Replace, augment, support: a nested task-level measure of generative-AI exposure for the full O*NET catalogue

Carles Gómez, María Jimena Contreras, Manuel Hidalgo and Jorge Galindo (EsadeEcPol)

Working paper (first draft, October 2026)

Abstract

We build a three-level nested measure of potential exposure to generative AI: whether AI can *replace* the output of a task at the quality of a typical worker (*Re*), whether it *augments* the worker enough to halve the time on the task or lift its quality into the top quartile (*Au*), and whether it merely *supports* the task with a usable, task-specific input (*Su*). A panel of three frontier models judges every task against a dated inventory of 33 AI capability benchmarks, and a Delphi round settles their disagreements. The measure covers all 24,087 task-activity pairs of the O*NET 31.0 catalogue, in 923 occupations. Weighted by the time workers spend on each task, the average occupation has 37.8% of its working time replaceable, 48.4% augmentable and 79.2% supportable. Agreement among the raters is high for replacement and lower for the other two layers. A separate layer measures occupation-level "shields" which could limit the implementation of the measured potential exposure (relational nexus, organisational responsibility, authorship). All labels, scores and shields are released at occupation level.

Keywords: generative AI, task exposure, O*NET, large language model annotation, Delphi, occupational crosswalks.
JEL: J23, J24, O33.

1 Introduction

Occupational exposure indices are where much of the policy work on AI and jobs starts (in Spain, Banco de España, 2026). We see four problems with them.

First, most of them compress a task's relation to AI into a single score (Felten et al., 2021; Webb, 2020; Eloundou et al., 2024; Gmyrek et al., 2023), which mixes substitution and complementarity: "AI can cut the time on this task by half" and "AI can do this task" can have different labour-market effects, yet both count as exposure (Acemoglu, 2025).

Second, when judgement comes from people, it covers only part of the catalogue and has expertise issues. Frey and Osborne (2017) hand-labelled 70 of 702 occupations with a group of machine-learning researchers and extended the labels to the rest with a classifier. The human annotators of Eloundou et al. (2024) rated every detailed work activity but only a subset of tasks, and the authors themselves note that a group familiar with language models but not

occupationally diverse may misjudge tasks in occupations it does not know. If professionals were used for each SOC, the gap would be the opposite: they would know the work but rarely follow the AI frontier. Rating all 24,087 task-activity pairs of the catalogue with three raters takes 344,874 first-round judgements; at that scale a human panel whose quality could be argued as uniform is prohibitively expensive.

Third, potential exposure moves with the technology: models released today do things that last year's could not. That is an issue on its own, which adds to the difficulty of any sort of time analysis. Methodologies followed in the past are rarely replicable, and are normally outdated for the present.

Fourth, an exposure label rarely comes with granularity. Without it, exposure cannot be read by kind of work: one cannot tell, for instance, whether AI takes over the tasks or raises what can be done.

Our design builds directly on Parshall and Lopez-Luzuriaga (2026), who classify a very similar universe (O*NET) of task-activity pairs,[1] read in the context of their occupation, with three models and a Delphi-style consensus round. We keep their unit, their panel design and, in adapted form, their first two axes. What we change is the exposure question (three nested layers with explicit thresholds, where they ask a single question about halving the time), the evidence the raters work from (a dated benchmark inventory), the raters themselves (frontier models, where their production run uses mid-tier ones) and the landing (official crosswalks to CNO-11 and a separate layer of shields). We drop their regulatory axis, which is built for the United States.

We try to address these four problems in our contribution. In the first, for every task-activity pair we ask whether frontier AI can replace the output (Re); where it cannot, whether it raises productivity past an explicit threshold, through time or through quality (Au); and where it does not, whether it can still supply a substantive input that belongs to the task (Su). Because each layer is asked only where the previous one is zero, the margins $Au - Re$ and $Su - Au$ can be read as objects of their own.

In the second, against partial and costly human judgement, we exploit the potential of agentic LLMs with three frontier models that rate every pair blind and a Delphi round that reconciles them, so the whole O*NET universe is classified, at a cost that is small compared to a human panel (Section 5). The raters read each task in the context of its occupation and judge what AI can do against the bible described next. Quality checks to avoid known limitations of the technology at this scale are carried out, and redundancy (three models) is favoured over cost effectiveness.[2]

The third, against measures that cannot be replicated or brought up to date, is the "bible" itself, a dated inventory of 33 AI capability benchmarks with the best score reached by 31 August 2026 and the human baseline when available. It does two jobs. Within a wave, it puts the three raters on the same evidence, which is not trivial, since otherwise their different training cutoff dates would bring up inconsistencies, and each would place the frontier in a

[1] Parshall and Lopez-Luzuriaga (2026) work with O*NET 30.2: 23,850 task-activity pairs from 18,796 tasks. We use release 31.0, with 24,087 pairs from 18,838 tasks; both cover 923 occupations.

[2] An argument could be made that the most capable model would give better results alone rather than having to arrive at a consensus with others.

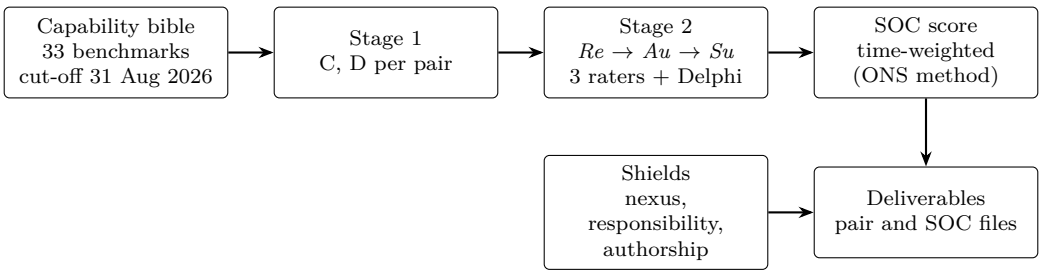


Figure 1: The measure in the main text, from the capability inventory to occupation scores. Appendix A carries it to the Spanish classification (Figure 3).

different spot. With a bible they start from the same one. Across waves, it makes the measure repeatable: updating the bible to a later date and re-running the panel with the same prompts and raters gives a second wave, and so a series of potential exposure. Parshall and Lopez-Luzuriaga (2026) also propose re-running their classification to follow the frontier; the bible adds a stated frontier at each date, so a change between waves can be traced to the evidence and not only to newer raters. The series is imperfect, because the bible has limits of its own and other changes also move labels (Section 9), and because the labour market evolves, and O*NET tries to do so too, with regular updates and tweaks to reflect the new reality, but with the frontier moving this fast an imperfect series is better than none.

In the fourth, against labels without granularity, before addressing AI exposure, we go through stage 1, which records how far the task can be written down as a procedure (codifiability, C) and how much it needs a body in the world (physical requirement, D), on ladders adapted from Parshall and Lopez-Luzuriaga (2026), together with the O*NET abilities it draws on. Exposure can then be read by kind of work. Of the pairs that call for judgement between defensible alternatives (C2), 50.5% are replaceable, against 13.7% of those that rest on tacit expert judgement (C3), where augmentation still reaches 47.7%. We then exploit that gap and build our measure of expert complementarity: the working time where AI raises what an expert can do without substitution (Section 7).

We also keep capability and protection apart, in a similar way to Parshall and Lopez-Luzuriaga (2026) with regulation, which they build because exposure is technical capability, and there are frictions that hold back capability from turning into substitution. We propose shields (relational ties, responsibility, authorship), which we measure in a separate layer and report beside exposure, as a vector. The labels, the occupation and CNO-11 scores and the shields are all released with their legends (Section 10).
In the average occupation, 37.8% of working time is replaceable, 48.4% augmentable and 79.2% supportable (Table 4).

This is a methods paper. The main text stops where the measure is complete, at O*NET level (Figure 1), and documents the apparatus we built on top of it. An appendix translates the scores to the Spanish classification through official crosswalks only (O*NET to ESCO, and INE's correspondence between CNO-11 and ISCO-08), averaging and logging every collision. The applications to labour-force microdata are in separate papers that rely on this one, and so is regulation, which can only be measured on Spanish occupational groups.

2 Related literature

We follow the task approach, in which occupations are bundles of tasks with different exposure to technology (Autor et al., 2003; Acemoglu and Autor, 2011).

