Working Paper

Beyond AI Exposure: Which Tasks are Cost-Effective to Automate with Computer Vision?

Maja S. Svanberg
Massachusetts Institute of Technology, svanberg@mit.edu

Wensu Li
Massachusetts Institute of Technology, wensu@mit.edu

Martin Fleming
The Productivity Institute, martin@fleming41.com

Brian C. Goehring
IBM's Institute for Business Value, goehring@us.ibm.com

Neil C. Thompson*
Massachusetts Institute of Technology, neil_t@mit.edu

Abstract.

The faster AI automation spreads through the economy, the more profound its potential impacts, both positive (improved productivity) and negative (worker displacement). The previous literature on "AI Exposure" cannot predict this pace of automation since it attempts to measure an overall potential for AI to affect an area, not the technical feasibility and economic attractiveness of building such systems. In this article, we present a new type of AI task automation model that is end-to-end, estimating: the level of technical performance needed to do a task, the characteristics of an AI system capable of that performance, and the economic choice of whether to build and deploy such a system. The result is a first estimate of which tasks are technically feasible and economically attractive to automate - and which are not. We focus on computer vision, where cost modeling is more developed. We find that at today's costs U.S. businesses would choose *not* to automate most vision tasks that have "AI Exposure," and that only 23% of worker wages being paid for vision tasks would be attractive to automate. This slower roll-out of AI can be accelerated if costs falls rapidly or if it is deployed via AI-as-a-service platforms that have greater scale than individual firms, both of which we quantify. Overall, our findings suggest that AI job displacement will be substantial, but also gradual – and therefore there is room for policy and retraining to mitigate unemployment impacts.

Funding: MIT-IBM Watson AI Lab provided funding for the project.

Key words: artificial intelligence, automation, AI, computer vision

Acknowledgements: We are grateful to Ling Yue Poon and Haoran Lyu for research assistance and to Nicholas Borge, Ben Tang, Anna Pastwa, Subhro Das, Eni Mustafaraj, and Christophe Combemale for their contributions. An early draft of this paper appeared as Svanberg (2023).*Corresponding author.

"Machines will steal our jobs" is a sentiment frequently expressed during times of rapid technological change. Such anxiety has re-emerged with the creation of large language models (e.g. ChatGPT, Bard, GPT-4) that show considerable skill in tasks where previously only human beings showed proficiency. A recent study found that about 50% of tasks could be at least partially automated with large language models (Eloundou et al. 2023). If task automation of that extent were to happen rapidly, it would represent an enormous disruption to the labor force. Conversely, if that amount of automation were to happen slowly then labor might be able to adapt as it did during other economic transformations (e.g. moving from agriculture to manufacturing). So, making good policy and business decisions depends on understanding how rapidly AI task automation will happen.

While there is already evidence that AI is changing labor demand (Fleming et al. 2019, Acemoglu et al. 2022), most anxieties about AI flow from predictions about "AI Exposure" that classify tasks or abilities by their potential for automation, as measured by various proxies (Arntz et al. 2017, Brynjolfsson et al. 2018, Felten et al. 2018, Webb 2019, Felten et al. 2021, Tolan et al. 2021, Meindl et al. 2021, Zarifhonarvar 2023, Felten et al. 2023). Importantly, nearly all these predictions are vague about the timeline and extent of automation because they do not directly consider the technical feasibility or economic viability of AI systems, but instead use measures of similarity between tasks and AI capabilities to indicate exposure. The only exception in the literature known to us is a McKinsey report (Ellingrud et al. 2023) that estimates AI adoption of between 4% and 55%. With such imprecise predictions, it is unclear what conclusions should follow. AI exposure models also conflate predictions about full task automation, which is more likely to displace workers, with partial automation, which could augment their productivity. Separating these effects is enormously important for understanding the economic and policy implications of automation (Acemoglu and Restrepo 2018).

In this paper, we address three important shortcomings of AI exposure models to construct a more economically-grounded estimate of task automation. First, we survey workers familiar with end-use tasks to understand what performance would be required of an automated system. Second, we model the cost of building AI systems capable of reaching that level of performance. This cost estimate is essential to understanding the deployment of AI, since technically-exacting systems can be enormously expensive. And third, we model the decision about whether AI adoption is economically-attractive. The result is the first end-to-end AI automation model.

A simple hypothetical example makes clear why these considerations are so important. Consider a small bakery evaluating whether to automate with computer vision. One task that bakers do is to visually check their ingredients to ensure they are of sufficient quality (e.g. unspoiled). This task could theoretically be replaced with a computer vision system by adding a camera and training the system to detect food that has gone bad. Even if this visual inspection task could be separated from other parts of the production process, would it be cost effective to do so? Bureau of Labor Statistics O*NET data imply that checking food quality comprises roughly 6% of the duties of a baker. A small bakery with five bakers making typical salaries ($48,000 each per year), thus has potential labor savings from automating this task of $14,000 per year. This amount is far less than the cost of developing, deploying and maintaining a computer vision system and so we would conclude that it is not economical to substitute human labor with an AI system at this bakery.

The conclusion from this example, that human workers are more economically-attractive for firms (particularly those without scale), turns out to be widespread. We find that only 23% of worker compensation "exposed" to AI computer vision would be cost-effective for firms to automate because of the large upfront costs of AI systems. The economics of AI can be made more attractive, either through decreases in the cost of deployments or by increasing the scale at which deployments are made, for example by rolling-out AI-as-a-service platforms (Borge 2022), which we also explore. Overall, our model shows that the job loss from AI computer vision, even just within the set of vision tasks, will be smaller than the existing job churn seen in the market, suggesting that labor replacement will be more gradual than abrupt.

The rest of the paper is structured as follows: Section 1 introduces our framework to estimate which tasks are economically attractive to automate, section 2 presents the results, section 3 discusses how labor-replacing AI could proliferate, section 4 discusses the relevance of computer vision automation for other parts of AI, and section 5 concludes.

1. Method

1.1. Overview

We develop a task-based approach to our analysis (Autor et al. 2003) that focuses on two key questions: (i) **Exposure:** might it possible to build an AI model to automate this task, and (ii) **Economically-attractive:** would it be more attractive to use an AI system for this task than to have human workers continue to do it. To assess exposure, we follow the literature (e.g., Brynjolfsson et al. (2018)) in evaluating task descriptions for whether it might be feasible for an AI systems to perform them. Our main contribution is the second part of the analysis: assessing the economic attractiveness of automation. For reasons discussed later, the economic attractiveness of human labor and AI systems is largely driven by the relative costs of each. Modeling the human cost is straightforward labor accounting. Modeling the cost of AI systems is more complicated, so we draw on the computer science literature on training and doing inference with deep learning,[1] as well as 35 case studies that we performed to gather additional data. Central to our comparison of human and AI costs is the concept of*minimum viable scale*, which occurs when the AI deployment's fixed costs are sufficiently amortized that the average cost of using the computer vision system is the same as the cost of human labor of equivalent capability (Borge 2022), as shown in figure 1. AI automation is cost-effective only when the deployment scale is larger than the minimum viable scale.

[1] The technique that has been the dominant source of AI progress since 2012.

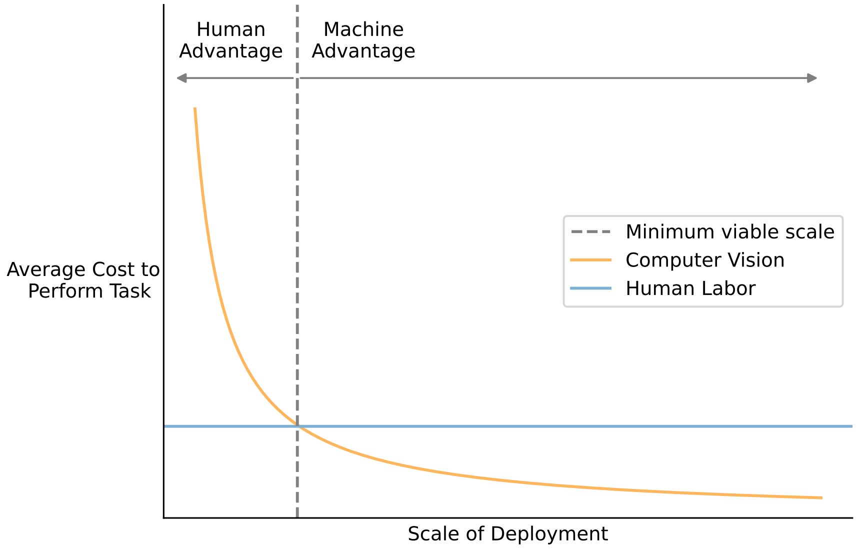


Figure 1 **The minimum viable scale for AI deployment.**

Concretely, we say that computer vision has an economic advantage $E^M_{i,j}$, where M is for *machine*, for deployment unit i, over human labor for a given task, j, when computer vision can be used for the task (i.e. it has *technology exposure,* T_j) and when the cost of the computer vision system, $C^M_{i,j}$, is less than the cost of *human* labor, $C^H_{i,j}$:

$$E^M_{i,j} = T_j \wedge (C^M_{i,j} < C^H_{i,j})$$

To get the fraction F of compensation that is at risk of being automated according to our model, we aggregate the cost of human labor for the economically feasible tasks over the total sum of compensation for all tasks J:

$$F_{I,J} = \sum_{i \in I} \sum_{j \in J} \frac{E^M_{i,j} \times C^H_{i,j}}{C^H_{i,j}}$$

Initially we consider the scope of deployment at the firm level, in other words, i denotes different firms. Later, we expand our analysis to broader scopes of deployment, namely, changing i to representing industry groups (4-digit NAICS), subsectors (3-digit NAICS), sectors (2-digit NAICS), or the entire U.S. Economy. In all cases, we focus on U.S. non-farm businesses to align our analysis with the data on firm sizes from the Statistics of U.S. Businesses (U.S. Census Bureau (2023)).

1.2. Exposure to Computer Vision (T_j)

Many tasks currently performed by workers could be carried out by a sufficiently sophisticated computer vision system, e.g., checking products for quality at the end of a factory assembly line or scanning medical imagery for anomalies. Other tasks have little use for vision technologies, e.g., negotiating the salary of subordinates.

j	O*NET task	Direct Work Activity (DWA)	Vision Task?
1	Operate diagnostic equipment, such as radiographic or ultrasound equipment, and interpret the resulting images.	Analyze test data or images to inform diagnosis or treatment.	Yes
2	Operate diagnostic equipment, such as radiographic or ultrasound equipment, and interpret the resulting images.	Operate diagnostic imaging equipment.	No
3	Examine trays to ensure that they contain required items.	–	Yes

Table 1 Computer Vision Exposure: Examples of O*NET Task - Direct Work Activity pairs, and their classification as vision tasks.

To assess whether a task j has technology exposure, T_j, we need to identify which tasks in the economy are *vision tasks* and which are not. While prior AI exposure analyses inspired our approach to this paper, we cannot use their data directly. Felten et al. (2018, 2023) base their method on linking AI progress to abilities, which does not translate into our task-based model of replacement. Webb (2019) uses a task-based model but does not allow for an easy distinction between computer vision and other AI domains, and Eloundou et al. (2023) only cover language tasks. Brynjolfsson et al. (2018) did produce task-level data with an indicator for whether they could be performed using image data. However, when filtering their results for highly scoring image-based tasks, the output contained many tasks for which we, upon manual inspection, could not see an obvious computer vision use case, e.g., “Analyze market conditions or trends” or “Dispose of biomedical waste in accordance with standards.” Therefore, we create our own data on computer vision exposure.

We take a manual approach to identifying T_j. Like Webb (2019), Eloundou et al. (2023), and Brynjolfsson et al. (2018), we rely on the O*NET Database 27.1 (U.S. Department of Labor 2023b). O*NET contains standardized characteristics of work and workers in the United States. By relying on this database, we assume that those tasks are an appropriate unit of replacement and that the initial technology deployment does not otherwise fundamentally change the task structure. The data contains descriptions of the nature of 1,016 occupations with 19,265 unique associated tasks, which are in turn mapped to 2,087 different direct work activities (DWAs) through a many-to-many relationship. Although the word “task” is a category in the O*NET schema, we find that the descriptions of the O*NET-tasks are too broad and include too many different capabilities. Therefore, we define a *task* for our purposes as the combination of an O*NET-Task and DWA, as shown in Table 1. This is a more detailed categorization than previous analyses that use just DWAs (Brynjolfsson et al. 2018), or just tasks (Webb 2019). O*NET-Tasks that do not have any associated DWAs are treated as one task. [2]

The large number of O*NET-Tasks makes manual identification of vision tasks challenging, but because of the lack of prior art on automatically identifying vision tasks, it was still our preferred approach. To

[2] To align the more-granular O*NET taxonomy with wage data that uses the Standard Occupational Classification (SOC), we truncate O*NET-SOC codes (see Appendix A.1).

classify the combinations of almost 20,000 tasks and 2,000 DWAs, we first identify 190 DWAs that indicate that they could be replaced by computer vision in some way. These included DWAs such as “Assess skin or hair conditions,” “Examine patients to assess general physical condition,” “Inspect items for damage or defects,” and “Monitor facilities or operational systems.” Filtering on these 190 DWAs yields 1,922 possible O*NET-Task-DWA combinations, which we also review manually to identify a total of 420 *vision tasks*, 414 of which exist in U.S. non-farm businesses. Additional details on task selection can be found in Appendix A.2.

1.3. Economic Attractiveness of Computer Vision ($E^M_{i,j}$)

To assess the economic attractiveness of using computer vision systems, it is important to consider both the benefits and cost of their deployment, as compared to the human workers currently doing those tasks. For the analysis that follows, our base case considers the building of AI systems with capabilities equivalent to the human workers doing the task - that is, we are modeling what Brynjolfsson called “Turing Trap” automation (Brynjolfsson 2022). By definition, this approach equalizes the benefits provided by the human and the AI system, and thus the key determinant of economic attractiveness becomes the costs of each. In reality, we are only matching on some headline capabilities, so there are likely other remaining advantages and disadvantages to using computer vision in place of human workers. For example, a computer vision system might scale more easily if a factory added an additional shift. We assume that, in the short to medium term, these effects are second order. Since our results are robust to even significant changes in the benefits of the AI systems, these secondary effects would have to be large to meaningfully change our conclusions.[3]

Implicitly, our modeling is making an important assumption about the type of automation that will happen first. In particular, we consider systems that have the same capabilities as the human workers that they are replacing. But why shouldn’t adoption first occur with systems that are less capable than human workers, or those that are more capable? We do not consider less capable systems because we focus on replacing the human doing a task. Since one of our thresholds for human capabilities is the point at which human workers would be fired (e.g. because they misdiagnosed too many x-rays), we judge that less capable systems would do too poor a job to replace human workers doing this task. Less capable systems might still be able to augment the human doing the work, which we consider in other work.

For more capable systems, the question for our analysis isn’t whether systems will be created that have better capabilities than human workers. This is already happening, for example in reading CAT scans (Agarwal et al. 2023). But, insomuch as these systems are economically-attractive to build *but so are systems with capabilities equal to human workers* (as is true in this case), then our modeling approach will correctly identify the extent and timing of automation. The challenge to our approach would occur if building a more

[3] Over the longer term, there could be adjustment strategies that are much more important, such as was seen with the adoption of electricity (David 1990). We do not attempt to model such changes.

capable system become economically-attractive *before* the equal-capabilities system. We argue that this is unlikely to be a common occurrance because improving the capability of AI systems results in an enormously rapid increase in the cost of these systems, as shown by Thompson et al. (2020) and as is consistent with foundational computer science work in this area (Kaplan et al. 2020, Henighan et al. 2020, Mikami et al. 2022).

