netwell

Measuring AI Exposure at the Task Level

Headlines assign risk to job titles. Work is not done in job titles — it is done in tasks, and AI meets them one at a time. This paper documents how Netwell scores that exposure across 923 occupations and 18,797 task statements, what the resulting number means, and — at equal length — what it does not.

PUBLISHED
August 2026

DATASET
onet-30.0/aei-2026-05/oasis-2025-v1.1/oai-tjf-2026-04/eloundou-2023

METHOD
heuristic-v0

1 Summary

Netwell publishes an **AI Exposure Score** — an *exposure index*: an estimate of how much of a person's work is the kind of work current AI systems can do. It is not a prediction that anyone will lose a job, and this paper is written to make that distinction hard to lose.

The index is built in two layers. The first is occupational and evidence-led: every task statement O*NET records for an occupation is scored for exposure, anchored where possible to observed AI usage, then combined into a single baseline weighted by how important each task is to the job. The second layer is personal — seniority, task mix, employment type, runway and a handful of other answers — and it is deliberately capped so it can refine the occupational estimate but never overwhelm it.

Every score derived from an O*NET occupation is published with a confidence band, and the band is wide: its median width across the dataset is 30 points on a 0–100 scale. (Where no occupation could be matched, there is no band to publish, and the score is labelled accordingly rather than given a false one.) That is not a defect of presentation. It is the honest width of what task-level data can currently support, and showing it is the difference between an index and a horoscope.

THE DATASET IN FIVE NUMBERS

Occupations scored	923
O*NET task statements scored	18,797
Occupations anchored to observed AI usage	718
Baseline exposure — lowest / median / highest	29 / 43 / 84
Confidence band width — narrowest / median / widest	19 / 31 / 44

2 Why the job title is the wrong unit

Public discussion of AI and work is conducted almost entirely in occupations. A study reports that some share of jobs is exposed; a headline converts the share into a list of professions; a reader finds their own on it, or does not, and takes that as the answer.

The unit is wrong in both directions. Two people with the same title can hold jobs with almost nothing in common — one paralegal spends the week drafting standard documents, another spends it interviewing witnesses and managing client relationships. The title they share predicts far less about their exposure than the tasks they do not.

It is also wrong about what AI actually does. Systems today are not replacing occupations; they are absorbing *tasks*, unevenly, and the tasks they absorb first are the ones that are linguistic, document-shaped

and reviewable. A job made mostly of those tasks changes quickly. A job with three such tasks out of twenty changes at the edges. Both outcomes are invisible to an occupation-level score.

So the unit of measurement here is the task. O*NET — the U.S. Department of Labor's occupational database — publishes task statements for each occupation together with a rating of how important each task is to the job. That gives a defensible inventory of what people in a role actually do, maintained by someone other than us, which is the right foundation for a claim about someone's livelihood.

---

3 What the number is, and is not

The AI Exposure Score estimates task overlap with current AI capability. A high score means a large share of the work — weighted by importance — is of a kind today's systems can perform or substantially assist.

> **STATED PLAINLY**
>
> A score is **not** a probability of job loss. It carries no timeline. It does not know your employer's plans, your budget, your union, your regulator, your reputation, or how good you are. Exposure is one input into a decision about your career, and the least personal one.

The gap between "this work is technically automatable" and "this person loses this job" is filled by economics, regulation, organisational inertia, liability, and the simple fact that firms adopt slowly. None of that is in the index, so the index cannot be read as though it were. Where the product says so, it says so in the same words: an exposure index, not a prediction.

---

4 Method

4.1 The task inventory

The base layer is the O*NET 30.0 database. For each occupation we take every task statement and its *importance* rating (O*NET scale `IM`, 1–5). That rating, normalised and then discounted for tasks O*NET marks Supplemental rather than Core, becomes the task's weight, and a task with no rating takes a neutral default — 997 of the 18,797 (5.3%) are weighted by that default or that discount rather than by a rating of their own. Nothing is sampled and nothing is dropped: the present dataset covers 923 occupations and 18,797 task statements.

