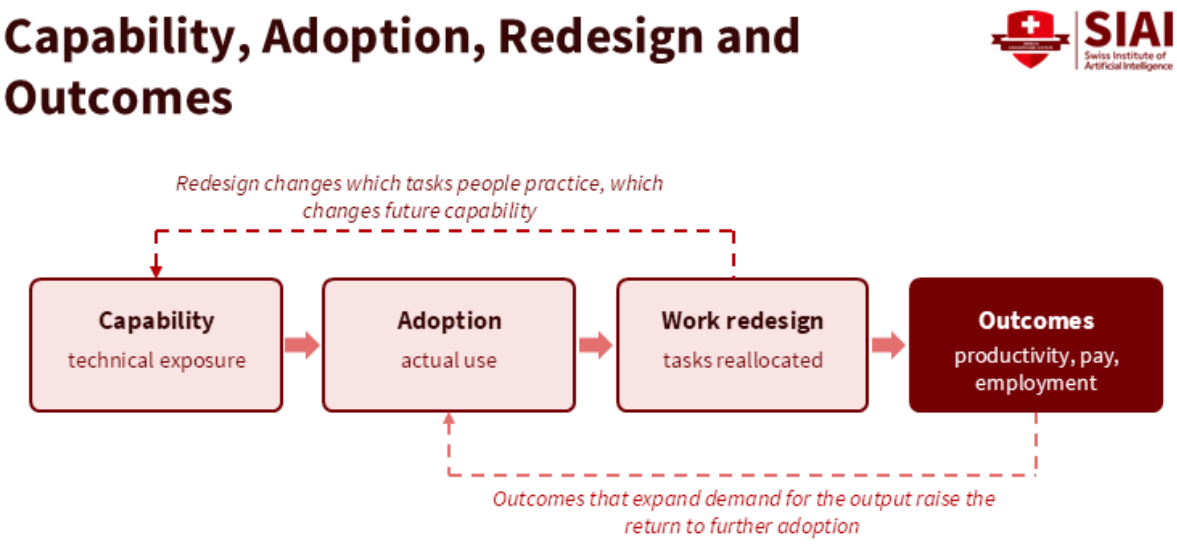


Figure 1: Deployment and work redesign mediate the path from AI capability to worker outcomes.

The three hypotheses correspond to different points in the framework. The first concerns the second filter, the second the third and the third the learning loop. The matching has practical consequences for the reading of the items: an experiment that measures the time to complete a stand-alone task checks the first and partly the second filter, while a study of administrative data on remuneration checks for the third without being able to discern which filter blocked the result. None of the studies considered here covers the whole chain and therefore the synthesis must keep the measures separate rather than merging them into a single estimate.

3 Evidence Review Method

The review includes studies published from 2023 to the cut-off date of 22 September 2026, addressing generative AI in cognitive and occupational work and reporting a measured outcome at the task, employee, or labor-market level for a defined population. Search terms combined ”generative AI” with ”occupational exposure,” ”productivity,” ”employment,” ”earnings,” ”skill formation,” ”customer support,” and ”software development.” Each study is classified as an experiment, quasi-experiment, descriptive data, model, self-reported research, or practitioner account and interpreted according to research design. Randomized experiments offer stronger internal causal identification under the study conditions, but for narrow populations and short horizons. Administrative data cover a broad population but rely on stronger assumptions about what would have happened without the technology and self-reports measure perception. Reports from consulting firms and companies developing AI models have an interest in the outcome and are not counted as independent confirmation. When a publication mentions a study, the original was searched for and used and where numbers differ between versions of the same study, the latest verifiable version is used unless an earlier version is explicitly identified for a specific comparison; the differences are noted in the text where relevant.

Two kinds of sources were excluded. The first is studies without a documented data source. The most well-known case is the pre-publication on AI in the discovery of new materials, which reported large innovation gains in a research lab and had been widely commented on by economists and the press. On May 16, 2025, MIT's economics department stated that it had no confidence in the origin, reliability, or validity of its data and requested its withdrawal.[14] The case serves as a reminder that a study can influence public debate long before it is evaluated. The second kind involves self-reports linked to time series despite changes in the questionnaire. The Federal Reserve of St. Louis revised its 2024 survey results in November 2025 due to a change in the order of the questions.[15] The estimates before and after the change are presented here separately and are not read as a trend.

Harmonisation is limited to what the data allows. Studies measure resolved requests by hour, time of completion, graded quality, correctness of solution, earnings, hours, acceptance of results without correction and employment by age. These percentages do not have a common unit and therefore no mean or common effect size is calculated. Table 1 presents them in the initial units, with the direction of the result in relation to the measure of each study. References to assignments follow the O*NET classification in version 29.2, the same as that used by the competency assessments in Section 4, so that the description of the assignments remains constant while the database is updated. The selection of the ten studies is not exhaustive. It includes those that measure any of the three filters with a transparent design and for each filter at least one with a zero or unfavorable result. Cross-study differences may also reflect unequal task difficulty, short experimental horizons, self-selection into adoption and changing tool versions rather than a common treatment effect.

For the task comparison, a separate 12-task dataset was constructed across customer support, professional knowledge work and software development. Each task was coded for AI exposure, verification cost, contextual knowledge requirements and the cost of a missed error using the descriptions and results reported in the underlying studies. The categories Low, Medium and High are ordinal qualitative judgments rather than calibrated quantitative scores. Figure 5 presents six representative tasks from this dataset; the complete coding is reported in Appendix B.