Since less capable systems are unlikely to be able to substitute for human workers, and more capable ones are likely to become economically-attractive only later, the modeling that will best predict the automation of human labor is the computer vision system with equivalent capabilities. And, because such a system provides similar benefits to the human doing that task (by definition), one can compare the economic attractiveness of these systems by comparing their costs.

1.3.1. The Cost of Computer Vision Systems ($C^M_{i,j}$) To estimate the cost of a computer vision system, we rely on prior work by Thompson et al. (2021, 2022, 2024) to break down and calculate individual cost components. In general, the cost for firm i to fine-tuning and deploying a computer vision system to perform a task, j, can be divided into three categories: fixed costs, performance-dependent costs, and scale-dependent costs. Figure 2 shows an overview of the different components.

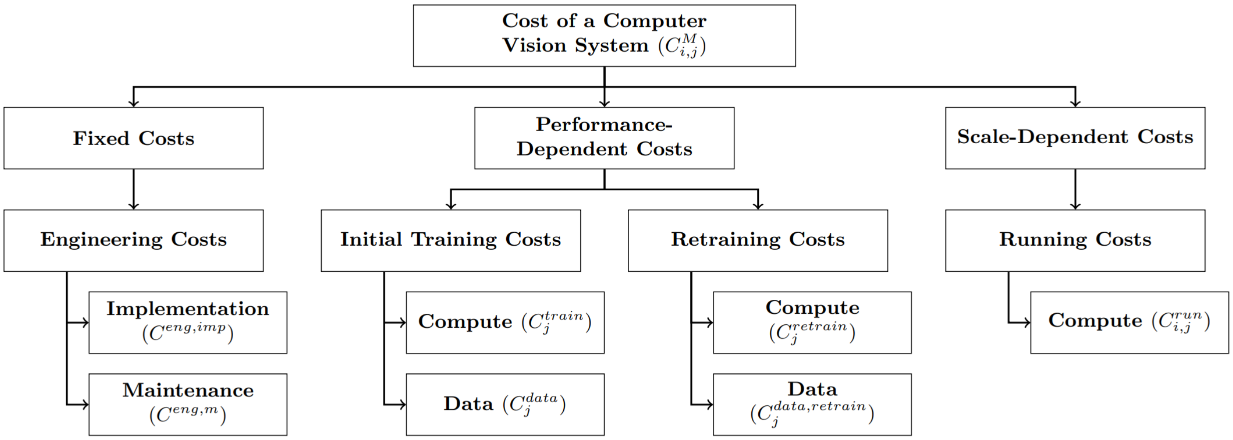

Cost of a Computer Vision System ($C^M_{i,j}$)
Fixed Costs
Performance-Dependent Costs
Scale-Dependent Costs
Engineering Costs
Initial Training Costs
Retraining Costs
Running Costs
Implementation ($C^{eng,imp}$)
Maintenance ($C^{eng,m}$)
Compute (C^{train}_j)
Data (C^{data}_j)
Compute ($C^{retrain}_j$)
Data ($C^{data,retrain}_j$)
Compute ($C^{run}_{i,j}$)

Figure 2 Cost drivers of a AI computer vision system

Fixed costs, or engineering costs,[4] includes implementation costs, $C^{eng,imp}$, and maintenance cost, $C^{eng,m}$. Performance-dependent costs are the costs that vary based on the system requirements, and include the cost of data, C^{data}_j, and the compute cost per training round, C^{train}_j. Finally, scale-dependent costs depend on the amount of work that the system needs to perform, i.e., running costs of the system, $C^{run}_{i,j}$.

[4] Although Thompson et al. (2022) also include *infrastructure cost* as a fixed cost, we ignore this since we assume the use of cloud computing.

To estimate the total cost of replacing human labor for a given task, we calculate the net present value of the cost of a system of a given lifespan. In addition to the initial round of fine-tuning, changes in the real world can lead to a decline in accuracy due to data drift, as explained by Moreno-Torres et al. (2012). To address this accuracy drop, the network must be retrained at regular intervals of K times per year. Denoting the discount rate as d, the yearly rate of decrease in computing costs as m, and the system lifespan as L, the total cost of building, maintaining, and running a computer vision system, $C^M_{i,j}$, to perform task j in firm (or NAICS code) i is as follows:

$$C^M_{i,j} = C^{eng,imp} + C^{data}_j + C^{train}_j + \sum_{t=0}^{L-1} \left(\frac{C^{eng,m} + (C^{data,retrain}_j \times K)}{(1+d)^t} + \frac{C^{run}_{i,j} + (C^{retrain}_j \times K)}{((1+d) \times (1+m))^t} \right)$$

Throughout the analysis, we assume a flat real discount rate of 5% applied across the economy, i.e., $d = 0.05$, corresponding to a conservative expected real stock market return based on historical values (Sullivan 2023). We assume that computing costs will decrease by 22% on an annual basis, i.e., $m = 0.22$ (Hobbhahn and Besiroglu 2022), and that the amount of system fine-tuning during retraining will be comparable to that during training (although our results are robust to significantly different assumptions). Furthermore, we assume that the system is going to be operational for $L = 5$ years, i.e., have a five-year lifespan, based on the custom software depreciation rate published by the U.S. Bureau of Economic Analysis (2003, p.31).

Performance-Dependent Costs

The cost of a computer vision system depends on the level of performance required. We use the regression proposed by Thompson et al. (2024), which models the cost of developing a deep learning system in terms of the required accuracy, the complexity of the task (as measured by entropy (Shannon 1948)), and the quantity and cost of the data needed. Importantly, the cost of such systems grows as a power law with the accuracy required, consistent with the broader deep learning scaling law literature (Thompson et al. 2020, Prato et al. 2021, Mikami et al. 2021, Zhai et al. 2022). Hence, we assume that, for each task j of type b_j (either classification or semantic segmentation), with entropy e_j, there exist a minimum level of accuracy a_j that a human worker must achieve to be deemed fit for the job.

If $f(a_j, e_j, b_j)$ is the number of datapoints required to achieve a_j and e_j for a system of type b_j (0 if classification and 1 if segmentation) according to Thompson et al. (2024), we model the total cost of data, C^{data}_j, as follows:

$$C^{data}_j = f(a_j, e_j, b_j) \times p^{data}_j$$

where p^{data}_j is the cost per datapoint and

$$f(a_j, e_j, b_j) = 10^{\frac{\log_{10}(1-a_j)+0.81-0.61\times\frac{\log_2(1/e_j)}{9.94}-0.62\times b_j}{-0.12}}$$

Unlike previous work that asks AI experts about applicability on tasks that they are unfamiliar with, we choose to gather information on the tasks from domain experts and then use their answers to calculate the AI applicability. In particular, we use an online survey to collect data on the performance needed to complete each task. By using a survey, we get those familiar with doing the task to provide the accuracy required and the cost that would be involved in gathering additional data points. Respondents are recruited from the online crowd-sourcing platform Prolific, which directs them to a survey based on Qualtrics. Respondents are guided to choose the job that they are familiar with and answer questions about all the vision tasks involved in the selected job. They can skip tasks that they are unfamiliar with. We drop answers that fail the attention checks or where the respondents are unsure. We aim to collect at least 5 valid answers for each vision task, and we use the mean of the responses as our measure. In practice, we collected an average of 9 responses per task, with 80% of the tasks having 5+ valid responses. For 33 tasks, where we were unable to find any users familiar with them, so we use the mean value of the other tasks as the inferred value.

We also attempted to gather information on the complexity (entropy) of applications from domain experts, but across multiple pilots surveys were unable to find a way to reliably get respondents to answer. Instead, we manually assess each task description to estimate the entropies. Since this data lacks the domain expert backing, we specifically test robustness to higher or lower entropy values. The details of the survey collection and entropy data are outlined in Appendix A.4.

To calculate the cost of compute, C_j^{train}, we use the number of datapoints implied by $f(a_j, e_j, b_j)$ and the following equation:

$$C_j^{train} = \frac{f(a_j, e_j, b_j) \times 2 \times \text{\# Model Connections} \times 3 \times \text{\# Epochs}}{\text{GPU FLOPs/h}} \times \frac{p^{GPUh}}{U}$$

Here, the numerator of the left factor is the number of floating-point operations (FLOPs) required to train the model, based on research by Sevilla et al. (2022). The denominator is the number of FLOPs a given graphics processing unit (GPU) can perform in 1 hour at peak utilization. The right factor is the price per GPU hour, p^{GPUh}, over the utilization, U, of that GPU.

We assume that computation is done on the cloud and that the cost for a hour of time for a 4 FP-32 TFLOPS GPU costs $p^{GPUh} = \$0.340$, based on AWS pricing.[5] We assume a GPU utilization of $U = 85\%$, which is consistent with the utilization when training large computer vision models, such as ResNet50 (Yeung et al. 2020). Finally, we assume that 50 epochs are used and that the foundation model used has the same parameter size as VGG-19, i.e., 1.44×10^8 parameters (Simonyan and Zisserman 2014), the largest architecture in the “sane list of the most-commonly used model architectures” provided by Thompson et al. (2024).

[5] `eia2.xlarge` pricing in U.S. East region on AWS `https://aws.amazon.com/machine-learning/elastic-inference/pricing/`, Accessed: 2023-04-09

Scale-Dependent Costs

There is a marginal cost of running the model, i.e., making inferences. While running costs of large AI models are frequently cited as very large, we are only interested in knowing the running costs at the minimum viable scale to calculate the economic advantage. To determine this, we consider running costs that are proportional to the amount of human labor being displaced. Machines have an advantage over human labor in terms of speed (Chui et al. 2016, Combemale et al. 2022). For instance, a 4 FP-32 TFLOPS GPU hour is enough to make approximately 50,000,000 inferences using VGG-19 (Simonyan and Zisserman 2014).[6] No human being could match this pace of almost 14,000 inferences per second. We therefore assume that fewer GPU hours than human hours are needed and thus making this calculation with the human hours will be an upper bound on the inference costs. We could make this more precise by estimating a relative factor between the two, but this would not meaningfully change our answer.

Therefore, the yearly running costs, $C^{run}_{i,j}$, for a computer vision system to perform task j within firm i are therefore modeled as follows:

$$C^{run}_{i,j} = \frac{p^{GPUh}}{U} \times 40 \times 50 \times v_j \times n_{i,j}$$

Here, 40 is the number of hours worked per week, and 50 is the number of weeks worked per year, v_j is the fraction of that task in the employees' duties, and $n_{i,j}$ is the number of employees in the firm that perform that task. Like for our training costs, we use a GPU hourly rate of $p^{GPUh} = \$0.34$ and assumed a GPU utilization of $U = 85\%$. Our method for finding v_j and $n_{i,j}$ based on publicly provided data is outlined in Section 1.3.2.

Fixed Costs

The engineering project for a computer vision system involves two phases: implementation ($C^{eng,imp}$) and maintenance ($C^{eng,m}$). We assume that the implementation and maintenance costs are the same for all tasks, reflecting the complexity of the engineering process rather than the complexity of individual tasks. To estimate these costs, we referred to the case study presented by Thompson et al. (2021), which describes a deep learning time series prediction project. The study reports an upfront implementation cost of $C^{eng,imp} = \$1,765,000$ for a 6-month project and a yearly maintenance cost of $C^{eng,m} = \$242,840$. A breakdown of these costs is shown in Appendix A.3. Importantly, these are the full cost of developing and deploying a production-ready system.

Alternative: Bare-Bones Setup

The costs outlined above assume that there is significant engineering work required to develop a computer vision system to replace the task at hand, but this is not always the case. There are instances where costs can be reduced or eliminated completely. For example, a foundation model might already be fit for the

[6] $(4 \times 10^{12} \times 3600)/(2 \times 1.44 \times 10^{8})$, where the nominator is FLOPs per hour and the denominator is FLOPs per inference.

task or sufficiently close to it that fine-tuning can be done with available data and hardware. Therefore, in addition to the setup above, we explore the possibility that the only cost of the system is that of a small engineering team, with an implementation cost of $C^{eng,imp,bb} = \$165,000$ and yearly maintenance cost of $C^{eng,m,bb} = \$122,840$ (see Appendix A.3).

Using the same assumption of discount rate, $d = 0.05$, and system lifespan, $L = 5$, as elsewhere in this paper, the total cost, $C^M_{i,j}$, of implementing this bare-bones computer vision system for a task, j, can, hence, be written as

$$C^M_{i,j} = C^{eng,imp,bb} + \sum_{t=0}^{L-1} \frac{C^{eng,m,bb}}{(1+d)^t}$$

1.3.2. Cost of Human Labor to Firm ($C^H_{i,j}$)

Compared to the cost of computer vision, human labor does not exhibit the same economies of scale. To a large extent, the cost of human labor is the same as the marginal cost of compensation per worker, and such we model it this way. For a given firm i and task j, where j can be accomplished by a computer vision system with a lifespan of L, we define the present value of labor cost to the firm as follows:

$$C^H_{i,j} = \sum_{t=0}^{L-1} \frac{w_{i,j} \times r \times v_j \times n_{i,j}}{(1+d)^t}$$

Here, $w_{i,j}$ is the mean wage of the occupation that performs task j within firm i. r is the wage to total compensation ratio, v_j is the fraction of an occupation's duties that makes up j, $n_{i,j}$ is the number of workers that perform j within firm i (in later sections we consider deployment scales larger than the firm) and d is the discount rate. Like in previous sections, we use a discount rate of $d = 0.05$ and assume that the lifespan of the system is $L = 5$. A diagram of the relationship between these factors can be seen in Figure 3.

To estimate wage costs $w_{o,i}$, we used the 2022 Occupational Employment and Wage Statistics (OEWS) data tables created by the U.S. Bureau of Labor Statistics (2022b), imputing missing employment and wage numbers in the more granular North American Industry Classification System codes (NAICS) (Murphy 1998). We narrow down the data to *U.S. non-farm businesses* by excluding NAICS codes not covered by the 2020 Statistics of U.S. Businesses produced by U.S. Census Bureau (2023).[7]

To convert employee wages to employer costs, we used the wage-to-compensation ratio for civilians of published by the U.S. Bureau of Labor Statistics (2022a, p.4), i.e., $r = 1.449$,[8]. To assign a fraction of an employee's duties, v_j, and thereby implicitly also a fraction of its wages, to a given task and calculate labor cost, we weight each task by its score on the O*NET-Task-Importance scale, following the examples of Brynjolfsson et al. (2018) and Webb (2019).

[7] NAICS codes excluded are Rail Transportation (482); Postal Service (491); Pension, Health, Welfare, and Other Insurance Funds (5251); Trusts, Estates, and Agency Accounts (525920); Offices of Notaries (541120); Private Households (NAICS 8111); and Public Administration (99), as well as public schools.

