We propose **occupation2vec**, a general approach to representing occupations, which can be used in matching, predictive and causal modeling, and other economic areas. In particular, we use it to score occupations on any definable characteristic of interest, say the degree of ‘greenness’. Using more than 17,000 occupation-specific descriptors, we transform each occupation into a high-dimensional vector using natural language processing. Similar, we assign a vector to the target characteristic and estimate the occupational degree of this characteristic as the correlation between the vectors. The main advantages of this approach are its universal applicability and verifiability contrary to existing ad-hoc approaches. We extensively validate our approach on several exercises and then use it to estimate the occupational degree of charisma and emotional intelligence (EQ). We find that occupations that score high on these tend to have higher educational requirements and projected employment growth. Turning to wages, highly charismatic occupations are either found in the lower or upper tail in the wage distribution. This is not found for EQ, where higher levels of EQ are correlated with higher wages.

**Keywords:** Text embeddings; natural language processing; deep learning

**JEL Classification:** J01; J24; C45

*The author is grateful to Elisa Macchi, Andrew Scott, Thyge Enggaard, Mathias Siggard for helpful comments and suggestions, and to the Department of Economics, Massachusetts Institute of Technology (MIT), the Center for Research in Econometric Analysis of Time Series (CREATES), the Dale T. Mortensen Center, Aarhus University, and the Danish Council for Independent Research (Grant 0166-00020B) for research support.

*Department of Economics, MIT, and CREATESE. Email: muhlbach@mit.edu.
I. Introduction

Designing optimal education and labor markets, policy-makers face an increasingly important yet difficult challenge of assessing critical aspects of occupations. For instance, Autor et al. (2003, 2006); Acemoglu and Autor (2011); Autor and Dorn (2013); Autor and Handel (2013) find the occupational degree of analytical, routine, and manual job tasks to be important in explaining changes in the labor market; Goldin (2014) examines the role of various occupational characteristics in explaining the gender wage gap; Deming (2017) consider the degree of social skills in occupations; Frey and Osborne (2017) examine the susceptibility of occupations to computerization; Brynjolfsson et al. (2018) estimate an occupational suitability index for machine learning (ML); Bowen et al. (2018) examine the ‘greenness’ of occupations; and Dingel and Neiman (2020) and Mongey et al. (2020) construct measures of the occupational feasibility of working at home and exposure to social distancing, respectively, to study the economic impact of social distancing in the perspective of COVID-19.

Common to these seminal studies is the need to measure fundamental characteristics of occupations. But all existing studies use case-by-case approaches that are deeply difficult to generalize and validate. Most on them rely on detailed data from the Occupational Information Network (O*NET) that provides measures of hundreds of occupational features. Although the richness of the data certainly allows for novel contributions, an inherent risk of overfitting lurks in the dark. That is, researchers might be tempted to select exactly those features of the data that makes the analysis fit the narrative. The real difficulty is, however, that it may be borderline impossible to validate and verify truly novel measures of occupational characteristics. Another challenge is that every new measure requires a new approach, method, study, etc., which is surely laborious.

We propose occupation2vec, a general approach to representing occupations as high-dimensional vectors, which can then be used in matching, comparative studies, predictive and causal modeling, and other economic areas of interest. Specifically, we demonstrate how the high-dimensional occupation vectors can be used to score occupations on any definable target characteristic, for instance, the occupational degree of ‘greenness’. At its core, our approach essentially transforms every occupation into a high-dimensional vector for which reason we call it an...
occupation2vec framework. The only input needed is an objective and reliable textual definition of the target characteristic. For instance, the U.S. Bureau of Labor Statistics defines green jobs as Definition 1:

**Definition 1** (Green jobs). "(A) Jobs in businesses that produce goods or provide services that benefit the environment or conserve natural resources. (B) Jobs in which workers’ duties involve making their establishment’s production processes more environmentally friendly or use fewer natural resources." (U.S. Bureau of Labor Statistics, 2021)

Using O*NET, we rely on 244 occupational attributes, 873 occupation descriptions, and 16,804 occupation-specific tasks to learn a vector representation of each occupation, leveraging a state-of-the-art natural language processing (NLP) algorithm. Once all occupations are embedded as high-dimensional vectors, we consider a given target characteristic of interest and assign to it a vector in the same vector space as the occupations. This allows us to estimate the occupational degree of the target characteristic as the standardized correlation coefficient between its vector and any occupation vector.

The advantage of this framework is that it is fully data-driven and universal, only requiring the user to provide a reliable and objective definition of the target characteristic. Hence, this is easily extendable to genuinely novel attributes of occupations. In principle, one could extend the framework to default to using an encyclopedia, e.g., Wikipedia, and then the approach would be truly automated. Another advantage is how easy it is to validate the framework in contrast to existing approaches that do not consider any ground truth. We validate occupation2vec in two ways. First, we visually inspect the occupation vectors by compressing them into two-dimensional vectors using principal component analysis (PCA) and t-distributed stochastic neighbor embedding (t-SNE). We plot all occupations on the two dimensions and show that occupations cluster according to major occupational groups and educational requirements. This indicates that even after compressing the information from the high-dimensional vectors into two dimensions, the low-dimensional occupation vectors still capture precisely

---

1In principal, any database containing detailed information on the universe of occupations could be used. To our knowledge, O*NET provides the most comprehensive information on occupations. In addition, O*NET provides both textual descriptions and numerical scores, which is necessary for our validation strategy. This will be thoroughly explained in Section IV.
differences between occupations. Second, we estimate the occupational degree of all of the 244 occupational attributes from O*NET and compare our estimates to the original O*NET scores. We find strong evidence that our estimates coincide with the original O*NET scores both between and within occupations and take this altogether as evidence that occupation2vec produces high-quality occupational vectors that capture essential features of the occupations. These vectors can then be used to match occupations more precisely, act as control covariates in regressions, etc. Specifically, we demonstrate how the vectors can be used to estimate novel characteristics of occupations.

Once the quality of the occupation vectors and the framework in general have been scrutinized and confirmed, we showcase the framework using several applications, which we divide into estimating well-studied versus novel characteristics. First, we revisit the popular task measures (abstract, manual, and routine tasks, respectively) by Autor et al. (2003, 2006); Acemoglu and Autor (2011); Autor and Dorn (2013); Autor and Handel (2013) and the innovative measures of artificial intelligence (AI) by Felten et al. (2018), Brynjolfsson et al. (2018), and Webb (2019). This illustrates how one could have used our framework had the other case-specific approaches not been invented. Note carefully that this is not another validation exercise, except under the assumption that the existing measures represent the ground truth. We feel more comfortable making this assumption for the original O*NET scores.

We document that our estimates of each of the task measures coincide well with the original measures. In particular, we find that performing a high amount of routine tasks is on average associated with lower educational attainment, lower wage, and lower projected employment growth. These occupations are often found within office and administrative support occupations. We find the same pattern for occupations that are characterized by performing manual, whereas the opposite holds for abstract tasks.

We find that our estimates of AI exposure generally behave as a mix of the measures of Felten et al. (2018) and Webb (2019) with no clear relationship to that of Brynjolfsson et al. (2018). The close similarity of Felten et al. (2018) and Webb (2019) in contrast to Brynjolfsson et al. (2018) was first noted by Acemoglu et al. (2020). Occupations that have a high degree of AI have similar profiles as those that score high on abstract tasks. That is, these occupations generally have
higher educational requirements and enjoy higher wages.

Second, we investigate the occupational degree of two novel characteristics, namely charisma and emotional intelligence (EQ). We use definitions from psychological outlets, e.g. the American Psychological Association. Both characteristics are intrinsically interesting and have been found to play a fundamental role in leadership skills and social affability. At the outset, charisma and EQ seem conceptually closely related, and we do find similarities empirically. The occupations that score high on both charisma and EQ are often found within community and social service, educational instruction, and arts, design, entertainment, sports, and media occupations, as well as educational instruction. For charisma, another frequent occupational group is sales. The occupations with high degrees of charisma and EQ also share similar wage profiles. Particularly, our estimates suggest occupations in the upper part of the wage distribution score high on both charisma and EQ. One difference appears in the lower part of the wage distribution, where occupations tend to score high on charisma but not on EQ. The projected employment growth is substantial for these occupations. Another difference that we detect at the aggregate level enters between occupations that require master’s degree versus doctoral degrees. Essentially, these occupations score similarly on charisma, whereas there is a dip in the level of EQ for occupations that require doctorates.

