Benchmarking the Future of Work: Mapping AI Progress to Occupational Exposure*

Kris Gulati[a,b]

[a] *Berkeley Haas*

[b] *University of California, Merced*

October 2, 2025

ABSTRACT

Artificial intelligence is advancing at a pace once thought unimaginable, yet we still lack clear tools to understand how these breakthroughs map onto the world of work and, in particular, how they shape an occupation's exposure to AI. We introduce a new measure of an occupation's exposure to AI that we call the Benchmark-based AI Occupational Exposure (BAIOE), which systematically links AI benchmark progress - the scoreboards that track frontier capabilities - to the occupational tasks that define human labor. Using O*NET tasks as a bridge, we connect benchmark trajectories across domains-including language, reasoning, vision, and multimodal tasks-to 52 human abilities, and translate these into occupation-level indices of AI exposure. The result is a dynamic, task-level methodology that allows us to track and forecast where automation pressures are likely to emerge. By repositioning benchmarks from technical scoreboards to economic indicators, this study offers a fresh lens for anticipating the future of work and shaping policy responses.

KEYWORDS: AI Benchmarks, Occupational Tasks, Automation Risk, Future of Work, Labor Market Exposure, Task-based Frameworks, Economic Impact of AI

*Kris Gulati thanks the Institute of Humane Studies under grant numbers: IHS018498, IHS018315, IHS01854, and Emergent Ventures. We thank Raymond Kim for his advice and feedback. All errors are our own.

"Benchmarking is the foundation of progress in AI. Without it, we cannot measure, compare, or improve." – GPT-5

1 Introduction

In just a few years, artificial intelligence has crossed thresholds that once seemed decades away: large language models now match or exceed human performance on professional exams, image models rival expert recognition, and reasoning systems solve math and logic problems once thought out of reach. At the center of these advances are benchmarks: public scoreboards that define progress in AI and signal when machines approach or surpass human capability.

In a world where policymakers, firms, and workers are urgently trying to anticipate which jobs are most at risk, benchmarks offer a powerful but underused tool. Prior work attempting to do this exercise has relied on patents Webb (2019), expert surveys Frey and Osborne (2017); Grace et al. (2018), or used a snapshot of AI performance Felten et al. (2021); Eloundou et al. (2024). In contrast, we believe and capitalize on AI benchmarks, which increasingly track advances in capability, capturing AI progress in close to real time.

This paper leverages that potential. We propose a framework that systematically links AI benchmark performance to occupational tasks using O*NET as a bridge. By treating benchmark trajectories as a dynamic proxy for AI capability, we create a method for tracking and forecasting how improvements in AI map onto the skills that underpin human work.

Our method proceeds in two steps. First, we assemble longitudinal data on AI benchmarks across domains such as language (e.g., MMLU, SuperGLUE), reasoning (e.g., GSM8K, ARC), vision (e.g., ImageNet, COCO), and multimodal tasks. Each benchmark provides a standardized, time-stamped measure of progress. We then map these benchmarks to the 52 O*NET abilities, which span cognitive, physical, psychomotor, and sensory domains. This mapping is established through an AI annotation process, ensuring that benchmarked capabilities are linked to the underlying skills required for occupational tasks.

Second, we translate benchmark progress into measures of task exposure. Following prior work Felten et al. (2021); Eloundou et al. (2024), we weight tasks by their importance and frequency in O*NET occupations. The result is an occupation-level index of AI exposure that reflects both the structure of work and the trajectory of AI performance. By aggregating these indices, we can analyze exposure across industries, as well as highlight domains where benchmarks align closely - or fail to align - with the realities of human work.

To the best of our knowledge, this is the first paper to directly use AI benchmarks to un-

derstand the impact of AI on O*NET tasks and occupations. This paper repositions AI benchmarks from a technical evaluation tool into a foundation for understanding and anticipating the economic consequences of AI progress.

