Who Decides Which Jobs AI Will Take? The Answer Will Surprise You

Date: June 01, 2026 | Model: anthropic-batch:claude-opus-4-7

Source: Screenshot (OCR via AI)

Contents

Note: the original article is provided as a separate file (attached to the email or downloadable from the website).

1. Reading Passage

Over the past year, dozens of studies have warned that AI is poised to upend white-collar work. Yet buried in the footnotes of nearly all of them is a detail that ought to give readers pause: the exposure scores driving the headlines are not calculated by economists. They are generated by an AI model — most often GPT-4, in a 2024 study by OpenAI — which reads occupation descriptions and decides how automatable each task is.

That methodology has now been stress-tested. Michelle Yin, a researcher at Northwestern University, took all 705 occupations in the US occupational coding scheme and ran the original analysis through four different models: the GPT-4 used in the OpenAI study, plus newer systems from OpenAI, Anthropic and Google. The results were jarring. Estimates of the share of jobs at risk swung from under 15 per cent when judged by Google's Gemini to 50 per cent when judged by Anthropic's Claude. On the share of jobs more than 10 per cent exposed, GPT-4 said around 50 per cent, its successor GPT-5 put the figure just above that, and Claude 4.5 — the newest model tested — landed at 80 per cent.

The disagreement doesn't just shift a number; it can flip the entire conclusion. Using GPT-4's scores, AI exposure had a weak negative effect on employment, suggesting modest job loss. But using Gemini's scores, the same regression produced a weak *positive* effect — meaning the jobs flagged as most exposed actually grew. Same data, same methodology, different judge, opposite story.

Why do the models diverge? Yin attributes part of the gap to newer systems 'knowing' more about their own expanded abilities and about emerging AI tools that didn't exist when GPT-4 was trained. But there is also a stylistic component: newer models are simply more confident, and confidence inflates exposure scores even where real capability has not changed. Claude, in particular, rated occupations from CEOs to factory-floor supervisors as highly exposed to automation. Gemini, asked the same question, treated those same roles as relatively safe.

The authors argue that the fix is straightforward in principle: any serious analysis of real-world AI impact should run its exposure measure through multiple models and compare the results. Where the models agree, conclusions can be trusted. Where they diverge — as they often do — the honest answer is uncertainty. They also point to a broader implication. The EU's GDPR already gives individuals the right to a 'human review' of consequential automated decisions, such as a denied loan or a rejected job application. A logical extension would be requiring a second or third AI 'opinion' as well — running the same decision through a different vendor's model to see whether the outcome holds. Hiring platforms like HireVue, which feed video interviews through proprietary models to score candidates, would be obvious test cases.

The deeper question the study raises is interpretive. Different AI systems are not just calibrated differently; they appear to be thinking about the labour market in fundamentally different ways. One views a CEO's job as a stack of automatable tasks; another sees it as judgement, leadership and risk-taking that a chatbot cannot replicate. Until that gap is understood, the confident percentages in newspaper headlines say less about the future of work than they do about the model that

produced them.

2. Explanation

> ***Dozens of headlines claim AI is coming for white-collar work — but the entire field rests on a secret most readers miss: the verdicts are written by the AI models themselves, and they wildly disagree.***

What's Going On?

Economists keep publishing studies estimating how 'exposed' jobs are to AI disruption. Buried in the footnotes of most of these studies is an awkward detail: the exposure scores aren't produced by humans. They're produced by an earlier AI model — usually GPT-4 from a 2024 OpenAI paper — which read job descriptions and rated how automatable each task is.

A new study by Michelle Yin at Northwestern ran the exact same methodology across four different models (the original GPT-4, plus newer ones from OpenAI, Anthropic and Google) on all 705 occupations in the US classification system. The models disagreed dramatically. Estimates of how many jobs are at risk ranged from under 15% (Gemini) to 50% (Claude). On GPT-4, the share of jobs more than 10% exposed was about 50%; on GPT-5, just over 50%; on Claude 4.5, a striking 80%.

How To Think About It

The issue here isn't really about AI — it's about what happens when a measurement instrument is also the thing being measured. A few parallels make the problem sharper:

- Imagine grading the SAT by asking four different teachers to score the same essays — and one gives a 14% pass rate while another gives 80%. The 'score' tells you more about the grader than the student.
- It's like asking restaurants to rate their own food. The newer, more confident models rate more tasks as AI-doable — partly because they genuinely can do more, but partly because they're trained to sound capable.
- Or think of pre-GPS election polls: same voters, same questions, but the methodology each pollster picks quietly determines the headline number.

Key Things To Know

- The foundational study most headlines trace back to is a 2024 OpenAI paper that used GPT-4 to score 705 occupations.
- When Yin re-ran the analysis using Gemini's scores, AI exposure had a weak *positive* effect on employment — the most-exposed jobs actually grew. Flipping to other models flips the story.
- Claude rated occupations from CEOs to factory-floor supervisors as highly exposed; Gemini rated the same jobs as low-exposure. The models are 'thinking' about the question in fundamentally different ways.
- The EU's GDPR already gives people the right to a 'human review' of consequential automated decisions (loans, job applications). A natural next step is requiring a second or third AI 'opinion' — running the same case through a different vendor's model.
- Most people assume one neat number ('X% of jobs at risk') reflects reality. It actually reflects which

model the researcher happened to query.