The rest of the paper is organized as follows. Section II presents the data used in occupation2vec, whereas Section III introduces the framework and our NLP method of choice. Section IV validates our framework and Section V considers our four applications on task measures, AI, charisma, and EQ, respectively. Section VI concludes. Appendices to data, method, and applications are found in Appendix A, B, and C–App:eq, respectively. Throughout the paper we will be using the notation $|A|$ for the cardinality of a generic set $A$ and $[a, b] = \{z \in \mathbb{Z} | a \leq z \leq b\}$ denotes the set of integers between $a$ and $b$.

II. Data

In this section, we describe the data used to construct the occupation embeddings in the occupation2vec framework. Starting with the data has two advantages. First, in our experience it is easier to understand the framework with a specific
source of data in mind, especially for readers less versed in NLP. Second, although any detailed occupational database could be used, O*NET provides the most comprehensive information to our knowledge. Specifically, the validation strategy requires the source of occupational information to contain both textual descriptions and numerical scores for numerous features of the entire universe of occupations, which we have not come across elsewhere.²

O*NET keeps track of hundreds of descriptors for each occupation and we use the 873 unique occupations available. Each occupation is described textually by a general description (e.g., writers generally originate and prepare written material, such as scripts, stories, advertisements, and other material) and by detailed tasks (e.g., bakers place dough in pans, molds, or on sheets), as well as measured numerically on a number of attributes (e.g., how important is critical thinking for surgeons). We denote the three sources of occupational information, i.e., descriptions, tasks, and attributes, collectively as occupational descriptors. We highlight this distinction because the majority of the existing approaches is only suited for using the numerical scores of the attributes, whereas our method easily incorporates purely textual data as the tasks and descriptions as well.³ Note that attributes are also defined textually and that tasks are also associated with a weight. Thus, all the occupational descriptors are both expressed as text and associated with an occupation-specific weight. For instance, the definitions of the attributes oral comprehension and depth perception follow from Table 1. These textual definitions naturally apply to all occupations, but each occupation is also assigned a specific score between zero and one for all attributes. For instance, being an astronomer is associated with a score of 0.044 on oral comprehension and of 0.015 on depth perception. Similar, Table 2 shows an example of an occupation description and well as a few tasks for astronomers. In addition, these descriptions are associated with a weight. We will return to the example of astronomers in Section III.

The descriptors represent ten categories; namely, description, tasks, abilities, interests, work values, work styles, skills, knowledge, work activities, and work

²Specifically, the O*NET® Content Model, which can be found here. We use O*NET version 25.3, which can be found here.
³One important exception is Webb (2019), who uses the tasks but not the attributes. To the best of our knowledge, Webb (2019) cannot be generalized to using the attributes as well.
### Table 1: Two examples of O*NET attributes

| Element ID | Element name          | Definition                                                                 |
|------------|-----------------------|-----------------------------------------------------------------------------|
| 1.A.1.a.1  | Oral Comprehension    | The ability to listen to and understand information and ideas presented through spoken words and sentences. |
| 1.A.4.a.6  | Depth Perception      | The ability to judge which of several objects is closer or farther away from you, or to judge the distance between you and an object. |

**Notes:** This table shows the definition of two occupational attributes from the O*NET® Content Model. The examples include the ability oral comprehension (link), and the ability depth perception (link).

### Table 2: Example of occupation description and tasks

| Category | Definition                                                                 | Weight |
|----------|-----------------------------------------------------------------------------|--------|
| Description | Observe, research, and interpret astronomical phenomena to increase basic knowledge or apply such information to practical problems. | 1      |
| Task     | Calculate orbits and determine sizes, shapes, brightness, and motions of different celestial bodies. | 0.053  |
| Task     | Measure radio, infrared, gamma, and x-ray emissions from extraterrestrial sources. | 0.061  |

**Notes:** This table shows the occupation description and two tasks of astronomers from the O*NET® Content Model. The summary report for astronomers can be found here.

context. The total number of occupational descriptors amounts to 17,048, which includes 873 occupation descriptions (one for each occupation), 16,804 unique tasks (some tasks are occupation-specific, while others are performed by a few occupations, leading to on average 20 tasks per occupation), and 244 attributes (common across occupations but with occupation-specific weights). Figure 1 shows the structure of the data.

We are exhaustive in the selection of occupational descriptors in the sense that we include all those for which meaningful textual descriptions are included. The

---

4Technically, the ten categories further represent four broader types of information about work: worker characteristics, worker requirements, occupational requirements, and occupation-specific information. Specifically, worker characteristics include abilities, interests, work values, and work styles; worker requirements include skills and knowledge; occupational requirements include work activities and work context; and occupation-specific information include description and tasks. In this paper, we focus on the ten categories and we do not distinguish further between the four broader types of occupational information.

5For instance, the ability stamina is defined by the ability to exert yourself physically over long periods of time without getting winded or out of breath, whereas the technological skill
Figure 1: Occupational descriptors and subcategories

Notes: This figure shows the occupational descriptors provided by O*NET that can be expressed meaningfully by both text and numeric scores. The 17,048 unique occupational descriptors consist of 873 occupation descriptions, 16,804 detailed tasks, and 244 attributes. Note that a small number of tasks may be core to some occupations while supplemental to others, meaning that the sum of core and supplemental tasks exceed the total number of unique tasks. We divide the descriptors into ten categories of which one category belongs to descriptions, one category belongs to tasks (core and supplemental tasks are grouped as one category), and eight categories belong to attributes.

occupation-specific weight that is associated with each descriptor is constructed based on at least one scale, e.g., importance, level, etc., and the realized value of the scale differs by occupations. The definition of each scale along with the set of categories that each scale applies to are presented in Table A.1 in Appendix A. Each scale has a minimum and maximum value and to enable comparison, all scales have been standardized to ranging from 0 to 1 following O*NET guidelines. In the case of multiple scales for a specific descriptor, we take a uniform average of the standardized scales.

III. Methodology

In this section, we present our general approach to measuring any definable target characteristic of an occupation by quantifying the similarity between two vectors that represent the occupation and the target characteristic, respectively. The occupation vectors are text embeddings generated by our preferred NLP method.
that heavily rely on the plethora of occupation-specific information provided by the O*NET. Similar, the vector of the target characteristic is a text embedding of its definition that can in principle be provided by any objective and reliable source of information, e.g., international standards or dictionaries. Intuitively, our approach transforms every occupation into a high-dimensional vector for which reason we call it an occupation2vec framework. The characteristic of interest, say the ‘greenness’ of occupations, is similarly transformed into a high-dimensional vector, which allow us to compute a measure of similarity of the two vectors; one vector representing a particular occupation and one vector representing the characteristic.

A. An illustration

We start by illustrating the core principle of our framework in Figure 2. Given a vector space produced by an NLP algorithm taking as its input a large corpus of text (e.g., all pages on Wikipedia), the first step is to assign a corresponding vector in the space to each occupational descriptor (all attributes, descriptions, and tasks) from O*NET. Once each source of occupational information is assigned a vector, all kinds of mathematical operations are in principle possible.

The second step aggregates all descriptor embeddings into one occupation-specific embedding that takes into account the occupational loadings on the descriptor expressed by the weights. Within each category, we use weights dictated by O*NET\(^6\) to aggregate the descriptor embeddings to category embeddings. The category embeddings are then further aggregated to occupational embeddings by a uniform average (within occupations). At this point, the occupation2vec framework has generated exactly one high-dimensional vector per occupation.\(^7\)

Having learned one representation of each occupation, the third step quantifies the degree of the target characteristic of interest, for which there exists a reliable definition. Likewise, we use the NLP algorithm to assign a high-dimensional vector to the target characteristic. Once all occupations and the target characteristic are quantified as embeddings, we measure the (semantic) similarity between each occupation and the target characteristic as the correlation coefficient between the

\(^6\)The weights are based on both the frequency, importance, and relevance of the descriptors, respectively.