Table 1. Evidence Matrix

STUDY	POPULATION	TASK	TECH PERIOD	DESIGN	OUTCOME	RESULT	EXTERNAL VALIDITY LIMIT
Brynjolfsson et al., 2025	5,172 support agents	Customer support	2020–21	Quasi-experiment	Requests/hour	+15%; larger novice gains	One firm/workflow
Noy & Zhang, 2023	453 professionals	Professional writing	Early 2023	RCT	Time; quality	∼40% time; +18% quality	Short online tasks
Dell'Acqua et al., 2026	758 consultants	Knowledge work	Jun 2023	RCT	Time; quality; accuracy	Gains inside frontier; −19 pp accuracy outside	Consultants; short horizon
Becker et al., 2025	16 developers; 246 tasks	Software development	Feb–Jun 2025	Task RCT	Completion time	**+19% time**	Experienced OSS developers
METR, 2026	57 developers; >800 tasks	Software development	From Aug 2025	Task RCT	Completion time	∼18% returning; −4% new; uncertain	Strong self-selection
Humlum & Vestergaard, 2026	25,000 Danish workers	11 occupations	2023–24	DiD + registers	Earnings; hours	**Null; >2% effects ruled out**	Denmark; broad outcomes
Bick et al., 2025	U.S. workers 18–64	Workplace GenAI	Nov 2024	Survey	Time saved	5.4% users; 1.4% all workers	Self-reported
Mertens et al., 2026	>60,000 worker evaluations	Occupational text tasks	2023–25 models	Capability assessment	Success without edits	Large task/job-family variation	Capability ≠ deployment
Shen & Tamkin, 2026	Developers	Learning/debuggin	2026	RCT	Learning; time	**−17% comprehension; no time gain**	Short learning horizon
Brynjolfsson et al., 2026	U.S. ADP payroll data	Employment by age/exposure	2022–26	Descriptive panel	Employment	∼19% youth gap	Descriptive association

The table does not sum up the results, but it does show a pattern that Section 4 examines in detail. Positive results are concentrated in stand-alone tasks where the quality of the result is quickly seen, either by the client or by the rater. The unfavorable ones occur where the correct solution depends on knowledge not contained in the instruction, or where the measure is learning and not the product. At the labor market level, the evidence is either null or descriptive. This pattern is compatible with the first hypothesis without proving it, because the studies differ simultaneously in population, tool, date and measure.

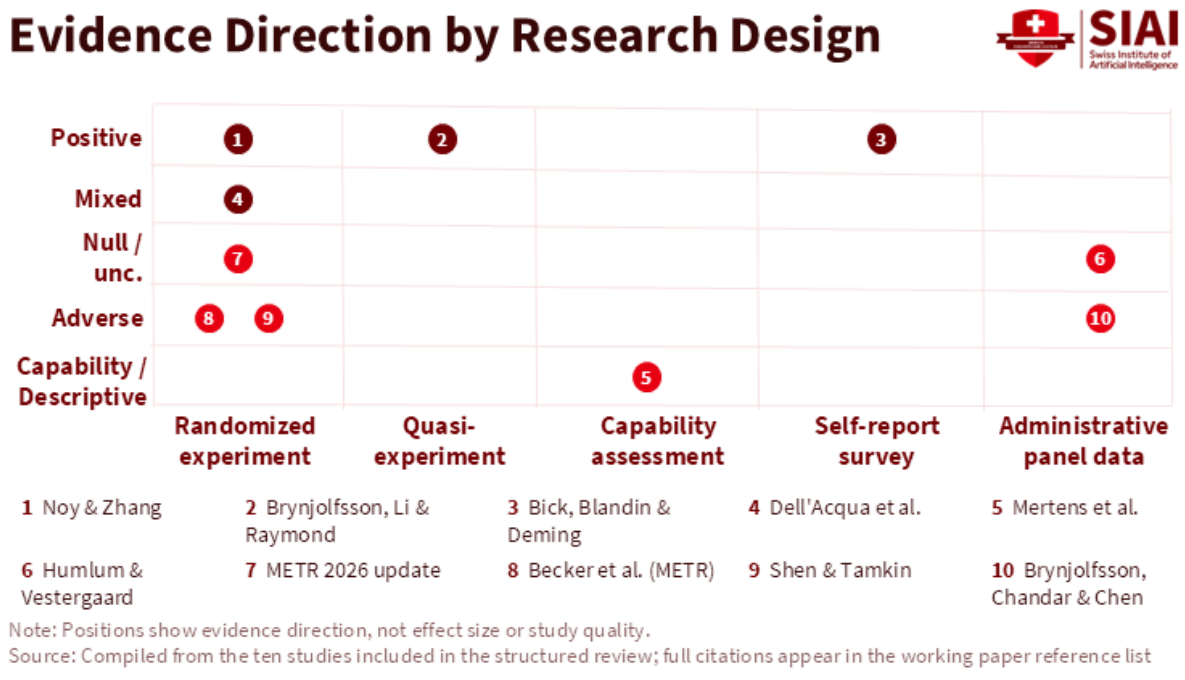
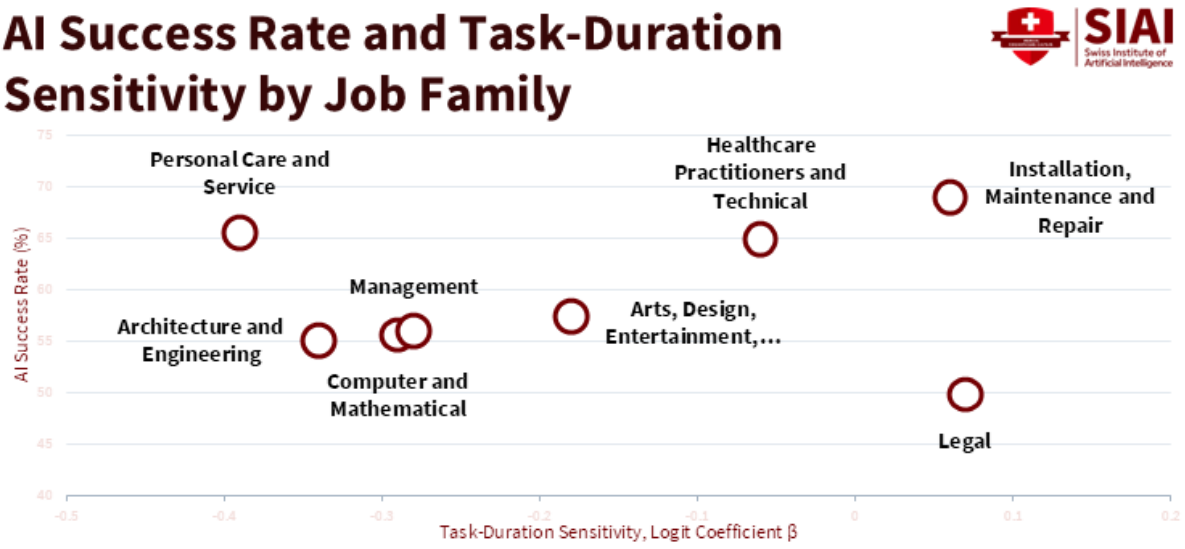


Figure 2: Evidence direction varies within as well as across research designs.

