Does the choice of AI-exposure measure matter? A dollar-weighted comparison, specified

Alessandro Conway

First draft 12 June 2026. This version September 2026.

Working paper. The comparison design and its statistic are fixed and the code is committed, the measures are loaded through one interface, and the computation is not yet run.

Extended summary

At least six serious measures score how exposed each occupation is to artificial intelligence, and a study that reports the share of jobs at risk picks one of them, usually without saying why. If the measures agree, the choice is harmless. If they disagree, then every ranking of at-risk occupations, every regional exposure map, and every policy brief built on one of them reflects the instrument as well as the labour market, and no reader can tell by how much. The companion paper (Conway, 2026) placed each measure at the level of the labour market where it was collected and found the capability measures closer to one construct than their agreement suggests. The Srinivasan automation score is the Eloundou construction re-run on a later release with a later model, so their correlation of 0.833 shows reproducibility, and the Srinivasan augmentation score is one minus a Herfindahl index over the same classification, so a quadratic in automation accounts for 77 per cent of it. The companion paper also showed that asking whether a capability measure predicts observed usage selects on the outcome, since the usage data record only tasks a deployed model has already been used for.

This note specifies the comparison that policy needs and fixes its statistic. Rank correlations, the usual comparison, weight an occupation of five hundred people and one of three million equally, and they are dominated by the tails, where every measure agrees that telemarketers are exposed and roofers are not. Policy questions are posed on people and on dollars, and they are posed in the middle of the distribution, where the marginal occupation is classified one way or the other. The statistic is therefore the share of the wage bill on which two measures disagree about whether an occupation is exposed, at a common cut, together with the agreement rate by decile of the pooled ranking, which separates the extremes from the interior. The design holds population, occupational base, vintage, and every downstream computation fixed and varies only the measure, which is possible because the model built alongside the companion paper takes each measure as an interchangeable input. Two downstream tests follow. The first is the short-run association of each measure with realised wage and employment changes from 2022 to 2025, framed as association and inferred by placebo years for the reason the companion paper gives, and the second is one adoption scenario computed under each measure, with the dispersion in the skill-exposure ranking it produces. The readings that decide whether the choice matters are written down before the numbers are.

1 The question

The measures differ in what they observe. Felten, Raj and Seamans (2021) score the fifty-two abilities of the Occupational Information Network from a survey linking ten AI applications to abilities, and convert the score to occupations through importance weights. Webb (2020) scores occupations by the overlap between patent text and task descriptions. Brynjolfsson, Mitchell and

Rock (2018) score tasks on a rubric of suitability for machine learning. Eloundou, Manning, Mishkin and Rock (2023) rate each of the Network's task statements on whether a language model could halve the time it takes, with human and machine raters. Srinivasan, Chen and Zakerinia (2024) re-run that construction on a later release and add an augmentation score. The Anthropic Economic Index (Handa et al., 2025) observes which tasks people bring to a deployed model and reports, for each, the share of use that is automation. Three of these are pre-language-model and three are post, and two are built from tasks, one from abilities, one from patents, and one from observed use.

Each of them is aggregated to the occupation level in a different way and then compared as though it described one object. The companion paper placed each at its native level and tested the step that converts a measure between levels, where that step could be tested. It established that the propagation from tasks to abilities discriminates among AI applications in the right order, reproducing the Felten ability scores at 0.667 for reading comprehension against 0.172 for image recognition. It did not make the comparison on the object that reaches a policy brief, the classification of occupations, and of the dollars they earn, as exposed or not, and that comparison is what this note specifies.

2 The statistic

Let ω_o be occupation o's share of the wage bill, employment times mean wage over the 829 occupations with observed wages in the Occupational Employment and Wage Statistics, which together earn 10.70 trillion dollars. For a measure a and a cut q, let $X_a(q)$ be the set of occupations that measure a ranks highest, taken from the top until their wage-bill shares sum to q. The disagreement between measures a and b at cut q is

$$D_{ab}(q) = \sum_o \omega_o \; \left| \; 1\left[o \in X_a(q)\right] - 1\left[o \in X_b(q)\right] \; \right|,$$

the share of the wage bill that one measure classifies as exposed and the other does not. It runs from zero, when the two measures pick the same dollars, to $2q$, when they pick disjoint sets. It is reported at cuts of a tenth, a quarter, and a half of the wage bill for every pair of measures, on the same 829 occupations, with a bootstrap interval from resampling occupations. Employment-weighted and unweighted versions are reported beside it, so that a reader can see how much of the disagreement comes from large occupations.

The second statistic locates the disagreement. Rank the occupations by the mean of their ranks under a and b, cut that pooled ranking into deciles of the wage bill, and within each decile report the share of dollars on which the two measures agree at the cut. Agreement that is high in the first and last deciles and low in the fourth to seventh is the pattern the extremes-agree hypothesis predicts, and it is the pattern that would make the choice of measure bind on the occupations that a targeting decision turns on. Agreement that is uniform across deciles would mean that the measures disagree about everything equally, which is a different finding and a simpler one.