Exposure indices. A first generation estimated automation risk for whole occupations (Frey and Osborne, 2017), and task-based estimates then showed that this overstates the risk, since tasks within an occupation differ (Arntz et al., 2016). Later measures map AI capabilities to occupational abilities (Felten et al., 2021), match patent text to task descriptions (Webb, 2020), or rate detailed work activities for their suitability for machine learning (Brynjolfsson et al., 2018). With generative AI came rubric-based task annotation by humans and LLMs (Eloundou et al., 2024), which was extended to global settings through ISCO tasks (Gmyrek et al., 2023) and refined with a worker survey and Delphi-style expert rounds (Gmyrek et al., 2025), together with a language-modelling version of the ability-based index (Felten et al., 2023) and a European mapping from benchmarks to abilities (Casas et al., 2026). We take the halving threshold of the E_1 label of Eloundou et al. (2024) for our Au time channel, and Au as a whole is closest to their β. Re is stricter than any of their labels, although they note that their halving threshold corresponded closely to whether a model could perform the core of a task or nearly all of it. We compare the two on the full set of occupations in Section 8.

Models as annotators. Language models can match or exceed crowd annotators on text classification (Gilardi et al., 2023), and a panel of diverse models can outperform a single model as judge (Verga et al., 2024), which supports the use of a panel. Parshall and Lopez-Luzuriaga (2026) split exposure into cognitive, deployment and regulatory axes and classify the full O*NET task-activity universe with a panel of three models and a Delphi-style round. We adapt their cognitive-complexity ladder in our stage 1, defined by a manual test (C), and we turn their deployment ladder into a physical-requirement axis (D), while we drop their regulatory axis, which is built for the United States, and does not address AI potential. Yin et al. (2026) show that exposure scores from a single model are unstable across models, so we report agreement axis by axis and how much the scores depend on the rater.

Productivity evidence. Field and laboratory studies measure what AI assistance does to the time and quality of real tasks: professional writing (Noy and Zhang, 2023), customer support (Brynjolfsson et al., 2025), consulting (Dell'Acqua et al., 2023) and coding (Peng et al., 2023). Expert-graded comparisons of model deliverables with professional work (Patwardhan et al., 2025) and the length of tasks models complete autonomously (Kwa et al., 2025) are among the benchmarks in our bible. We use these results to set the scale of the augmentation thresholds (Section 4).

Complementarity and job boundaries. Pizzinelli et al. (2023) and Cazzaniga et al. (2024) cross exposure with a complementarity index built from O*NET work contexts and job zones. Garicano et al. (2026) model occupations as task bundles held together by coordination costs, and argue that AI dissolves weak bundles. Acemoglu (2025) stresses that exposure is not

automation, since only a fraction of exposed tasks is profitable to automate. We take the first idea for our shields, built from O*NET descriptors, and the second by keeping capability and protection apart, since there are frictions that hold back capability from turning into substitution.

Observed use and labour-market effects. Usage data such as the Anthropic Economic Index (Handa et al., 2025; Massenkoff et al., 2026) and Copilot conversations (Tomlinson et al., 2025) measure where AI is used, which is a lower bound on where it could be used. Linked survey and register data so far find no effect of chatbot adoption on earnings or recorded hours, ruling out effects larger than 2% (Humlum and Vestergaard, 2025). We use usage data only as a check on our labels (Section 8).

Time weighting. We weight tasks within occupations by time shares derived from O*NET task ratings, following Martin and Monahan (2022), instead of equal weights or the importance-and-relevance weights of Tomlinson et al. (2025).

3 Data

Task universe. The unit of analysis is the O*NET task-activity pair: a task statement read in the context of one of its Detailed Work Activities (DWA). O*NET 31.0 (August 2026; U.S. Department of Labor, Employment and Training Administration, 2026b) links 18,838 tasks to 2,085 DWAs in 923 occupations, giving 24,087 pairs, and we classify all of them.[3] A task has 1.28 pairs on average, and 23.8% of tasks have more than one.

The bible. The bible lists 33 benchmarks in 3 blocks (language, knowledge and reasoning; multimodal and perception; code, agentic work and computer use). For each one it records what the benchmark measures, the best score by a model released on or before 31 August 2026 (with the model and its date), the metric, the source and the human baseline when available. We built it in several passes and had it audited externally before use, and we release it with the data (Section 10).

Occupation descriptors. Work Context, Work Activities and Task Ratings come from the same O*NET release. Task Ratings (relevance and frequency) feed the time weights, and Work Context and Work Activities feed the shields. For 12 occupations both Work Context and Work Activities are missing in 31.0, so we estimate their descriptors (Section 7.2).

[3] We first ran the panel on the previous release (29.x task list), and the pairs that 31.0 left unchanged keep those labels, which come from the same bible, prompts and raters. Only the 999 new or reworded pairs were asked again. DWA identifiers are matched by title, and the one title reworded between releases is matched by hand.

Table 1: The three layers of exposure

Layer	Question put to the rater	Asked on
Re replacement	Could AI produce this task's output at the quality of a typical worker in the occupation?	all pairs
Au augmentation	*Time*: does access to AI cut the worker's time on the task *as a whole* by at least half, at reference quality? *Quality*: in the same time, would a peer grading blind rank the AI-assisted output above three of every four unassisted outputs? $Au = \text{time} \vee \text{quality}$.	$Re = 0$
Su support	Can AI perform a distinct, task-specific action normally required for the output or as a direct input to its decisions, incorporable without redoing it, with the outcome at typical quality? No threshold.	$Au = 0$

4 Measuring exposure

4.1 Stage 1: what the task requires

The first step of our procedure is to give each task-DWA pair two ordinal requirements, judged independently of AI and adapted from the ladders of Parshall and Lopez-Luzuriaga (2026). **C** (codifiability) asks whether a complete how-to document could let a trained worker perform the task correctly every time. It runs from C0 (self-evident) and C1 (fully specifiable) to C2 (judgement between defensible alternatives), C3 (tacit expert judgement, verifiable only by a peer) and C4 (produces transferable new knowledge). **D** (physical requirement) asks whether a disembodied executor could complete the task. D0 is purely informational, D1 needs presence to sense or move, and D2 to D4 need manipulation of objects in increasingly variable environments. Stage 1 also records the O*NET abilities and capability families the task draws on. These parameters are used as input for stage 2 and carry no presumption about exposure. They give the measure its granularity, and help improve the raters' performance in our experience. Of all pairs, 15.5% are C3 or above and 37.7% need manipulation (D2+).

4.2 Stage 2: three nested layers

Table 1 has the three questions, which address exposure. The rater pictures a typical worker in the occupation, with AI, the usual tools of the task and a laptop, and judges what frontier AI could do on 31 August 2026 (the models released by then, and the software and agents built on them). The bible is the first source of evidence, and background knowledge comes in only when the benchmarks leave the question open, in which case it has to be declared and cannot reach past the cut-off.

An example. The tasks of chefs and head cooks show how the layers split one occupation. "Record production or operational data on specified forms" is replaceable. "Arrange for equipment purchases or repairs" is not, but AI cuts the time it takes by at least half. "Determine how food should be presented and create decorative food displays" gains through the quality channel instead: in the same time, a blind peer would rank the assisted result above three of every four unassisted ones. "Check the quantity and quality of received products" gets support only, an input the worker can use, with no large gain in time or quality. "Prepare and cook

foods of all types, either on a regular basis or for special guests or functions" falls outside all three layers.

Why these thresholds. We set both thresholds high on purpose and based on existing literature. Halving the time is the threshold of the E_1 label of Eloundou et al. (2024), who call it relatively high, and it lies above most measured effects: about 15% more resolutions per hour in customer support (Brynjolfsson et al., 2025), a quarter less time in consulting tasks (Dell'Acqua et al., 2023) and 40% less in professional writing (Noy and Zhang, 2023), with more than a halving only in a well-specified coding task (Peng et al., 2023). On quality, blind grading by professionals from the same occupation found gains of about 0.4 standard deviations (Noy and Zhang, 2023). Under normality that places the median assisted output near the 66th percentile of unassisted ones, below the 75th that our quality channel requires. We want Au to catch large gains only, and modest ones do not count.