2 Literature Review

Scholars have long sought to measure how advances in technology affect labor markets. One prominent approach infers exposure from patents and innovation outputs. Webb (2019) uses patent text to measure the overlap between new inventions and occupational descriptions, arguing that patents provide a forward-looking proxy for displacement risk. While effective for many general-purpose technologies, this method has limited traction in AI, where many of the most transformative advances originate in private labs that patent sparingly. Moreover, patents are inherently lagged indicators, reflecting inventions only after formal filings and approvals, often well after the underlying technological progress has occurred.

A second line of work relies on expert surveys. Frey and Osborne (2017) famously elicited expert judgments about the susceptibility of occupations to automation, while Grace et al. (2018) surveyed AI researchers to forecast timelines to human-level performance across a wide set of tasks. These studies generate valuable expectations but are inherently subjective, costly, and updated intermittently.

A third, more recent, and directly relevant stream of literature leverages task-level data from O*NET. Felten et al. (2021) creates an "AI Occupational Exposure" index by linking expert assessments of AI capabilities to O*NET tasks, while Eloundou et al. (2024) extend this approach by connecting the abilities of large language models (LLMs) to O*NET tasks, showing that knowledge-intensive occupations are particularly exposed. These studies move beyond broad occupation-level analysis by grounding exposure in tasks and underscore the value of transparent, task-level frameworks for translating technological performance into economic exposure. Benchmarks, by contrast, offer dynamic, transparent, and continuously updated signals of AI progress, capturing advances in near real time and providing a direct way to link evolving capabilities to human skills and tasks.

In a similar vein, Tolan et al. (2021) links AI benchmarks to O*NET tasks by way of cognitive skill proxies. While this approach usefully connects technical progress to labor market data, it does so indirectly: task relevance is inferred from broad cognitive categories, and benchmark performance is approximated through research intensity rather than observed technical outcomes. By contrast, our framework establishes a direct mapping from benchmark results to ONET abilities through AI-driven annotation. This yields a more fine-grained and empirically

grounded assessment of which human skills are most exposed to AI advances and highlights where significant capability gaps remain.

More recently, Handa et al. (2025) take a complementary approach by analyzing millions of real-world conversations from Claude mapped to O*NET tasks. Their framework offers an empirical snapshot of how AI is currently being used across occupations, with especially high usage in software development and writing tasks but little penetration into roles requiring physical manipulation. While this usage-based evidence provides valuable insight into present-day diffusion, benchmarks serve as an essential complement in two respects. First, they provide forward-looking and transparent signals of frontier capabilities, offering a clearer view of where AI progress is heading. Second, because access to model prompts and usage data is often restricted, benchmarks remain one of the few systematic and openly available measures through which researchers can track advances.

Study	**Data/Approach**	**Mapping to Work**
Webb (2019)	Patent text	Overlap between patents and O*NET tasks
Frey & Osborne (2017)	Expert forecasts	Judgments on occupations' automation risk
Grace et al. (2018)	Expert forecasts	Researcher predictions of AI reaching human-level tasks
Felten et al. (2021)	EFF AI Progress project	10 AI progress areas → O*NET tasks
Eloundou et al. (2024)	LLM evaluations	LLM abilities → O*NET tasks
Tolan et al. (2021/2024)	Research intensity in AI fields	Research intensity in AI fields → Cognitive categories → O*NET tasks
Handa et al. (2025)	Real-world usage data (Claude conversations)	AI usage patterns → O*NET tasks
This paper	AI benchmarks (e.g., MMLU, GSM8K, ImageNet)	Benchmark performance → O*NET tasks

Table 1: Overview of prior approaches to measuring AI exposure and this paper's contribution.