4.2 The observed-usage anchor

A lexicon alone would only ever encode our own assumptions about which work AI can do. So each occupation is anchored, where data exists, to what AI is *observed* doing — the Anthropic Economic Index,

which reports for each occupation how much of real AI usage maps to it and how much of that usage is automating rather than assisting.

```
anchor  =  0.60 · percentile-rank(usage share)  +  0.40 · automation share
```

Both inputs are noisy exactly where they matter least — an occupation with 0.2% of observed usage can rank high in a distribution full of zeroes, and its automation share may rest on a handful of conversations. Treating that as evidence would manufacture precision. Both terms are therefore shrunk toward a prior in proportion to how much usage was actually observed:

```
λ     =  u / (u + 0.6)                    u = usage share, %
rank  =  λ · percentile-rank  +  (1 − λ) · 40   prior 40
auto  =  λ · automation share  +  (1 − λ) · 50  prior 50

anchor  =  0.60 · rank  +  0.40 · auto          shrink first, then blend
```

The rank prior is 40 rather than 50 on purpose. Near-zero observed usage is itself weak evidence of low exposure, so an occupation nobody is using AI for should not be pulled all the way back to the population median. 718 of the 923 occupations carry an anchor of their own; the rest fall back to the median of their SOC major group, and pay for that in a wider band (§4.5). One asymmetry in that arrangement is worth stating plainly, because it flatters us: 209 of those 718 record zero usage in the latest month, so their shrinkage weight is zero and their anchor is entirely the prior — carrying no more occupation-specific evidence than an occupation with no anchor entry at all. They nonetheless escape the unanchored widening in §4.5, and the gap is visible in the output: those 209 average a band 29.6 points wide against 29.2 for the 509 with real observations, while the 205 unanchored occupations average 39.8. A reader should treat a narrow band on a zero-usage occupation as a statement about our formula rather than about the evidence behind that occupation.

4.3 The task-text modifier

The anchor describes an occupation. Tasks within one occupation differ, so each task statement is then adjusted by a lexicon of matched patterns, accumulating to at most ±28 points. There are 47 rules; the families below are the bulk of them, and their direction is capability:

THE MAIN PATTERNS, AND WHICH WAY THEY MOVE A TASK

RAISES EXPOSURE	LOWERS EXPOSURE
Drafting, editing, summarising and translating documents; data entry and reconciliation; classification and coding; scheduling; text- and phone-mediated customer contact; analysis of data and trends; writing and testing software.	Physical and manual work; operating vehicles and equipment; hands-on inspection; direct patient and personal care; supervising and training people; negotiation; enforcement and custody; live performance; in-person hospitality, retail-floor and facility work.

The lexicon is a first version and is labelled as one: the dataset's method field reads `heuristic-v0`. A one-time language-model refinement pass is built and has **not been run** — **0 of the 18,797** tasks carry a refined score today. When it runs, its contribution is recorded per task, so the confidence band can narrow only where refinement actually happened rather than everywhere at once (§4.5). Until then every task is scored by the lexicon alone, and every band pays the full unrefined penalty. The runtime performs no model calls in either case — a score is a lookup and a weighted average, and the same inputs against the same dataset always give the same number. The dataset is one of those inputs, and it is versioned precisely because it will change (§8): what is fixed is the computation, not the world it measures.

4.4 From tasks to an occupation baseline

The occupation's baseline is the importance-weighted mean of its task exposures. A task that is central to the job moves the number more than one performed occasionally, which is the entire reason for using O*NET's importance ratings rather than counting tasks.

```
baseline  =  Σ ( wᵢ · eᵢ )  /  Σ wᵢ        w = importance × task-type factor
```

That weighted mean is one estimate, and since August 2026 it is no longer the only one. Two further bodies of evidence are read for every occupation: OpenAI's *AI Jobs Transition Framework*, whose task-exposure ratings come from Eloundou et al.'s published rubric and whose classification records whether a person is still needed for what remains; and *OaSIS*, the Canadian government's occupational descriptor system, from which we derive an automatability composite over 102 skill, ability, work-activity and work-context ratings. Each is placed on this scale by matching its mean and spread against the occupations where our own anchor is strongest — two parameters, no more, because a richer fit would make an outside signal a prediction of ours and its agreement would then mean nothing.

```
consensus  =  Σ ( uⱼ · xⱼ )  /  Σ uⱼ        u from precision, discounted where two
sources share evidence
```

The weights are not equal and are not chosen. Each estimate is weighted by its own precision, and the weighting is then discounted for correlation between estimates, because two sources that read the same underlying task descriptions are not two independent opinions. Measured on the shipped data, our baseline and the OpenAI framework agree at $\rho = 0.73$, and OaSIS — which ESDC itself built partly from O*NET — at 0.67. Counting those as three independent votes would let one body of evidence outvote everything else, so it is measured and discounted instead.