Why It Matters

If you're choosing a college major or thinking about which career path is 'AI-proof,' the advice you're getting is built on shaky foundations. Policymakers writing retraining programmes, companies planning layoffs, and journalists writing scary headlines are all leaning on numbers that swing 5x depending on which chatbot was asked. The honest answer to 'will AI take this job?' is closer to 'it depends who you ask — including which AI you ask.'

The Bigger Picture

This is an early example of a problem that will define the next decade: AI systems are increasingly the judges, scorers and gatekeepers in decisions that affect real lives — hiring, lending, medical triage. The Yin study suggests a future where 'model disagreement' becomes its own field, and where regulations might require decisions to be cross-checked across competing AI systems before they stick. Watch for the first lawsuit where someone argues they were denied a job by one model that another model would have approved.

3. Key Terms Glossary

AI exposure

An estimate, usually expressed as a percentage of an occupation's tasks, of how much of a job could plausibly be done by AI. It's a measure of *potential* automation, not actual job losses.

Large language model (LLM)

An AI system trained on huge amounts of text to predict and generate language. Examples include GPT-4, GPT-5, Claude and Gemini. In this article they're being used as judges that score jobs.

Occupational coding scheme

A standardised government list that divides the labour market into discrete jobs — the US version contains 705 occupations and is used to track employment statistics.

Methodology

The specific recipe a study follows — what data it uses, how it measures things, what assumptions it makes. Two studies can use the 'same' methodology but reach opposite conclusions if one ingredient changes.

GDPR

The European Union's General Data Protection Regulation, which gives individuals legal rights over how their personal data is used, including the right to demand human review of important automated decisions.

Non-deterministic

A system that can give different answers to the same input on different runs. Most modern LLMs are non-deterministic, which is part of why their exposure scores vary.

Devil's advocate

Someone who argues a position they don't necessarily believe in, in order to stress-test the opposite view. Used in the article to introduce a counter-argument about why newer models' higher scores might still be useful.

4. Reading Comprehension Quiz

Circle the best answer for each question.

Q1. The passage most directly argues that:

- ◯ **A)** AI will replace roughly half of all white-collar jobs within a decade.
- ◯ **B)** Estimates of AI's impact on jobs depend heavily on which AI model produced them.
- ◯ **C)** Newer AI models are more accurate than older ones at predicting automation.
- ◯ **D)** Governments should ban the use of AI in hiring and lending decisions.

Q2. According to the passage, what did Michelle Yin's study find when the original OpenAI exposure analysis was re-run using Gemini's scores?

- ◯ **A)** AI exposure had a weak negative effect on employment.
- ◯ **B)** AI exposure had a weak positive effect on employment.
- ◯ **C)** AI exposure had no measurable effect on employment.
- ◯ **D)** AI exposure caused sharp job losses in white-collar work.

Q3. Which choice best states the central idea of the passage?

- ◯ **A)** AI's labour-market impact is best measured by economists, not by AI itself.
- ◯ **B)** Claude is the most accurate large language model for predicting job automation.
- ◯ **C)** Widely cited AI job-exposure figures rest on an unstable, model-dependent foundation.
- ◯ **D)** The EU's GDPR should be expanded to require human review of all AI decisions.

Q4. As used in the passage, the word 'exposed' most nearly means:

- ◯ **A)** publicly revealed
- ◯ **B)** physically uncovered
- ◯ **C)** vulnerable to being automated
- ◯ **D)** experienced in a topic

Q5. As used in the passage, the word 'bullish' most nearly means:

- ◯ **A)** stubborn and aggressive
- ◯ **B)** confident and optimistic
- ◯ **C)** physically powerful
- ◯ **D)** financially risky

Q6. Which statement about newer AI models can most reasonably be inferred from the passage?

- ◯ **A)** They produce more reliable predictions because they're trained on more data.
- ◯ **B)** Their higher exposure scores may reflect both real capability gains and overconfidence.
- ◯ **C)** They have been independently verified as accurate by Northwestern researchers.
- ◯ **D)** They consistently agree with one another on which jobs are at risk.

Q7. The passage suggests that the practical fix Yin's findings point toward is:

- ◯ **A)** abandoning AI exposure research entirely.
- ◯ **B)** using only the newest and most capable model available.
- ◯ **C)** running the same analysis across multiple models and comparing results.
- ◯ **D)** letting human workers self-report whether AI could do their job.

Q8. The author's tone in discussing widely cited AI exposure studies is best described as:

- ◯ **A)** alarmed and outraged
- ◯ **B)** skeptical but constructive
- ◯ **C)** neutral and detached
- ◯ **D)** dismissive and contemptuous

Q9. Which inference about hiring software like HireVue is best supported by the passage?