4 Comparative Evidence

If technical competence were the narrow point, the most recent data would show that models fail in most professional text tasks. They do not show it. MIT FutureTech started with 18,786 O*NET 29.2 tasks, kept 11,768, or 62.6%, in which a language model was estimated to save at least 10% of the time and asked workers with at least six months of experience in the respective occupation to score responses of more than 40 models in realistic work scenarios. In the 17,205 evaluations of the preliminary sample, 60% of the responses were deemed sufficient for a supervisor to accept them without any correction, with a range from 46.8% in legal work to 72.5% in installation, maintenance and repair work.[16] The July 2026 revision expands the evidence base to more than 6,000 text-based tasks and over 60,000 experienced-worker evaluations; the occupational-family breakdown discussed here comes from the preliminary sample. The authors caution that the percentages do not correspond to a share of automatable positions: the sample overrepresents professions that are easier to research, each scenario contained all the information the model needed, which in a real-world environment requires costly integration and it was not considered whether development is economically advantageous. The inverse of the same number is equally worthy of attention. Four out of ten responses needed corrections or re-execution and someone with knowledge of the profession had to identify which ones.



Note: Success rate is the share of outputs accepted without edits; β shows sensitivity to task duration.
Source: Recreated from Mertens et al. (2026), Crashing Waves vs. Rising Tides: Findings on AI Automation from Thousands of Worker Evaluations of Labor Market Tasks, Table 2, July 2026 revision

Figure 3: AI success is substantial across job families, but sensitivity to task duration differs sharply.

The dynamics of the ability reinforce the same conclusion from another direction. In the same data, success decreases with the duration of the task, but only mildly, since a tenfold increase in duration reduces the logarithm of the odds of success by 0.31, which corresponds to about 7.6 percentage points when the initial success rate is 60%. The improvement of the models is rapid. The time at which the failure rate of the pioneer models is halved is estimated at 2.19 years for five-minute tasks and gradually rises to 2.76 years for 24-hour tasks.[17] Two observations are important for verification. The longer the task, the slower its errors recede and at a

success rate of 80%, the duration of the tasks that the models could complete did not exceed about five minutes throughout the observation period. Average ability rises rapidly, but the almost complete reliability required by tasks with little tolerance for error is still several years away, according to the authors themselves.

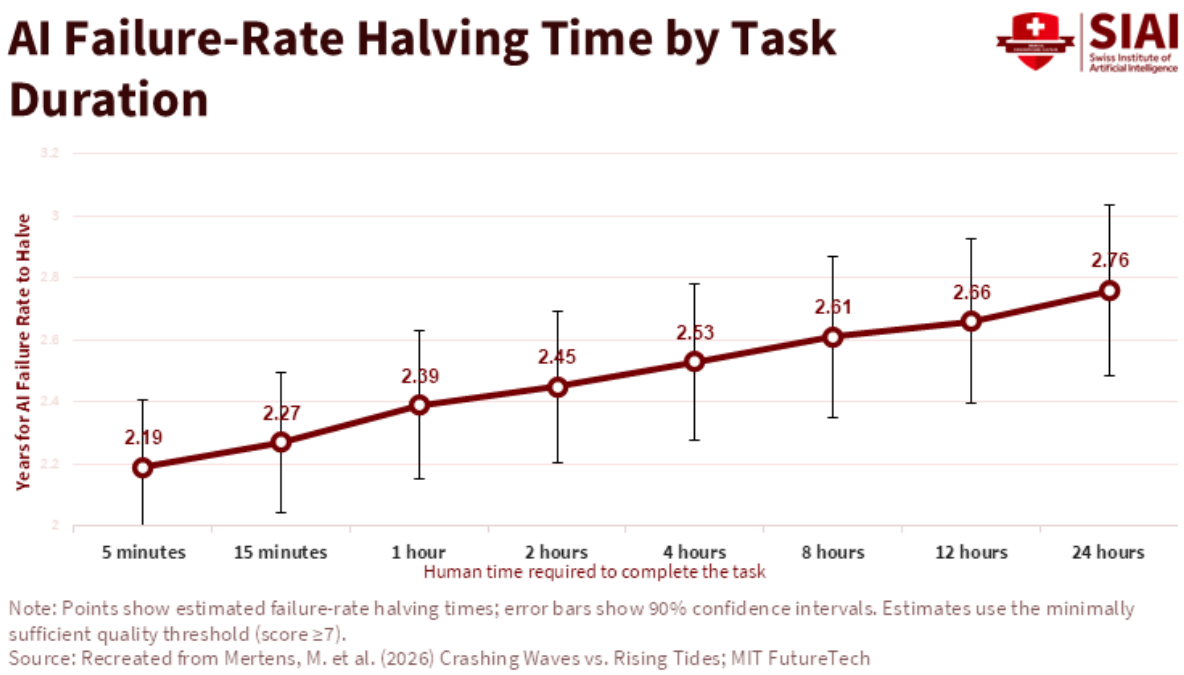


Figure 4: AI failure rates improve rapidly, but more slowly for longer tasks.

Field experiments show that high exposure translates into measured task-level productivity gains. In the customer support business, the assistant was trained on successful conversations from the same company, its suggestions appeared within the workflow and the employee could accept, modify or ignore them at the time of the conversation, while the quality of the response was immediately shown in the customer's reaction. The 15% increase in resolved requests per hour came mainly from less experienced and less competent employees, while more experienced ones saw small speed gains and small quality losses.[18] In the experiment with 453 professionals and written tasks such as press releases, short reports and sensitive emails, access to a language model reduced time by 40% and increased graded quality by 18%.[19] In the experiment with 758 consultants, for 18 realistic tasks within the model's capacity limit, those who had access completed 12.2% more tasks, 25.1% faster and with a quality higher by over 40%. In a task that was deliberately chosen outside the limit, they were 19 percentage points less likely to provide a correct solution than those who worked without AI.[20] Consultants did not know in advance which side of the limit each task was on and that uncertainty is the cost of verification in its purest form.