\(^7\)The high-dimensional space to which the vectors belong belong is produced by the NLP algorithm and the training data, which we explain below.
Step 1: Descriptor-specific Embedding

- Description
  - Description #1
  - Description #2

- Tasks
  - Task #1
  - Task #2

- Abilities
  - Ability #1
  - Ability #2

Step 2: Occupation-specific Aggregation

- Occupational embedding

Figure 2: Illustration of occupation2vec

Notes: This figure shows how the occupation2vec framework represents each occupation as a vector based on the rich occupational information from O*NET. → represents the text embedding by an NLP algorithm, × represents the multiplication between the vector and the occupation-category specific weights, dotted lines represent the weighted average of descriptor embeddings within categories, and dashed lines represent a uniform average (with weights 1/10) across category embeddings.

two high-dimensional vectors. The last step standardizes the similarities to zero mean and unit variance. These standardized similarities then represent the degree of the characteristic in question and they have an ordinal interpretation.

As a specific example, imagine that we are interested in the job-level degree of oral comprehension, which covers the ability to listen to and understand information as well as depth perception, which covers the ability to judge the distance between objects (see Table 1 in Section II for the specific O*NET definitions). We choose these particular characteristics because they are included as ability attributes in O*NET. We compare two occupations from both tails of the attribute distribution, namely arbitrators and roof bolters. For arbitrators, a common task is to prepare written opinions or decisions regarding cases whereas a common task for roof bolters is to drill bolt holes into roofs at specified distances from ribs or adjacent bolts. The estimated degrees of each ability are shown in Table 3. The estimates
Table 3: Examples of estimates of occupational characteristics

|                     | Arbitrators | Roof Bolters |
|---------------------|-------------|--------------|
| Oral Comprehension  | 2.31        | -2.08        |
| Depth Perception    | -1.95       | 2.13         |

Notes: This table shows the description of two occupational descriptors from the O*NET® Content Model. The examples include the ability oral comprehension (link), and the ability depth perception (link).

follow closely our expectations for the chosen mix of occupations and attributes. In Section IV, we will take additional steps to validate our framework.

B. General framework

Index occupations by $i$ for $i = 1, \ldots, n$ and assume access to a collection of $K$ sets, $G = \{G_1, \ldots, G_K\}$. In the O*NET context, $G$ represents the set of ten categories available. The distinction between categories allows us to assign different weights to descriptors within and between categories if desirable. For each $k = 1, \ldots, K$, $G_k$ contains an occupation-specific set, $G_k = \{G_{1,k}, \ldots, G_{i,k}, \ldots, G_{n,k}\}$. We think of $G_{i,k}$ as containing occupation-specific information of the $k$th category, such as textual descriptions and numeric weights. All the occupation-specific sets are of the form

$$G_{i,k} = \{(W_{i,k,1}, T_{i,k,1}), \ldots, (W_{i,k,j}, T_{i,k,j}), \ldots, (W_{i,k,N_{i,k}}, T_{i,k,N_{i,k}})\}$$  (1)

where $j = 1, \ldots, N_{i,k}$ indexes the $j$th occupational descriptor out of $N_{i,k}$ descriptors in category $k$, $W_{i,k,j} \in \mathbb{R}$ represents the weight, and $T_{i,k,j} \in \mathcal{T}$ is the textual definition of descriptor $j$ in category $k$ for occupation $i$ with $\mathcal{T}$ containing the universe of unique descriptor definitions. We scale the weights to sum to one within occupation $i$ and category $k$, that is $\sum_{j=1}^{N_{i,k}} W_{i,k,j} = 1$. Note that within each category $k$, the textual information $T_{i,k,j}$ is allowed but not required to vary across occupations, which also explains why the number of definitions within a category, $N_{i,k}$, varies with $i$. This is the case for both the category Description and Tasks, respectively, because occupation descriptions and tasks differ between occupations. For other categories, e.g., Abilities, $T_{i,k,j} = T_{i',k,j}$ for all $i, i' = 1, \ldots, n$. The number of texts to be embedded is $N = |\mathcal{T}|$, which in our case equals 17,048.
The goal of the NLP algorithm of choice is to turn each $T_{i,k,j}$ into a vector $X_{i,k,j}$ of dimensionality $d$, which is then used to generate the occupational embeddings, $Y_i$ for $i = 1, \ldots, n$, via (2), that is

$$Y_i = \frac{1}{K} \sum_{k=1}^{K} \sum_{j=1}^{N_{i,k}} W_{i,k,j} X_{i,k,j},$$

(2)

where $Y_i \in \mathbb{R}^d$ represents occupation $i$ as a $d$-dimensional vector. Note that $d$ is typically several hundreds and we use $d = 1024$. We introduce the NLP algorithm of choice next.

### B.1. Text embedding

Embedding more than 17,000 textual descriptors into vector spaces, a plethora of methods exist. Common to all is the assumption of the Distributional Hypothesis due to Harris (1954), claiming that words that occur in the same contexts tend to have similar meanings and also popularized by Firth (1957) as “a word is characterized by the company it keeps”. Some of the natural choices to consider are Word2Vec (Mikolov et al., 2013), its document version Doc2Vec/Paragraph2Vec (Le and Mikolov, 2014), GloVe (Pennington et al., 2014), Fasttext (Bojanowski et al., 2016; Joulin et al., 2016a,b), or BERT (Devlin et al., 2018). In this paper, we use an optimized version of BERT called RoBERTa (Liu et al., 2019), and thus our foundation is BERT. Our framework is not limited to specific embedding techniques, and our findings are robust to the choice of NLP methods. We will not review all details of BERT/RoBERTa as this is out of scope. Instead, we provide an intuitive explanation of BERT/RoBERTa and refer readers to Appendix B for more details or Rogers et al. (2021) for an excellent review.

Let $T_m$ be a random piece of text (e.g., a sentence or a paragraph) from $\mathcal{T}$. Each $T_m$ is essentially a sequence of tokens (e.g., words or subwords), $t_{m,1}, \ldots, t_{m,N_m}$. Further, let $X_m \in \mathbb{R}^d$. We assume that there is a function $f$ approximating the relationship between $X_m$ and $Y_m$ as

$$X_m = f(T_m) + \varepsilon_m,$$

(3)

where $\varepsilon_m$ is noise. Conceptually, any text embedding model may be viewed as
a nonparametric estimate $\hat{f}$ of the function $f$. The challenge is that $X_m$ are latent variables that are not observed. The various methods consider different approaches to circumventing this.

**BERT—Deep Bidirectional Transformers** BERT stands for Bidirectional Encoder Representations from Transformers and is a revolutionary technique to learning general-purpose language representations, using a deep bidirectional Transformer encoder based on the original Transformer encoder-decoder from Vaswani et al. (2017). The bidirectional aspect of BERT means that it simultaneously learns a language representation from left to right and from right to left compared to either a left-to-right model or the shallow concatenation of a left-to-right and a right-to-left model. Upon release, BERT achieved state-of-the-art results on a range of NLP tasks. Because the use of Transformers has become common, we will omit an exhaustive description of the attention architecture and refer readers to Vaswani et al. (2017).

In general, BERT consists of two steps, namely a pre-training and a fine-tuning step. First, the pre-training step encodes a lot of semantics and syntactic information about the language by training the model on massive amounts of unlabeled textual data drawn from the web. The objective is to detect generic linguistic patterns in the text, and it is mathematically represented as the sum of two separate objectives, that is, the Masked Language Model (MLM) objective and the Next Sentence Prediction (NSP) objective. The intuition behind MLM is that deep language knowledge is required to be able to filling in the blanks within a sentence with missing words. This in return requires one to read the entire sentence and that is why the bidirectional conditioning is important. The MLM is essentially a cloze procedure (Taylor, 1953) that masks a random sample of tokens by replacing the given token with the $[\text{MASK}]$ token. The MLM objective is then a cross-entropy loss on predicting the masked tokens, where the cross-entropy loss for a specific masked token, $t_0$, is

$$H (t_0) = - \sum_{c=1}^{|V|} y_{t_0,c} \log p_{t_0,c},$$  \hspace{1cm} (4)$$

where $y_{t_0,c} = 1\{c = \text{max}_{c \in V} p_{t_0,c}\}$ is a binary indicator that indicates if the (predicted) class $c$ is correct for token $t_0$, and $p_{t_0,c}$ is the (predicted) probability
that token $t_0$ is of $c$.\footnote{BERT uniformly selects 15\% of the input tokens and perform one of three replacements on the selected tokens; (1) 80\% are replaced with $[\text{MASK}]$, (2) 10\% are left unchanged, and (3) 10\% are replaced with a randomly selected token from $\mathcal{V}$.}

In theory, BERT performs MLM once in the beginning, although in practice, the masks are not the same for every sequence because the data are duplicated for parallel training.