The consensus is not allowed to move a score far. The revision is capped at ± 2 points, and that cap is derived rather than chosen: it is half the tightest slack in the ordering relationships our own calibration tests pin, so no new dataset can quietly reorder the occupations this paper claims to rank. Across all 923 occupations the measured effect is that 490 fell 2 points, 300 rose 2, 97 moved 1, and 36 did not move at all. Where the sources disagree the disagreement is not averaged away; it widens the band in §4.5 instead. That is the honest direction: more evidence buys more confidence only when the evidence agrees.

4.5 **The confidence band**

Every baseline ships with a band, and the band is computed from the specific reasons this particular estimate might be wrong rather than applied as a uniform decoration:

```
half-width  =  clamp( 5
                    +  0.4 · sd(task exposures)   weighted sd — the job varies
internally
                    +  5   if unanchored          no observed-usage data
                    +  3 · (1 − λ)                anchor rests on thin usage
                    +  4 · (1 − refined share),   tasks not individually refined
                    6, 22 )
```

An occupation whose tasks disagree with each other, or whose anchor is thin, is reported less confidently — as it should be. One caveat about reading that formula: because no task has been refined yet (§4.3), the last term currently sits at its maximum for every occupation, so today it is a flat +4 rather than something that varies. Across the dataset the resulting bands span 19 to 44 points wide, with a median of 31. A product that wanted to look authoritative would hide that. Publishing it is the point: the width is the finding, and part of that width is ours to earn back.

4.6 **The personal layer**

Two people in one occupation are not equally exposed, so the assessment adjusts the baseline using the person's own answers. Each factor contributes a fixed signed number of points.

PERSONAL MODIFIERS, IN POINTS

RAISES EXPOSURE	+	LOWERS EXPOSURE	−
Mostly routine, rule-based work	20	Leadership responsibility	10
Freelance, no institutional protection	20	Strategic, relationship-led work	10
Early career	15	Actively building AI skills	10
Recent headcount cuts on the team	15		
Part-time	15		
Team already uses AI daily	10		
Under three months of runway	10		
On contract	10		
Not yet building AI skills	10		
Large enterprise (5,000+)	8		
Partly routine workload	8		
Mid-career	5		
Sector investing heavily in AI	5		
Highest salary band	5		
Very worried about AI	5		

Not every modifier below can fire from today's questionnaire, and the shortfall is larger than a first reading suggests. Company size and salary band are collected by nothing at all. Three more — recent headcount cuts (+15), under three months of runway (+10) and being very worried about AI (+5) — are collected only by the retired questionnaire, not by the one now live, so 30 further points in the table below cannot presently fire either. Employment type is the reverse case: no screen asks it, but the founder pathway sets it, so its +20 row does fire for those members. In total six of the rows below describe the engine rather than a live assessment. They are listed because the engine applies them the moment those questions exist, and because a table that quietly omitted them would be a worse description of how a score is made.

The net adjustment is then clamped to ±20 points. This is the most consequential design decision in the personal layer, and it is a decision about whose evidence counts: the occupational baseline rests on 18,797 task statements and observed usage data, while the personal layer rests on a short questionnaire. Uncapped, a determined set of answers could swamp the measured part of the estimate and push any job to either extreme. Capped, the answers can move a score meaningfully and can never overwrite what the task data says.

When the cap binds, the per-factor breakdown shown to the member is scaled proportionally so the parts still sum exactly to the whole, by largest-remainder rounding. Signs and reasons stay true; individual

magnitudes become approximate.

One channel is deliberately not capped, and the argument above should not hide it. The task-mix step lets a member re-weight their own tasks, and because that re-weights the *measured* task list rather than asserting an opinion about it, it moves the baseline itself with no ceiling. Measured across all 913 occupations carrying at least eight tasks, choosing the three most- versus least-exposed tasks moves the baseline by a median of 13 points (mean 12.8), reaching 19 at the ninetieth percentile and 28 at the most task-sensitive occupation in the set. So for a typical occupation a member's own answers carry rather more weight than the ±20 cap alone implies, and in the most task-sensitive ones they carry considerably more. We give the distribution rather than a single figure because the two numbers answer different questions, and because the cap is two-sided while this swing is measured end to end — adding them into one headline number would combine unlike quantities. That is defensible, because task mix is the most informative thing a person can tell us about their own job. It is stated here because a cap presented as the whole answer would be the more flattering half of one.