[8] $1/0.69$

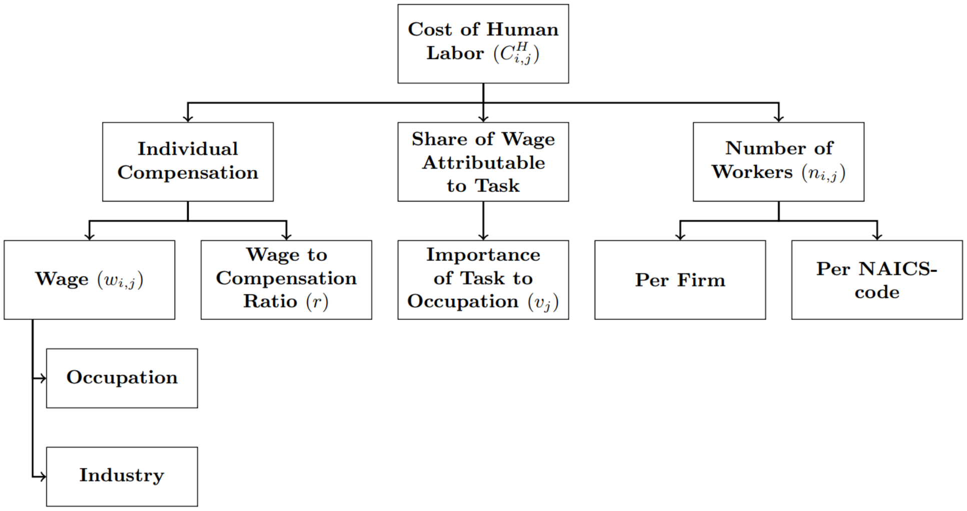

Cost of Human Labor ($C^H_{i,j}$)
Individual Compensation
Share of Wage Attributable to Task
Number of Workers ($n_{i,j}$)
Wage ($w_{i,j}$)
Wage to Compensation Ratio (r)
Importance of Task to Occupation (v_j)
Per Firm
Per NAICS-code
Occupation
Industry

Figure 3 **Cost of human labor performing a task**

To calculate the number of employees of a given occupation per firm, we say that

$$n_{i,j} = \text{size}_i \times \text{occ}_{i,j}$$

where size_i is the size of the firm and $\text{occ}_{i,j}$ is the fraction of that firm's employment base that is of a given occupation. Since $\text{occ}_{i,j}$ is not directly visible in aggregate data, we estimate it using the method outlined in Appendix A.2.2.

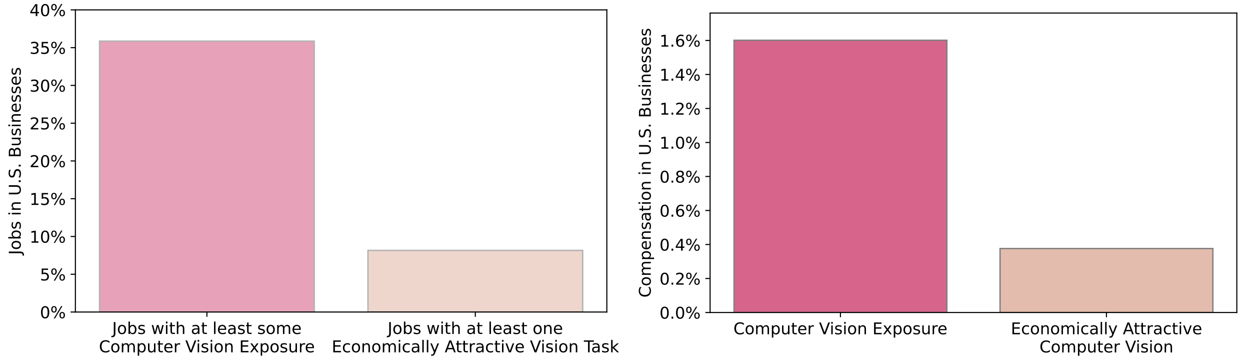
2. Firm-level Results

We first report the results of applying our framework to AI adoption at the firm-level, the natural decision making point for market economies. Later, we consider the results if we allow for consolidation of economic activity via large shifts in market shares or the creation of AI-as-a-service platforms.

2.1. Key Findings

There is a dramatic difference between the vision tasks that are exposed to AI and those that firms would find economically-attractive to automate. While 36% of jobs in U.S. non-farm businesses have at least one task that is exposed to computer vision, only 8% (23% of them) have a least one task that is economically attractive for their firm to automate (Figure 4a).

Since only a small fraction (2%-30%) of any occupation are vision tasks, the more relevant metric is the share of compensation. Figure 4b aggregates the compensation per task, and presents the cost-comparison as a percentage of U.S. labor compensation instead of percentage of jobs. We find that vision tasks comprise 1.6% of U.S. non-farm compensation, where only 0.4% (again 23% of total) is attractive to automate with AI.



(a) Jobs

(b) Worker compensation

Figure 4 Comparison of AI exposure and firm-level economic attractiveness for computer vision.

These results are driven by the cost of AI system deployments. Figure 5 shows the share of vision task compensation that firms could profitably replace by an AI computer vision system of a given cost. Even if a system only costs $1,000, there are tasks that are not economically attractive to replace (tasks in occupations with low wages, many different tasks per occupation, working in small firms). By contrast, for other tasks, there are sufficient labor cost savings to justify investments in systems costing more than $100 million (tasks with high wages, few tasks per occupation, firms with many workers doing the task). We have annotated the graph with the median estimated system cost to illustrate the impact of system costs on automation.[9] As this graph shows, system costs are enormously important to economic attractiveness. This graph also shows an important result for the future dissemination of AI systems: exponential decreases in cost are needed for linear increases in the share of tasks attractive to automate. In Appendix B, we show our key findings by sector.

2.2. Sensitivity Analysis

Because our model of AI automation is end-to-end, it necessarily includes a range of technical and economic assumptions. To ensure that our results are robust to reasonable deviations in these parameters, we consider three types of sensitivity tests (i) changes to cost parameters, (ii) changes to the benefits provided by switching to AI systems, and (iii) a "bare-bones" scenario that tests when many costs are lower.

2.2.1. Sensitivity to cost assumptions

Table 2 provides an overview of the cost parameters that we test sensitivity for. In each case, we compare low and high cost cases to our baseline results. We model the low and high cases for the needed accuracy as polynomials to ensure that the range of possible values remains between 0 and 1. Figure 6a shows the results of varying our assumptions according to Table 2, one variable at a time. Many of the changes to have relatively small impact on the results. Changes to the required accuracy, data costs, or engineering costs are more consequential, although they only increase the

[9] Note: the share of compensation attractive to automate at the median cost is *not* the same as the percentage of compensation in Figure 4b, because the latter comes from calculating benefits and costs across the distribution of firms

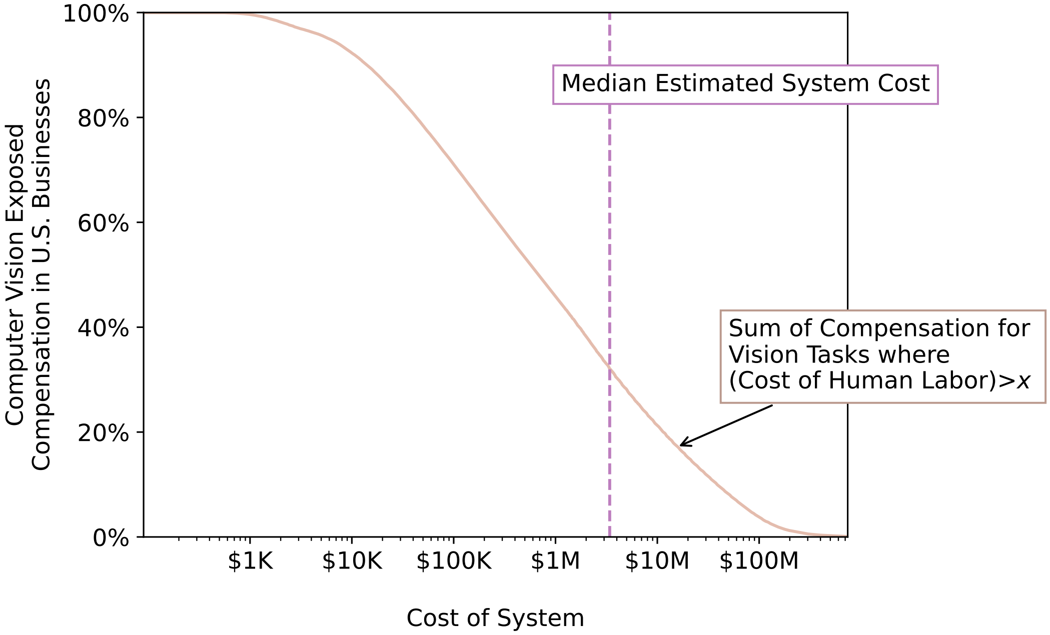


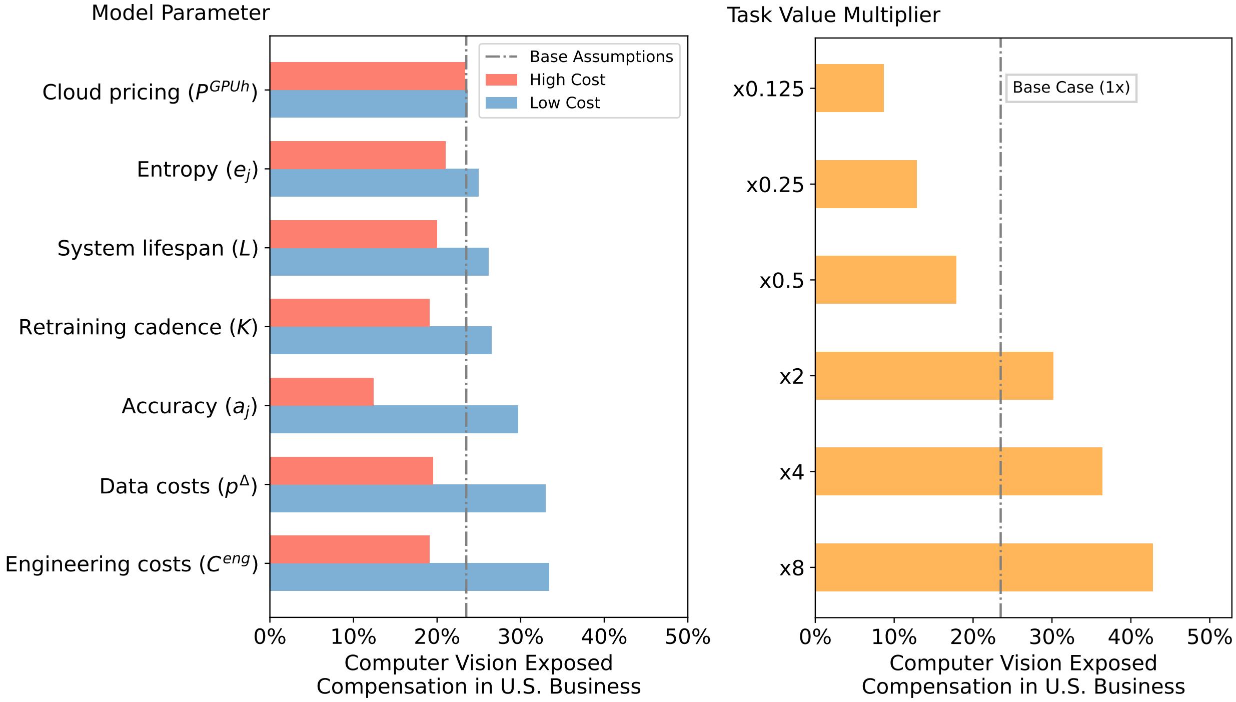
Figure 5 **Relationship between the cost of AI computer vision systems and the share of human vision task compensation that would be attractive for firms to automate.**

		Low Cost	Base	High Cost
C^{eng}	Engineering costs	$0.2 \times C^{eng}$	C^{eng}	$2 \times C^{eng}$
p^{data}	Data costs	$0.5 \times p_j^{data}$	p_j^{data}	$2 \times p_j^{data}$
p^{GPUh}	Cloud pricing	$0.1	$0.34	$1
L	System lifespan	10 years	5 years	2 years
K	Retraining cadence	Never	1 year	2 months
a_j	Accuracy	a_j^2	a_j	$\sqrt{a_j}$
e_j	Entropy	†	e_j	‡

Table 2 **Base case parameter values and sensitivity analysis. †, ‡ see figure 14 in Appendix A.4**

share of automation from 23% to 33% of compensation, at most. In addition to testing robustness to these specific costs, we also test for robustness in extrapolating costs to high levels of accuracy. This check is important because the estimates from (Thompson et al. 2024) are only observed over a limited range of accuracies and then we must extrapolate to higher levels. If we instead extrapolate using curves estimated by others, we do not see any substantial changes to our results.

2.2.2. Sensitivity to benefit assumptions As discussed in section 1.3, our model assumes that building AI computer vision systems with the same task accuracy capabilities as human workers means that they will provide similar benefits. However, this assumption could easily miss other dimensions of performance that would increase or decrease the value of such systems. For example, it would likely be easier to add a third shift at a factory by using a computer vision system for more hours per day than it



(a) Costs (b) Benefits

Figure 6 **Sensitivity of automation results to different cost and benefit assumptions.**

would be to hire an additional worker. On the other hand, if the volume of work being done is variable, human workers might be better equipped to do other tasks during slow times.

To explore the sensitivity of automation to the value created by an AI system, we consider how the adoption decision would change if, for example, it generates $2\times$ the value of human worker. To estimate this empirically, we start from a typical economic assumption: that workers are paid their marginal product (Clark 1908). Figure 6b tests how much the economics of AI adoption change if the AI system delivers some multiple of that human marginal product (as measured by the person's wage). We find that an AI system that doubles the benefit of the human worker would increase the share of compensation that is attractive to automate from 23% to 30%. Our analysis here also echoes the finding from Figure 5, showing that exponential changes in benefits are needed for linear changes in automation share.

Importantly, this analysis should only be thought of as applying to the short-to-medium adoption decisions. Over the longer-term, firms can adapt their production more fundamentally. Estimating the scale of benefits that could be derived from such deeper structural changes to production are beyond the scope of this article.

2.2.3. Bare-bones implementation Thus far, we have considered sensitivity to univariate changes in our parameters. In this analysis, we consider a bare-bones development setup described in Section 1.3.1. This assumes free data, free compute, and only minimal engineering effort is required. Even with those extremely aggressive assumptions, the amount of economically attractive firm-level automation only increases to 49% (0.79% of compensation), as shown in Figure 7. This reflects the extremely fragmented distribution of tasks in the economy, which can make even moderate costs of development prohibitive. This result will be important for our later generalization to AI-as-a-service platforms because it likely also means that automating many of these tasks would likely require an extensive sales/coordination effort which would slow these efforts.

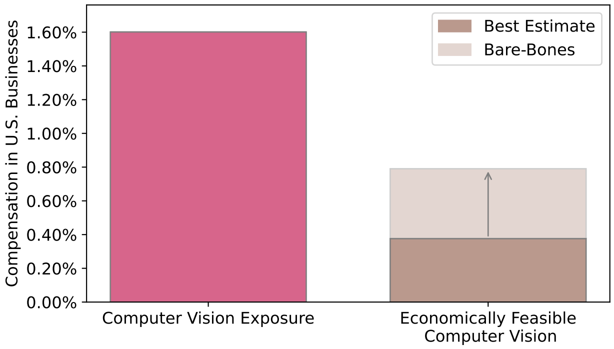


Figure 7 Impact of "bare-bones" cost assumptions on which vision tasks are economically attractive to automate.

2.3. Predictive Power for Historical Labor Outcomes

Over the past decade, many studies have investigated the relationship between technology exposure and the susceptibility of various occupations to labor replacement, see Table 3. The success of these measures in predicting historical labor outcomes was explored by Frank et al. (2023), who measured unemployment risk by calculate the probability of receiving unemployment benefits, using data from each state's unemployment benefits office. To assess the power of our results against these other measures, we recreate Frank et al.'s analysis. Since their data is at the 2-digit SOC code, we aggregate our measure to (i) the share of tasks in that area that are computer vision using our AI exposure variable, and (ii) a composite score for how close tasks are from having an economic advantage. For each task, we calculate the cost difference between AI computer vision completing the task and a human completing it at the same level of proficiency. We then aggregate to the 2-digit SOC level.