Nesting. If AI does the whole task there is a productivity gain, and if there is a substantial gain there is support. So $Re = 1 \Rightarrow Au = 1 \Rightarrow Su = 1$, and each layer is asked only where the previous one is zero:

$$Au = Re \vee Au^{\text{measured}}, \qquad Su = Au \vee Su^{\text{measured}}. \tag{1}$$

The step from Re to Au is imposed by design, and we know it is a generous reading, since once review and correction of model output are counted, measured speed-ups on professional deliverables can be far smaller (Patwardhan et al., 2025). The published value of each layer includes the previous ones. Au measured is positive for 18.0% of the 15,143 pairs with $Re = 0$ (time channel 7.6%, quality channel 15.5%). Su measured is positive for 60.5% of the 12,411 pairs with $Au = 0$.

How the bible plugs in. The rubric asks the rater to read each benchmark by how it is built and not only by its score, and a benchmark counts as positive evidence only if it is at least as demanding as the task. When no benchmark covers the task, or there is no human baseline, that counts as evidence neither way. Physical tasks have their own test: observation that commodity hardware can capture (cameras, microphones, drones) is judged as interpretation, while presence itself and manipulation need evidence of deployment in environments at least as unstructured as the task's. Scores obtained in sandboxes and multi-step tasks are checked against the documented autonomy horizon (Kwa et al., 2025).

What exposure excludes. Regulation, legal responsibility and a preference for a human provider do not make a task technically impossible, so the rubric leaves them out. They are

frictions, and we deal with them in other parts of our analysis: responsibility and the preference for a person enter in the shields layer (Section 7), and regulation is left to the applied work.

5 The rater panel

Raters. Three frontier models (Claude Opus 5, GPT-5.6 and Gemini 3.1 Pro) rate every pair blind and independently, in batches of about twenty pairs, returning a label, a short reasoning, the benchmarks used and a confidence level. We use three models and not only the most capable one, favouring redundancy over cost, since exposure scores from a single model are unstable across models (Yin et al., 2026), and Section 8 shows how much each of ours moves the scores.

Delphi reconciliation. Pairs that are not unanimous on an axis go to a single Delphi-style round (Dalkey and Helmer, 1963) in a fresh context, where each rater sees the reasoning of the other two with their labels hidden, so that no label can anchor the second answer, and decides again. The final label is the majority of that round. We take the round from Parshall and Lopez-Luzuriaga (2026) and run it on our own raters, prompts and questions. When there is no majority of at least two votes, a calibrated tie-break gives the label of the rater that agreed most with the majority in the two-to-one splits of round 1 of the same run, and with fewer than two valid votes the pair is flagged as unresolved.

Quality checks and the results bank. One known limitation of the technology at this scale is that models drop rows (they omit the label column or mistype an identifier), and the loss is invisible unless measured. We check every leg against its request file before anything is parsed, and a leg that comes back short is asked again for the missing rows only. Final labels are stored in a bank keyed by a hash of the rubric and the models, so that no pair is paid for twice and a label is only reused under the same instructions.

Agreement. Table 2 reports round-1 agreement on the universe run, as unanimity and as chance-corrected agreement (Cohen, 1960; Fleiss, 1971). In percentage terms the raters agree on most pairs, and on every axis about two thirds of the pairs or more are unanimous in round 1. Agreement is highest where the question is most objective: 85.3% of pairs are unanimous on *Re* (Fleiss' $\kappa = 0.79$; the raters' positive rates range from 37.1% to 42.4%) and 71.3% on D. Corrected for chance, it is lowest for the threshold channels of *Au* (Fleiss' κ of 0.34 for time and 0.43 for quality), where the raters also differ most in their base rates, and for *Su* (unanimity 64.6%; the three raters' positive rates range from 38.0% to 65.9%). This is the cost of asking a quantitative threshold question, and it is why we reconcile the two channels separately and report them both apart and combined.

What the Delphi round changes. Among the 3,404 *Re* pairs that reached the second round with a two-to-one split, the final label differs from the round-1 majority in 22.4% of cases. The share is similar for the *Au* channels (21.7% for time, 25.4% for quality), C (23.1%)

Table 2: Inter-rater agreement and Delphi reconciliation, universe run

			Round 1			Delphi round		
Axis	Pool	Pairs	Unanimous (%)	Fleiss κ	Cohen κ range	Sent (%)	Converged (%)	Tie-breaks
C codifiability	all pairs	21,011	66.1	0.64	0.62 to 0.65	33.9	53.5	53
D physical requirement	all pairs	21,011	71.3	0.71	0.70 to 0.73	28.7	54.0	182
Re replacement	all pairs	23,233	85.3	0.79	0.78 to 0.81	14.7	46.6	3
Au time channel	Re = 0	14,572	77.7	0.34	0.24 to 0.47	22.3	64.7	0
Au quality channel	Re = 0	14,572	78.0	0.43	0.39 to 0.47	22.0	46.0	0
Su support	Au = 0	11,956	64.6	0.53	0.45 to 0.64	35.4	42.4	0

Universe run on the 29.x task list, before the 31.0 update. Stage 1 excludes the pairs carried over from the 50-occupation calibration run. "Sent" = share of pairs not unanimous in round 1 and sent to the Delphi round. "Converged" = share of those on which all three raters agree in the second round. Cohen κ range over the three rater pairs.

and D (23.9%), and highest for *Su* (34.4%). So the minority carries the pair often enough to matter, as the design intends.

6 From pairs to occupations

Pair to task. Since for an occupation the same task may appear several times under different DWAs, and we base our analysis on task-DWA pairs, we then need a reconciliation. A task's score is the mean over its pairs, that is, the share of its activity readings that are positive. As in Eloundou et al. (2024), we aggregate activity labels to tasks before occupations, and a task's weight is split equally among its n pairs and never multiplied by them (Tomlinson et al. (2025), on the other hand, give a task's full weight to every activity it maps to).

Task to occupation: time weights. Following Martin and Monahan (2022), each task's relevance RT (share of incumbents who rate the task as relevant to their job) and frequency distribution FT_c over seven categories are turned into a share of working time. With the implicit "never" category at weight zero, the revised frequency share is $FT_c \cdot RT/100$. The composite time score is

$$s_t = \sum_{c=1}^{7} \frac{FT_c\, RT}{100}\, w_c, \qquad w = (0.02,\, 0.03,\, 0.06,\, 0.08,\, 0.10,\, 0.13,\, 0.16), \tag{2}$$

from "yearly or less" to "hourly or more", and the time share of task t in occupation o is $\pi_t = s_t / \sum_{t' \in o} s_{t'}$. The occupation's layer score is

$$L_o = \sum_{t \in o} \pi_t\, \bar{L}_t, \qquad L \in \{Re, Au, Su\}, \tag{3}$$

Table 3: Task weighting schemes, 923 occupations

Weighting	*Re* mean (%)	*Au* mean (%)	*Su* mean (%)	Rank corr. of *Re* with headline	Largest *Re* shift (points)
ONS time shares (headline)	37.8	48.4	79.2	1.000	0.0
Annual occurrence counts	37.5	47.8	79.4	0.974	27.4
Task importance	37.1	48.1	78.8	0.988	19.4
Equal weights	37.0	48.0	78.5	0.983	19.0

Means across occupations. All schemes give tasks without ratings an equal share within their occupation.

Table 4: Exposure in the O*NET 31.0 universe

Layer	Per pair (%)	Time-weighted, mean of SOC (%)	Median SOC (%)	SOC with more than half of time
Replacement (*Re*)	37.1	37.8	35.5	259 (28%)
Augmentation (*Au*)	48.5	48.4	48.8	447 (48%)
Support (*Su*)	79.7	79.2	91.5	763 (83%)

24,087 pairs, 923 occupations. Layers are cumulative ($Re \subseteq Au \subseteq Su$). *Au* read through one channel only: time 42.0%, quality 46.9% of working time.

the share of the occupation's working time that is replaceable, augmentable or supportable. Tasks without ratings get an equal share within their occupation, and occupations without any ratings get equal weights throughout. The ONS estimate the frequency weights by regression and call them fairly arbitrary (Martin and Monahan, 2022), so we set them against three alternatives in Table 3: annual occurrence counts (Tomlinson et al., 2025), task importance and equal weights. The choice barely moves the means (less than a point) or the ranking by *Re* (rank correlation with the headline of at least 0.974), but a single occupation can move by more than a quarter of its working time, so weights matter when we look at one occupation, and much less when we look at all of them. And we see value in the ONS time-share method.