In contrast, NSP is used for understanding the relationship between sentences and it similarly a classification problem. But where MLM considers multiclass classification, however, NSP uses a binary classification loss for whether one sentence is adjacent to another in the original texts. BERT optimizes the sum of the two loss functions using ADAM\footnote{In principal, the pre-training step is sufficient to construct the text embeddings but Reimers and Gurevych (2019) demonstrate that these embeddings are not always meaningful. For this reason, we add the fine-tuning step.} (Kingma and Ba, 2015) and the parameters can be found in Table B.1 in Appendix B.

Second, fine-tuning is used to optimize performance for a specific downstream task, e.g., question answering, text summarization, or sentence embedding, and BERT is thus an example of transfer learning that can use that language representation learnt in the pre-training step across many specific language tasks. Fine-tuning is often performed by adding one or more layers on top of the deep network obtained under pre-training. We follow Reimers and Gurevych (2019) and fine-tune BERT/RoBERTa to yield useful sentence embeddings by training the model on labeled sentence pairs and minimizing the mean-squared-error (MSE) loss between sentence embeddings.\footnote{RoBERTa—A robustly optimized BERT Liu et al. (2019) find that the original implementation of BERT was significantly undertrained and propose RoBERTa, a robustly optimized version of BERT. RoBERTa builds on the same Transformer model as BERT and also uses the MLM objective, where it learns to predict intentionally hidden pieces of texts. The architectural difference is that RoBERTa drops the NSP objective and updates the MLM strategy to be dynamic within training epochs. In addition, RoBERTa is trained on more data and modifies some hyperparameters in the training process compared to BERT, e.g., larger batches and step sizes, leading to 4-5 times longer training times. The advantage is a 2-20\% improvement over BERT on several benchmark tasks for NLP algorithms. In essence, RoBERTa outperforms BERT and could, at the}
time of release, match or exceed every NLP model published after BERT in all individual tasks on several benchmarks, which is the reason we choose RoBERTa as our preferred NLP algorithm. We will omit repeating unnecessarily the details of the implementation of RoBERTa and refer readers to Liu et al. (2019).

IV. Validation

In this section, we validate the occupation2vec framework by performing two exercises. The first exercise is exploratory and visual, where we show that occupation vectors can be well separated by major occupational groups and required educational attainment. The second exercise is more formal and statistical, where we use various metrics to assess the performance of our estimates of the same occupational attributes that O*NET provides. Note that if one is interested in characteristics that appear in the O*NET database (any of the 244 occupational attributes), there is no reason to use our approach. In this case, one should use the O*NET scores directly. Our method is relevant when one is considering novel characteristics that are not included in O*NET. Taking the O*NET scores as the ground truth simply allows for an transparent validation of our approach.

A. Visualizing occupations

Having outlined how we quantify an occupation by transforming it to a high-dimensional vector, we illustrate the occupations in a two-dimensional space and color them by major occupational groups and educational requirements in Figure 3 and 4, respectively. Reducing the 1024 dimensions of the embeddings to only two dimensions, we first use PCA to reduce the number of dimensions to 50. Then, we use t-SNE (van der Maaten and Hinton, 2008) to further reduce the dimensionality to two.

From Figure 3, it follows that the 873 occupations can robustly be separated into the 22 major occupational groups. Note that the different colors have no

---

10Benchmark include the General Language Understanding Evaluation (GLUE), the Stanford Question Answering Dataset (SQuAD), and the ReAding Comprehension from Examinations (RACE) benchmarks

11This is a popular tool to visualize high-dimensional data, which works by converting similarities between data points to joint probabilities and minimizing the Kullback-Leibler divergence between the joint probabilities of the low-dimensional embedding and the high-dimensional data.
Figure 3: Occupation embeddings by major occupation groups

Notes: This figure shows the occupational embeddings in a two-dimensional space, where each dot represents a unique occupation. We apply PCA and \( t \)-SNE to reduce the dimensionality of the embeddings to two dimensions. The occupations are color-coded by the corresponding major occupational group.

Interpretation as groups are sorted alphabetically. For instance, both \textit{Production Occupations} in purple, \textit{Construction and Extraction Occupations} in green, and \textit{Installation, Maintenance, and Repair Occupations} in blue can be found in the lower right corner, and despite the differences in colors, these groups represent a cluster related to physical and manual occupations.

Likewise, Figure 4 indicates that educational requirement is an important separator of occupations. Especially the two extremes, i.e., \textit{No formal educational credential} and \textit{Doctoral or professional degree} occupy two different positions in the two-dimensional space. Note that the colors do have an interpretation is this figure, because we can order educations by level. The two figures support the idea that occupations are quantifiable and separable and this is despite the fact that we compress all the information in the high-dimensional space (\( d = 1024 \)) to only two dimensions. When comparing the figures, note that they would be identical without the coloring because we only perform the dimension reduction once.

B. Comparing estimates of O*NET attributes

Recall that O*NET provides 244 occupational attributes for each of the 873 occupations. Principally, we may estimate the occupational degree of each of these
attributes the same way as we would for any target characteristic of interest. That is, we compare the occupational embeddings to each of the attribute embeddings, thereby quantifying the degree of each attribute. This gives us 244 estimates of occupational attributes for each of the 873 occupations, totaling 213,012 estimates we can readily use to validate our framework as we know the ground truth given as the O*NET numerical scores that are also associated with each attribute alongside the textual definition. Naturally, we cannot do this for the 873 occupation descriptions nor the 16,804 occupation-specific tasks both because these descriptors not comparable across occupations.

**Between and within occupations** First, we assess the ability to correctly estimate an occupational attribute by considering correlations between and within occupations. The reason we care about a high correlation between occupations is that for a specific attribute, we should be able to rank the occupations accordingly (e.g., how do judges compare to carpenters on critical thinking?). To assess between-occupations validity, we use the 213,012 attribute estimates and compute the Pearson correlation coefficient between the O*NET score and our estimate by attribute, yielding 244 correlation estimates.

A high correlation within occupations is equally important because for a
### Table 4: Correlation between O*NET measure and our generalized measure

| (a) Between occupations | (b) Within occupations |
|-------------------------|------------------------|
| $\rho = \rho_0$ | $\rho = \rho_0$ |
| $t$-stat | $t$-stat |
| $p$-value | $p$-value |
| 0.000 | 0.000 |
| 32.826 | 118.870 |
| 0.000 | 0.000 |
| 0.500 | 0.530 |
| 1.461 | 1.473 |
| 0.145 | 0.141 |

**Notes:** This table shows the results from several $t$-tests that examine various null hypotheses of the form $H_0 : \rho = \rho_0$ on the correlation between the O*NET measures and our generalized measures. For each table, the first row covers the null hypothesis that the correlation is not significantly different from zero ($\rho_0 = 0$), whereas the second row highlights the first null hypothesis that leads to a $p$-value above 5% in the testing sequence $\rho_0 = 0.01, \ldots, 0.99$. Likewise for each table, the first column is the postulated null hypothesis ($\rho$), the second column is the $t$-statistics, and the third column is the corresponding $p$-value. Table 4a considers the correlation between occupations, whereas Table 4b considers the correlation within occupations.

Specific occupation, we should be able to rank multiple attributes (e.g., how does static strength compare to dynamic strength for firefighters?). To assess within-occupations validity, we again use the 213,012 attribute estimates and compute the Pearson correlation coefficient but this time by occupation, yielding 873 correlation estimates.

Obtaining two sets of correlations, we determine if each mean correlation differs significantly from zero using classic $t$-tests. In addition, we test if each mean correlation differs significantly from hypothesized values ordered from 1% to 99% and report the first postulated correlation for which we fail to reject the null hypothesis of no significant difference at the 5% level. The results follow from Table 4, where Table 4a shows the between-occupations tests and Table 4b shows the within-occupations tests.