In Table 3, we re-state the findings from Frank et al. (2023) in models 1-8, and then compare them to the predictiveness of our computer vision measures for predicting the risk of unemployment across different occupations from 2010-2020 (model 9).

As these results show, our method is as good or better at explaining unemployment risk as any of the other measures, explaining 10.9% of variance as compared to less than 3% for most others and 8.9% and 10.7% respectively for the 3 Webb measurements and Arntz et al. Our high predictive power relative to these other

Study	Variable	Description
Acemoglu and Autor (2010)	Computer Usage	Assess occupations on computer usage.
Acemoglu and Autor (2010)	Routine Cognitive	Assess occupations on routineness and cognitive.
Acemoglu and Autor (2010)	Routine Manual	Assess occupations on manual requirements.
O*NET Education	O*NET %college	The fraction of workers in an occupation holding a bachelor's degree.
Frey and Osborne (2017)	auto	Probability of computerization that combines a subset of occupation skills with subjective assessments of occupation automation levels.
Arntz et al. (2016)	auto2	Job automatibility risk in OECD countries based on an occupation task-based approach.
O*NET Degree of Automation	O*NET Deg.Auto.	Level of automation integrated into the tasks and responsibilities of a particular job or occupation.
Brynjolfsson et al. (2018)	SML	Suitability for machine learning at the task level within various job categories.
Felten et al. (2018)	AI2	Links AI advancements to occupational abilities and aggregates them at the occupation level.
Webb (2019)	% AI Exposure	Compared technology patents with occupation tasks to measure the exposure of occupations to AI.
Webb (2019)	% Robot Exposure	Compared technology patents with occupation tasks to measure the exposure of occupations to robots.
Webb (2019)	% Software Exposure	Compared technology patents with occupation tasks to measure the exposure of occupations to software.
Our study	% Computer Vision Exposure	The percentage of compensation in U.S. businesses that are vision tasks.
Our study	Economic Attractiveness	A composite score for how close vision tasks are to having an economic advantage.

Table 3 Studies estimating AI exposure by occupation.

Dependent Variable: $\log_{10}$ Unemployment Risk by Occupation, Month, & State									
Variable	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6	Model 7	Model 8	**Model 9**
Computer Usage	0.000								
Routine Cognitive	-0.096***								
Routine Manual	0.137***								
O*NET % college		-0.134***							
auto			0.024***						
auto2				0.327***					
O*NET Deg. Auto.					0.152***				
SML						0.082***			
AI2							-0.109***		
% AI Exposure								0.332***	
% Robot Exposure								0.412***	
% Software Exposure								-0.294***	
% Computer Vision Exposure									**0.134*****
Economic Attractiveness									**0.166*****
R2	0.028	0.018	0.001	0.107	0.023	0.007	0.012	0.089	**0.109**
adj. R2	0.028	0.018	0.001	0.107	0.023	0.007	0.012	0.089	**0.109**
$p_{val} < 0.1^{*}$, $p_{val} < 0.01^{**}$, $p_{val} < 0.001^{***}$									

Table 4 Extension of the regression analysis in Frank et al. (2023) between technology exposure and unemployment risk in occupations. Bolded results are our analysis

analyses is particularly noteworthy since we only consider computer vision, whereas they consider many more types of automation. This suggests that extending our approach to other forms of automation could be even more explanatory.

3. Paths to AI Proliferation

Section 2 reveals an important limitation for the proliferation of labor-replacing AI: with today's technology, many of these systems are unattractive for firms to adopt.

The economic attractiveness of AI could be substantially increased in two important ways. The first is deployment scale, finding ways for AI systems to automate more labor per system. The second is development costs, inventing less expensive ways to build AI systems. Here, we explore how these changes would affect the pace of AI deployment. In particular, to test the impact of scale, we estimate the impact of firms getting larger and of AI-as-a-service being used for more tasks. We also explore the hypothetical pace of computer vision proliferation based on different rates of cost decreases.

3.1. Human Labor is Replaced at Larger Scales

In addition to firms custom-building their own systems, there is a possibility of achieving the minimum viable scale by aggregating human labor across firms. While this could hypothetically be done by one firm winning market share from its competitors because of increased efficiency, vision tasks are a small part of firm costs, so an advantage in vision is unlikely to generate large differences in competitive advantage at the firm level in many areas.

We believe that aggregating demand for AI solutions is more likely to happen through AI-as-a-service business models – where, for example, one firm develops the AI system for a task and others that also need the system outsource. Real-world examples of this include a diamond classification tool built by NavTech (Thompson 2021a) and a self-driving platform collaboration by NVIDIA (Thompson 2021b). We define economic advantage for AI-as-a-service as when the aggregate compensation paid to workers performing tasks in a given NAICS code is strictly larger than the cost of developing and running a computer vision system.

We use the same definition of $C^H_{i,j}$ as above, with the only difference being that i is a either the overall U.S. private non-farm economy, a sector, subsector, or industry group. We obtain $n_{i,j}$ directly using the imputed OEWS data described earlier.

We find that the median employee works in a firm where close to none of the vision tasks are cost-effective to automate.[10] Even a firm with 5,000 employees, i.e., larger than 99.9% of firms in the United States, could only cost-effectively automate less than one tenth of their existing vision labor at the current cost structure. This finding helps explain results from McElheran et al. (2023), fewer than 6% of firms use AI-related technologies but that these are disproportionately large firms, representing 18% of employment.

At the extreme end of the firm size spectrum, even a hypothetical firm as big as Walmart lacks the scale to make automating 15% of their vision tasks attractive.[11] As shown in our sensitivity analysis, large

[10] The median employee works in a firm with between 500-749 employees. We gathered this statistic and the other ones mentioned in the paragraph from the SUSB Data Tables released by U.S. Census Bureau (2023)

[11] 1,600,000 U.S. employees according to `https://corporate.walmart.com/about`, Accessed: 2023-05-08

differences in value creation would be needed to substantially change our results. As such, we expect only minor changes in computer vision attractiveness because of changes in firm size distribution or occupational concentration. Indeed, even a perfectly concentrated economy with exactly one occupation per firm would only have a tenfold multiplier on task value compared to our base assumption.

Most firms are, and likely will remain, too small to cost-effectively develop computer vision to replace their existing workers. But, if labor costs for a given task can be aggregated across multiple firms, the economics of automation become much more attractive. If systems could be deployed at the national level – a single system doing all instances of that task across the entire economy – then AI already has an economic advantage for 88% of vision task compensation. Most of the scale needed is already present at the industry group level (i.e. NAICS 4-digit level), rather than the national level, as shown in Figure 8. These results suggest that business models that offer AI-as-a-service will likely be an important driver of AI automation, since they can provide a scale that makes many more tasks attractive to automate.

But, while AI-as-a-service has the potential for much greater automation, there are important technical and economic reasons to doubt whether this industry-level automation can be achieved in the short-to-medium run. The technical challenge is whether systems designed for particular tasks generalize to the industry level. For example, building on the tasks shown in Table 1, O*NET groups the interpretation of radiographic and ultrasound results. But a system for interpreting x-rays for broken bones may not generalize to interpreting ultrasounds for cancer. And while AI systems have indeed increasingly shown an ability to generalize across tasks (Tu et al. 2023), these have also been accompanied by rapidly rising costs (Cottier 2023).

The economic challenge of deploying AI-as-a-service is the cost of coordination: getting many disparate firms onto single platforms is expensive. Whether that coordination comes in the form of salespeople pitching clients, or advertising to get clients to opt themselves in, we would expect only partial adoption of platforms. There could be many reasons for this. For example, Hannan and Freeman (1984) describe how inertia, i.e., resistance to change, is a powerful force within companies, and Walsh (1991) illustrates how worker resistance plays a role in avoiding the automation of existing tasks. It is thus perhaps not surprising that closing rates are as low as 20% in Enterprise IT sales, according to a HubSpot Sales blogger Fuchs (2022). Combining these insights, we find it unlikely that any third-party vendor could capture more than a fraction of the total market. For all these reasons, we expect it will be hard and time consuming to capture the scale that AI-as-a-service could offer, even though we expect many start-ups and venture capitalist to actively pursue these opportunities.

Access to data for fine-tuning is another obstacle to the proliferation of AI-as-a-service at scale. The reason we need fine-tuning is to be able to incorporate knowledge about objects and situations that the system needs to be able to handle gracefully, as discussed further in Section 4.2.1. This data can be most easily be collected within firms where these tasks are carried out. But those firms might have important reasons not to

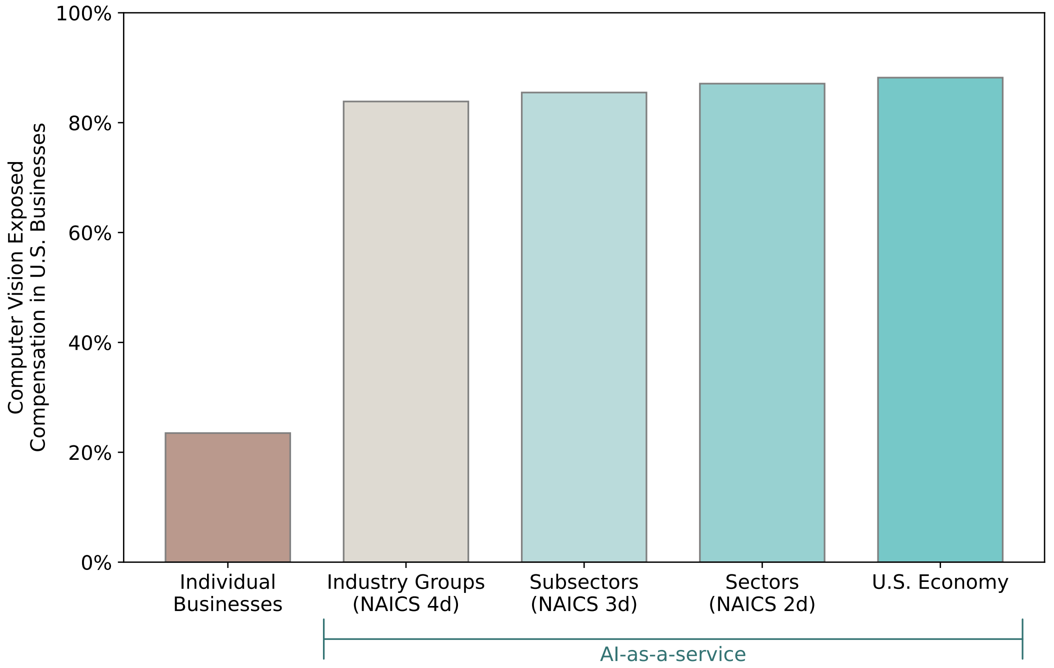


Figure 8 Fraction of vision task compensation economically-attractive to automate if single systems are deployed at this scope.

release this data outside the company. This provides a barrier to the creation of AI-as-a-service offerings by third-parties. Some industry actors have come together to establish data-sharing agreements where a third party could not otherwise collect the required data, such as the NVIDIA Drive collaboration described by Thompson (2021b). Governments and regulatory bodies could also accelerate or hinder platform offerings through rules on data sharing.

To quantify the effects of AI-as-a-service adoption, we perform a simulation exercise that models diffusion as increases in the effective deployment scale that firms get when they make the automation decision. For example, because of coordination, sales, etc. a firm might have market size X in year 0, but that could grow by a factor g to Xg in year 1, Xg^2 in year 2, etc. until the entire market is covered. Firms adopt if and when this additional scale is sufficient to justify their costs. Because there is ambiguity in the adoption order of firms, we assume an *anti-trust ordering*, such that the large firms must start with selling to smaller firms, rather than their largest competitors. Figure 9 shows the results of these simulations based on various growth rates of these platforms.

As these results show, very rapid platformization across all vision tasks ($+20\%$ per year) would result in significant automation within the next decade, whereas a more gradual platformization ($+5\%$ per year) would require decades to get to the full platform automation potential.

Given the challenges in developing AI-as-a-service for *every* vision tasks, including the significant industry restructuring needed to outsource so many tasks, we expect that much of the proliferation of computer

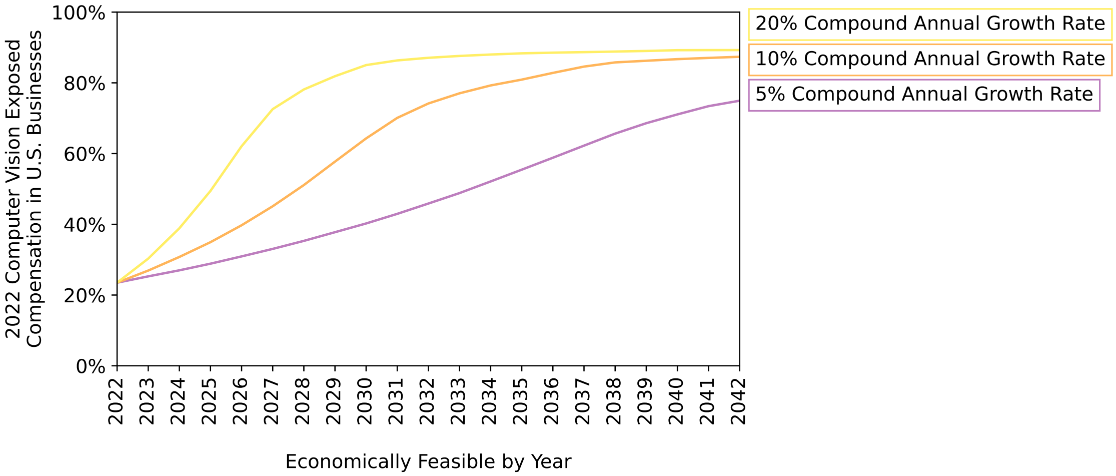


Figure 9 Simulation results: Computer vision automation if AI-as-a-platform offerings allow deployment sizes to grow

vision will come not through platformization but through AI systems becoming less expensive through technological change.

3.2. AI Deployment Becomes Cheaper Through Technological Change

As long as there is a need to customize AI to specific applications (e.g., through fine-tuning), the costs required will affect how it proliferates. Because computer vision, as it stands today, only has an economic advantage in 23% of vision tasks at the firm-level and barriers to AI-as-a-service deployments exist, there will most likely need to be a sharp reduction in cost for computer vision to replace human labor.

Figure 10 simulates what will happen to the amount of economic advantage computer vision will have in vision tasks over time, if we keep other aspects of the model constant but have annual system cost decreases ranging from a 10% to 50%. Even with a 50% annual cost decrease, it will take until 2026 before half of the vision tasks have a machine economic advantage and by 2042 there will still exist tasks that are exposed to computer vision, but where human labor has the advantage. At a 10% annual system cost decrease, computer vision market penetration will still be less than half of exposed task compensation by 2042.