This paper adds to this literature by proposing a new measure for occupational exposure to AI, which we term the Benchmark-based AI Occupational Exposure (BAIOE). We capitalize on the rapid improvements in AI benchmarks, which are expanding in scope, becoming more comprehensive, and increasingly aligned with real-world tasks. Benchmark suites now cover a wider range of domains - from language and reasoning to vision and multimodal capabilities - allowing for more granular assessment of technical progress. Because benchmark results are continuously updated and openly reported, they provide near real-time signals of frontier performance, offering a transparent and replicable foundation for measuring AI exposure. Ultimately, we believe that leveraging AI benchmarks represents the best available proxy for

tracking the progress of AI models, and in doing so, provides a foundation for understanding how occupations in the future may be at risk of AI exposure.

3 Conceptual Framework

Our framework builds on the task-based approach to technological change and labor markets Acemoglu and Restrepo (2018). In this view, automation and AI affect employment not at the level of whole occupations, but at the level of tasks-the discrete units of human activity that together constitute work. The displacement and complementarity effects of new technologies therefore hinge on how technical progress maps onto specific human abilities. A central challenge in this literature has been measurement: while task-level models are theoretically compelling, empirical proxies for technological capability have typically relied on lagged or indirect indicators such as patents, R&D expenditures, or expert surveys.

We conceptualize AI benchmarks as a new measurement instrument that overcomes these limitations. Benchmarks provide standardized, replicable, and continuously updated records of AI performance across a wide array of tasks, from language understanding to motor control. By linking these benchmarks to O*NET's taxonomy of abilities, we embed frontier AI progress directly into the task-based framework of labor markets. This approach situates benchmarks alongside patents and publications in the broader innovation-measurement tradition, but extends the paradigm by providing real-time, task-level indicators of capability. In doing so, our framework offers a systematic way to track how technological progress unfolds across the building blocks of human work.

4 Methodology

Our methodological approach of mapping AI benchmark performance into occupational exposure builds on two recent papers.

First, Felten et al. (2021) map the Electronic Frontier Foundation (EFF) AI Progress Measurement project to ONET tasks through a structured labeling approach. They begin by selecting ten AI applications tracked by the EFF - such as image recognition, reading comprehension, language modeling, and speech recognition - and then link each application to the 52 workplace abilities defined in O*NET. To establish these links, they run a large survey on Amazon Mechanical Turk, asking respondents (gig workers) to rate how related each AI application is to each O*NET ability. This produces an application-ability relatedness score between 0 and 1 for every pairing. The scores are organized into a matrix that systematically

connects the EFF applications to O*NET abilities. These ability-level exposures are then aggregated to the occupational level by weighting them with O*NET's measures of ability importance and prevalence, resulting in the AI Occupational Exposure (AIOE) index. This labeling procedure provides a transparent way of mapping progress in frontier AI applications to specific occupational tasks and abilities

Second, Eloundou et al. (2024) map large language model (LLM) capabilities to O*NET tasks using a structured labeling approach built around a new "exposure rubric." They begin with O*NET's Detailed Work Activities (DWAs) and tasks, and then assess whether access to an LLM (e.g., via ChatGPT) or to LLM-powered software could reduce the time needed to complete a task by at least 50% without loss of quality. Each task is assigned one of three exposure categories: E0 (no exposure), where LLMs provide minimal benefit or degrade quality; E1 (direct exposure), where the LLM alone can reduce task time by half; and E2 (LLM+ exposure), where the LLM by itself is insufficient, but complementary software or tools built on top of it could achieve that threshold. This study used human annotators to apply the exposure rubric to O*NET tasks, but AI annotators (GPT-4) were found to perform just as well, producing comparable task-level classifications.

Building on these approaches, my framework maps AI benchmarks directly to O*NET tasks using AI-driven annotations. Instead of relying on patents, expert surveys, or one-off model evaluations, I treat benchmark performance-on datasets such as MMLU, GSM8K, ImageNet, and other standardized scoreboards - as dynamic signals of frontier AI capability. Each benchmark is systematically linked to O*NET abilities through an AI annotation process, which rates how well benchmarked skills translate into real-world occupational tasks and how AI performance compares to human ability. These benchmark-to-task links are then aggregated to the occupation level, weighted by O*NET's measures of task importance and prevalence, to produce an index of Benchmark-based AI Occupational Exposure (BAIOE).