Both types of validity tests strongly reject the null hypothesis of zero correlation. In fact, it is not until reaching a postulated correlation coefficient 50% (i.e., 50% and 53% for between and within occupations, respectively) that we fail to reject the null hypothesis. Note that the $t$-statistic is more than three times higher comparing the null hypothesis of zero correlation from within-occupation to between-occupation. This indicates that our framework is relatively more capable of ranking attributes for a specific occupation. Together, the validation exercises suggest a significant non-zero mean correlation in the range of 50% between the O*NET scores and our estimates.
Table 5: Explainability by regression

| (1) | (2) | (3) | (4) | (5) | (6) | (7) | (8) |
|-----|-----|-----|-----|-----|-----|-----|-----|
| O*NET score on our estimate | 0.5233** | 0.5233** | 0.5233** | 0.5206** | 0.5233** | 0.5206** | 0.5206** |
| (0.002) | (0.002) | (0.002) | (0.002) | (0.002) | (0.002) | (0.002) | (0.002) |
| Job dummies included | False | True | False | False | True | False | True |
| Descriptor dummies included | False | False | True | False | True | False | True |
| Category dummies included | False | False | False | False | True | False | True |
| Adj. R² | 0.274 | 0.274 | 0.274 | 0.274 | 0.385 | 0.385 | 0.385 |
| N | 213012 | 213012 | 213012 | 213012 | 213012 | 213012 | 213012 |

Notes: This table shows the results from regressing the O*NET score on our estimate using eight specifications that each represent a column. Column (1) represents the baseline specification with no control, whereas column (8) represents the full specification with occupation, descriptor, and category dummies included, respectively. Columns (2)-(7) represent the specifications in between the baseline and full specification as detailed in rows 2-4. Coefficients estimates are shown in the first row alongside standard errors in parentheses and t-statistics in brackets. Superscripts ***, **, and * indicate statistical significance based on a (two-sided) t-test using heteroskedasticity-robust standard errors at significance levels 1%, 5%, and 10%, respectively.

Overall explainability As a second validation exercise, we run various regressions of the O*NET scores on our estimates and a constant, using the entire set of 213,012 attribute estimates. The specifications we consider are the eight combinations that may be generated using occupation, attribute, and category dummies, respectively. That is, the baseline specification includes no control dummies, whereas the full specification includes all three sets of dummies. In Table 5, we report the results from all eight regressions. The first row shows the coefficient on our measure, the standard error in parentheses, and the t-statistic in brackets. Rows 2-4 details the specification and the last two rows show the adjusted R² and the number of observations, respectively.

It follows from Table 5 that the coefficient under scrutiny is highly significant and strongly robust across all specifications. In fact, the conditional correlation varies narrowly from 52.1% to 52.3%. This further supports the consistency, reliability, and validity of our proposed occupation2vec framework.

V. Applications

In this section, we estimate the occupational degree of both some well-known characteristics (task measures and AI measures, respectively), and some novel characteristics (charisma and EQ, respectively). We show how our occupation2vec framework ranks occupations according to the characteristics and how our estimates correlate with various occupational statistics, e.g., educational requirements.
and wage. All occupational statistics are sourced from the U.S. Bureau of Labor Statistics.\textsuperscript{12} For the well-known characteristics, we also assess the similarity of our estimates to those already established in the literature.

A. Well-known characteristics

A.1. The occupational degree of task measures

We revisit the seminal work of Autor et al. (2003, 2006); Acemoglu and Autor (2011); Autor and Dorn (2013); Autor and Handel (2013) to study tasks measures, namely abstract, manual, and routine tasks. In Table C.1 in Appendix C, we outline the individual O*NET attributes that comprise the three tasks measures and refer to Acemoglu and Autor (2011) for details on precisely how to construct the tasks measures. We will focus on the degree of routine tasks and refer to Appendix C for similar analyses of the degree of abstract and manual tasks, respectively.

Routine tasks We start by defining routine tasks in Definition 2, relying heavily on Acemoglu and Autor (2011). Recall how we feed this definition into our NLP method, which returns an embedding in the same vector space as the occupations. We then measure the occupational degree of routine tasks by the standardized correlation coefficient between each occupational embedding and the embedding of routine tasks.

**Definition 2** (Routine tasks). "Routine tasks are cognitive or physical tasks, which follow closely prescribed sets of precise rules and well-defined procedures and are executed in a well-controlled environment. These routine tasks are increasingly codified in computer software and performed by machines or sent electronically to foreign worksites to be performed by comparatively low-wage workers. Routine tasks are characteristic of many middle-skilled cognitive and production activities, such as bookkeeping, clerical work and repetitive production tasks."

In Table 6, we list the top and bottom 10 occupations on the degree of routine tasks. The results are as expected and the NLP algorithm even picks up the specific examples in Definition 2 such as clerical work (e.g., rank 2) and bookkeeping.\textsuperscript{12}Table 1.7 Occupational projections, 2019–29, and worker characteristics, 2019, accessible here.
Table 6: Top and bottom 10 occupations on degree of routine tasks

| Top 10 occupations                                      | Score | Bottom 10 occupations                                      | Score |
|--------------------------------------------------------|-------|-----------------------------------------------------------|-------|
| Word Processors and Typists                           | 2.587 | Social Work Teachers, Postsecondary                      | -2.626|
| Office Clerks, General                                | 2.576 | Philosophy and Religion Teachers, Postsecondary          | -2.546|
| Maids and Housekeeping Cleaners                       | 2.211 | Communications Teachers, Postsecondary                  | -2.243|
| Bookkeeping, Accounting, and Auditing Clerks          | 2.209 | Public Relations Specialists                           | -2.140|
| Data Entry Keyers                                     | 2.179 | Law Teachers, Postsecondary                           | -2.084|
| Secretaries and Administrative Assistants, Except Legal, Medical, and Executive | 2.151 | Psychologists                                            | -2.100|
| Cleaning, Washing, and Metal Pickling Equipment Operators and Tenders | 2.144 | Sociology Teachers, Postsecondary                | -2.296|
| Dishwashers                                           | 2.124 | Political Science Teachers, Postsecondary          | -2.292|
| Machine Feeders and Offbearers                        | 2.115 | English Language and Literature Teachers, Postsecondary | -2.275|
| Office Machine Operators, Except Computer             | 2.100 | Psychology Teachers, Postsecondary                     | -2.271|

Notes: This table shows the top and bottom 10 occupations on degree of routine tasks estimated by the occupation2vec framework as the standardized correlation coefficient between occupation embeddings and the embedding of routine tasks.

(e.g., rank 4). All top 10 occupations are truly characterized by accomplishing codifiable tasks that can be specified as a series of instructions to be executed by a machine. In stark contrast, the tasks of psychiatrists are by no means codifiable.

Given the ability of the framework to separate occupations by major occupational groups and educational requirements, we show the degree of routine tasks by these two attributes in Figure 5 and 6, respectively. Figure 5 confirms that occupations that accomplish a higher degree of routine tasks are to be found in office and administrative support occupations, where the opposite can be found in community and social service occupations as well as educational instruction and library occupations. Figure 6 further shows that routine tasks are characteristic of occupations that require low-to-medium educational attainment, where occupations that require bachelor’s, master’s or doctoral degrees rarely involve routine tasks.

We next highlight how our estimates largely coincide with the original measures by Autor and Handel (2013). Essentially, we estimate a smoothed polynomial regression of both measures of routine tasks for each occupation against its rank in the wage distribution as well as its rank in the employment growth distribution. The results are shown in Figure 7a and 7b, respectively. Overall, the two approaches agree nearly perfectly on the relationship between the degree of routine tasks and the wage rank and employment rank. Figure 7a shows how occupations in the upper part of the wage distribution tend to be characterized by smaller amounts of routine tasks, whereas the opposite holds for the lower part of the wage distribution where mainly occupations with many routine tasks reside. An identical pattern holds for employment growth. Those occupations
**Figure 5:** Routine tasks by major occupational group

*Notes:* This figure shows a boxplot of the occupational degree of routine tasks by major occupational groups.

**Figure 6:** Routine tasks by education

*Notes:* This figure shows a boxplot of the occupational degree of routine tasks by educational requirement.

From which we can expect the highest growth are not characterized by performing many routine tasks. A trend that has been evidenced by many (see, e.g., Acemoglu and Autor (2011)).