Whatever the weights, an occupation's score depends on how O*NET writes its task list. For chefs and head cooks, cooking is one of 21 tasks and gets 4% of working time, while the other tasks are supervisory and administrative. This matters most in supervisory occupations, whose scores are better read together with the task list behind them, and the released pair file allows it.

Behind the means (Figure 2), replacement rises smoothly across occupations, and where it is low but the work is informational, augmentation and support cover most of the remaining time. The least replaceable quarter of occupations holds 63% of all working time outside the three layers, and 79% of their task-activity pairs require manipulation, compared to 8% in the most replaceable quarter.

7 The complementarity apparatus

At the current frontier almost every informational task comes out exposed: 99.9% of D0 pairs are at least supportable (Table 8). Exposure is technical capability, and there are frictions that hold back capability from turning into substitution, which exposure leaves out on purpose. We propose shields to measure some of them, built around what ties the work to a person, together

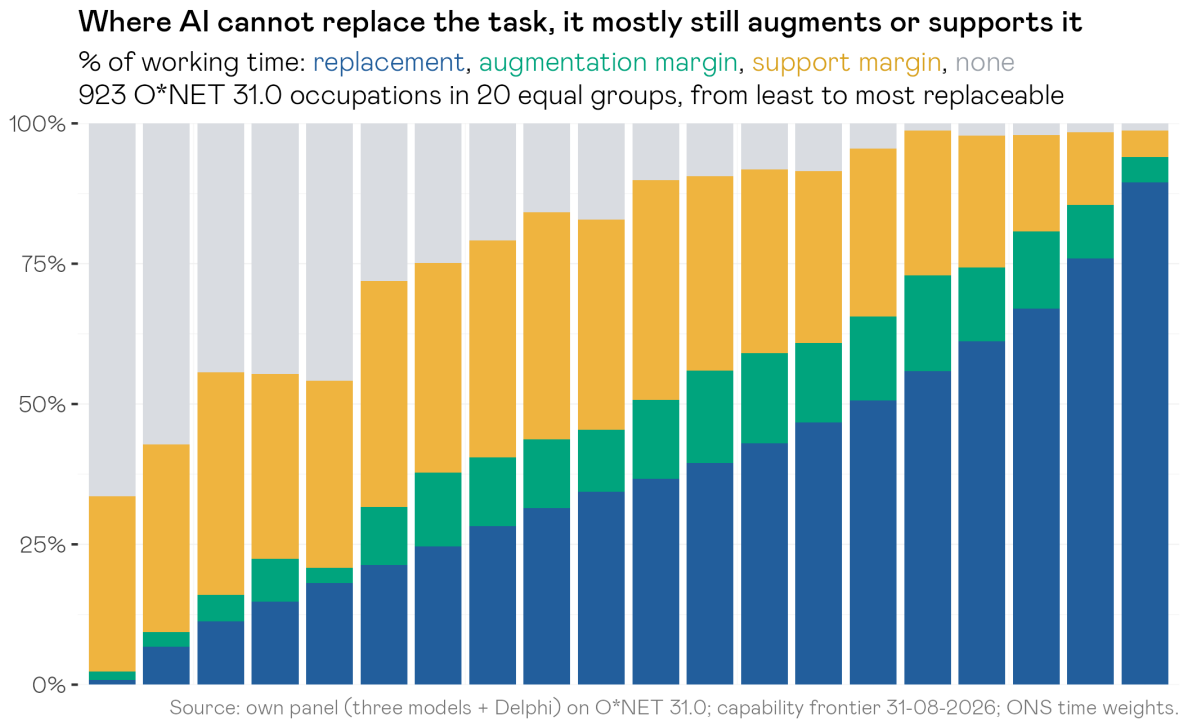


Figure 2: Time-weighted layers across occupations. Occupations sorted by their replaceable share and grouped into 20 bins of equal size; each bar is the bin mean.

with one more measure, expert complementarity, for the work where AI helps an expert without replacing them. We report them as a vector and never average them into exposure. The shields reach CNO-11 through the same bridge as the exposure layers (Appendix A).

7.1 Shields from O*NET descriptors

We compute three shields from eighteen O*NET descriptors, with no model annotation (Table 5). Each descriptor is rescaled to $[0, 1]$ with $(x - 1)/4$ from its scale of 1 to 5. Channels that are alternative routes to protection are combined with a power mean,

$$M_p(x_1, \ldots, x_k) = \Big(\frac{1}{k}\sum_{i=1}^{k} x_i^p\Big)^{1/p}, \qquad p = 4, \tag{4}$$

which sits close to the maximum, since one high channel is enough to tie the work to a person and a low channel should not average it away.

- *Relational nexus.* Friction in volume, FRIC $= F \times V$, with

  $$F = \tfrac{1}{2}\,\text{conflict} + \tfrac{1}{2}\max(\text{unpleasant},\ \text{violent})$$

  and V the mean of the three contact items. The product keeps friction without volume, or volume without friction, low. Presence, PRES, is the mean of the three items on what the client buys from the person. Nexus $= M_4(\text{FRIC}, \text{PRES})$.

- *Organisational responsibility.* The cost of nobody answering for the result: Resp $=$

Table 5: O*NET descriptors behind the shields

Shield	Channel	O*NET descriptor	Domain
Nexus	Friction F	Conflict Situations	WC
		Dealing With Unpleasant, Angry, or Discourteous People	WC
		Dealing with Violent or Physically Aggressive People	WC
	Volume V	Contact With Others	WC
		Deal With External Customers or the Public in General	WC
		Work With or Contribute to a Work Group or Team	WC
	Presence	Performing for or Working Directly with the Public	WA
		Assisting and Caring for Others	WA
		Establishing and Maintaining Interpersonal Relationships	WA
Responsibility	Criticality	Consequence of Error	WC
	Discretion	Freedom to Make Decisions	WC
		Impact of Decisions on Co-workers or Company Results	WC
	Authority	Coordinate or Lead Others in Accomplishing Work Activities	WC
		Work Outcomes and Results of Other Workers	WC
		Health and Safety of Other Workers	WC
Authorship		Determine Tasks, Priorities and Goals	WC
		Thinking Creatively	WA
		Level of Competition (only × Thinking Creatively)	WC

WC: Work Context (context scale). WA: Work Activities (importance scale). Both rescaled from a scale of 1 to 5 to $[0, 1]$.

$M_4(\mathrm{CRIT}, \mathrm{DISC}, \mathrm{AUT})$, with criticality the consequence of error, discretion the mean of the two decision items, and authority the mean of the three items on other workers.

- *Authorship.* $\mathrm{Auth} = \frac{1}{3}(\mathrm{determine} + \mathrm{creative} + \mathrm{competition} \times \mathrm{creative})$. The competition item enters only through creativity, so that commercial pressure alone does not load on the index.

We also express each shield as a percentile across the 923 occupations, since the aggregators have different ranges and percentiles can be compared across shields.

7.2 Occupations without descriptors

In O*NET 31.0, 12 occupations have none of the eighteen descriptors, and three model raters estimate them. Each rater receives the item definitions with their scale anchors, four reference occupations with their tasks and observed ratings, and the target occupation's title and tasks. Two references are chosen to cover the low, middle and high bands of the items, and two are the closest occupations by work content, and every estimate cites the reference it is placed against. The final value is the mean of the three estimates, and we flag an item for review when the estimates span two points or more or any rater reports low confidence (48 of the 216 adopted items).

To check the estimates we ran a pilot on three occupations held out with their 54 observed scores. The mean absolute error of the panel mean on the scale of 1 to 5 was 0.29 with this mixed reference set, 0.30 with the four closest occupations and 0.34 with four band-covering ones, compared to 0.31 for the plain mean of the closest references. Three occupations are too few to say that the panel beats the simpler rule (the gap between the two is small), but the pilot does show an error of about a third of a scale point. Rows built on estimates carry a

Table 6: Rank correlation between shields and exposure layers, 923 occupations

Shield	*Re*	*Au*	*Su*
Relational nexus	0.06	0.04	0.12
Organisational responsibility	-0.13	-0.12	-0.03
Authorship	0.28	0.40	0.50

source flag in the released file. The point is not so much for these 12 occupations, but rather for future ones in future O*NET releases, which may become an issue if the number gets larger.