We strongly agree with the proposition that computer vision costs will drop over time, albeit not as predictably as some might suggest. Ford (2015) argues that this will happen rapidly because of Moore's law. More directly relevant, Thompson et al. (2020), Erdil and Besiroglu (2023) measure the annual cost decrease in the cost of computing on GPUs. We use the more recent estimate of a 22% annual cost decrease from Hobbhahn and Besiroglu (2022). As described in section 1, the costs of data and engineering must also be accounted for. These are likely to decrease but not as predictably. Data might become cheaper with increasing digitization, e.g., if the data needed is already collected and labelled for other purposes. Improved developer tools and the spread of machine learning engineering expertise might reduce staffing costs for the engineering team. Foundation models might improve, reducing the need for fine-tuning. There might

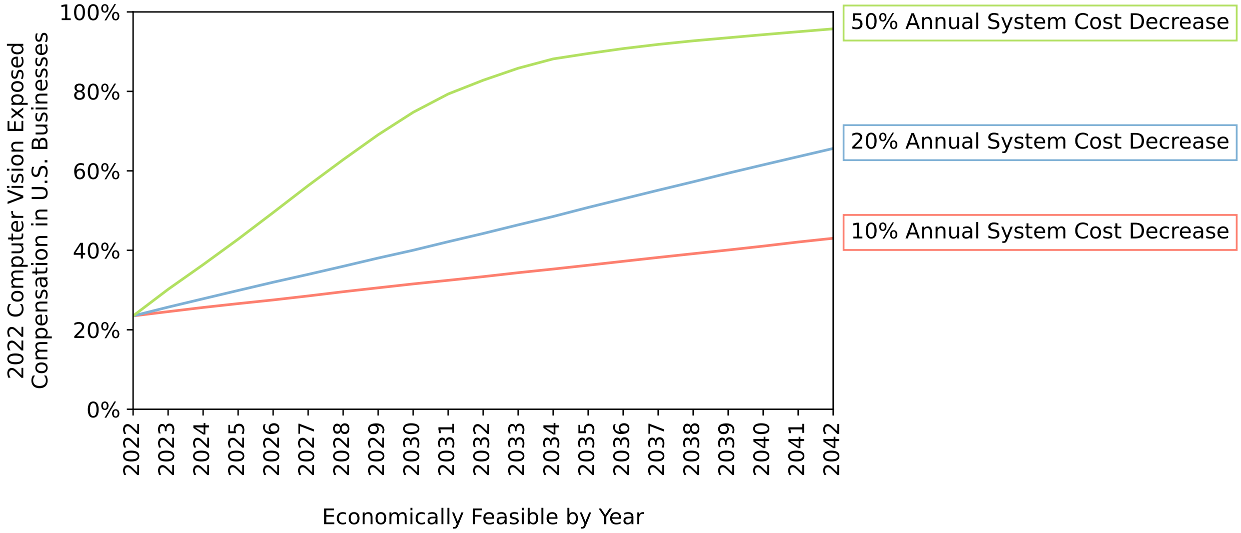


Figure 10 Simulation results: Computer vision automation if AI system costs drop

be a paradigm shift in the way that we fine-tune models that will drastically reduce costs.[12] Importantly, the overall pace of improvement will likely be set by the bottleneck of all of these – meaning that the cost components that improve rapidly become less and less important to the pace of improvement.

4. Discussion

4.1. Impacts on Worker Displacement

One of the most important AI policy discussions is how much worker displacement will occur, and therefore how much retraining, social support, or other intervention might be required. Acemoglu et al. (2022) show correlations between hiring plans in AI and elsewhere. They document a rapid takeoff in AI hiring attempts starting in 2010 and significantly accelerating around 2015-16. They find that AI-exposed establishments reduce their non-AI and overall hiring at the establishment level.[13] Labor effects are uneven, with Grennan and Michaely (2020) showing that security analysts with high exposure to AI are more likely to leave the profession and that departing analysts leave for non-research jobs that require management and social skills.

According to data from the U.S. Census Bureau (2021), on average 11% of jobs in private sector establishments were destroyed annually between 2017 and 2019.[14] However, with substantial job creation, there was still a net gain of on average 1.6% over the period.

Initially, we should expect a significant shock to the labor market as 23% of vision compensation tasks have only recently become attractive to automate, and thus we expect automation attempts to be scaling up.

[12] One example of this would be training methods such as Andrew Ng's Landing AI's Visual Prompting tools (Dey 2023), but their performance in industrial applications is so far unknown to us.

[13] But, interestingly, not at the occupation or industry level.

[14] We purposefully excluded data from the pandemic since it caused extraordinary churn in the labor market.

In subsequent years, once this initial wave of automation occurs, the incremental automation falls to well below this existing job destruction rate.

Figure 11 shows that if there is a 50% annual computer vision cost decrease and if we assume that all vision tasks for which machines gain economic advantage on the firm-level do get automated the same year, the percentage of vision task compensation that is lost every year will be 6-8% in the peak years.

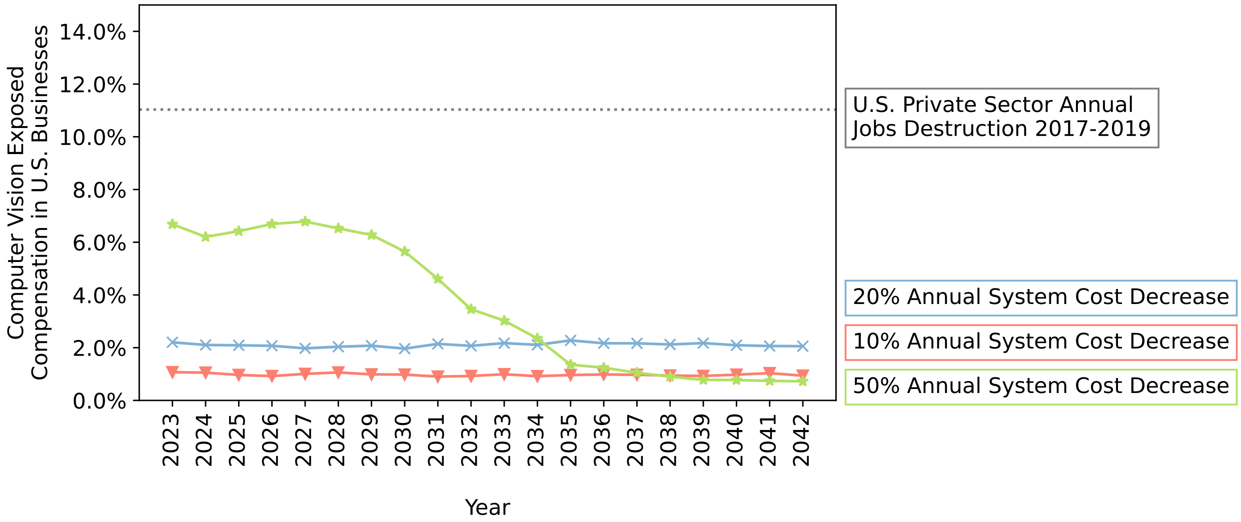


Figure 11 Simulation: Yearly human task destruction as a share of total vision task compensation.

Similarly, Figure 12 shows that task automation from the platformization of AI will only peak above the overall job destruction rate briefly, if at all.

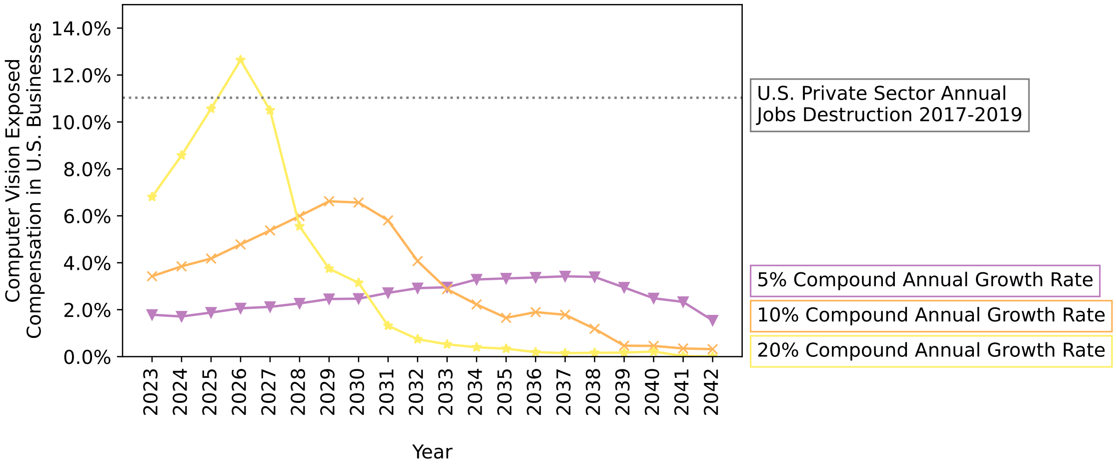


Figure 12 Simulation: Effect of AI-as-a-service platform growth on job destruction

Our results suggest that we should expect the effects of AI automation to be smaller than the existing job automation/destruction effects already seen in the economy. Whether adding AI automation to these existing effects will substantially increase overall job destruction is unclear. We would expect at least some increase, but we also find it likely that a substantial fraction of the AI task automation will happen in areas where traditional automation is occurring. Hence the two types will substitute for each other, at least in part, and the net effect will be less than the sum of each.

4.2. Foundation models and Automation outside Computer Vision

In this section, we discuss how our findings on computer vision fit into the larger landscape of AI, including foundation models and generative AI. In particular, we consider how our modeling of computer vision automation can inform automation using other AI techniques, such as language modeling (e.g., ChatGPT). While there are important differences, we believe the economics of AI as described in this paper will be broadly applicable.

4.2.1. Foundation Models Bommasani et al. (2021) define foundation models as deep learning models that are "trained on broad data at scale and are adaptable to a wide range of downstream tasks".The existence of foundation models in no way undermines the method of this paper; in fact, the underlying economic model for predicting the cost of computer vision used in this paper presupposes that a foundation model exists to fine-tune (Thompson et al. 2024). Our cost estimates, although sometimes substantial, result from specializing pre-existing models to suit a specific task.

Foundation models could impact our results if, as they improve, tasks that workers are doing can increasingly be replaced by the foundation model without fine-tuning. This would reduce the costs of such implementations, making them more economically-attractive. A smaller, but still relevant, effect could occur if better foundation models just reduce the amount of fine-tuning data needed. Hence, improvements in foundation models have a role in reducing costs as described in Section 3.2.

We find it unlikely that foundation models will be able to entirely displace specialized models for two reasons: data availability and slowing progress in foundation models. Because many human vision tasks are not monitored by cameras, and because the data from those tracked by cameras might be sensitive or proprietary, it is likely that data on many tasks will not be shared with the creators of foundation models. This lack of available data will limit the ability of computer vision systems to generalize. For example, consider one of the case studies done in preparation for this paper. In that case, a industrial parts manufacture wanted to use a computer vision system to identify which of their proprietary parts needed to be shipped based on customers sending in pictures of broken ones. The idea that foundation model providers would have enough data to correctly label each of that firm's products (never mind those of all firms) seems unlikely.

A second restriction on foundation models' impact is that progress in improving them may slow. This slowing would arise because the cost of training foundation models grows extremely rapidly, with Thompson et al. (2020) finding that costs for vision models would escalate rapidly into billions of dollars to even improve vision models by small increments. We suspect that these escalations in cost will slow the progress of foundation models and indeed there is already evidence of such slowing (Lohn 2023).

4.2.2. Generative Language Models

We believe that our economic model of AI adoption will still apply to generative language setting, although there are a few notable differences that are important.

First, language, to a much greater extent than vision, seems to generalize across contexts. One potential reason for this could be that the amount of language data available for the training of foundation models is more comprehensive than the image data available; another is that language models are much better than vision models at taking advantage of unlabeled data (LeCunn and Misra 2021). However, much of this shared utility from language will still run into challenges because of specialized knowledge – and so fine-tuning, e.g., to know about specific product information, will still be needed. An important piece of future work will be the need to quantify the fraction of language tasks that do *not* require fine-tuning and thus can easily be automated.

Another important difference better vision and language automation is the cost of data. Firms often have substantial stores of text data, and it is often easier to gather than photos or video. For example, text from customer support chats, email exchanges, and internal knowledge hubs may make language model fine-tuning cheaper than that of image models.

4.3. Limitations

4.3.1. Automation versus Augmentation

In our paper we consider the automation of tasks. This provides only a partial view of AI adoption, since AI can also be deployed augment human labor or to make new products entirely. One survey cites 83% of executives believe AI will augment human labor rather than automate it (IBM Institute for Business Value 2023, p.8). Bessen et al. (2018) found that only 50% of AI startups help customers reduce labor costs, whereas 98% build products to enhance capabilities. While augmentation is not addressed in this article, we are addressing it elsewhere.

Our article similarly does not consider *new* tasks that are created as a consequence of the roll-out of AI, which could easily include tasks for AI itself but could also include other complementary tasks done by human workers.

4.3.2. Cost estimates

For our cost estimates, we rely heavily on the work by Thompson et al. (2024), which predicts the cost of developing a computer vision system ahead of time. However, as it stands, using their model for our purposes has two important limitations. The first is that the input data for their model does not contain any datapoints with accuracies higher than 95%. In other words, for around 40% of vision tasks where higher accuracy is needed, we are extrapolating using their costs function. Since these extrapolations are power laws, they are sensitive to differences in the estimated coefficient. To address this, we consider several alternate scaling laws for these extrapolations (see Appendix B.1). These extrapolations can have more substantial effects on our estimates, underscoring the importance of future research on scaling laws that can provide more precise estimates.

The second limitation is that their analysis uses a limited number of models as sources for transfer learning. An implementation could potentially save resources by starting with a larger foundation model that starts with higher accuracy. One might imagine that a model that was pre-trained closer to the target domain might also be better, although they surprisingly find a limited impact of data distance. We are currently doing work to further test their assumptions.

4.3.3. Survey We use an online survey to identify respondents with expertise in the selected vision tasks and collect information about the required accuracy and the cost per data point for each task. Although this enables us to account for the variation in different occupations and job tasks, this suffers from the limitation that respondents usually lack the knowledge of AI to give precisely the information that is needed for our analysis. For example, accuracy is a rigorously defined concept for a vision classification task which could be difficult for workers from regular occupations to understand. Respondents may only have a vague impression of how often making mistakes is acceptable when performing the task. They might mistakenly include errors that are from other tasks carried out simultaneously with the vision tasks we defined when reporting the error rate. To address these issues, we design survey questions in a way that is easy for respondents to understand and then infer the information we need for our research. This is necessarily a compromise versus the ideal case where we can find experts who know both AI knowledge and work details for each of the 400+ tasks.

4.3.4. Task data and equivalence There are limitations to using O*NET data for our purposes. O*NET was developed as a dictionary of occupations within the United States to serve a wide range of purposes relating to understanding the nature of work. There are important aspects of a potential mismatch between O*NET and our method, and data developed specifically for our purpose would look very different. The first mismatch is that O*NET-Task and DWA combinations, what we simply have called *tasks* above, are unlikely to be perfect units of automation. It might not be the case that the tasks are separable from each other or that a task could be fully automated by computer vision. Additionally, it is unclear whether tasks are similar enough across firms and NAICS codes such that they can be codified and automated at scale. Conversely, there is also the possibility that O*NET agglomerations are too small and that tasks in different occupations could be automated by a single system, e.g., a doctor and a nurse reading an X-ray.

Even if the tasks we derive from O*NET are perfect units of automation, there are limits to assigning value to individual tasks. It is difficult to determine whether the time spent on a task, the skill required for that specific task, or the effort spent on the task determines what compensation an employee receives, especially at the scale of all occupations and tasks in the U.S. economy. In line with this, Autor and Handel (2013) state that there is an unclear mapping between the value of the skills a worker brings to the table and the tasks they perform. Our team settled on using O*NET-Task-Importance. Furthermore, even where tasks are separable, the relationship between automation and worker compensation is complex and non-linear. Combemale et al. (2022) describe how the effect of automation on skill demand, and therefore wages, differs depending on the nature of automation, a complexity that our framework treats as abstract.

There are also limitations in the method for selecting vision tasks. We chose to use a manual way to surface the vision tasks from the set of over 20,000 possible O*NET-Task and DWA combinations. When filtering DWAs in the first step, we may have eliminated possible vision tasks from consideration. However, multiple members of our team validated the list of 414 vision tasks for feasibility, so we believe that we have a low number of false positives.