- ◯ **A)** Such systems are illegal under current EU law.
- ◯ **B)** Such systems always reach the same decision regardless of which AI is used.
- ◯ **C)** Such systems' decisions could change if a different underlying model were used.
- ◯ **D)** Such systems are more accurate than human interviewers.

Q10. Which choice provides the best evidence for the answer to the previous question?

- ◯ **A)** The mention that GDPR gives people the right to demand human review.
- ◯ **B)** Sarah's suggestion that running an application through a Google and Anthropic model could reach a different decision.
- ◯ **C)** John's note that Claude rated CEOs as highly exposed to AI.
- ◯ **D)** The description of HireVue as a major provider of hiring software.

My Score: _____ / 10

5. Answer Key with Explanations

Q1. The passage most directly argues that:

Answer: B

*The passage's central claim is that different LLMs produce wildly different exposure scores using the same methodology, so the numbers reflect the judge as much as the jobs. A overstates a specific figure; the passage gives a range, not a verdict. SAT Tip: For 'primarily argues' questions, pick the option that captures the *whole* passage's argument, not a single fact mentioned in one paragraph.*

Q2. According to the passage, what did Michelle Yin's study find when the original OpenAI exposure analysis was re-run using Gemini's scores?

Answer: B

The passage states that swapping to Gemini's scores flipped the result to a weak positive effect, with the most-exposed jobs actually seeing growth. A describes the GPT-4 result, not Gemini's — that's the main trap (Trap B: passage vocabulary in the wrong combination). SAT Tip: When a passage compares multiple cases, underline which result belongs to which case before you look at the answer choices.

Q3. Which choice best states the central idea of the passage?

Answer: C

The passage's spine is that the numbers shaping public debate are unstable because they shift dramatically depending on which model produced them. D is a real-world position someone might hold but is not the passage's central claim (Trap C: true-sounding but unsupported). SAT Tip: 'Central idea' answers are usually broader than any single example in the passage but narrower than a sweeping generalisation.

Q4. As used in the passage, the word 'exposed' most nearly means:

Answer: C

In context, 'exposed' describes how susceptible a job is to AI doing its tasks — i.e. vulnerable to automation. A is the everyday meaning of 'exposed' (revealed) and is the main trap (Trap B: common meaning vs. context-specific meaning). SAT Tip: For vocab-in-context, substitute each option for the word in the original sentence — the right answer keeps the sentence's meaning intact.

Q5. As used in the passage, the word 'bullish' most nearly means:

Answer: B

Sarah uses 'bullish' to describe newer models being more confident in claiming jobs can be automated. A describes a stubborn personality (Trap B: common alternative meaning of 'bullish'). SAT Tip: Financial and market vocabulary ('bullish,' 'bearish,' 'leverage') often shows up on the SAT — learn the figurative meaning, not just the literal one.

Q6. Which statement about newer AI models can most reasonably be inferred from the passage?

Answer: B

John replies that newer models scoring more tasks as exposed could reflect genuine new abilities, but he adds caveats — the very fact they disagree suggests it isn't all real capability. D directly contradicts the passage, which emphasises disagreement (Trap A: right scope, wrong direction). SAT Tip: Inference answers should be a small logical step from the text — not a leap, and not a restatement of the text.

Q7. The passage suggests that the practical fix Yin's findings point toward is:

Answer: C

The passage explicitly proposes that exposure analyses should be run using several different models' assessments so that findings reflect the labour market, not the quirks of one AI. B is tempting because newer models sound better, but the passage warns the newest model isn't necessarily right (Trap C: plausible but unsupported). SAT Tip: When a passage names a problem and a solution, the solution usually appears as a direct sentence — look for words like 'solution,' 'fix,' or 'one approach is.'

Q8. The author's tone in discussing widely cited AI exposure studies is best described as:

Answer: B

*The passage questions the reliability of headline figures (skeptical) but proposes a concrete fix — running analyses across multiple models (constructive). D is too strong; the authors take the studies seriously enough to suggest improvements (Trap A: right scope, wrong intensity). SAT Tip: Tone questions reward looking at *how* the author talks about the topic — verbs like 'glossed over' or 'flips' signal scepticism without hostility.*

Q9. Which inference about hiring software like HireVue is best supported by the passage?

Answer: C

HireVue is mentioned to illustrate that hiring AI uses proprietary models — and given the passage's whole argument that different models reach different conclusions, the natural inference is that swapping the underlying model could change the outcome. B contradicts the passage's core finding (Trap A: right topic, wrong direction). SAT Tip: For inference questions, the right answer is almost always a quieter, more cautious claim than the dramatic-sounding distractors.

Q10. Which choice provides the best evidence for the answer to the previous question?

Answer: B

Sarah explicitly hypothesises that running the same job application through a different underlying model could yield a different decision — directly supporting the inference in Q9. D names HireVue but doesn't speak to whether different models would disagree (Trap C: related fact, not evidence). SAT Tip: On evidence-pairing questions, first re-find the sentence that made you choose your previous answer — then pick the option that quotes or paraphrases that exact sentence.