As a final comparison check, we compute various correlation coefficients between our estimates of abstract, manual, and routine tasks and the original ones. We show the coefficients in Table 7, where we include the Pearson’s correlation as a
Figure 7: Comparing estimates of routine tasks on wage and growth

Notes: This figure shows a smoothed polynomial regression of both standardized measures of routine tasks in each 6-digit SOC occupation against its rank in the wage distribution (7a) and the employment growth distribution (7b).

Table 7: Correlation coefficients between measures

|                   | Pearson’s correlation | Kendall’s rank correlation | Spearman’s rank correlation |
|-------------------|-----------------------|----------------------------|----------------------------|
| Abstract tasks    | 0.2476                | 0.1548                     | 0.2307                     |
| Manual tasks      | 0.8287                | 0.6453                     | 0.8470                     |
| Routine tasks     | 0.6756                | 0.4819                     | 0.6747                     |

Notes: This table shows various correlation coefficient from comparing our occupational estimates of task measures to the ones of Autor et al. (2003). The correlation coefficients include Pearson’s correlation, Kendall’s rank correlation, and Spearman’s rank correlation.

default measure of linear relationships as well as two non-parametric alternatives, namely Kendall’s and Spearman’s rank correlation, respectively. For all three tasks measures (and all three correlation coefficients), the degree of association between the two approaches appears strong and significant, particularly for manual and routine tasks.

A.2. The occupational degree to artificial intelligence

Turning to the occupational degree of AI, we use the Definition 6 to 14 presented in Appendix D. Similar, all figures and tables for this application is deferred to Appendix D. We start by considering the top and bottom 10 occupations on degree of AI from Table D.1. In top 10, we find e.g., data scientists and statisticians,
which is aligned with our expectation. These occupations involve tasks that require various forms of e.g., predictive and prescriptive analytics, which are integral parts of AI. In the bottom 10, we find e.g., manicurists, pedicurists, and stonemasons.

Considering the degree of AI by major occupational groups and educational requirement, we show Figure D.1 and Figure D.2, respectively. As expected, the top major occupational groups computer and mathematical occupations, which coincides with the top occupations from Table D.1. The educational requirements tend to be higher for occupations with high degrees of AI, which makes sense as AI is difficult to skill to acquire, but the association is not monotonically increasing in education.

Next, we compare our estimates of occupational degree of AI to the already established measures, namely those of Felten et al. (2018), Brynjolfsson et al. (2018), and Webb (2019). In Figure D.3, we repeat the previous analysis and estimate a smoothed polynomial of the measure for each occupation against its rank in the wage and employment growth distribution. On both statistics of the labor market, our measure of the occupational degree of AI is to be found in the middle between the measures of Felten et al. (2018) and Webb (2019), meaning that its relationship to occupational wage and employment growth is a mix of those of these well-established measures. Interestingly, our measure of occupational AI tends to agree relatively more with the one of Felten et al. (2018) for the high-wage occupations and agree relatively more with the one of Webb (2019) for the low-wage occupations. In fact, the correlation coefficient between the measures is 0.38 and 0.28, respectively, as shown in Table D.2. Specifically, occupations receiving higher wages tend to have a higher degree of AI compared to the occupations that receive less. The picture is less clear when looking at the relationship between the degree of AI and employment growth, and no significant pattern appears. In contrast, our measure does not align well with the one of Brynjolfsson et al. (2018), which seems to be picking up something different. The fact that the three established measures are picking up different aspects of AI is already found by Acemoglu et al. (2020), who also find that the measures of Felten et al. (2018) and Webb (2019) tend to coincide and be different than the one of Brynjolfsson et al. (2018).
B. Novel characteristics

The second part of this section considers two novel characteristics of occupations, namely the degree of charisma and EQ. To economize on space, all figures and tables for the EQ application are deferred to Appendix F.

B.1. The occupational extent of charisma

The first systematic treatment of charisma is due to Weber (1947) and popularized by Dow (1969). Charisma is at the very center of effective leadership (Avolio and Yammarino, 2013) and is continuously being studied (see, e.g., von Hippel et al. (2016)). It has been recognized as one of the main explanations for why certain leaders, i.e., charismatic leaders, develop emotional attachment with followers and other leaders that eventually foster performance that surpass expectations. Likewise, charisma is associated with being considerate, inspirational, visionary and intellectually stimulating (for a review, see, e.g., Spencer (1973) or Turner (2003)). One definition of charisma follows from Definition 3, which is part of the definitions we use to construct the charisma embedding.\(^{13}\)

**Definition 3** (Charisma). "The special quality of personality that enables an individual to attract and gain the confidence of large numbers of people. It is exemplified in outstanding political, social, and religious leaders." (American Psychological Association, 2021a)

We begin our analysis by highlighting the top and bottom 10 occupations on the degree of charisma in Table 8. Actors, fundraisers, and sociologists are both among the top occupations that associate with charisma, because both occupations require both the ability to catch the attention of the audience and the knowledge of group behavior and dynamics, societal trends and influences. Similar for teachers (particularly within special education and psychology) who must be knowledgeable of human behavior and be capable of assessing individual differences in personality and interests. An interesting yet obvious finding is that clergy are among the top occupations on charisma as these formal religious leaders must be able to speak to the public and gain followers by presiding over specific rituals and teaching their religion’s doctrines. Generally, occupations

\(^{13}\)The other definitions that are included in the construction of the embedding of charisma follow in Appendix E.
Table 8: Top and bottom 10 occupations on degree of charisma

| Top 10 occupations | Score | Bottom 10 occupations | Score |
|--------------------|-------|------------------------|-------|
| Public Relations Specialists | 2.982 | Woodworking Machine Setters, Operators, and Tenders, Except Sawing | -2.079 |
| Actors | 2.905 | Professional and Related Occupations | -2.037 |
| Telemarketers | 2.472 | Soybean Planters, Except Farm Operators | -1.902 |
| Clergy | 2.180 | Industrial Machinery Mechanics | -1.809 |
| Fruitickers | 2.205 | Farms, Nurseries, and Greenhouse Operators | -1.876 |
| Door-to-Door Sales Workers, News and Street Vendors, and Related Workers | 2.271 | Retail Salespersons, Cashiers, and Gasoline Attendants | -1.872 |
| Directors, Religious Activities and Education | 2.204 | Retail Salespersons, Cashiers, and Gasoline Attendants | -1.846 |
| Broadcast Announcers and Radio Disc Jockeys | 2.249 | Textile Weaving, Twisting, and Dyeing Occupations | -1.835 |
| Social workers | 2.249 | Farmworkers and Laborers, Crop, Nursery, and Greenhouse | -1.798 |

Notes: This table shows the top and bottom 10 occupations on degree of charisma estimated by the occupation2vec framework.

Figure 8: Charisma by major occupational group

Notes: This figure shows a boxplot of the occupational degree of charisma by major occupational groups.

within arts, design, entertainment, sports, media, education and sales score high on charisma. On the other side of the spectrum, we mostly find occupations within construction, installation, maintenance, and production. These occupations have less of a need to be able to inspire and motivate large numbers of people. This finding is confirmed by Figure 8, showing the occupational degree of charisma by major occupational groups.

Interestingly, an almost monotonic pattern appears between education requirements and degree of charisma as shown in Figure 9. The extent of charisma is comparable for occupations that generally do not require higher degrees, whereas occupations with educational requirements of bachelor’s degrees or higher are increasingly associated with charisma.

Last, we examine how charisma is connected to wage and employment growth in Figure 10 by the means of a smoothed polynomial regression. Figure 10a
Figure 9: Charisma by education

Notes: This figure shows a boxplot of the occupational degree of charisma by educational requirement.

indicates that there is a non-monotonic relationship between charisma and wage. Essentially, in the lowest part of the wage distribution, occupations tend to have high degrees of charisma. These occupations are, e.g., actors (presumably non-Hollywood actors), teachers, and social workers. In contrast, the occupations with the lowest score on charisma are on average found between the first and second quartile of the wage distribution, whereas occupations in the upper part of the wage distribution can be characterized by high degrees of charisma. The latter would be the occupations that require advanced degrees and leadership capabilities. Considering employment growth in Figure 10b, the relationship is almost a linear and positive, meaning that the employment outlook for occupations with high degrees of charisma is bright.