These tasks all have high exposure to any indicator. They differ in characteristics that the indicator does not capture, namely how difficult it is to control the outcome, how much contextual knowledge is required beyond the command and how much it costs to make a mistake that goes unnoticed. Figure 5 qualitatively classifies representative tasks across the review by AI exposure, verification cost, contextual knowledge required and the cost of a missed error. The classification is qualitative, is based on the descriptions of the studies and is not a measurement. The pattern is consistent with the first hypothesis: positive outcomes are concentrated where

control is cheap or integrated into the workflow and the work is self-contained, while unfavorable outcomes occur where the error is plausible and the correct response depends on implicit knowledge of the project or organization.

Exposure and Operational Constraints by Task Type

Task	Exposure	Verification cost	Context knowledge needed	Cost of a missed error
Customer support reply	High	Low	Low	Low
Professional writing draft	High	Low	Medium	Low
Consulting task, in frontier	High	Medium	Medium	Medium
Consulting task, out of frontier	High	High	High	High
Code in a mature repository	High	High	High	Medium
New library, learning phase	High	High	Medium	Medium

Note: High, Medium and Low are qualitative assessments based on the task descriptions in the underlying studies. They are ordinal categories, not calibrated quantitative scores.
Source: Adapted from Brynjolfsson et al. (2025); Noy and Zhang (2023); Dell'Acqua et al. (2026); Becker et al. (2025); Shen and Tamkin (2026)

Figure 5: Similar AI exposure can involve very different verification, context, and error costs.

The most rigorous testing of the first hypothesis comes from programming, where exposure is among the highest in any index. In a randomized trial by METR, 16 experienced open-source developers worked on 246 tasks between February and June 2025, in large repositories they had known for years, averaging over 22,000 stars and over a million lines of code. Each task was randomized to a condition where the use of AI tools was allowed or prohibited. Before starting, the developers predicted that AI would reduce the completion time by 24% and upon completion they estimated that it had reduced it by 20%. The measurement showed that time increased by 19%.[21] The distance between perception and measurement is 39 points in a time index where work without AI equals 100. This discrepancy is relevant because it shows that time spent checking, correcting and adapting generated code to implicit project requirements may be missed in self-reports, so perceived time savings cannot replace measurement.

The result that would further weaken the first hypothesis comes from the continuation of the same study. In a new test from August 2025, with 57 developers, 143 repositories and more than 800 tasks, METR estimated an 18% reduction in time, with a confidence interval ranging from a decrease of 38% to a 9% increase, for the ten developers who had also participated in the first study and a decrease of 4%, with a range from a decrease of 15% to an increase of 9%, for new entrants; For the first study, it reports an increase of 2% to an increase of 39%.[22] If tools based on autonomous agents reduced the cost of verification in programming within one year, then these costs may be transitory and the first hypothesis describes a phase rather than a permanent mechanism. The finding must be taken seriously. METR itself, however, considers the estimate unreliable. Between 30% and 50% of participants said they avoided submitting tasks they did not want to do without AI,

an increased number refused to participate for the same reason, pay was reduced from 150 to 50 per hour and measuring time became difficult when developers were handling multiple agents in parallel, which is why METR considers estimation a lower limit of actual acceleration. A cautious interpretation is that verification costs in programming have likely decreased, that the magnitude of the reduction is unknown and that the inability to measure itself is a finding: when employees refuse to work without the tool, comparison with and without it ceases to be feasible at the moment when it would have the greatest value.

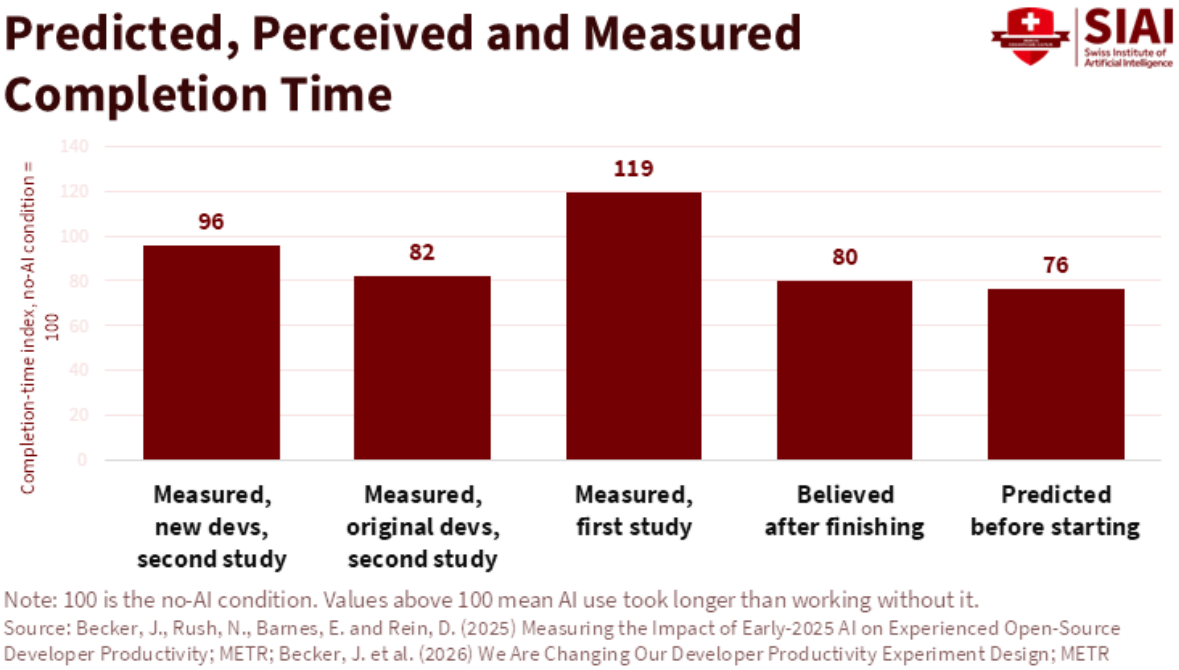


Figure 6: Developers' expectations and perceptions diverged sharply from measured productivity effects.