4.1 AI Annotation

Our methodology proceeds in a structured sequence designed to connect AI benchmarks to O*NET abilities and, ultimately, to occupational exposure measures. The process consists of six main steps, which were done by the AI.

4.1.1 Step 1: Benchmark Selection

For each O*NET ability, we reviewed from a list of candidate benchmarks available online from papers and selected the single most relevant one that best captured the underlying skill

or ability. Selection prioritized benchmarks with documented performance results available as of 2025. When no appropriate benchmark was available, we recorded "none" and provided a brief justification. This ensured transparency about both inclusions and omissions.

4.1.2 Step 2: Benchmark Validation and Updating

We conducted systematic searches to verify whether more accurate or recent benchmarks existed for each O*NET ability. Candidate benchmarks were considered valid replacements only if they (i) directly measured the underlying skill, (ii) offered superior construct validity or coverage, (iii) had documented 2025 performance, (iv) used standardized and comparable evaluation metrics, and (v) showed evidence of community adoption. If no benchmark satisfied these criteria, we defaulted to the original selection. A fallback search strategy was employed to guard against overlooking emerging benchmarks.

4.1.3 Step 3: Benchmark Metadata Compilation

For each selected benchmark, we compiled a structured summary of its key attributes. This included its name and abbreviation, domain (e.g., NLP, vision, reasoning, multimodal, robotics), task format, release year and updates, intended purpose, descriptive scope, and coverage across skills or domains. These metadata provided the foundation for linking benchmarks to O*NET abilities in a transparent and replicable way.

4.1.4 Step 4: Benchmark-to-Task Translation

We then rated how directly each benchmark translated into real-world human use of the associated O*NET ability. Ratings were assigned on a 0 - 10 scale, where 0 indicated no meaningful connection and 10 represented a near-direct match to workplace application. Each rating was accompanied by a justification outlining the aspects of alignment and any gaps that remained.

Specifically, we rated how directly the benchmark translates into real-world human use of the O*NET task on a 0-10 scale, where 0 indicates no meaningful connection, 5 reflects only partial translation, and 10 indicates a direct match. Each rating is accompanied by a brief justification (1-3 sentences) that explains how well the benchmark captures the way this ability is actually used in work settings, noting both the areas of alignment and the gaps that remain.

4.1.5 Step 5: Performance Synthesis

We synthesized the most recent performance of AI systems on each benchmark as of 2025. This step included both quantitative results (leaderboard scores, accuracy rates, etc.) and qualitative analysis of strengths, weaknesses, and error patterns. We also examined the sources of limitations, such as architectural constraints, training data biases, or benchmark design features, and interpreted what these meant for the benchmark's ability to approximate real-world capability.

4.1.6 Step 6: Human-Comparative Rating

Finally, we rated benchmarked AI systems relative to human ability on the corresponding O*NET task using a standardized 0-10 scale, where 0 indicates near-zero competence, 5 reflects useful but clearly sub-human performance, 7-9 corresponds to human-level capability, and 10 represents expert-level or superhuman proficiency. Each score is defined to translate benchmark outcomes into a common interpretive frame directly tied to occupational skills: a 0 means the system cannot meaningfully perform the task; a 5 reflects sub-human but practically useful contributions; 7-9 indicates roughly human-level performance with variation depending on task difficulty; and 10 represents consistent expert-level or superhuman proficiency. This rating scheme ensures that benchmark results can be interpreted in terms of real-world human capabilities.

4.2 Aggregation

To construct our measure of AI exposure, we integrate two complementary task-level scores into a single index. The first derives from step 4 (benchmark-to-task translation), and the second from step 6 (human-comparative rating). Both raw scores are normalized onto a common $[0, 1]$ scale and then combined through a weighted average. This yields a composite exposure score for each O*NET task.