B.2. The occupational extent of emotional intelligence

Although the term gained popularity by Goleman (1995), the concept of EQ has been well-studied before starting with Maslow (1950); Beldoch (1964); Leuner (1966). Generally speaking, EQ refers to the ability to intelligently perceive, understand, and manage emotions, and people with high EQ use this ability to guide thinking and behavior. The occupational degree of EQ is particularly interesting as EQ has been shown to correlate positively with many desirable
Figure 10: Estimates of charisma on wage and growth

Notes: This figure shows a smoothed polynomial regression of our standardized measure of charisma in each 6-digit SOC occupation against its rank in the wage distribution (10a) and the employment growth distribution (10b).

outcomes. In summary, a recent review by Mayer et al. (2008) documents that higher EQ correlates with better social relations, better perceptions by others, better academic achievement, and better general well-being, e.g., higher life satisfaction.

We begin the analysis with Table F.1, tabulating the top and bottom ten occupations on EQ, and Figure F.1 showing EQ by occupational groups. All top 10 occupations are either therapists, counselors, psychologists or teachers, typically within social service or health care, and these occupations truly require the ability to carry out reasoning about emotions to support others in various needs. Contrary in the bottom 10, these occupations, e.g., pipe-layers, plumbers, or carpenters, require less emotional knowledge and more arm-hand steadiness and manual dexterity. This applies more generally to occupations within construction, maintenance, installation, and repair.

An interesting tail effect is found when considering EQ by educational requirements in Figure F.2. We find a monotonic relationship from low education and low EQ to high education and high EQ, but this monotonicity stops once we consider occupations that on average require a doctorate. This indicates that occupations that require doctorates rather than people with a master’s degree need very specialized employees that may not have to engage too often with a larger group of coworkers rather than focusing deeply on very complex problems.
Last, we consider the usual smoothed regression against ranks in the wage and employment growth distributions and plot the results in Figure F.3. The EQ-wage relationship from Figure F.3a is concave up to the median wage percentile and linear hereafter. This means that while occupations in the very low part of the wage distribution generally score higher on EQ compared to occupations around the first quartile, we generally find that increasing levels of EQ are associated with higher-wage occupations. That EQ should be positively correlated with wage at the individual level is found by Rode et al. (2017), however, the full distribution tells us that on aggregate relatively high level of EQ is also to be found in very low-wage occupations. Recall that we find a fully concave relationship between occupational charisma and wage in Figure 10a, but we cannot recover this pattern for EQ. This tentatively suggests that EQ is on average a more difficult characteristic to achieve compared to charisma in terms of wage returns. However, we emphasize that these relationships are mere correlations and we do not attempt to draw any causal inference. The employment outlook for occupations with high degrees of EQ is bright, as documented by Figure F.3b.

VI. Conclusion

It is inherently interesting to study occupational characteristics and how they associate with descriptive statistics of the labor market, e.g., employment and wage growth, educational attainment, labor force participation rates, etc. These studies inform policy-makers on how to optimally design education and labor markets. But estimating the occupational degree of certain target characteristics is challenging for many reasons. Most importantly, it may be severely difficult to validate and verify novel measures of occupational characteristics once constructed. This implicitly leaves too much room for research creativity with regards to the selection of occupational features that enter the composition of new measures.

To accommodate this issue, we propose occupation2vec as a fully data-driven and general approach to quantifying occupations, which enable the measurement of any occupational characteristic of interest in a transparent and verifiable way. Using natural language processing, we embed more than 17,000 occupation-specific descriptors sourced from the O*NET as vectors and combine them into unique occupation vectors of high dimensions. Using an objective and reliable
definition of a given target characteristic of interest, we also embed this into a high-dimensional vectors and measure the occupational degree of the given characteristic as standardized correlation coefficients between the vectors.

Using the actual scores from O*NET on 244 attributes common to the universe of occupations as the ground truth, we extensively validate our approach by comparing our estimates to those scores. We find that our estimates explain more than 50% of the variation in the original scores and, joint with other validation exercises, we take this as evidence that our framework is capable of producing high-quality occupation vectors that accurately capture detailed aspects of occupations.

We apply occupation2vec to four applications, where we study the occupational degree of two known characteristics as well as two novel. As known, or previously studied, characteristics, we consider the popular task measures (abstract, manual, and routine tasks), and exposure to AI. Our estimates of the task measures largely match the original ones. For instance, we also find that occupations involving many routine or manual tasks tend to appear in the bottom of the wage distribution. The opposite holds for abstract tasks. Regarding exposure to AI, our measure also broadly agrees with the already-proposed measures and can in fact be seen as a mix of the two most widely used ones. As novel characteristics, we consider the occupational degree on charisma and EQ. We find that occupational that score high on these attributes tend to be found within community and social service occupations, arts, entertainment, sports, and media occupations, and educational instruction occupations. High scores occur most often for occupations that require advanced degrees and in the top of the wage distribution. The most striking difference in found for occupations in the bottom of the wage distribution, where scores tend to be high on charisma but not on EQ. This suggests that wage returns to charisma and EQ could be very different.

In summary, this paper proposes a data-driven and universal approach to representing occupations as mathematical objects. This has many interesting economic applications, and particularly, we demonstrate how one may advance from purely qualitative definitions to quantitative scores at the occupational level. The occupation2vec framework, therefore, opens many doors for future analyses that study how novel occupational characteristics that have previously been inaccessible to researchers matter for labor market outcomes.
References

ACEMOGLU, D. AND D. AUTOR (2011): “Skills, Tasks and Technologies: Implications for Employment and Earnings,” in Handbook of Labor Economics, Elsevier, 1043–1171.

ACEMOGLU, D., D. AUTOR, J. HAZELL, AND P. RESTREPO (2020): “AI and Jobs: Evidence from Online Vacancies,” Working Paper 28257, National Bureau of Economic Research.

AMERICAN PSYCHOLOGICAL ASSOCIATION (2021a): “Definition of charisma,” Dictionary of Psychology, Retrieved August 26, 2021, https://dictionary.apa.org/charisma.

——— (2021b): “Definition of emotional intelligence,” Dictionary of Psychology, Retrieved August 26, 2021, https://dictionary.apa.org/emotional-intelligence.

AUTOR, D. H. AND D. DORN (2013): “The growth of low-skill service jobs and the polarization of the U.S. labor market,” American Economic Review, 103, 1553–97.

AUTOR, D. H. AND M. J. HANDEL (2013): “Putting Tasks to the Test: Human Capital, Job Tasks, and Wages,” Journal of Labor Economics, 31, S59–S96.

AUTOR, D. H., L. F. KATZ, AND M. S. KEARNEY (2006): “The polarization of the U.S. labor market,” American Economic Review, 96, 189–194.

AUTOR, D. H., F. LEVY, AND R. J. MURNANE (2003): “The Skill Content of Recent Technological Change: An Empirical Exploration*,” The Quarterly Journal of Economics, 118, 1279–1333.

AVOLIO, B. J. AND F. J. YAMMARINO, eds. (2013): Transformational and Charismatic Leadership: The Road Ahead 10th Anniversary Edition, Emerald Group Publishing Limited.

BELDOCH, M. (1964): Sensitivity to expression of emotional meaning in three modes of communication, The Communication of Emotional Meaning, Davitz, Joel R. (Ed.) McGraw-Hill.
Bojanowski, P., E. Grave, A. Joulin, and T. Mikolov (2016): “Enriching Word Vectors with Subword Information,” CoRR, abs/1607.04606.

Bowen, A., K. Kuralbayeva, and E. L. Tiptoe (2018): “Characterising green employment: The impacts of ‘greening’ on workforce composition,” Energy Economics, 72, 263–275.

Brynjolfsson, E., T. Mitchell, and D. Rock (2018): “What Can Machines Learn, and What Does It Mean for the Occupations and Industries,” in AEA Papers and Proceedings.

Cambridge University Press (2021a): “Definition of charisma,” Dictionary, Retrieved August 26, 2021, https://dictionary.cambridge.org/dictionary/english/charisma.

——— (2021b): “Definition of emotional intelligence,” Dictionary, Retrieved August 26, 2021, https://dictionary.cambridge.org/dictionary/english/emotional-intelligence.

Deming, D. J. (2017): “The Growing Importance of Social Skills in the Labor Market*,” The Quarterly Journal of Economics, 132, 1593–1640.

Devlin, J., M. Chang, K. Lee, and K. Toutanova (2018): “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding,” CoRR, abs/1810.04805.