A major alternative explanation for the set of discrepancies is temporal. According to it, the results differ because jobs, businesses and countries are at different stages of adoption, not because different mechanisms work and as capacity rises and the cost of use falls, verification costs will shrink and results will converge toward the positive findings of the experiments. This explanation has empirical support. MIT FutureTech projections, which the authors themselves consider a ceiling, give success rates of 80% to 95% in most text tasks by 2029.[23] The model of Lindenlaub and colleagues predicts that the share of German workers using AI will rise from 44% to about 81% within three years, mainly due to a reduction in the cost of use rather than an improvement in models.[24] The Federal Bank of St. Louis also finds that those who have been using AI for at least six months apply it to more tasks, which is consistent with a history of costly learning that pays for itself over time.[25] For the pace of adoption, the temporal explanation is convincing and the paper accepts it.

Its limits are seen when the debate moves from adoption to outcomes. The elasticity of demand and the cost of an error in a legal opinion or a patient-to-patient directive are not functions of time and the near-complete reliability that such work requires is, according to the same projections, far from 2029. In Denmark, zero effects on earnings also applied to those who used AI daily, reported large profits, or worked in businesses that encouraged and financed the use and the authors attribute the effect to the small size of the gains combined with their weak transmission to wages.[26] These groups are already ahead on the time scale assumed by the

alternative explanation and the effect was no different. A recent work by SIAI concludes that access is not integration and that the economic value of AI increasingly arises from its connection to data, workflows, systems and decision-making processes.[27] The temporal explanation describes how access is propagated; it does not explain when access becomes production, nor what happens to the demand for labor when it occurs.

The scenario in Figure 7 makes the mechanism of the first hypothesis visible in numbers. A written task that without AI requires 60 minutes is assumed. The 40% reduction in production time is the value measured in the written assignment experiment described above and corresponds to a gross saving of 24 minutes. The rest of the parameters are assumptions. The control time ranges from zero to 30 minutes per task and with a probability of zero to 30% the test detects an error that requires the task to be redone from the beginning, at a full cost of 60 minutes. The net saving is equal to 24 minutes minus the control time minus the product of the rework probability multiplied by 60. With a 10% rework probability, the gain is zero when the check lasts 18 minutes, 20% when it lasts 12 and 30% when it lasts just six minutes. The scenario does not contain an estimate for any specific task. It shows that a measured gain of 40% is within a parameter range where moderate control and relatively rare errors are enough to make it disappear, which is compatible with the METR result for 2025.

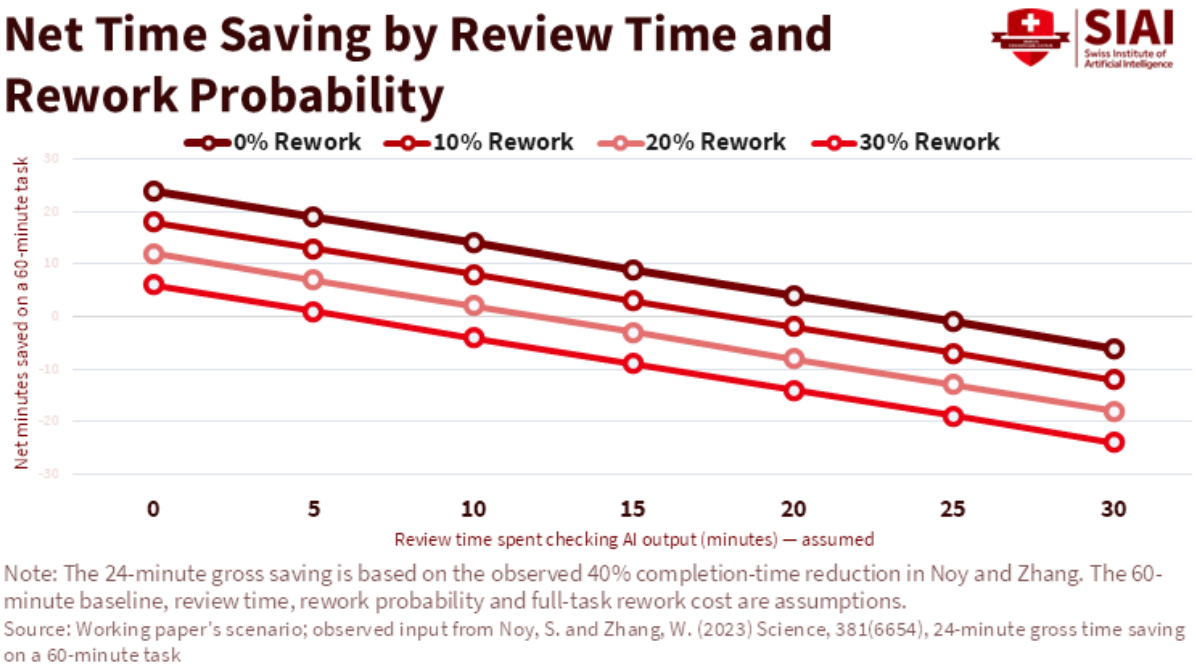


Figure 7: Review and rework can erase a large gross time saving.

The same logic explains why self-reports consistently paint a positive picture. In November 2024, American workers who had used generative AI in the previous week reported saving an average of 5.4% of their working hours, with 20.5% reporting four hours or more, while for all employees the savings corresponded to 1.4% of the hours.[28] The question asks the employee to imagine how many extra hours they would need without AI, so they record the perception of gross savings rather than the net time after checking and reworking. The second assumption concerns the next step and here the evidence is less and indirect. An analysis by The Economy argues that AI is thinning positions before eliminating them because businesses are leaving vacancies, merging entry-level roles, or stopping replacing those leaving and none of these moves show up as layoffs in the monthly

statistics.[29] The interpretation fits the second hypothesis, but it comes from articles, not measurement. The assumption remains consistent with the data but is poorly tested, because no study in the review measures the productivity of a job and the elasticity of demand for its product at the same time.