Following Felten et al. (2021), we then map task-level exposure into occupational exposure. Each task is weighted by O*NET's measures of importance and prevalence, ensuring that frequently performed and central activities exert greater influence on an occupation's overall score. Aggregating in this way links advances in AI benchmarks to the heterogeneous distribution of exposure across the labor market.

$$\tilde{T}_\tau = \frac{T_\tau}{10}, \qquad \tilde{H}_\tau = \frac{H_\tau}{10}, \qquad 0 \le T_\tau, H_\tau \le 10 \tag{1}$$

Here, T_τ and H_τ are rescaled from their original 0-10 range to a common 0-1 range, so that both measures can be compared directly.

$$X_\tau = \lambda\, \tilde{T}_\tau + (1-\lambda)\, \tilde{H}_\tau, \qquad 0 \le \lambda \le 1 \tag{2}$$

We then take a convex combination of the two normalized measures. The parameter λ governs their relative weight, producing a single exposure score X_τ for each task.

$$\bar{w}_{\tau,o} = I_{\tau,o}\, F_{\tau,o}, \tag{3}$$

$$w_{\tau,o} = \frac{\bar{w}_{\tau,o}}{\sum_{\tau' \in \mathcal{T}_o} \bar{w}_{\tau',o}} \tag{4}$$

Each task τ within occupation o is weighted by its O*NET importance $I_{\tau,o}$ and frequency $F_{\tau,o}$. Normalization ensures the weights sum to one, so they reflect each task's relative contribution within the occupation.

$$\text{Exposure}_o = \sum_{\tau \in \mathcal{T}_o} w_{\tau,o}\, X_\tau \tag{5}$$

Finally, the exposure of occupation o is computed as the weighted average of its task-level scores, meaning tasks that are more central or more frequent dominate the occupation's exposure index.

5 Results

5.1 Preliminary Results

The preliminary results in Tables 2 and 3 reveal a clear divide between the least and most exposed occupations. At first glance, the jobs identified as least exposed-such as dishwashers, cleaners, and food service workers-align with expectations, since these roles depend heavily on physical, manual, or social interaction tasks that current AI benchmarks do not capture well. By contrast, the list of most exposed occupations, which includes airline pilots, physicians, and firefighters, appears less intuitive. This unexpected outcome likely reflects some methodological issues. Because the exposure index aggregates across all tasks, the cumulative effect of smaller, less relevant tasks may overshadow the more central abilities that actually define these jobs. To address these concerns, the next subsections will revisit the data using an alternative aggregation strategiy that better accounts for task importance and relevance.

Table 2: Top 20 Least Exposed Jobs

Rank	Least Exposed
1	Models
2	Graders and Sorters, Agricultural Products
3	Locker Room, Coatroom, and Dressing Room Attendants
4	Dishwashers
5	Pressers, Textile, Garment, and Related Materials
6	Manicurists and Pedicurists
7	Shampooers
8	Fast Food and Counter Workers
9	Maids and Housekeeping Cleaners
10	Slaughterers and Meat Packers
11	Packers and Packagers, Hand
12	Food Preparation Workers
13	Telemarketers
14	Cleaners of Vehicles and Equipment
15	Dining Room and Cafeteria Attendants and Bartender Helpers
16	Cooks, Fast Food
17	Ushers, Lobby Attendants, and Ticket Takers
18	Janitors and Cleaners, Except Maids and Housekeeping Cleaners
19	Food Servers, Nonrestaurant
20	Funeral Attendants

Notes: Rankings are ordered from 1 (least exposed overall) to 20 (relatively more exposed within this group).

These occupations are identified as having the lowest exposure scores in the dataset.