7.3 Expert complementarity

We build a fourth measure from the panel itself, exploiting the gap that stage 1 shows along the codifiability ladder (Table 9): replacement peaks at C2 (50.5%) and falls to 13.7% at C3 and 0.0% at C4, while augmentation stays high (74.4% at C4, a band of only 219 pairs). Expert complementarity is the share of working time in pairs that require tacit expert judgement or produce new knowledge (C3 or C4), are not replaceable ($Re = 0$) and are still supportable ($Su = 1$), that is, the work where AI raises what an expert can do without substitution. It is very uneven across occupations: it is zero in 43% of them and has a median of 3.8% of working time, against a mean of 12.1% (90th percentile 36.9%) and a peak of 80.1% for Biochemists and Biophysicists.

7.4 How the shields relate to exposure

Across the 923 occupations, the nexus and responsibility are almost uncorrelated with exposure (Table 6). Authorship, on the other hand, correlates positively with every layer ($\rho = 0.40$ with Au), so we report the quadrant index with and without it.

7.5 A quadrant typology

In a similar way to Pizzinelli et al. (2023), the vector can be summarised in four quadrants, with replacement on one axis and a shield index on the other, each cut at its median across occupations, where the index is the mean of the shield percentiles. At O*NET level there are no employment weights, so Table 7 counts occupations. Median cuts put half of them on each side of each axis, and since the index is nearly uncorrelated with replacement, each quadrant ends up with close to a quarter. We use the table to show how much the classification depends on what enters the index.

Authorship correlates with exposure (Table 6) and is a marker of creative work. It has IP issues, and it is tricky, so we treat it with caution. Dropping it moves the rank correlation of the index with *Re* from 0.08 to -0.04, and 15% of occupations change quadrant, so whenever we use the quadrants we say which shields enter the index.

For policy, the quadrant that matters most is high exposure with a low shield. It holds 217 occupations with the full index, and 86% of them stay there when authorship is dropped. The ones with the most replaceable time are clerical and analytical: insurance claims and policy

Table 7: Quadrants of replacement and shields, share of 923 occupations (%)

Quadrant	Index: nexus, responsibility, authorship	Index: nexus, responsibility
High exposure, low shield	23.5	25.2
High exposure, high shield	26.4	24.7
Low exposure, high shield	23.5	25.2
Low exposure, low shield	26.5	24.8

Replacement is the time-weighted *Re*; the index is the mean of the shield percentiles. Both axes are cut at their median across occupations.

Table 8: Layers by physical band (%)

D	Pairs	Share	*Re*	*Au*	*Su*
D0	12,067	50.1	70.5	85.3	99.9
D1	2,935	12.2	14.1	32.9	90.9
D2	3,751	15.6	0.5	6.5	59.1
D3	4,673	19.4	0.0	3.6	44.4
D4	661	2.7	0.0	0.0	25.7

Table 9: Layers by codifiability band (%)

C	Pairs	Share	*Re*	*Au*	*Su*
C0	54	0.2	11.1	11.1	22.2
C1	10,700	44.4	33.7	35.6	62.6
C2	9,596	39.8	50.5	62.8	93.4
C3	3,518	14.6	13.7	47.7	93.7
C4	219	0.9	0.0	74.4	100.0

processing clerks (100%); data scientists (100%); insurance underwriters (100%); payroll and timekeeping clerks (99%); customer service representatives (96%). Among those that leave it, the three with the most replaceable time are office and booking jobs, whose index rises once their low authorship no longer counts: medical secretaries and administrative assistants (92%); reservation and transportation ticket agents and travel clerks (89%); human resources assistants, except payroll and timekeeping (88%).

8 Validation

Construct gradients. If the rubric works, replacement should fall where the task needs physical manipulation, and augmentation should outlast replacement where judgement is tacit, and both patterns appear (Tables 8 and 9). *Re* falls from 70.5% of D0 pairs to 14.1% at D1 and 0.5% at D2, while *Su* remains at 59.1% at D2.

Comparison with Eloundou et al. We correlate our unweighted occupation means with the replication package of Eloundou et al. (2024) on the 923 occupations common to both (Table 10). *Au* tracks their β closely ($r = 0.89$; $r = 0.87$ against their human annotators' β) and their γ ($r = 0.90$), *Re* tracks β at 0.83, and all three layers correlate weakly with α, the direct-exposure measure. Levels differ, as intended, since *Re* (37.0%) is stricter than any reading of exposure that counts partial time savings, and *Su* is laxer than γ.

Observed use. The June 2026 release of the Anthropic Economic Index (Massenkoff et al., 2026; Anthropic, 2026) is keyed on O*NET 30.2 task identifiers and can be joined to our tasks. Usage is a lower bound on capability, and a missing row only means that the cell was not published (usage there may still be positive), so the check only works in one direction: we expect usage to fall almost entirely on tasks with $Su = 1$, and a used task with $Su = 0$ is a

Table 10: Correlation with Eloundou et al. (2024), occupation level

	Mean (%)	α (GPT-4)	β (GPT-4)	γ (GPT-4)	β (human)
Re (unweighted)	37.0	0.56	0.83	0.79	0.78
Au (unweighted)	48.0	0.50	0.89	0.90	0.87
Su (unweighted)	78.5	0.35	0.78	0.84	0.78

Pearson correlations, unweighted task means per SOC. Eloundou et al. means over the same occupations: α 14.2%, β 34.5%, γ 54.8% (GPT-4); human β 30.3%. We follow the replication package in calling the broadest measure γ; the paper calls it ζ.

candidate labelling error. We join the 2,823 tasks with a published row in Claude.ai and the 2,427 in the first-party API (April and May 2026, global),[4] and a task carries a label when all its pairs do. Use falls where we expect it, 99.9% of it on tasks with $Su = 1$ (99.7% in the API) and 84.7% on tasks with $Re = 1$, and coverage follows the layers: 30.8% of the tasks we label replaceable have a row, 18.7% of those only augmentable, 5.5% of those only supportable and 0.5% of those outside all three layers. The API, mostly used in automated pipelines, puts more of its use on tasks with $Re = 1$ (88.3% against 84.7% in Claude.ai), as we would expect, but the gap is small, since at least 90% of its use is automation whatever the label of the task. The 3,665 tasks with $Su = 0$, 19% of all our tasks, get 0.1% of the use (0.3% in the API), and only 20 of them have a row. Of these, four are tasks of actors, where role-play makes our label arguable, and the rest are greetings, reassurance and physical routines such as weighing ingredients, which look like conversations about the task that the index files under it, more than AI doing the task. Since a missing row is not a zero, the check rules out false negatives at scale (tasks where AI is used and we say it cannot help), but tells us nothing about false positives.

Human validation. A blind human audit of 50 tasks on an earlier, Eloundou-style rubric agreed with the model labels on 94% of tasks (Cohen's $\kappa = 0.88$). We also validated the current rubrics with human review of samples of pairs at each stage and of the resulting occupation scores, and corrected them where the review found systematic errors.

Sensitivity to the rater. We never sampled pairs, since the universe is classified in full, so there is no sampling error at O*NET level. What remains is measurement error from the raters, and the question raised by Yin et al. (2026) of how much an occupation's score depends on the model that rates it, which is also the question behind using three models instead of the most capable one alone. We recompute the time-weighted Re of each of 923 occupations three times, each with one model's round-1 labels alone, and compare them with the reconciled score. The single-model means range from 37.5% to 43.2% against 37.8% for the consensus, and each single-model version correlates at least 0.93 with it. For an individual occupation, however, the three versions differ by a median of 9.4 points and by 25.1 or more for one occupation in ten. The consensus falls inside the single-model band for 92% of occupations, close to its lower

[4] 3 of the Claude.ai tasks and 2 of the API ones are not in O*NET 31.0 and drop out.

end. The ordering of occupations barely depends on the rater, but the levels do: the mean moves by several points from one rater to another and single occupations move by more, and we intend to release that band with the scores. We think the results are quite consistent.