Our model assumes full substitutability of labor for capital, which although is not unreasonable in many cases, it is in others. We select tasks where computer vision has the potential to replace human labor but we do not know whether technology would automate or augment work. Some use cases will likely strictly reduce the labor needed, like technology for automating sorting process or quality checks at the end of an assembly line. Others would augment human productivity in their other areas of work: one likely example is athletes being aided by computer vision when watching replays of their competitions to improve future performance.

4.3.5. Taxation When comparing the cost of human labor to the cost of developing a system for automation, we did not consider the difference in tax rates between the two. Acemoglu et al. (2020) argue that the tax system favors automation. However, even with a tax on labor of 25–30%, the change to our overall results would be rather small. As we show in Section 2, to have a linear change in the amount of human labor that could be profitably replaced by machines, an exponential change must occur in the underlying costs.

5. Conclusion

In this paper, we develop the first end-to-end model of AI automation that evaluates: the level of proficiency needed for a task, the cost of achieving that proficiency via human workers or AI systems, and the economic decision by firms whether to adopt. Looking at computer vision, where the cost estimates for AI systems are more developed, we find that most systems are cost effective to deploy when single systems can be used across entire sectors or the whole economy. Conversely 77% of vision tasks are *not* economical to automate if a system can only be used at the firm-level. This contrast makes it clear that the cost-effectiveness of AI models will likely play an important role in the proliferation of the technology.

Over time, changes in the cost of AI systems or the scale at which they are deployed have the potential to increase automation. Scale can be gained either by firm getting larger (e.g. through more market share) or through the formation of AI-as-a-service operations. The former effect is unlikely to be meaningful in the short term, because it would require too great a redistribution of firm sizes in the economy. The latter, where AI system development costs could be offset by deploying the system across many firms, would make many more systems economically attractive, but it would likely require industry collaborations or policy initiatives to enable data sharing across companies. If this were to happen, it would also imply a major restructuring of industries, as tasks are separated out from firm operations to third-party providers. The economic advantage of machines will also improve as computer vision deployments become cheaper. But even with rapid decreases in cost of 20% per year, it would still take decades for computer vision tasks to become economically efficient for firms.

The slower diffusion of AI implied by our model mitigates the scale of labor displacement that we should expect. This is certainly true for computer vision, since vision tasks represent only 1.6% of wages. But even if we consider the impact of computer vision just within vision tasks, we find that the rate of job loss is lower than that already experienced in the economy.

Our results emerge because of the need to customize models for specialized tasks. In the same way that database systems are useful broadly, but often require costly customization, AI vision systems will be useful deployed broadly but at substantial cost. We envision that our framework will be useful even beyond computer vision, because many other AI deployments (e.g., in language) will still require customization to adapt them to firm-specific characteristics (e.g. fine-tuning on the specific products being offered). Thus, our results point to a notably different path for AI automation than previously explored in the literature - one where the pace is more in-line with traditional job churn and more amenable to traditional policy interventions and where the cost-effectiveness of systems is crucial to determining their spread.

References

Acemoglu D, Autor D (2010) Skills, tasks and technologies: Implications for employment and earnings. *National Bureau of Economic Research, Inc, NBER Working Papers* 4.

Acemoglu D, Autor DH, Hazell J, Restrepo P (2022) Artificial intelligence and jobs: Evidence from online vacancies. *Journal of Labor Economics* 40(S1):S293–S340.

Acemoglu D, Manera A, Restrepo P (2020) Does the us tax code favor automation? Technical report, National Bureau of Economic Research.

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

Agarwal N, Moehring A, Rajpurkar P, Salz T (2023) Combining human expertise with artificial intelligence: Experimental evidence from radiology. *NBER Working Paper Series* .

Arntz M, Gregory T, Zierahn U (2016) The risk of automation for jobs in oecd countries. Technical report.

Arntz M, Gregory T, Zierahn U (2017) Revisiting the risk of automation. *Economics Letters* 159:157–160.

Autor DH, Handel MJ (2013) Putting tasks to the test: Human capital, job tasks, and wages. *Journal of Labor Economics* 31(S1):S59–S96.

Autor DH, Levy F, Murnane RJ (2003) The skill content of recent technological change: An empirical exploration. *The Quarterly Journal of Economics* 118(4):1279–1333.

Axtell RL (2001) Zipf distribution of u.s. firm sizes. *Science* 293(5536):1818–1820.

Bessen JE, Impink SM, Reichensperger L, Seamans R (2018) The business of AI startups. *Boston Univ. School of Law, Law and Economics Research Paper* (18-28).

Bloom N, Guo A, Lucking B (2020) Outsourcing, occupational and industrial concentration Working paper, American Economics Association Annual Meeting. 2020.

Bommasani R, Hudson DA, Adeli E, Altman R, Arora S, von Arx S, Bernstein MS, Bohg J, Bosselut A, Brunskill E, et al. (2021) On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258* .

Borge NJ (2022) *Deep pockets: The Economics of Deep Learning and the Emergence of New AI Platforms*. Master's thesis, Massachusetts Institute of Technology.

Brynjolfsson E (2022) The turing trap: The promise peril of human-like artificial intelligence.

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

Chui M, Manyika J, Miremadi M (2016) Where machines could replace humans-and where they can't (yet). *McKinsey Quarterly* .

Clark JB (1908) *The distribution of wealth: a theory of wages, interest and profits* (Macmillan).

Combemale C, Ales L, Fuchs ER, Whitefoot KS (2022) How it's made: A general theory of the labor implications of technological change. Working Paper.

Cottier B (2023) Trends in the dollar training cost of machine learning systems. `https://epochai.org/blog/trends-in-the-dollar-training-cost-of-machine-learning-systems`, Accessed: 2024-01-04.

David PA (1990) The dynamo and the computer: An historical perspective on the modern productivity paradox. *The American Economic Review* 80(2):355–361.

Deng J, Dong W, Socher R, Li LJ, Li K, Fei-Fei L (2009) ImageNet: A Large-Scale Hierarchical Image Database. *2009 IEEE Conference on Computer Vision and Pattern Recognition*, 248–255 (IEEE).

Dey V (2023) Andrew Ng's Landing AI makes it easier to create computer vision apps with Visual Prompting. `https://venturebeat.com/ai/andrew-ngs-landing-ai-makes-it-easier-to-create-computer-vision-apps-with-visual-prompting/`, Accessed: 2023-07-17.

Ellingrud K, Sanghvi S, Madgavkar A, Chiu M, White O, Hasebe P (2023) *Generative AI and the Future of Work in America* (McKinsey Global Institute).

Eloundou T, Manning S, Mishkin P, Rock D (2023) GPTs are GPTs: An Early Look at the Labor Market Impact Potential of Large Language Models. *arXiv preprint arXiv:2303.10130* .

Erdil E, Besiroglu T (2023) Algorithmic progress in computer vision.

Felten EW, Raj M, Seamans R (2018) A Method to Link Advances in Artificial Intelligence to Occupational Abilities. volume 108, 54–57 (AEA Papers and Proceedings).

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

Felten EW, Raj M, Seamans R (2023) How will Language Modelers like ChatGPT Affect Occupations and Industries? *arXiv preprint arXiv:2303.01157* .

Fleming M, Clarke W, Das S, Phongthiengtham P, Reddy P (2019) The Future of Work: How New Technologies Are Transforming Tasks. *MIT-IBM Watson AI Lab* .

Ford M (2015) *Rise of the Robots: Technology and the Threat of a Jobless Future* (Basic Books).

Frank M, Ahn YY, Moro E (2023) Ai exposure predicts unemployment risk. *arXiv preprint arXiv:2308.02624* .

Frey CB, Osborne MA (2017) The Future of Employment: How Susceptible are Jobs to Computerisation? *Technological Forecasting and Social Change* 114:254–280.

Fuchs J (2022) How Close Rates are Shifting in 2023 [New Data]. `https://blog.hubspot.com/sales/new-sales-close-rate-industry-benchmarks-how-does-your-close-rate-compare`, Accessed: 2023-04-04.

Grennan J, Michaely R (2020) Artificial intelligence and high-skilled work: Evidence from analysts. *Swiss Finance Institute Research Paper* (20-84).

Hannan MT, Freeman J (1984) Structural Inertia and Organizational Change. *American Sociological Review* 149–164.

Henighan T, Kaplan J, Katz M, Chen M, Hesse C, Jackson J, Jun H, Brown TB, Dhariwal P, Gray S, et al. (2020) Scaling laws for autoregressive generative modeling. *arXiv preprint arXiv:2010.14701* .

Hobbhahn M, Besiroglu T (2022) Trends in GPU price-performance. `https://epochai.org/blog/trends-in-gpu-price-performance`, Accessed: 2023-03-04.

IBM Institute for Business Value (2023) Enterprise Generative AI; State of the market. *IBM Institute for Business Value* `https://www.ibm.com/thought-leadership/institute-business-value/en-us/report/enterprise-generative-ai`, Accessed: 2023-07-18.

Kaplan J, McCandlish S, Henighan T, Brown TB, Chess B, Child R, Gray S, Radford A, Wu J, Amodei D (2020) Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361* .

LeCunn Y, Misra I (2021) Self-supervised Learning: The Dark Matter of Intelligence. *Meta AI Research* `https://ai.facebook.com/blog/self-supervised-learning-the-dark-matter-of-intelligence/`, Accessed: 2023-04-04.

Lohn A (2023) Scaling ai. Technical report, Center for Security and Emerging Technology.

McElheran K, Li JF, Brynjolfsson E, Kroff Z, Dinlersoz E, Foster LS, Zolas N (2023) Ai adoption in america: Who, what, and where. Technical Report 31788, National Bureau of Economic Research.

Meindl B, Frank MR, Mendonça J (2021) Exposure of occupations to technologies of the fourth industrial revolution.

Mikami H, Fukumizu K, Murai S, Suzuki S, Kikuchi Y, Suzuki T, Maeda Si, Hayashi K (2021) A scaling law for synthetic-to-real transfer: How much is your pre-training effective? *arXiv preprint arXiv:2108.11018* .

Mikami H, Fukumizu K, Murai S, Suzuki S, Kikuchi Y, Suzuki T, Maeda Si, Hayashi K (2022) A scaling law for syn2real transfer: How much is your pre-training effective? *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, 477–492 (Springer).

Moreno-Torres JG, Raeder T, Alaiz-Rodríguez R, Chawla NV, Herrera F (2012) A Unifying View on Dataset Shift in Classification. *Pattern Recognition* 45(1):521–530.

Murphy JB (1998) Introducing the North American Industry Classification System. *Monthly Lab. Rev.* 121:43.

Prato G, Guiroy S, Caballero E, Rish I, Chandar S (2021) Scaling laws for the few-shot adaptation of pre-trained image classifiers. *arXiv preprint arXiv:2110.06990* .

Sevilla J, Heim L, Ho A, Hobbhahn M, Besiroglu T, Villalobos P (2022) Estimating training compute of deep learning models. Technical report, Tech. rep, `https://epochai.org/blog/estimating-training-compute`, Accessed: 2023-07-24.

Shannon CE (1948) A mathematical theory of communication. *The Bell System Technical Journal* 27(3):379–423.

Simonyan K, Zisserman A (2014) Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556* .

Sullivan B (2023) Average stock market return. `https://www.forbes.com/advisor/investing/average-stock-market-return/`, Accessed: 2023-05-08.

Svanberg MS (2023) The economic advantage of computer vision over human labor, and its market implications .

Thompson NC (2021a) Navtech: The new platforms being created by deep learning http://www.neil-t.com/wp-content/uploads/2022/01/Navtech-case-study-2021-12-14.pdf, Accessed: 2023-07-18.

Thompson NC (2021b) NVIDIA: Building a Compute and Data Platform for Self-Driving Cars http://www.neil-t.com/wp-content/uploads/2022/01/NVIDIA-case-study-2021-12-14.pdf, Accessed: 2023-04-04.

Thompson NC, Borge NJ, Pande A, Fleming M (2021) Demand Forecasting with A.I.: Building the Business Case.

Thompson NC, Fleming M, Das S, Goehring B, Borge NJ (2022) Where is it Cost Effective To Deploy AI: MIT-IBM Industry Showcase.

Thompson NC, Fleming M, Tang BJ, Pastwa AM, Borge N, Goehring BC, Das S (2024) A Model for Estimating the Economic Costs of Computer Vision Systems that use Deep Learning. *In Proceedings of the 38th Annual AAAI Conference on Artificial Intelligence (Forthcoming)* .

Thompson NC, Greenewald K, Lee K, Manso GF (2020) The computational limits of deep learning. *arXiv preprint arXiv:2007.05558* .

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

Tu T, Azizi S, Driess D, Schaekermann M, Amin M, Chang PC, Carroll A, Lau C, Tanno R, Ktena I, Mustafa B, Chowdhery A, Liu Y, Kornblith S, Fleet D, Mansfield P, Prakash S, Wong R, Virmani S, Semturs C, Mahdavi SS, Green B, Dominowska E, y Arcas BA, Barral J, Webster D, Corrado GS, Matias Y, Singhal K, Florence P, Karthikesalingam A, Natarajan V (2023) Towards generalist biomedical ai.

US Bureau of Economic Analysis (2003) Fixed Assets and Consumer Durable Goods in the United States, 1925–97. https://www.bea.gov/node/24441, Accessed: 2023-07-24.

US Bureau of Labor Statistics (2022a) Employer Costs for Employee Compensation - SEPTEMBER 2022. https://www.bls.gov/news.release/pdf/ecec.pdf, Accessed: 2023-04-04.

US Bureau of Labor Statistics (2022b) Occupational Employment and Wage Statistics. https://www.bls.gov/oes/, Accessed: 2023-07-17.

US Census Bureau (2021) 2021 business dynamics statistics. https://www.census.gov/programs-surveys/bds.html, Accessed: 2023-12-20.

US Census Bureau (2023) 2020 SUSB Annual Data Tables by Establishment Industry. https://www.census.gov/data/tables/2020/econ/susb/2020-susb-annual.html, Accessed: 2023-07-24.

US Department of Labor (2023a) Crosswalk O*NET-SOC 2019 to 2018 SOC. https://www.onetcenter.org/taxonomy/2019/soc.html, Accessed: 2023-04-04.

US Department of Labor (2023b) O*NET. `https://www.onetcenter.org/overview.html`, Accessed: 2023-04-04.

Walsh JP (1991) The social context of technological change: The case of the retail food industry. *Sociological Quarterly* 32(3):447–468.

Webb M (2019) The Impact of Artificial Intelligence on the Labor Market. *Available at SSRN 3482150* .

Yeung G, Borowiec D, Friday A, Harper R, Garraghan P (2020) Towards GPU Utilization Prediction for Cloud Deep Learning. 6–6 (Proceedings of the 12th USENIX Conference on Hot Topics in Cloud Computing).

Zarifhonarvar A (2023) Economics of chatgpt: A labor market view on the occupational impact of artificial intelligence. *Available at SSRN 4350925* .

Zhai X, Kolesnikov A, Houlsby N, Beyer L (2022) Scaling vision transformers. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 12104–12113.

Appendix A: Detailed Methodology

A.1. O*NET-SOC to SOC

O*NET offers a crosswalk from occupations listed in the O*NET-SOC 2019 taxonomy to the 2018 SOC U.S. Department of Labor (2023a). However, due to the vast difference in tasks between O*NET occupations that are mapped to the same SOC code, e.g., 11-1011.00 Chief Executive Officer (CEO) and 11-1011.03 Chief Sustainability Officer (CSO), we exclude occupations with a non ".00" decimal notation for which a corresponding ".00" exists. This has the effect that the compensation for all 11-1011 Chief Executive Officers in OEWS is allocated to CEO tasks instead of the average of the tasks of the CEOs and CSOs. Because of this logic, 149 out of 1016 occupations are discarded. Although this is a large number, we assume that their absence in the 2018 SOC occupations speaks to their relative small size in the economy. For other cases where multiple O*NET-SOC occupations map to the same SOC occupation, i.e., where there is no canonical ".00" code, we aggregate all tasks into the same occupation.