Table 3: Top 20 Most Exposed Jobs

Rank	Most Exposed
1	Airline Pilots, Copilots, and Flight Engineers
2	Physicists
3	Commercial Pilots
4	Emergency Medicine Physicians
5	Air Traffic Controllers
6	Ophthalmologists, Except Pediatric
7	Firefighters
8	Dentists, General
9	Molecular and Cellular Biologists
10	Biochemists and Biophysicists
11	Robotics Engineers
12	Oral and Maxillofacial Surgeons
13	Police and Sheriff's Patrol Officers
14	Marine Engineers and Naval Architects
15	Captains, Mates, and Pilots of Water Vessels
16	Urologists
17	Manufacturing Engineers
18	Aircraft Mechanics and Service Technicians
19	Nanosystems Engineers
20	Radiologists

Rankings are ordered from 1 (most exposed) to 20 (relatively less exposed within this high-exposure group).

These occupations are identified as having the highest exposure scores in the dataset.

5.2 Excluding Less Relevant Tasks

In this subsection, we refine the exposure rankings by excluding tasks that are less relevant to the core functions of each occupation. This adjustment reduces distortions caused by peripheral or incidental activities that may artificially inflate or deflate exposure scores without capturing the true nature of the work. The updated rankings reveal a sharper divide: low-exposure occupations remain concentrated in manual, service, and cleaning roles, while high-exposure occupations are consistently clustered in scientific, technical, and professional domains. For instance, graders, dishwashers, and packagers persist among the least exposed, whereas physicists, engineers, and physicians dominate the most exposed group. By narrowing the focus to central tasks, this approach enhances the robustness of the framework, producing rankings that more closely align with intuitive expectations of AIâs impact and drawing attention to the skills most essential to each occupation.

Table 4: Top 20 Least Exposed Jobs (Excluding Less Relevant Tasks)

Rank	Least Exposed
1	Graders and Sorters, Agricultural Products
2	Locker Room, Coatroom, and Dressing Room Attendants
3	Models
4	Pressers, Textile, Garment, and Related Materials
5	Manicurists and Pedicurists
6	Dishwashers
7	Fast Food and Counter Workers
8	Packers and Packagers, Hand
9	Shampooers
10	Food Preparation Workers
11	Slaughterers and Meat Packers
12	Cleaners of Vehicles and Equipment
13	Maids and Housekeeping Cleaners
14	Dining Room and Cafeteria Attendants and Bartender Helpers
15	Telemarketers
16	Ushers, Lobby Attendants, and Ticket Takers
17	Crossing Guards and Flaggers
18	Painting, Coating, and Decorating Workers
19	Janitors and Cleaners, Except Maids and Housekeeping Cleaners
20	Cooks, Fast Food

Notes: Rankings are ordered from 1 (least exposed overall) to 20 (relatively more exposed within this group).

These occupations are identified as having the lowest exposure scores in the dataset.

Table 5: Top 20 Most Exposed Jobs (Excluding Less Relevant Tasks)

Rank	Most Exposed
1	Physicists
2	Emergency Medicine Physicians
3	Airline Pilots, Copilots, and Flight Engineers
4	Biochemists and Biophysicists
5	Molecular and Cellular Biologists
6	Mathematicians
7	Robotics Engineers
8	Nanosystems Engineers
9	Manufacturing Engineers
10	Marine Engineers and Naval Architects
11	Microbiologists
12	Urologists
13	Ophthalmologists, Except Pediatric
14	Astronomers
15	Oral and Maxillofacial Surgeons
16	Dentists, General
17	Epidemiologists
18	Bioinformatics Scientists
19	Geneticists
20	Air Traffic Controllers

Rankings are ordered from 1 (relatively less exposed within this high-exposure group) to 20 (most exposed overall).

These occupations are identified as having the highest exposure scores in the alternative dataset.

6 Limitations

This study faces several important limitations. First, for each O*NET ability we relied on a single "best" benchmark that most closely captures the underlying skill. While this choice provides clarity and tractability, it inevitably excludes alternative benchmarks that might capture different dimensions of the same ability. Multi-benchmark aggregation could help smooth idiosyncrasies in benchmark design and reduce sensitivity to the particularities of any single evaluation.