9 Limitations and roadmap

- **US task content.** O*NET describes US jobs. For our field of interest, we need crosswalks to land it on the Spanish labour market.
- **Thresholds at occupation level.** Table 4 counts occupations above half of their time, and since the cut is a convention, Figure 2 shows the full distribution.
- **Unweighted bridge.** Collisions in the bridge (Appendix A) are averaged without employment weights. Weighting by US employment (OES) is a candidate refinement. Inside some cultural groups the crosswalk introduces visible noise (for example, collection managers mapped to debt collectors).
- **Frozen at one date.** The layers are fixed at the 31 August 2026 frontier, so any series built on them is compositional. We do not see that as a defect, since we expect AI to show up in survey data first as a change in the occupational mix within sectors, before it shows in sector headcounts. Re-running the panel with an updated bible and the same prompts and raters is how we would measure capability change, but changes in prompts, raters and coverage also move labels, and so does O*NET, which updates its task lists regularly. A comparison over time has to hold as much as possible fixed, and the raters can be held fixed only while their model versions remain available. Using open-weight models could solve that, but they are not state of the art, so for method validation, we opted for closed ones, where availability is an issue.
- **A ceiling on capability.** The layers measure technical feasibility at the cut-off, which is a ceiling, as in other task-based measures (Brynjolfsson et al., 2018). How much of it is put into practice depends on adoption, which varies across firms and sectors (a start-up and a bank do not move at the same speed) and is not measured here; early evidence finds no effect so far on earnings or recorded hours, although tasks are already being reorganised (Humlum and Vestergaard, 2025).
- **What the bible covers.** Benchmarks cover informational work much more densely than physical work, which the bible only reaches through simulation. Some best scores are reported by the developers themselves, and a score obtained in a sandbox says little about open interaction with people. An updated bible keeps these gaps, and a new benchmark can move labels by filling one, so part of the change between two waves will come from the inventory and not only from the frontier.
- **Rater lag.** Where benchmarks do not settle a question, raters fall back on background knowledge, which may lag the frontier, and if so labels are biased downward. The check against observed use (Section 8) is a one-sided test of this.

Table 11: Released datasets

Dataset	Unit	Rows	Content
Capability inventory	benchmark	33	what each benchmark measures, best score and model by 31 August 2026, metric, source, human baseline
Pair labels	task-DWA pair	24,087	C, D, *Re*, *Au* time and quality channels, *Au*, *Su*, with provenance
Occupation scores	O*NET-SOC	923	time-weighted *Re*, *Au*, *Su*, *Au* by channel, expert complementarity, alternative weightings
Shields	O*NET-SOC	923	nexus, responsibility, authorship, their channels and percentiles, observed or estimated
ESCO scores	ESCO occupation	3,220	layers, with the SOC behind each mean
CNO-11 scores	CNO-11, 3 digits	169	layers, channels and shields, with coverage diagnostics

- **Imputed shields.** 12 occupations have estimated shields, pending O*NET data collection. Tasks without ratings take equal time shares within their occupation.

10 Data availability

We release all results, and Table 11 lists the files. Each comes with a legend that defines every variable, its unit and its source. Labels keep their provenance (unanimous round 1, Delphi majority, tie-break, carried over from the earlier task list), and rows built on estimated descriptors are flagged.

11 Conclusion

We ask three nested questions of every task-activity pair in O*NET and answer them against capability evidence dated to a single cut-off, with a Delphi round where the three model raters disagree. Since the evidence is dated, the same apparatus can be run again with a later bible to give a series, within the limits of Section 9. Its first use is to separate what a single exposure score mixes: replacement is most frequent among tasks that call for judgement between defensible alternatives and rare where the judgement is tacit and expert, while augmentation stays high there (although the most knowledge-producing band rests on few pairs).

The measure is a ceiling on capability, and it does not forecast adoption or displacement, since there are frictions that hold back capability from turning into substitution, some of which we measure separately as shields (relational, organisational and authorial, reported as a vector). For *Re*, the ordering of occupations holds under other weighting schemes and other raters, while levels move more, above all for single occupations, and we intend the release to carry that band.

We release all labels and scores at pair, occupation and CNO-11 level, so that applied work on Spanish microdata can use the measure without rebuilding it.

References

Daron Acemoglu. The simple macroeconomics of AI. *Economic Policy*, 40(121):13–58, 2025. doi: 10.1093/epolic/eiae042. Working paper version: NBER Working Paper 32487, May 2024, doi:10.3386/w32487.

Daron Acemoglu and David Autor. Skills, tasks and technologies: Implications for employment and earnings. In Orley Ashenfelter and David Card, editors, *Handbook of Labor Economics*, volume 4B, chapter 12, pages 1043–1171. Elsevier, Amsterdam, 2011. doi: 10.1016/S0169-7218(11)02410-5. Working paper version: NBER Working Paper 16082 (2010).

Anthropic. Anthropic economic index, release 2026-06-26 (data and documentation). Hugging Face dataset `Anthropic/EconomicIndex`, 2026. URL `https://huggingface.co/datasets/Anthropic/EconomicIndex/tree/main/release_2026_06_26`. Data under CC-BY. Monthly aggregates for April and May 2026; task nodes keyed by the numeric O*NET Task ID (column `node_external_id`); O*NET version 30.2 as stated in the report appendix.

Melanie Arntz, Terry Gregory, and Ulrich Zierahn. The risk of automation for jobs in OECD countries: A comparative analysis. OECD Social, Employment and Migration Working Papers 189, OECD Publishing, Paris, 2016. URL `https://www.oecd.org/content/dam/oecd/en/publications/reports/2016/05/the-risk-of-automation-for-jobs-in-oecd-countries_g17a27d8/5jlz9h56dvq7-en.pdf`.

David H. Autor, Frank Levy, and Richard J. Murnane. The skill content of recent technological change: An empirical exploration. *Quarterly Journal of Economics*, 118(4):1279–1333, 2003.

doi: 10.1162/003355303322552801. URL https://academic.oup.com/qje/article/118/4/1279/1925105.

Banco de España. *Annual Report 2025*. Banco de España, Madrid, June 2026. doi: 10.53479/43567. URL https://www.bde.es/f/webbe/SES/Secciones/Publicaciones/PublicacionesAnuales/InformesAnuales/25/InfAnual_2025_En.pdf. Published 18 June 2026. Chapter 3, section 4.2 and Chart 3.18 (pp. 237–238); methodological annex, "Measurement of the degree of exposure of youth employment in Spain to GenAI" (pp. 247–248). Spanish edition: Informe Anual 2025.

Erik Brynjolfsson, Tom Mitchell, and Daniel Rock. What can machines learn, and what does it mean for occupations and the economy? *AEA Papers and Proceedings*, 108:43–47, 2018. doi: 10.1257/pandp.20181019. URL https://www.aeaweb.org/articles?id=10.1257/pandp.20181019.

Erik Brynjolfsson, Danielle Li, and Lindsey Raymond. Generative AI at work. *Quarterly Journal of Economics*, 140(2):889–942, 2025. doi: 10.1093/qje/qjae044. URL https://academic.oup.com/qje/article/140/2/889/7990658. Earlier version: NBER Working Paper 31161 (2023), doi:10.3386/w31161.

Pablo Casas, Enrique Fernández-Macías, Fernando Martínez-Plumed, Emilia Gómez, Ignacio González Vázquez, and Simone Salotti. Revisiting the occupational impact of AI in the generative AI era. JRC Working Papers Series on Labour, Education and Technology 2026/02, European Commission, Joint Research Centre, Seville, March 2026. URL https://publications.jrc.ec.europa.eu/repository/handle/JRC145832. JRC145832.

Mauro Cazzaniga, Florence Jaumotte, Longji Li, Giovanni Melina, Augustus J. Panton, Carlo Pizzinelli, Emma J. Rockall, and Marina Mendes Tavares. Gen-AI: Artificial intelligence and the future of work. IMF Staff Discussion Note SDN/2024/001, International Monetary Fund, January 2024.

Jacob Cohen. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46, 1960. doi: 10.1177/001316446002000104.

Norman Dalkey and Olaf Helmer. An experimental application of the Delphi method to the use of experts. *Management Science*, 9(3):458–467, 1963. doi: 10.1287/mnsc.9.3.458.

Fabrizio Dell'Acqua, Edward McFowland III, Ethan Mollick, Hila Lifshitz-Assaf, Katherine C. Kellogg, Saran Rajendran, Lisa Krayer, François Candelon, and Karim R. Lakhani. Navigating the jagged technological frontier: Field experimental evidence of the effects of AI on knowledge worker productivity and quality. Working Paper 24-013, Harvard Business School, September 2023. URL https://web.archive.org/web/2024id_/https://www.hbs.edu/ris/Publication%20Files/24-013_d9b45b68-9e74-42d6-a1c6-c72fb70c7282.pdf.