4.7 **Bands**

0–30	31–50	51–70	71–100
LOW	MEDIUM	HIGH	CRITICAL

Band labels are a reading aid, not a finding. Because the median confidence band is 30 points wide, a score near a boundary should be read as sitting between two labels rather than firmly inside one — which is precisely why the number and its band are always shown together, and why the band moves with the score rather than being dropped once the personal layer has been applied.

---

5 Exposure is only half the picture

Exposure describes the work. It says nothing about the person’s capacity to absorb a change in it — and two people with identical scores can be in entirely different trouble. One has nine months of savings, transferable skills and a permanent contract; the other has three weeks and a freelance invoice. Reporting only exposure would treat them as the same case.

So a second, independent score is computed from four pillars, each worth up to 25 points. Two of them are ahead of the assessment: **financial buffer and employment stability are not asked by the current questionnaire**, so today they carry a fixed value rather than a measurement, and the resilience total is not yet comparable to its own 100-point scale. They are described here as designed, and flagged here as pending, because the alternative is a page that reads like four measurements when it is two.

RESILIENCE PILLARS

PILLAR	WHAT IT ASKS	MAX
Financial buffer	How long could you cover your costs without income?	25
Skill readiness	Are you building capability that survives the change?	25
Role durability	How much of your value is hard to substitute?	25
Employment stability	What protection does your arrangement give you?	25

The two numbers are kept apart on purpose and never averaged into a single “score”. They answer different questions — *how much of my work is exposed* and *how well could I absorb it* — and a single figure would hide the one that is actionable. Of the two, resilience is the one a person can move within a quarter.

6 A worked example

A mid-career paralegal at a large firm, on a permanent contract, whose work is a mix of routine and non-routine, whose team has begun using AI daily, and who is not yet building AI skills.

HOW THE SCORE IS BUILT

STEP		POINTS
Occupation baseline	importance-weighted mean of the role’s task exposures (SOC 23-2011.00, band 36–72)	56
Partly routine workload	raises exposure	+8
Team already uses AI daily	raises exposure	+10
Not yet building AI skills	raises exposure	+10
Large enterprise	raises exposure	+8
Mid-career	raises exposure	+5
Sector investing heavily in AI	raises exposure	+5
Raw personal adjustment		+46
Clamped to ±20	the task data keeps its weight	**+20**
Score	Critical band; reported with its confidence band	**76**

Note what the cap did. Six true statements about this person summed to +46 — more than twice what the cap allows — and uncapped they would have pushed a mid-range occupation to the ceiling on the strength of a questionnaire. The reported 76 is a serious score, and it still rests on the measured baseline rather than replacing it. Note also what changes it: two of the six factors — the routine share of the work, and whether

they are building AI skills — are things this person can act on within a quarter, which is the difference between a diagnosis and a verdict.

7 What we test, and what we refuse to

There is no ground truth to validate against. Nobody knows which jobs AI will displace, so any claim to have measured accuracy against that outcome would be false. What can be tested is whether the model orders occupations sensibly, and that is tested continuously: a calibration gate holds twenty roles and fails the build if the ordering breaks.

The gate is deliberately built of two different kinds of assertion. Absolute expectations are *wide* — a band roughly 30 points across for each role. Relative ones are *sharp*: analysts must score at least 32 points above nurses, translators 24 above truck drivers, technical writers 20 above janitorial roles — and, because pairs can all hold while a population drifts together, the median of the five most text-mediated roles must stay at least 20 points above the median of the five most embodied. A separate assertion requires the whole dataset to span at least 40 points, so the model cannot collapse toward a single number and still pass.

WHY THE BANDS STAY WIDE

Narrowing the absolute bands would make the gate look more rigorous while manufacturing a precision the data does not support. The ordering is what the product actually claims; the absolute number is an index. A model that ranked nurses above analysts fails today — that is the property worth protecting, and tightening the wrong half would quietly trade it away.

8 Limits

These are the ones we would want to read first if this were somebody else's paper.

- **Nothing has been refined yet.** The method field says `heuristic-v0` and means it: every one of the 18,797 task scores is the lexicon's, unreviewed. The refinement pass exists and has not been run, so the bands here are the widest this design produces.
- **The lexicon encodes a view.** The task-text modifier is a set of hand-written rules calibrated against automatability research. Reasonable people would write different rules and get different scores. It is versioned (`heuristic-v0`) so that disagreement has something to attach to.
- **Observed usage is a sample of one ecosystem.** The anchor reflects AI usage as it appears in one provider's data. It is the best observational evidence available to us and it is not the whole market.
- **Capability moves.** A task judged out of reach today may not be next year. The dataset is versioned and dated for exactly this reason; a score is a reading taken at a moment, not a constant.