5 Expertise and Career Development

The Section 4 experiments agree on one point: AI squeezes performance differences between employees. In customer service, less experienced and less competent employees improved speed and quality, while more experienced employees hardly benefited. In the consultant experiment, those who were below the average performance on the baseline metric improved by 43% and those above it by 17%, compared to their own initial scores.[30] A plausible interpretation is that the tool transfers to younger people practices that were previously mastered by the best and that this is how younger people move up the experience curve faster. This interpretation carries weight, because it means that AI can broaden access to expertise that until now required years of working alongside experienced colleagues and this is evidence consistent with faster access to expertise for younger employees.

The third hypothesis challenges the step from performance convergence to capability convergence. Performance with the tool and proficiency without it are different quantities and the Section 4 experiments only measure the former. The only randomized review study that measures the second comes from Anthropic, a company that develops AI models and should be read with the caveat befitting a source with an interest in the subject, although its finding does not favor the company's product. Programmers learning a new AI-assisted Python library scored 17% lower on a conceptual-comprehension test administered a few minutes later, with no statistically significant acceleration of work and the largest difference appeared in debugging. The researchers attribute the result to the fact that the non-AI team encountered and corrected more bugs on its own. How the tool was used mattered: those who asked the model for conceptual explanations scored between 65% and 86%, while those who delegated writing or debugging scored between 24% and 39%.[31]

The study measures learning of a few minutes, not career years and cannot support a conclusion about long-term skill loss. But it does show the mechanism described in the third hypothesis. The action that directly improves the outcome, i.e. assigning the difficult task to the model, removes the error and its correction, which constitute the material of learning. Experiments with support workers and counsellors do not refute this conclusion, because neither measured the competence of the participants after the tool was removed. Direct convergence of performance is therefore compatible with both faster learning and dependency and the available data do not establish which mechanism is dominant. Both mechanisms may operate for different workers, depending on whether they use the tool to understand or to deliver.

At the occupation level, the Autor and Thompson framework described in Section 2 explains why this affects wages and entry into the profession. The jobs that businesses assign to young people are usually the least demanding of the profession. If they are assigned to models, the remaining work becomes more skilled and better-paid, but also more difficult as an entry point. US payroll data shows a pattern compatible with this mechanism. In absolute terms, the employment of 22- to 25-year-olds in the two most exposed quintile

occupations fell about 11% between November 2022 and June 2026, while in the three least exposed ones it increased by about 10%. The relative gap widened from 15% in the July 2025 data to 19% in the June 2026 data, the adjustment was mainly through fewer hires rather than more departures and the reductions are concentrated in occupations where the use of AI tends to automate rather than assist work. The August 2026 revision adds a distinction that matches the learning loop: youth employment decreased in occupations based on codified knowledge, taught through manuals and processes, while the employment of experienced people increased in occupations based on tacit knowledge, acquired through practice, guidance and repeated exposure to real cases.[32]

The authors themselves characterise the patterns as descriptive rather than causal. The gaps narrow when education is taken into account; some different trends are visible before the spread of generative AI and the gaps are larger in the ADP sample than in the national surveys, with the difference concentrated in education, health and public administration.[33] The result is therefore consistent with the mechanism in the third hypothesis as a descriptive pattern, not as evidence that AI caused the decline in recruitment.

Boston Consulting Group has formulated an organizational hypothesis that connects the level of work to the level of the profession. Based on interviews with tech executives and its experience with clients, it argues that traditional pyramids give way to smaller teams where experienced staff work directly with AI, that new hires are expected to contribute at a higher level from day one and that businesses need to train employees to oversee the work that AI produces instead of producing it themselves from scratch.[34] The case comes from an interested source and interviews with tech companies, so the question is whether it extends beyond them. Payroll data provides a partial answer, since the gap for young people remains when technology companies and IT professions are excluded.[35] But the case contains a contradiction that it does not resolve. Overseeing the work of AI requires the judgment gained by performing the same task and an organization that hires fewer young people and asks them to supervise from day one relies on a stock of expertise that does not renew.

This contradiction has an economic form. For a single company, replacing introductory work with AI is often rational, because the cost of training a new employee is borne by it while the benefit of their expertise can be reaped by their next employer. When multiple companies make the same choice, the supply of future auditors decreases and the cost of verification, which the first hypothesis faces as an obstacle to deployment, increases for everyone with a delay. None of the available data yet tracks a cohort of young workers who have been learning with AI for years and without such data the third hypothesis remains plausible but untested over the horizon that matters.

6 Distribution and Implications

Exposure is unevenly distributed and this inequality is the starting point of the distributional analysis, not its conclusion. According to the International Labour Office, 4.7% of global female employment belongs to the highest exposure rating, compared to 2.4% for men and 5.7% compared to 3.1% in the immediately preceding one. In high-income countries, 41% of female employment is exposed, with 14.4% in the third and 9.6% in the fourth grade, while for men the corresponding overall figure is 28%, with 3.5% in the fourth grade. The

difference is mainly due to the concentration of women in office positions, which remain the most exposed, such as data entry clerks and accounting clerks.[36]