Tyna Eloundou, Sam Manning, Pamela Mishkin, and Daniel Rock. GPTs are GPTs: Labor market impact potential of LLMs. *Science*, 384(6702):1306–1308, 2024. doi: 10.1126/scienc

e.adj0998. Working paper version: "GPTs are GPTs: An Early Look at the Labor Market Impact Potential of Large Language Models", arXiv:2303.10130, v5, 21 August 2023.

European Commission, Directorate-General for Employment, Social Affairs and Inclusion and U.S. Department of Labor, Employment and Training Administration. The crosswalk between ESCO and O*NET (Technical Report). Technical report, European Commission, Brussels, 2022. URL https://esco.ec.europa.eu/en/about-esco/publications/publication/crosswalk-between-esco-and-onet-technical-report. Published 30 September 2022; one table updated December 2022. Mapping table: https://esco.ec.europa.eu/en/use-esco/other-crosswalks.

Florentino Felgueroso, Rafael Doménech, Juan Ramón García, Marcel Jansen, and Analía Viola. Observatorio trimestral del mercado de trabajo, 2T2026. Technical report, BBVA Research and Fedea, Madrid, September 2026. URL https://www.bbvaresearch.com/wp-content/uploads/2026/09/Nota_OTMT19.pdf. 17 September 2026. Section 2, "IA generativa, estructura ocupacional y entrada al mercado de trabajo", pp. 5–8.

Edward Felten, Manav Raj, and Robert Seamans. Occupational, industry, and geographic exposure to artificial intelligence: A novel dataset and its potential uses. *Strategic Management Journal*, 42(12):2195–2217, 2021. doi: 10.1002/smj.3286.

Edward Felten, Manav Raj, and Robert Seamans. How will language modelers like ChatGPT affect occupations and industries?, 2023. URL https://arxiv.org/abs/2303.01157. arXiv:2303.01157, v2, 18 March 2023 (the paper prints the first author as "Ed Felten").

Joseph L. Fleiss. Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76(5):378–382, 1971. doi: 10.1037/h0031619.

Carl Benedikt Frey and Michael A. Osborne. The future of employment: How susceptible are jobs to computerisation? *Technological Forecasting and Social Change*, 114:254–280, 2017. doi: 10.1016/j.techfore.2016.08.019.

Luis Garicano, Jin Li, and Yanhui Wu. Weak bundle, strong bundle: How AI redraws job boundaries. CEPR Discussion Paper 21453, Centre for Economic Policy Research, 2026. URL https://ideas.repec.org/p/cpr/ceprdp/21453.html.

Fabrizio Gilardi, Meysam Alizadeh, and Maël Kubli. ChatGPT outperforms crowd workers for text-annotation tasks. *Proceedings of the National Academy of Sciences*, 120(30): e2305016120, 2023. doi: 10.1073/pnas.2305016120. Preprint: arXiv:2303.15056.

Paweł Gmyrek, Janine Berg, and David Bescond. Generative AI and jobs: A global analysis of potential effects on job quantity and quality. ILO Working Paper 96, International Labour Organization, Geneva, August 2023. URL https://www.ilo.org/publications/generative-ai-and-jobs-global-analysis-potential-effects-job-quantity-and.

Paweł Gmyrek, Janine Berg, Karol Kamiński, Filip Konopczyński, Agnieszka Ładna, Balint Nafradi, Konrad Rosłaniec, and Marek Troszyński. Generative AI and jobs: A refined global

index of occupational exposure. ILO Working Paper 140, International Labour Organization, Geneva, May 2025. URL https://www.ilo.org/publications/generative-ai-and-jobs-refined-global-index-occupational-exposure.

Kunal Handa, Alex Tamkin, Miles McCain, Saffron Huang, Esin Durmus, Sarah Heck, Jared Mueller, Jerry Hong, Stuart Ritchie, Tim Belonax, Kevin K. Troy, Dario Amodei, Jared Kaplan, Jack Clark, and Deep Ganguli. Which economic tasks are performed with AI? Evidence from millions of Claude conversations, 2025. URL https://arxiv.org/abs/2503.04761. arXiv:2503.04761.

Anders Humlum and Emilie Vestergaard. Still waters, rapid currents: Early labor market transformation under generative AI. NBER Working Paper 33777, National Bureau of Economic Research, 2025. URL https://www.nber.org/papers/w33777. Issued May 2025, revised March 2026. Earlier versions circulated as "Large Language Models, Small Labor Market Effects" (also Becker Friedman Institute Working Paper 2025-56, https://bfi.uchicago.edu/working-paper/large-language-models-small-labor-market-effects/).

Institute for Structural Research (IBS). Occupation classifications crosswalks: O*NET-SOC to ISCO. Institute for Structural Research (IBS), Warsaw, April 2016. URL https://ibs.org.pl/app/uploads/2016/04/onetsoc_to_isco_cws_ibs_en1.pdf. Description of the crosswalk files and Stata do-file, posted 6 April 2016 (https://ibs.org.pl/?p=2207). Users are asked to cite Hardy, Keister and Lewandowski (2018).

Instituto Nacional de Estadística. Clasificación nacional de ocupaciones 2011 (CNO-11): correspondencias teóricas para fines estadísticos, CNO-11 con CIUO-08. INE, Madrid, 2011. URL https://www.ine.es/daco/daco42/clasificaciones/corr_cno11_ciuo08.xls. Listed on the INEbase CNO page, https://www.ine.es/dyngs/INEbase/es/operacion.htm?c=Estadistica_C&cid=1254736177033&menu=ultiDatos&idp=1254735976614. CNO-11 approved by Real Decreto 1591/2010, de 26 de noviembre, BOE núm. 306, 17 December 2010, BOE-A-2010-19389.

Thomas Kwa, Ben West, Joel Becker, Amy Deng, Katharyn Garcia, Max Hasin, Sami Jawhar, Megan Kinniment, Nate Rush, Sydney Von Arx, Ryan Bloom, Thomas Broadley, Haoxing Du, Brian Goodrich, Nikola Jurkovic, Luke Harold Miles, Seraphina Nix, Tao Lin, Chris Painter, Neev Parikh, David Rein, Lucas Jun Koba Sato, Hjalmar Wijk, Daniel M. Ziegler, Elizabeth Barnes, and Lawrence Chan. Measuring AI ability to complete long software tasks. In *Advances in Neural Information Processing Systems 38 (NeurIPS 2025)*, pages 102382–102435, 2025. doi: 10.52202/085713-3086. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/85069585133c4c168c865e65d72e9775-Abstract-Conference.html. Preprint: arXiv:2503.14499, "Measuring AI Ability to Complete Long Tasks".

Josh Martin and Ellys Monahan. Developing a method for measuring time spent on green tasks. Methodology article, Office for National Statistics, March 2022. URL https://www.ons.gov.uk/economy/environmentalaccounts/articles/developingamethodformeasuringtimespentongreentasks/march2022. Released 7 March 2022.

Maxim Massenkoff, Eva Lyubich, Szymon Sacher, Zoe Hitzig, Shaoyi Zhang, Ryan Heller, and Peter McCrory. Anthropic economic index report: Cadences. Anthropic, June 2026. URL https://www.anthropic.com/research/economic-index-june-2026-report. Published 26 June 2026. Appendix: https://cdn.sanity.io/files/4zrzovbb/website/03ed1410f74a65ae4cc2a27120d0875e1e569535.pdf.

Antoni Mestre, Xavier Naya, Manoli Albert, and Vicente Pelechano. Inteligencia artificial y empleo en España: una aproximación territorial y de género a la exposición laboral, 2025. URL https://arxiv.org/abs/2512.23059. arXiv:2512.23059, v3, 5 May 2026. In Spanish.

Shakked Noy and Whitney Zhang. Experimental evidence on the productivity effects of generative artificial intelligence. *Science*, 381(6654):187–192, 2023. doi: 10.1126/science.adh2586. Effect sizes read in the accepted manuscript, https://conference.nber.org/conf_papers/f184551.pdf.