- **O*NET describes the typical job, not yours.** The task-mix step lets a member re-weight their own tasks, which moves the baseline — but the inventory itself is still a national average of a role.
- **205 occupations have no usage anchor, and they are not evenly spread.** They fall back to their major group's median and carry a wider band — but sorting the dataset by baseline shows where they sit: **124 of the 230 least-exposed occupations have no anchor of their own, against 0 of the 233 most exposed.** Read those groups with their boundaries in view: the lower quartile tops out at 39 and the upper begins at 48, inside a median band of 31, so "least" and "most" here are ends of a fairly tight distribution rather than two different worlds. The low end of this index is therefore where the evidence is thinnest, which is the honest way to say that "we observe little AI usage for this work" and "this work has low exposure" are harder to separate than the number suggests. The 40 prior in §4.2 exists to stop that becoming a free pass, and the wider bands say the rest. Read a low score with its band, and an unanchored low score as coarser still.
- **The personal layer is self-reported** and unverifiable, which is the other half of the argument for capping it at ±20.

9 Data, licences and provenance

Every dataset underlying this work is openly licensed, and every one of them requires credit. That credit is carried wherever the derived numbers are displayed, not only here.

SOURCES

SOURCE	USED FOR	LICENCE
O*NET 30.0 Database `onetcenter.org`	Task statements and importance ratings for 923 occupations	CC BY 4.0
Anthropic Economic Index `anthropic.com/research`	Observed AI usage and automation share, used as occupation anchors	CC BY
OpenAI AI Jobs Transition Framework `openai.com`	Task-exposure ratings (Eloundou et al. 2023) and the human-necessity classification, as one consensus estimate	CC BY 4.0 · MIT
OaSIS 2025 `Employment and Social Development Canada`	Occupational descriptor ratings, from which an automatability composite is derived as a second consensus estimate	Open Government Licence – Canada
NOC ↔ SOC concordances `Statistics Canada`	Mapping Canadian NOC 2021 occupations onto the O*NET-SOC codes the rest of the model uses	Statistics Canada Open Licence

O*NET is a trademark of the U.S. Department of Labor, Employment and Training Administration. Netwell is not affiliated with or endorsed by the Department of Labor. OaSIS material contains information licensed under the Open Government Licence – Canada; concordance tables are adapted from Statistics Canada, NOC 2016 V1.3 ↔ SOC 2018 (US) and NOC 2016 V1.3 ↔ NOC 2021 V1.0, which does not constitute an endorsement by Statistics Canada of this product. Neither Employment and Social Development Canada nor OpenAI has reviewed, approved or endorsed this work or the figures derived from their data. Derived scores, the task-text lexicon, the descriptor weights, the personal modifier weights, the consensus method and the confidence model are Netwell's own work and are described in full above rather than held back.

One caveat about independence, stated because it cuts against us. OaSIS is not a wholly separate reading of the world: ESDC records that it used O*NET data as an input when assigning its descriptor ratings. So the agreement between our O*NET-anchored baseline and the OaSIS composite is partly inherited rather than earned, which is exactly why the correlation between them is measured and used to discount their combined weight rather than being read as corroboration. The one genuinely independent observation in the model remains the Anthropic Economic Index, because it records what people actually asked an AI to do rather than what an analyst judged a job to contain.

Versioning

Every score is produced from a dated dataset (`onet-30.0/aei-2026-05/oasis-2025-v1.1/oai-tjf-2026-04/eloundou-2023`) by a named method (`heuristic-v0`), and previously issued scores are never rewritten when either changes — stored scores are immutable, so a member's history is a record of readings rather than a moving number. **Storing that version alongside each score is not yet implemented**, so the provenance of an individual old reading has to be reconstructed from its date. Saying otherwise would describe an audit trail this paper cannot point at. A change to the method is a change to the paper.

Reading a score responsibly

Read the band before the number. Read the ordering before the band. Treat a score near a boundary as sitting between two labels. And treat the whole thing as one input into a decision that also includes everything we do not know about your employer, your field and you — which is most of it.

This paper documents the method behind the assessment at `netwell.ai/start`. Figures cited here are read from the shipping dataset and engine rather than transcribed, and the paper is rebuilt when they change.