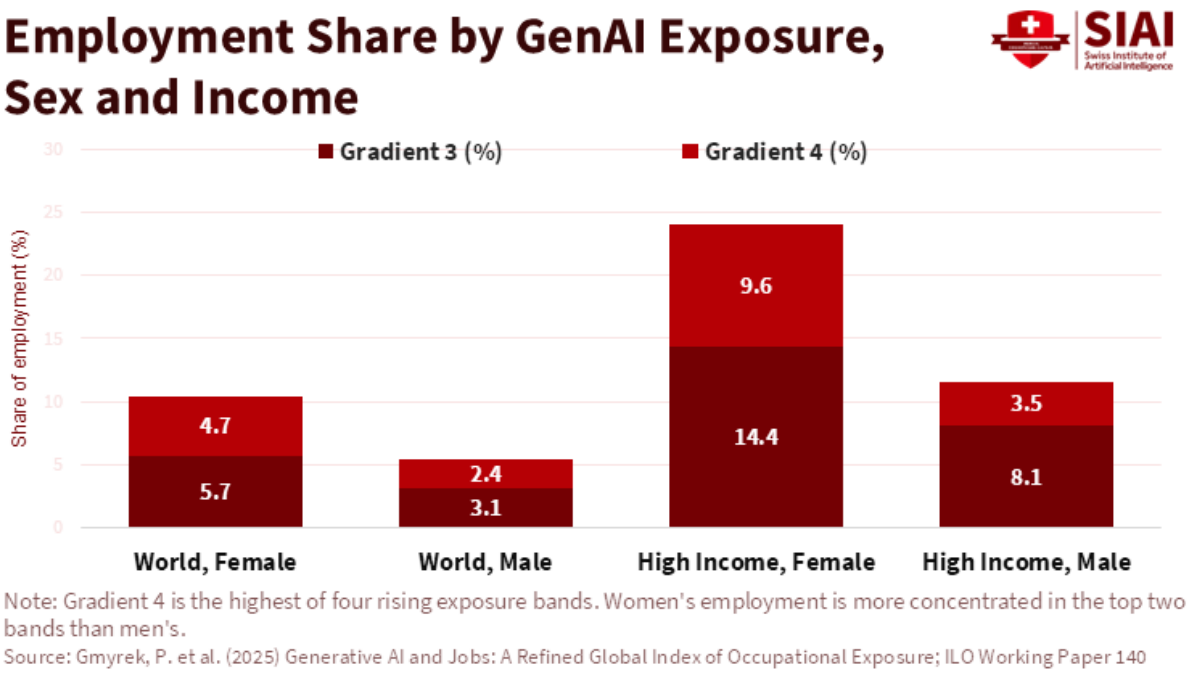


Figure 8: High GenAI exposure is more concentrated in female employment, especially in high-income economies.

Based on the context of the document, this breakdown describes only the first filter. If the uptake in office positions handling sensitive files falls short of exposure, such as in medical secretaries, the immediate effect will be less than the indicator implies. But if these jobs are self-contained, cheap to control and with inelastic demand, they have exactly the characteristics for which the second hypothesis predicts a reduction in labour demand without a corresponding increase in output. The groups most likely to lose ground are therefore those where high exposure, low verification costs and limited demand coincide and are not necessarily the same as the most exposed ones. An indicator that combines exposure with control costs and with the elasticity of demand per job would be more useful for policy than any improvement in the exposure indicator alone.

The practical consequences vary by institution. For workers in the learning phase, the code-learning experiment shows that the use of AI for explanations sustains learning, while delegating execution reduces it and this distinction can become an explicit work practice. For managers, the conclusion is metric: the net time after the audit, the percentage of results accepted without correction and the frequency of rework are the metrics that indicate whether an assignment is worthwhile and self-reports do not record them. For businesses, verification costs should be included in the development decision before purchasing licenses and keeping a part of the introductory work as a learning space should be treated as an investment in the future pool of auditors, not as inactivity. A practical testimony in Forbes, written by a remuneration software company executive, points to a risk linked to the above: when hiring, paying and evaluation decisions are accelerated with AI, the technology escalates flawed processes that were previously absorbed without much cost.[37]

Professional bodies can define who is responsible for an AI-generated result and what level of control is required per job category, so that the cost of verification is not implicitly passed on to younger people or those

who do not have the experience to bear it. Universities can separately assess performance with the tool and competence without it, because the former does not certify the latter. For governments, public intervention is justified in two respects. The first is measurement, since recruitment to entry-level positions by occupation and age shows the adjustment much earlier than unemployment and no business has an incentive to publish them. The second is the possible underinvestment in transferable skills, when the benefit of young people's education is spilled over to other employers. Whether this underinvestment is real and whether co-financing adds training rather than subsidizing what would be done anyway remains an empirical question.

The limits of the document are specific. The experiments cover a few populations, mostly in the US, in standalone tasks and over a horizon of hours or months. Administrative data cover broad populations, but they cannot distinguish which filter blocked an outcome and payroll data for young people are descriptive. No study directly measures verification costs across different jobs, which is why Figure 7 relies on assumptions rather than data. No dataset tracks the skill development of those starting their careers with AI. Results in programming are changing so quickly that a 2025 study can already describe a past situation.

7 Conclusion - From AI Exposure to the Value of Human Work

The value of human labor is judged on two points that lie beyond technical ability: whether assigning a task to a model is advantageous when the control is calculated and what happens to demand when the work becomes cheaper. The evidence in the review is consistent with this account. Gains occur in stand-alone jobs where control is cheap and error is visible; losses or reversed where the error is plausible and the correct answer depends on implicit knowledge and at the labor-market level they coexist with zero average changes in earnings in Denmark and descriptive evidence of fewer young hires in highly exposed U.S. occupations where AI use tends toward automation. The rapid improvement of models explains the pace of adoption, but not its transformation into outcomes for workers.

The long-term tension lies in the connection between verification and learning. Controlling the work of AI requires expertise and expertise is formed in the tasks that are now assigned first to models. If the reliability of models improves quickly enough, the need for auditors will decrease before their shortage is felt. If not, businesses will find themselves with cheaper production and more expensive control and employees with fewer paths to the expertise that will be requested. The available data do not establish which of the two will occur first.

The policy implication is narrow. Statistical offices and businesses need two measures that are currently missing: net time saved after auditing and reworking by job category and recruitment to entry-level positions by occupation and age. The first would show where adoption is actually working. The second would show where the training of the next generation of auditors has started to dwindle, long before it appears in unemployment statistics.

References

[1, 18] Brynjolfsson, E., Li, D. and Raymond, L. (2025) 'Generative AI at Work', *The Quarterly Journal of Economics*, 140(2), pp. 889–942.

[2, 26] Humlum, A. and Vestergaard, E. (2026) *Still Waters, Rapid Currents: Early Labor Market Transformation under Generative AI.* NBER Working Paper No. 33777, revised March 2026. Cambridge, MA: National Bureau of Economic Research.