Daniel Parshall and Andrea Lopez-Luzuriaga. Measuring AI's economic reach: A multidimensional task taxonomy. Working Paper 2026-005, The George Washington University, Center for Economic Research, March 2026. URL https://econpapers.repec.org/paper/gwcwpaper/2026-005.htm. Also distributed by the Canary Institute, https://canaryinstitute.ai/papers/Parshall_Lopez-Luzuriaga_TaskExposure_2026.pdf (same text, without the CER cover page); replication package https://github.com/danparshall/reproduction-task-exposure.

Tejal Patwardhan, Rachel Dias, Elizabeth Proehl, Grace Kim, Michele Wang, Olivia Watkins, Simón Posada Fishman, Marwan Aljubeh, Phoebe Thacker, Laurance Fauconnet, Natalie S. Kim, Patrick Chao, Samuel Miserendino, Gildas Chabot, David Li, Michael Sharman, Alexandra Barr, Amelia Glaese, and Jerry Tworek. GDPval: Evaluating AI model performance on real-world economically valuable tasks, 2025. URL https://arxiv.org/abs/2510.04374. arXiv:2510.04374.

Sida Peng, Eirini Kalliamvakou, Peter Cihon, and Mert Demirer. The impact of AI on developer productivity: Evidence from GitHub Copilot, 2023. URL https://arxiv.org/abs/2302.06590. arXiv:2302.06590.

Carlo Pizzinelli, Augustus J. Panton, Marina Mendes Tavares, Mauro Cazzaniga, and Longji Li. Labor market exposure to AI: Cross-country differences and distributional implications. IMF Working Paper 2023/216, International Monetary Fund, October 2023.

Francisco Rodríguez-Fernández. Inteligencia artificial y mercado de trabajo en españa. exposición ocupacional, efectos sobre el empleo y adopción empresarial. Investigaciones de Funcas 33/2026, Funcas, Madrid, April 2026. URL https://www.funcas.es/documentos_trabajo/inteligencia-artificial-y-mercado-de-trabajo-en-espana-exposicion-ocupacional-efectos-sobre-el-empleo-y-adopcion-empresarial/. ISSN 3045-8994. English version: "Artificial intelligence and the labor market in Spain: Occupational exposure and estimated effects on employment", Funcas SEFO 15(3), May 2026, pp. 77–83.

Kiran Tomlinson, Sonia Jaffe, Will Wang, Scott Counts, and Siddharth Suri. Working with AI: Measuring the applicability of generative AI to occupations, 2025. URL https://arxiv.org/abs/2507.07935. arXiv:2507.07935, v6, 22 December 2025.

U.S. Department of Labor, Employment and Training Administration. Crosswalk files: European skills, competences, qualifications and occupations (ESCO) to O*NET-SOC 2019. O*NET Resource Center, 2026a. URL https://www.onetcenter.org/crosswalks.html#esco. File https://www.onetcenter.org/crosswalks/esco/ESCO_to_ONET-SOC.xlsx. Accessed 15 October 2026. Neither the page nor the file states the ESCO version.

U.S. Department of Labor, Employment and Training Administration. O*NET 31.0 database. O*NET Resource Center, August 2026b. URL https://www.onetcenter.org/database.html. Release date from https://www.onetcenter.org/db_releases.html. Used under the CC BY 4.0 license.

Pat Verga, Sebastian Hofstätter, Sophia Althammer, Yixuan Su, Aleksandra Piktus, Arkady Arkhangorodsky, Minjie Xu, Naomi White, and Patrick Lewis. Replacing judges with juries: Evaluating LLM generations with a panel of diverse models, 2024. URL https://arxiv.org/abs/2404.18796. arXiv:2404.18796.

Michael Webb. The impact of artificial intelligence on the labor market. Working paper, Stanford University, January 2020. URL https://www.michaelwebb.co/webb_ai.pdf. Also SSRN 3482150, doi:10.2139/ssrn.3482150.

Michelle Yin, Hoa Vu, and Claudia Persico. How (un)stable are LLM occupational exposure scores? Evidence from multi-model replication. NBER Working Paper 35110, National Bureau of Economic Research, April 2026. URL https://www.nber.org/papers/w35110.

A From O*NET to the Spanish classification

In this appendix we carry the occupation scores (Section 6) and the shields (Section 7) to three-digit CNO-11, the classification used in Spanish labour-force microdata. Other estimates for Spain import US or global indices through classification mappings (Banco de España, 2026; Felgueroso et al., 2026; Rodríguez-Fernández, 2026; Mestre et al., 2025), while we carry our own scores, and only through official crosswalks.

Crosswalks. We use three official files for the bridge: the ESCO to O*NET-SOC 2019 crosswalk published by the O*NET Resource Center (U.S. Department of Labor, Employment and Training Administration, 2026a), the ESCO and O*NET crosswalk of the European Commission and the US Department of Labor with its match types (exact, broad, narrow, close) (European Commission, Directorate-General for Employment, Social Affairs and Inclusion and U.S. Department of Labor, Employment and Training Administration, 2022), and INE's official correspondence between CNO-11 and ISCO-08 (Instituto Nacional de Estadística, 2011), and we build no mapping by hand.

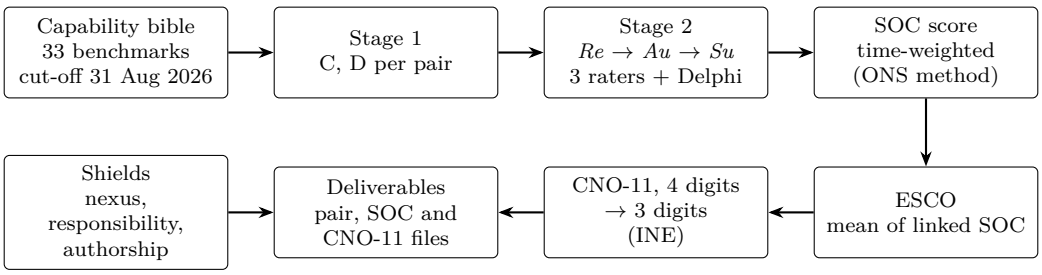


Figure 3: The measurement pipeline, from the capability inventory to the Spanish classification.

Occupation to CNO-11. We read the bridge from the destination (Figure 3). Each ESCO occupation takes the mean score of all O*NET-SOC occupations the official crosswalk links to it, used as published and without selecting by the match types (exact, broad, narrow, close) of European Commission, Directorate-General for Employment, Social Affairs and Inclusion and U.S. Department of Labor, Employment and Training Administration (2022). The ISCO-08 group is read from the ESCO code. INE's correspondence maps ISCO-08 to CNO-11 at four digits, and the only truncation in the chain is CNO-11 from four to three digits, on the Spanish side. Every collision is resolved by an unweighted mean, as in the IBS crosswalk files (Institute for Structural Research (IBS), 2016), and we keep the detail behind each mean. The Banco de España uses weighted averages for the same step from ISCO-08 to CNO-11 (Banco de España, 2026), and weighting is a candidate refinement (Section 9).

Two lessons shaped this design: truncating ISCO to three digits before the jump merges siblings (police and firefighters share ISCO 541), and the hand-built mappings we tried introduced errors of a kind the official files avoid, although the official files carry noise of their own (Section 9).

Coverage is close to complete. The bridge reaches 3,220 ESCO occupations, 39% of them from a single SOC, and all 169 CNO-11 groups get a score, resting on a median of 15 SOC and 298 O*NET tasks. In 167 every four-digit unit is covered. In the other 2, 589 (5 of 6 four-digit units), 770 (8 of 9 four-digit units), one unit has no linked SOC and the group score rests on the rest. The only SOC the crosswalk does not reach is 35-9011 (dining room attendants), which we report and leave without a score.

B Rubric excerpts

The replacement judgement, verbatim from the rater prompt:

> E = 1 (replaceable): AI as of {CAPABILITY_CUTOFF} could produce this task's output at the quality of a typical worker in the occupation. E = 0 (not replaceable): it could not.

The two augmentation channels, verbatim:

> time = 1 if access to AI cuts the worker's time on this task as a whole by at least half, at the quality of the reference. A saving confined to one part of the task counts only if it halves the time of the whole task.
>
> quality = 1 if, in the same time, a peer from the occupation grading blind would rank the

output the worker produces with AI above at least three of every four outputs that typical workers produce without it.

The full rater prompts are not part of the release.