In addition, when comparing the datasets, we found multiple occupations without a corresponding SOC-code in the target OEWS dataset, where we created a manual mapping to an occupation with a similar occupation title (Table 5).

O*NET	O*NET Title	SOC	SOC Title
13-1021	Buyers and Purchasing Agents, Farm Products	13-1020	Buyers and Purchasing Agents
13-1022	Wholesale and Retail Buyers, Except Farm Products	13-1020	Buyers and Purchasing Agents
13-1023	Purchasing Agents, Except Wholesale, Retail, and Farm Products	13-1020	Buyers and Purchasing Agents
13-2022	Appraisers of Personal and Business Property	13-2020	Property Appraisers and Assessors
13-2023	Appraisers and Assessors of Real Estate	13-2020	Property Appraisers and Assessors
21-1011	Substance Abuse and Behavioral Disorder Counselors	21-1018	Substance Abuse, Behavioral Disorder, and Mental Health Counselors
21-1014	Mental Health Counselors	21-1018	Substance Abuse, Behavioral Disorder, and Mental Health Counselors
25-2055	Special Education Teachers, Kindergarten	25-2052	Special Education Teachers, Kindergarten and Elementary School
25-2056	Special Education Teachers, Elementary School	25-2052	Special Education Teachers, Kindergarten and Elementary School
25-9042	Teaching Assistants, Preschool, Elementary, Middle, and Secondary School, Except Special Education	25-9045	Teaching Assistants, Except Postsecondary
25-9043	Teaching Assistants, Special Education	25-9045	Teaching Assistants, Except Postsecondary
25-9049	Teaching Assistants, All Other	25-9045	Teaching Assistants, Except Postsecondary
29-2011	Medical and Clinical Laboratory Technologists	29-2010	Clinical Laboratory Technologists and Technicians
29-2012	Medical and Clinical Laboratory Technicians	29-2010	Clinical Laboratory Technologists and Technicians
31-1121	Home Health Aides	31-1120	Home Health and Personal Care Aides
31-1122	Personal Care Aides	31-1120	Home Health and Personal Care Aides
39-7011	Tour Guides and Escorts	39-7010	Tour and Travel Guides
39-7012	Travel Guides	39-7010	Tour and Travel Guides
45-3031	Fishing and Hunting Workers	–	*Not in OEWS*

47-4091	Segmental Pavers	47-4090	Miscellaneous Construction and Related Workers
47-4099	Construction and Related Workers, All Other	47-4090	Miscellaneous Construction and Related Workers
51-2022	Electrical and Electronic Equipment Assemblers	51-2028	Electrical, Electronic, and Electromechanical Assemblers, Except Coil Winders, Tapers, and Finishers
51-2023	Electromechanical Equipment Assemblers	51-2028	Electrical, Electronic, and Electromechanical Assemblers, Except Coil Winders, Tapers, and Finishers
51-2092	Team Assemblers	51-2090	Miscellaneous Assemblers and Fabricators
51-2099	Assemblers and Fabricators, All Other	51-2090	Miscellaneous Assemblers and Fabricators
51-1042	First-Line Supervisors of Helpers, Laborers, and Material Movers, Hand	53-1047	First-Line Supervisors of Transportation and Material Moving Workers, Except Aircraft Cargo...
51-2043	First-Line Supervisors of Material-Moving Machine and Vehicle Operators	53-1047	First-Line Supervisors of Transportation and Material Moving Workers, Except Aircraft Cargo...
51-1044	First-Line Supervisors of Passenger Attendants	53-1047	First-Line Supervisors of Transportation and Material Moving Workers, Except Aircraft Cargo...
51-1049	First-Line Supervisors of Transportation Workers, All Other	53-1047	First-Line Supervisors of Transportation and Material Moving Workers, Except Aircraft Cargo...

Table 5: Mapping of O*NET-SOC to SOC codes for cases where truncating the decimal point fails.

A.2. Selection of tasks

Naturally, instances occurred where tasks are ambiguous or we lack context or knowledge to determine their computer vision exposure. Many tasks could only partially be replaced by computer vision, or they could be replaced using computer vision and some complementary technology. Other work, including the work by Eloundou et al. (2023), resolves this issue by including multiple categories of tasks. We, instead, used the heuristic that if there is a **use case** for computer vision, and image recognition in particular, it is a vision task. Furthermore, we took advantage of the multiple DWA components that many O*NET-Tasks consist of. Where the computer vision use case only applied to one of the DWAs, we only labeled one of them as exposed. For example, for the O*NET-Task "Diagnose fractures using X-rays," which consists of the DWAs "Diagnose conditions" and "Analyze medical data," we only considered the latter DWA exposed. We assume the healthcare professional will still make the official diagnosis based on the output of the computer vision system. An important exception is that if a task requires prohibitively complex supplementary systems, e.g., "Piloting aircraft" or "Driving ground vehicles," we did not consider it exposed even if it can be done with computer vision.

Other tasks may suggest that vision is an important component of carrying out the work, yet there are reasons why it still would not make sense to replace human labor with machines for those tasks. One such instance is when tasks are not repetitive, i.e., if a task needs new criteria each time it is performed. For example, a costume attendant "check[ing] the appearance of costumes on stage or under lights to determine whether desired effects are being achieved" only does so once per production, requiring new standards every time, so it would not be exposed to computer vision. Another instance is when there are cheaper ways to replace human labor, i.e., when computer vision would be used to read gauges that could instead be directly digitally encoded, or when the task mentions a different technology for carrying out the task, e.g., GIS. Similarly, if the vision part of the task comes for free when a worker carries out all the other components, it is not exposed. As an example, the DWA "Locate suspicious objects or vehicles" is not exposed in the context of "Search prisoners and vehicles and conduct shakedowns of cells for valuables and contraband, such as weapons or drugs," although in other contexts it is, e.g., "Locate suspicious bags pictured in printouts sent from remote monitoring areas, and set these bags aside for inspection."

Finally, when evaluating the exposure of a task, we did not consider the ethics of replacement nor the ethics of the camera surveillance that is implicit in many applications. We asked whether human labor *could* be replaced for a task, not whether it *should* be.

A.2.1. Task-weight scores For O*NET-Tasks that have multiple DWAs associated with them, we distributed the score of the O*NET-Task equally among the DWAs, giving us a weight for our definition of tasks described in Section 1.2. In the hypothetical example of an occupation called "baking assistant" with only the two O*NET-Tasks *knead dough* and *mix dough*, with an Importance of 5 and 3 respectively, *knead dough* would account for 5/8 of compensation and *mix dough* 3/8. If *mix dough* had two DWAs, each of them would account for 3/16 each. In general, for an O*NET-Task Y (associated with a task j), with an O*NET-Task-Importance score Φ_Y, and q_Y associated DWAs, we find the task-weight scores using the following formula:

$$v_j = \frac{\Phi_{Y(j)}/q_{Y(j)}}{\sum_{\text{Q}\in\text{O*NET-Tasks in Occupation}} \Phi_{\text{Q}}}$$

A.2.2. Firm's occupation fraction First, we outline our method for finding $\text{occ}_{i,j}$. Not all firms contain employees of all occupations, and not all occupations exist in all firms. Bloom et al. (2020) estimate that the average number of SOC-5 occupations for a given establishment is 5.5 (since we are using SOC-6 occupations, and the ratio of SOC-5 to SOC-6 is 1.8, we say that there are 10 SOC-6 occupations per establishment). The same paper also found that if an enterprise has multiple establishments, these establishments tend not to have the same occupational makeup. Following the assumption that the more establishments an enterprise has, the more likely the establishments are to contain similar employees,[15] we create a function to estimate the occupation concentration within one firm, occ_i, as follows:

$$\text{occ}_{i,j} = \frac{1}{10 \times \lambda_i^{1/4}}$$

Here, λ_i is the number of establishments within i. The fourth root is chosen to ensure that the number of occupations could make sense for firms with many establishments, e.g., according to this function, a coffee shop chain with 35,000 establishments would employ 136 different occupations. Note that since we do not have any data that distinguishes the concentration of different occupations, we have to assume that all occupations have the same concentration, i.e., $\text{occ}_{i,j} = \text{occ}_i$, although this obviously fails for occupations such as CEOs.

We find the distribution of firm sizes, size_s, the number of establishments per firm, λ_i, for each NAICS code in the 2020 SUSB Annual Data Tables by Establishment Industry (U.S. Census Bureau 2023). We increase the granularity of the histogram using an approach outlined in Appendix A.5, and multiply the employment in each firm by a factor of 1.07 corresponding to the growth in the U.S. non-farm economy between 2020 and 2022 to ensure that the data is comparable to the 2022 wage and task data. Where we find that the cost of human labor for a task is higher for firms in a more detailed NAICS than in the parent NAICS, we aggregate the value of those tasks into the calculations for the parent NAICS code.

[15] As an example, we imagine that for a retail firm, the second or third establishment is likely to be a corporate headquarters, whereas the 600th establishment is likely to be another retail location.

A.3. Fixed Costs

Worker Type	#	Utilization	Monthly/Worker	Monthly Total
IBM Engineers	6	100%	$40,000	$240,000
Client Engineers	4	80%	$16,666	$53,333
Subject Matter Experts	1	5%	$16,666	$833
			Monthly Cost:	$294,166
			Total Cost:	$1,765,000

Table 6 Implementation Team Costs ($C^{eng/imp}$) (Thompson et al. 2021).

Worker Type	#	Utilization	Monthly/Worker	Yearly Total
IBM Engineers	0	0%	$40,000	$0
Client Engineers	4	30%	$16,666	$240,000
Subject Matter Experts	1	1.42%	$16,666	$2,840
			Total Cost:	$242,840

Table 7 Maintenance Team Costs ($C^{eng/main}$) (Thompson et al. 2021).

Worker Type	#	Utilization	Monthly/Worker	Monthly Total
Engineers	2	80%	$16,666	$26,666
Subject Matter Experts	1	5%	$16,666	$833
			Monthly Cost:	$27,500
			Total Cost:	$165,000

Table 8 Bare-Bones Implementation Team Costs ($C^{BB/eng/imp}$).

Worker Type	#	Utilization	Monthly/Worker	Yearly Total
Engineers	2	30%	$16,666	$120,000
Subject Matter Experts	1	1.42%	$16,666	$2,840
			Total Cost:	$122,840

Table 9 Bare-Bones Maintenance Team Costs ($C^{BB/eng/main}$).

A.4. Data Collection for Accuracy, Entropy, Data Cost

A.4.1. Accuracy To determine the minimum required accuracy, a_j, we used the survey platform Prolific to find participants who are familiar with each of the occupations exposed to computer vision and ask them "What's the worst error rate that a worker doing this task could have and still be considered qualified to do it? (i.e. any worse and the work would be given to someone else)". We aimed to collect at least 5 responses per task and use the mean of the response for a_j. In practice, we collected an average of 9 responses per task. 80% of the tasks with more than 5 responses, and 92% of the tasks with at least 3 responses. For the 33 tasks where we could not find any Prolific users familiar with the occupation, we used the mean accuracy of the other tasks, 92%. Figure 13 shows a histogram of the required accuracy for all exposed tasks.

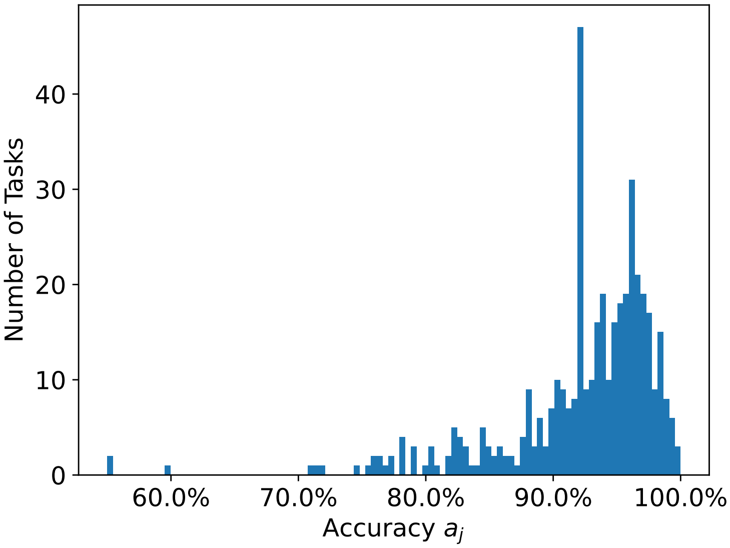


Figure 13 **This figure shows the distribution of minimum required accuracy according to our data.**

A.4.2. Entropy To find the entropy level, e_j, we manually labelled the 414 vision tasks with data on lower and higher bounds for the number of classes a computer vision system would need to recognize in order to perform the task. This can be thought to reflect either the complexity of the specific context, whether a task is deployed on a small or large scope, or the implementation details of the system. We assume that the highest possible number of classes needed is 1000. We then used the geometric mean of these bounds, and for simplicity assume that the classes are of equal size to determine the task entropy. The upper and lower bounds on number of classes, as well as their geometric mean, for each task can be seen in Figure 14. The source entropy of the foundation model is also relevant; we assumed that pre-training was done using the ImageNet dataset (Deng et al. 2009).

A.4.3. Cost per Datapoint We obtain the cost per datapoint by asking the following questions of our Prolific survey respondents:

- "Are there images/video recordings of the *inputs* needed for doing this task?"
  - —No, there is no image/video recording. (E.g., no camera is capturing the worker's view when they check whether cargo is properly loaded)

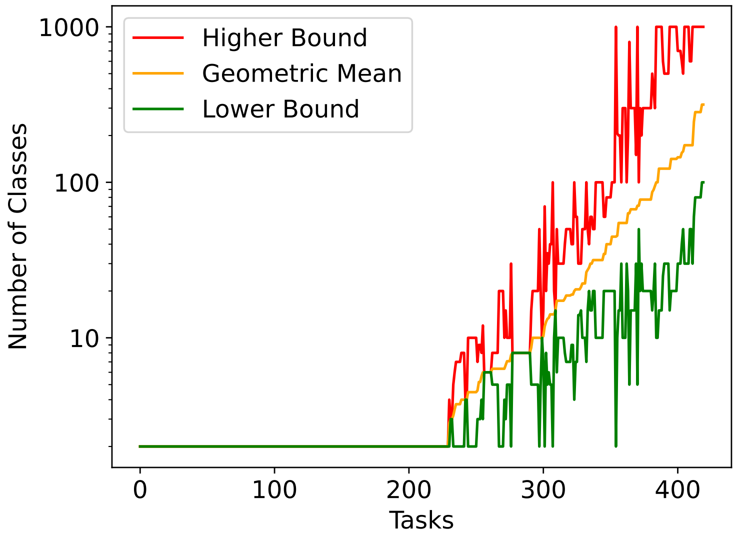


Figure 14 **This figure shows the assumed number of classes for each of our studied vision tasks. We use this to calculate entropy, e_j.**

— Yes, there is image/video recording. (E.g., the X-ray image is stored after the diagnosis.)

- “Are the *outcomes* for the task recorded in a way that they could be connected to the visual inputs?”

— No, it is NOT recorded (E.g., the hairdresser does not record the style of a client’s new haircut.)

— Yes, it is recorded (E.g., the x-ray diagnosis is recorded with the doctor’s marking of the abnormal area.)