3 Design

The comparison is clean only if the measure is the only thing that changes, and three sources of discretion are fixed for that reason. The occupational base is the 829 occupations with observed wages in the 2025 release, on the 2018 classification, and a measure published on an earlier

classification is mapped across by the Network's own crosswalk with the allocation logged. The vintage of each measure is its published release, and a measure with several releases enters at each, so that drift between releases of one measure is reported in the same units as disagreement between measures. The aggregation from a measure's native level to occupations is the same for every task-level measure, the importance-weighted mean over the occupation's tasks, so that no measure gains or loses from the way it was aggregated.

The computations downstream of the measure are written so that each exposure source enters through the same definition, a capital share and a substitution elasticity for a given occupation and task, and nothing downstream depends on which source is in use. The same skill exposure, role resilience ranking, or scenario can therefore be recomputed under each measure and the spread reported. Six of the measures are already in that form, those of Felten, Webb, the machine-learning suitability scores, the two Srinivasan scores, and the Anthropic index, together with a blend that averages whichever have data for a task, and the Eloundou ratings enter directly from their published label set. The harmonisation choices are themselves a result of this note. The log of every crosswalk allocation, every rescaling, and every release is published with the comparison, because a reader who wants to reproduce a disagreement needs to know which of those choices produced it.

4 Downstream tests

The first test concerns which measure tracks what happened. For each measure, the change in log mean wage and in log employment from 2022 to 2025 across the 829 occupations is regressed on the measure's exposure score, on the balanced panel the companion paper assembled, and the coefficients are reported side by side. The companion paper established the limits of this test. On the stock of employment every deviation from the pre-2020 path begins in 2020 or 2021, before language models were released, and no skill group moved by more than 1.51 times its own year-to-year volatility after 2022. The association is therefore reported as association, inference comes from the rank of the treated years among placebo designations of the panel's own transitions rather than from conventional standard errors, and a measure that tracks the 2022 to 2025 change is one that tracks a period the pandemic dominates. The test is still worth running because the measures will differ in it, and a measure whose score is uncorrelated with a realised change that another measure's score predicts is informative about what it observes.

The second test concerns whether the choice of measure changes results further downstream. One adoption scenario, a proportional rise in the substitution elasticity of every exposed task, is computed under each measure in turn, and the output is a ranking of twenty-five skill groups by the share of the wage bill they lose. The spread across measures in that ranking, measured as the range of each skill group's rank and as the share of the wage bill whose skill-exposure sign flips between measures, shows whether the choice of measure reorders a leaderboard or moves a structural quantity. Reordering the top ten of a list would be a nuisance, while a change in which skills lose share under the same scenario would be a finding about the measures.

5 Prespecification

The inputs are the six measures at their published releases, the Eloundou label set, the 2025 wage and employment panel, the Network crosswalk, and the model at the version that produced the

companion paper. The readings are fixed as well. If D_{ab} at a quarter of the wage bill is below 0.10 for every pair, the choice is harmless, and the note reports it as such. If the pairs fall into two families, capability against capability below 0.15 and capability against usage above 0.30, then the measures observe two objects, capability and use, and the default that follows is to report one measure from each family. If the decile profile shows agreement above 0.9 in the tails and below 0.6 in the interior, the choice binds where targeting happens, and a study that reports one measure owes its readers the interior. The practical output is one table giving, for each measure, what it observes, at what date, at what level, and which question it can therefore answer, so that a reader who has to pick can pick by construction rather than by habit.

6 Conclusion

A study that reports a share of jobs exposed to artificial intelligence reports the view of one measure, and the measures observe different things at different dates and at different levels. The comparison that matters for policy is on dollars, at a common cut, in the middle of the distribution, with everything but the measure held fixed, and the model built alongside the companion paper makes that comparison possible. The statistic, the design, and the readings are set out here, and the computation follows.

References

Brynjolfsson, E., Mitchell, T. and Rock, D. (2018). What can machines learn, and what does it mean for occupations and the economy? *AEA Papers and Proceedings*, 108, 43–47.

Conway, A. (2026). AI exposure at three levels: occupations, tasks and skills as one wage bill. Working Paper AC-WP-2026-04.

Eloundou, T., Manning, S., Mishkin, P. and Rock, D. (2023). GPTs are GPTs: an early look at the labor market impact potential of large language models. arXiv:2303.10130.

Felten, E., Raj, M. and Seamans, R. (2021). Occupational, industry, and geographic exposure to artificial intelligence: a novel dataset and its potential uses. *Strategic Management Journal*, 42(12), 2195–2217.

Handa, K., Tamkin, A., McCain, M. et al. (2025). Which economic tasks are performed with AI? Evidence from millions of Claude conversations. Anthropic, arXiv:2503.04761.

Srinivasan, S., Chen, D. and Zakerinia, S. (2024). Occupational exposure to generative AI: automation and augmentation scores. Working paper.

Webb, M. (2020). The impact of artificial intelligence on the labor market. Stanford University, working paper.