Second, our approach relies on AI-driven annotation to link benchmarks to O*NET tasks, and assumes that this process provides a valid proxy for real-world task capability. Prior work Eloundou et al. (2024) has shown that AI annotation can generate mappings that meaningfully capture occupational exposure, lending credibility to this method. Nonetheless, systematic validation remains limited. We are currently pursuing efforts to validate these benchmark-

to-task mappings through human annotation, which would provide an important check on construct validity and help ensure that exposure measures reflect actual workplace relevance.

Third, our framework does not account for the costs of deploying AI systems in real-world settings. Benchmarks capture technical capability, but they do not reflect the economic, organizational, or regulatory frictions that often determine whether a technology is adopted. High infrastructure costs, compliance requirements, or integration challenges can significantly delay or prevent deployment, even when benchmark performance suggests technical readiness. Ignoring these costs means our exposure estimates may overstate the immediacy of impact in certain occupations.

Fourth, AI benchmarks themselves are still developing. Many are evolving rapidly in scope, coverage, and methodology, and new and better benchmarks are created on a continuous basis. While this study uses the most appropriate benchmarks available as of 2025, the framework should be seen as iterative and updatable, with future work incorporating emerging benchmarks to reflect frontier capabilities more accurately.

7 Conclusion

This paper has introduced a novel framework for linking AI benchmark progress to occupational tasks, repositioning benchmarks from narrow technical scoreboards into forward-looking indicators of economic impact. By systematically mapping benchmark trajectories to O*NET abilities, we provide a dynamic method for assessing how advances in AI translate into task- and occupation-level exposure.

Building on prior work, we argue that AI benchmarks represent the most reliable instrument currently available for tracking AI progress. Whereas earlier studies relied on patents, expert surveys, or one-off evaluations that are often lagged, subjective, or static, benchmarks offer continuously updated measures of capability. In this sense, our approach extends and strengthens the existing literature by transforming benchmarks into a tool for capturing the evolving boundaries between human and machine skills.

By reframing benchmarks as economic indicators, this study contributes to a deeper understanding of the future of work and offers a new lens for guiding research, policy, and organizational strategy. As AI systems continue to advance across domains, our framework provides a systematic foundation for forecasting where automation pressures may arise and for designing responses that harness innovation while safeguarding workers.

References

Acemoglu, D. and P. Restrepo (2018). The race between man and machine: Implications of technology for growth, factor shares, and employment. *American economic review 108*(6), 1488–1542.

Eloundou, T., S. Manning, P. Mishkin, and D. Rock (2024). Gpts are gpts: Labor market impact potential of llms. *Science 384*(6702), 1306–1308.

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

Frey, C. B. and M. A. Osborne (2017). The future of employment: How susceptible are jobs to computerisation? *Technological forecasting and social change 114*, 254–280.

Grace, K., J. Salvatier, A. Dafoe, B. Zhang, and O. Evans (2018). When will ai exceed human performance? evidence from ai experts. *Journal of Artificial Intelligence Research 62*, 729–754.

Handa, K., A. Tamkin, M. McCain, S. Huang, E. Durmus, S. Heck, J. Mueller, J. Hong, S. Ritchie, T. Belonax, K. K. Troy, D. Amodei, J. Kaplan, J. Clark, and D. Ganguli (2025). Which economic tasks are performed with ai? evidence from millions of claude conversations. *arXiv preprint arXiv:2503.04761*.

Tolan, S., A. Pesole, F. Martínez-Plumed, E. Fernández-Macías, J. Hernández-Orallo, and E. Gómez (2021). Measuring the occupational impact of ai: tasks, cognitive abilities and ai benchmarks. *Journal of Artificial Intelligence Research 71*, 191–236.

Webb, M. (2019). The impact of artificial intelligence on the labor market. *Available at SSRN 3482150*.