[3, 32, 33, 35] Brynjolfsson, E., Chandar, B. and Chen, R. (2026) *Canaries in the Coal Mine? Six Facts about the Recent Employment Effects of Artificial Intelligence.* Revised August 2026. Stanford, CA: Stanford Digital Economy Lab.

[4, 12] Ben-Ishai, G. and Thompson, N.C. (2026) *Workforce Policy for the Age of AI: Recommendations from the Economic Literature.* Washington, DC: Brookings Institution, 15 September.

[5, 16, 17, 23] Mertens, M., Kuzee, A., Harris, B.S., Lyu, H., Li, W., Rosenfeld, J., Anto, M., Fleming, M. and Thompson, N.C. (2026) *Crashing Waves vs. Rising Tides: Preliminary Findings on AI Automation from Thousands of Worker Evaluations of Labor Market Tasks.* Revised July 2026. Cambridge, MA: MIT FutureTech.

[6, 11, 25] Bick, A., Blandin, A., Deming, D.J. and Schumacher, T. (2026) 'What Work Does Generative AI Do?', *On the Economy*, Federal Reserve Bank of St. Louis, 1 September.

[7] Eurostat (2026) *The Use of Artificial Intelligence Technologies in the European Union: Key Results, 2026 Edition.* Luxembourg: Publications Office of the European Union.

[8, 36] Gmyrek, P., Berg, J., Kamiński, K., Konopczyński, F., Ładna, A., Rosłaniec, K. and Troszyński, M. (2025) *Generative AI and Jobs: A Refined Global Index of Occupational Exposure.* ILO Working Paper 140. Geneva: International Labour Organization.

[9, 24] Lindenlaub, I., Oh, R., Rodríguez, M.A. and Veldkamp, L. (2026) *Beyond Exposure: Predicting AI Adoption Based on Comparative Advantage.* NBER Working Paper No. 35271. Cambridge, MA: National Bureau of Economic Research.

[10] Svanberg, M.S., Li, W., Fleming, M., Goehring, B.C. and Thompson, N.C. (2024) *Beyond AI Exposure: Which Tasks Are Cost-Effective to Automate with Computer Vision?* Working paper. Cambridge, MA: MIT FutureTech.

[13] Autor, D. and Thompson, N.C. (2025) 'Expertise', *Journal of the European Economic Association*, 23, pp. 1203–1271.

[14] MIT Department of Economics (2025) *Assuring an Accurate Research Record.* Cambridge, MA: Massachusetts Institute of Technology, 16 May.

[15, 28] Bick, A., Blandin, A. and Deming, D.J. (2025) 'The Impact of Generative AI on Work Productivity',

On the Economy, Federal Reserve Bank of St. Louis, 27 February.

[19] Noy, S. and Zhang, W. (2023) 'Experimental Evidence on the Productivity Effects of Generative Artificial Intelligence', *Science*, 381(6654), pp. 187–192.

[20, 30] Dell'Acqua, F., McFowland III, E., Mollick, E.R., Lifshitz-Assaf, H., Kellogg, K., Rajendran, S., Krayer, L., Candelon, F. and Lakhani, K.R. (2023) *Navigating the Jagged Technological Frontier: Field Experimental Evidence of the Effects of AI on Knowledge Worker Productivity and Quality.* Harvard Business School Working Paper No. 24-013.

[21] Becker, J., Rush, N., Barnes, E. and Rein, D. (2025) *Measuring the Impact of Early-2025 AI on Experienced Open-Source Developer Productivity.* arXiv:2507.09089. Berkeley, CA: METR.

[22] Becker, J., Rush, N., Cunningham, T., Rein, D. and Mahamud, K. (2026) 'We Are Changing Our Developer Productivity Experiment Design', *METR Blog*, 24 February.

[27] Lee, K. (2026) 'From AI Access to Organizational Capability: Pricing the Corporate AI Transition', *SIAI Research*, Swiss Institute of Artificial Intelligence, 9 August.

[29] The Economy Editorial Board (2026) 'AI Productivity Gains Will Thin Jobs Before They Erase Them', *The Economy AI Review*, 1 July.

[31] Shen, J.H. and Tamkin, A. (2026) *How AI Impacts Skill Formation.* arXiv:2601.20245. San Francisco, CA: Anthropic.

[34] Bedard, J., Lucero, E., Ebeling, R., Breitling, F., Garcia-Garcia, C. and Sakhuja, A. (2025) *AI Is Moving Faster Than Your Workforce Strategy. Are You Ready?* Boston, MA: Boston Consulting Group, 15 September.

[37] Colacurcio, M. (2026) 'AI Strategy Is Workforce Strategy, So HR's Involvement Is Vital', *Forbes Human Resources Council*, 11 June.

Appendix B. Task Comparison Dataset

FUNCTION	TASK	SOURCE	AI EXPOSURE	VERIFICATION COST	CONTEXT KNOWLEDGE NEEDED	COST OF A MISSED ERRO
Customer support	Customer support reply	Brynjolfsson et al. (2025)	High	Low	Low	Low
Professional/consulting knowledge work	Professional writing draft	Noy and Zhang (2023)	High	Low	Medium	Low
Professional/consulting knowledge work	Press release	Noy and Zhang (2023)	High	Low	Medium	Low
Professional/consulting knowledge work	Short report	Noy and Zhang (2023)	High	Medium	Medium	Medium
Professional/consulting knowledge work	Sensitive email	Noy and Zhang (2023)	High	Medium	High	Medium
Professional/consulting knowledge work	Consulting task, inside frontier	Dell'Acqua et al. (2026)	High	Medium	Medium	Medium
Professional/consulting knowledge work	Consulting task, outside frontier	Dell'Acqua et al. (2026)	High	High	High	High
Software development	Bug fix in mature repository	Becker et al. (2025)	High	High	High	Medium
Software development	Feature implementation in mature repository	Becker et al. (2025)	High	High	High	Medium
Software development	Code refactor in mature repository	Becker et al. (2025)	High	High	High	Medium
Software development	New-library coding task	Shen and Tamkin (2026)	High	High	Medium	Medium
Software development	New-library debugging task	Shen and Tamkin (2026)	High	High	Medium	Medium