For each “No”, we asked the follow-up question:

- “How expensive would it be to collect image/video recordings of the inputs needed for doing this task?” or “How expensive would it be to create a record of the outcome of this task?”

— Very cheap. (E.g., install a camera above the conveyor belt to automatically capture all finished products for observation of any defects.)

— Cheap.

— Neither cheap nor expensive. (E.g., workers must manually take photos of cargo for observation of proper loading.)

— Expensive.

— Very expensive. (E.g., new specialty equipment needed.)

For each input and outcome, we then score these answers from 0 to 5 (from 0 if data already exists to 5 if very expensive). Half of the 5th power of this score is the number of cents per input or outcome (0=$0, 1=$0.005, 2=$0.16, 3=$1.215, 4=$5.12, 5=$15.625), and the cost of the datapoint is the sum of the cost of the input and outcome. If survey respondents disagree, we use the mean of the total data cost.

This gives us a maximum cost per datapoint, i.e., image and label pair, of \$31.25. We believe this is in line with the price of capturing satellite data,[16] X-rays,[17] and some survey data,[18] making it a high yet reasonable price for “very expensive” data collection.

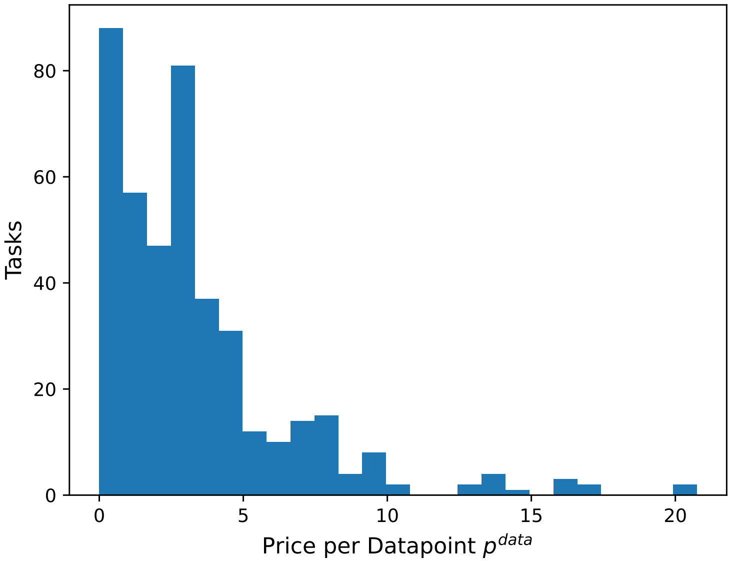


Figure 15 The distribution of our price per datapoint. We derive a distribution of price per datapoint to be between 0 and 21 dollars depending on the task. This reflects the ready availability of data for some tasks compared to the steep cost of acquiring speciality cameras or paying high-income labor to manually and diligently label data for others. We obtain this distribution using online surveys of the availability and relative cost of data per task, and then applying a function to give this a dollar value (see Appendix A.4.3).

[16] As much as \$10 per square km for images alone according to https://www.arlula.com/pricing/, Accessed: 2023-09-04

[17] If an MD earns \$100 per hour and it takes 10 minutes to diagnose and label an X-ray image, labeling would cost \$17, not including the price of the camera, or the wage of nurses and technicians.

[18] In expensive surveys, 25 datapoints might be collected for the cost of \$1000, i.e., \$40 per datapoint.

A.5. Imputing Firm Size Data

We impute the approximate firm size distributions in each NAICS-code of the U.S. economy using the 2020 SUSB Annual Data Tables by Establishment Industry U.S. Census Bureau (2023). The data is published as a histogram binned by total U.S. firm size by employment with a catch-all bin for firms larger than 5,000 employees. Our goals for imputing the data are to (i) approximate a continuous distribution by making the histograms more granular, (ii) add an upper bound to the catchall bin for firms larger than 5,000, and (iii) ensure that the employment and number of firms in smaller NAICS codes logically aggregate into parent NAICS codes.

Our first step is to make the finite bins more granular. We do that by splitting the original bins into 10 bins of the same log of the difference between the upper and lower bound. We then distribute the number of firms in the original bin evenly across the 10 bins. We first assume that the average number of employees per firm in each of the new bins is the average of the upper and lower bound of that bin. Then, we scale that by a factor such that the sum of the employment in each of the 10 bins is equal to the employment in the original bin. We also assume that the average number of establishments per firm is the same across the entire original bin. We repeat this for all finite bins in all NAICS codes.

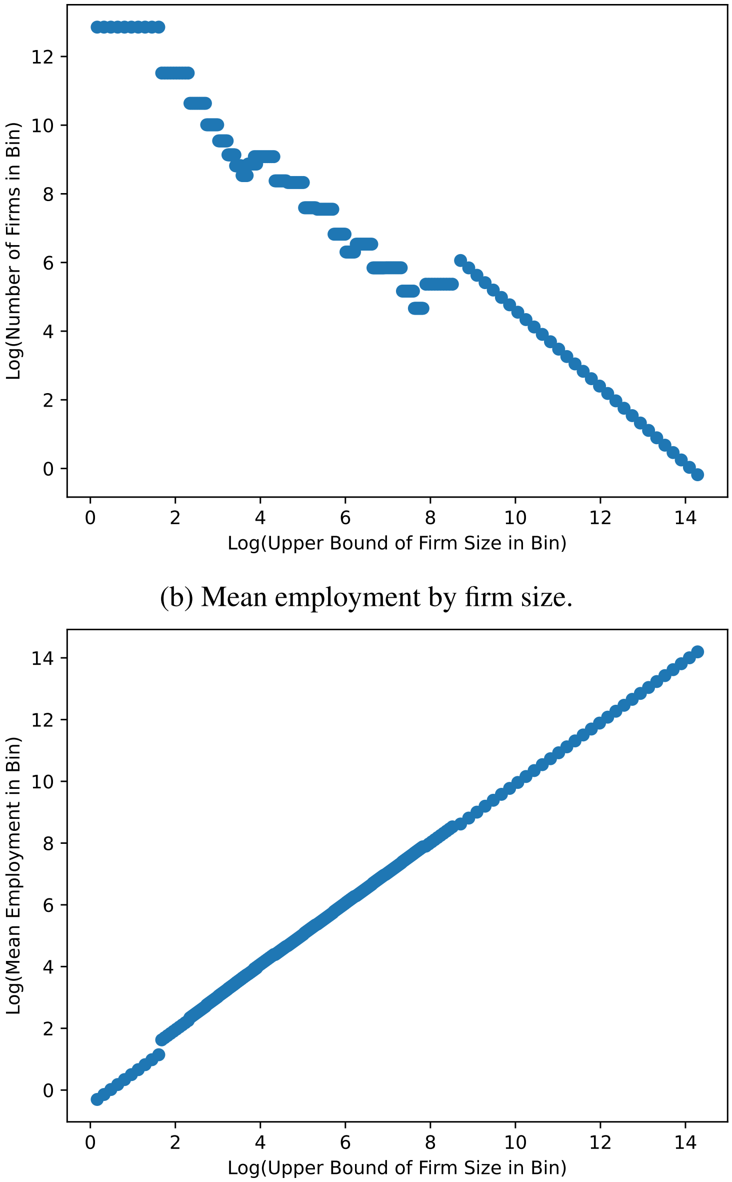
Our second step is to estimate the distribution of firms with a size larger than 5,000 employees. We impose an upper bound on the bin of 1,600,000, i.e., the size of the largest private employer in the United States.[19] We then split the original bin into 30 smaller bins of the same log of the difference between the upper and lower bound. However, instead of assigning firms to bins evenly as we did for the finite bins, we apply a power-law assumption, based on findings that Zipf's law approximates firm size distributions (Axtell 2001).[20] We minimize a loss function of the mean squared error of the employment predictions as well as the firm number predictions, weighing the firm loss by 10000 since we correct the employment in post-processing. We start by doing this for the top NAICS code in the dataset (the U.S. private non-farm economy), without any constraints on the parameters other than a non-negative Y-intercept and a negative slope. For each child NAICS code, we add the constraints that the Y-intercept of the slope must be no larger and the slope no less steep than its parent NAICS code, ensuring the number of firms in each of the imputed bins is always smaller than or equal to the parent NAICS code.

We assume that all firms in each newly imputed bin have the same number of employees. In other words, the data is still in a histogram shape, but the steps are much smaller. In total, there can be up to 250 different bins per NAICS code. Figure 16 shows the imputed values for number of Firms and Mean Employment. We note that the smallest original bin, with firm sizes below 5, has slightly lower mean employment than the middle of the range of the bin. This stems from the large number of one-person firms, but has no effect on our results since no tasks have the minimum viable scale at this firm size.

[19] Largest global U.S. employer: `https://en.wikipedia.org/wiki/List\_of\_largest\_United\_States\%E2\%80\%93based\_employers\_globally`, Accessed: 2023-05-08, 1,600,000 U.S. employees according to `https://corporate.walmart.com/about`, Accessed: 2023-05-08

[20] The reason we do not do this for the finite bins is that that data is not strictly decreasing.

(a) Number of firms by firm size.



(b) Mean employment by firm size.

Figure 16 This figure shows our imputed data for U.S. private non-farm economy.

Appendix B: Additional Results

B.1. Alternative Scaling Laws

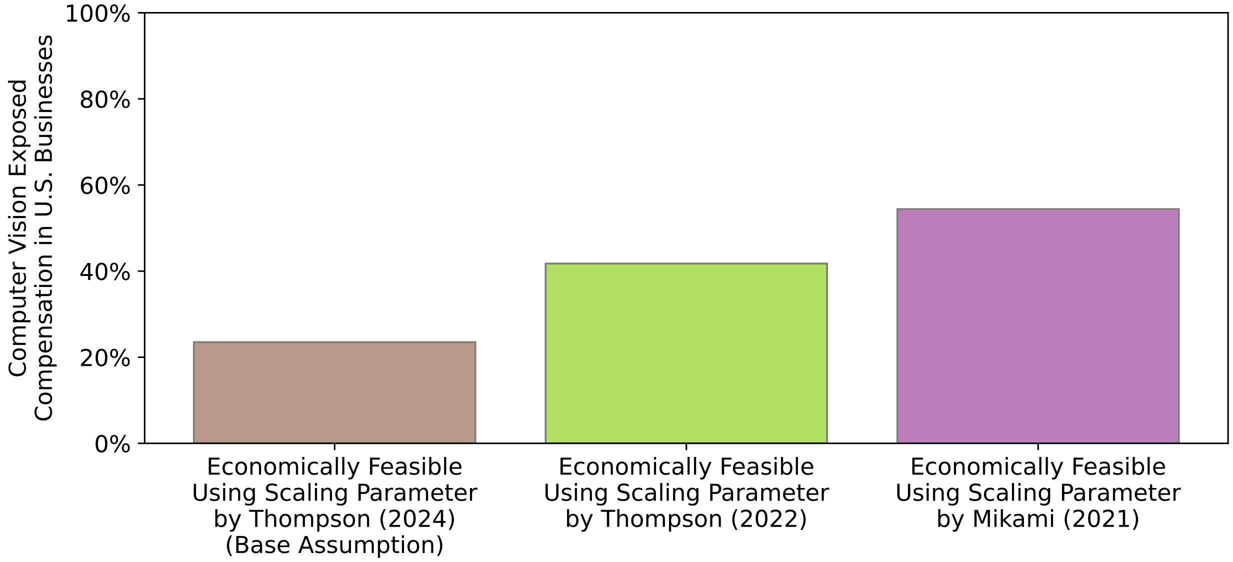


Figure 17 Sensitivity of economically feasible vision automation if model costs scale according to other other scaling laws (Thompson et al. 2020, Mikami et al. 2021).

B.2. Sectors

Economic advantage for computer vision is concentrated in only a handful of sectors. Figure 18 illustrates how each of the sectors contributes to the overall makeup of vision tasks in the U.S. non-farm economy. It also shows the economic advantage within that sector. It is not surprising to find a pattern that sectors with the highest contribution of vision tasks in the economy, in particular *44-45: Retail*, *62: Health Care and Social Services*, and *48-49: Transportation and Warehousing*, also have the most economic advantage for computer vision. Additionally, one could imagine that these sectors (especially Retail and Transportation) have significant returns to scale, containing some of the largest firms in the economy, including Walmart and Amazon.

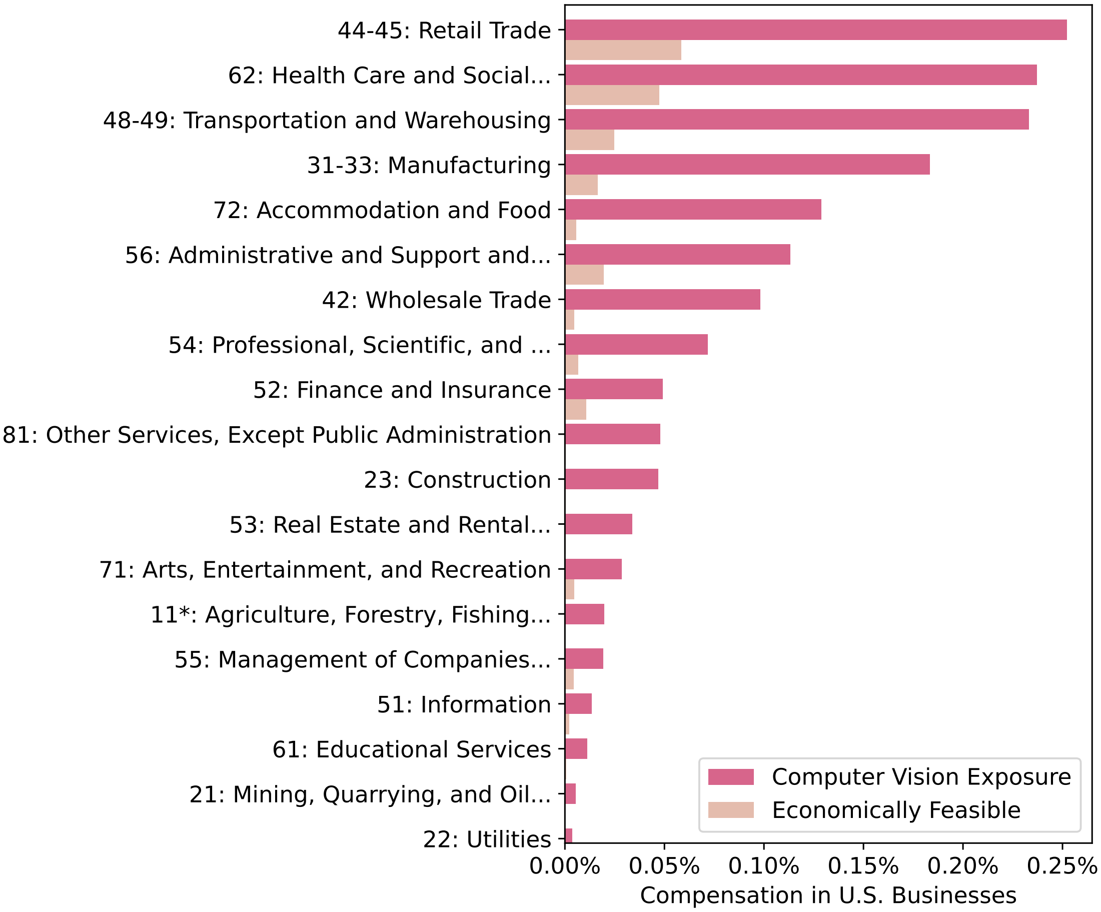


Figure 18 **Computer vision exposure and economic feasibility for compensation within individual sectors (NAICS 2d)..**