Dingel, J. and B. Neiman (2020): “How Many Jobs Can be Done at Home?” National Bureau of Economic Research Working Paper.

Dow, T. E. (1969): “The Theory of Charisma,” The Sociological Quarterly, 10, 306–318.

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

Firth, J. (1957): “A Synopsis of Linguistic Theory 1930–1955,” in Studies in Linguistic Analysis, Oxford: Philological Society, reprinted in Palmer, F. (ed. 1968) Selected Papers of J. R. Firth, Longman, Harlow.
Frey, C. B. and M. A. Osborne (2017): “The future of employment: How susceptible are jobs to computerisation?” Technological Forecasting and Social Change, 114, 254–280.

Goldin, C. (2014): “A Grand Gender Convergence: Its Last Chapter,” American Economic Review, 104, 1091–1119.

Goleman, D. (1995): Emotional intelligence, New York: Bantam Books.

Harris, Z. S. (1954): “Distributional Structure,” WORD, 10, 146–162.

Holland, J. L. (1985): Making vocational choices: A theory of vocational personalities and work environments, Englewood Cliffs, NJ: Prentice-Hall.

Joulin, A., E. Grave, P. Bojanowski, M. Douze, H. Jégou, and T. Mikolov (2016a): “FastText.zip: Compressing text classification models,” CoRR, abs/1612.03651.

Joulin, A., E. Grave, P. Bojanowski, and T. Mikolov (2016b): “Bag of Tricks for Efficient Text Classification,” CoRR, abs/1607.01759.

Kingma, D. P. and J. Ba (2015): “Adam: A Method for Stochastic Optimization,” in 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings, ed. by Y. Bengio and Y. LeCun.

Le, Q. and T. Mikolov (2014): “Distributed Representations of Sentences and Documents,” in Proceedings of the 31st International Conference on Machine Learning, ed. by E. P. Xing and T. Jebara, Beijing, China: PMLR, vol. 32.2 of Proceedings of Machine Learning Research, 1188–1196.

Leuner, B. (1966): “Emotionale Intelligenz und Emanzipation. Eine psychodynamische Studie über die Frau. [Emotional intelligence and emancipation. A psychodynamic study on women],” Praxis der Kinderpsychologie und Kinderpsychiatrie. [Practice of Child Psychology and Child Psychiatry], 15, 196–203.

Liu, Y., M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov (2019): “RoBERTa: A Robustly Optimized BERT Pretraining Approach,” CoRR, abs/1907.11692.
MASLOW, A. (1950): *Social theory of motivation*, Twentieth century mental hygiene: New directions in mental health, Shore, Maurice J. (Ed.) New York, Social Science Publishers.

MAYER, J. D., R. D. ROBERTS, AND S. G. BARSADE (2008): “Human Abilities: Emotional Intelligence,” *Annual Review of Psychology*, 59, 507–536.

MIKOLOV, T., K. CHEN, G. S. CORRADO, AND J. DEAN (2013): “Efficient Estimation of Word Representations in Vector Space,” in *ICLR*.

MONGEY, S., L. PILOSSOPH, AND A. WEINBERG (2020): “Which Workers Bear the Burden of Social Distancing?” *National Bureau of Economic Research Working Paper*.

OXFORD UNIVERSITY PRESS (2021a): “Definition of charisma,” Dictionary, Retrieved August 26, 2021, https://www.oxfordreference.com/view/10.1093/oi/authority.20110803095603124.

——— (2021b): “Definition of emotional intelligence,” Dictionary, Retrieved August 26, 2021, https://www.oxfordreference.com/view/10.1093/oi/authority.20110803095749954.

PENNINGTON, J., R. SOCHER, AND C. D. MANNING (2014): “Glove: Global Vectors for Word Representation,” in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, October 25-29, 2014, Doha, Qatar, A meeting of SIGDAT, a Special Interest Group of the ACL*, ed. by A. Moschitti, B. Pang, and W. Daelemans, ACL, 1532–1543.

PSYCHOLOGY TODAY (2021a): “Definition of charisma,” Dictionary, Retrieved August 26, 2021, https://www.psychologytoday.com/us/basics/charisma.

——— (2021b): “Definition of emotional intelligence,” Dictionary, Retrieved August 26, 2021, https://www.psychologytoday.com/us/basics/emotional-intelligence.

REIMERS, N. AND I. GUREVYCH (2019): “Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks,” in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*, Association for Computational Linguistics.
Rode, J. C., M. Arthaud-Day, A. Ramaswami, and S. Howes (2017): “A time-lagged study of emotional intelligence and salary,” *Journal of Vocational Behavior*, 101, 77–89.

Rogers, A., O. Kovaleva, and A. Rumshisky (2021): “A Primer in BERTology: What We Know About How BERT Works,” *Transactions of the Association for Computational Linguistics*, 8, 842–866.

Spencer, M. E. (1973): “What Is Charisma?” *The British Journal of Sociology*, 24, 341.

Taylor, W. L. (1953): “The Cloze Procedure: A New Tool for Measuring Readability,” *Journalism Quarterly*, 30, 415–433.

Turner, S. (2003): “Charisma Reconsidered,” *Journal of Classical Sociology*, 3, 5–26.

U.S. Bureau of Labor Statistics (2021): “Measuring Green Jobs,” Green Jobs, Retrieved November 2, 2021, [https://www.bls.gov/green/home.htm](https://www.bls.gov/green/home.htm).

Van der Maaten, L. and G. Hinton (2008): “Visualizing Data using t-SNE,” *Journal of Machine Learning Research*, 9, 2579–2605.

Vaswani, A., N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin (2017): “Attention Is All You Need,” *CoRR*, abs/1706.03762.

Von Hippel, W., R. Ronay, E. Baker, K. Kjelsaas, and S. C. Murphy (2016): “Quick Thinkers Are Smooth Talkers: Mental Speed Facilitates Charisma,” *Psychological Science*, 27, 119–122, pMID: 26621964.

Webb, M. (2019): “The Impact of Artificial Intelligence on the Labor Market,” *SSRN Electronic Journal*.

Weber, M. (1947): *The theory of social and economic organization*. Translated by A. M. Henderson and Talcott Parsons; edited with an Introduction by Talcott Parsons, New York, Oxford University Press.

Wikipedia (2021a): “Article on charisma,” Encyclopedia, Retrieved August 26, 2021, [https://en.wikipedia.org/wiki/Charisma](https://en.wikipedia.org/wiki/Charisma).
Wu, Y., M. Schuster, Z. Chen, Q. V. Le, M. Norouzi, W. Macherey, M. Krikun, Y. Cao, Q. Gao, K. Macherey, J. Klingner, A. Shah, M. Johnson, X. Liu, L. Kaiser, S. Gouws, Y. Kato, T. Kudo, H. Kazawa, K. Stevens, G. Kurian, N. Patil, W. Wang, C. Young, J. Smith, J. Riesa, A. Rudnick, O. Vinyals, G. Corrado, M. Hughes, and J. Dean (2016): “Google’s Neural Machine Translation System: Bridging the Gap between Human and Machine Translation,” CoRR, abs/1609.08144.
## A. Data appendix

### Table A.1: O*NET scales

| Scale           | Category                              | Definition                                                                 |
|-----------------|---------------------------------------|-----------------------------------------------------------------------------|
| Importance      | Tasks, Knowledge, Skills, Abilities, Work Activities, and Work Styles | Rating indicates the degree of importance a particular descriptor is to the occupation. |
| Level           | Knowledge, Skills, Abilities, and Work Activities | Rating indicates the degree to which a particular descriptor is required or needed to perform the occupation. |
| Relevance       | Tasks                                 | Rating indicates the degree to which a particular task is relevant to perform the occupation. |
| Frequency       | Tasks                                 | Rating indicates how frequent a particular task is performed within a given time period. |
| Occupational Interest | Interests          | Rating indicates the degree to which a particular interest (Holland, 1985) matches the occupation. |
| Extent          | Work Values                          | Rating indicates the degree to which a particular work value affects the nature of the occupation. |
| Context         | Work Context                         | Rating indicates the degree to which a particular work context influences the nature of the occupation. |

_Notes:_ This table shows scales associated with each descriptor within categories. All scales are standardized to a scale ranging from 0 to 1. See O*NET site on scales, ratings, and standardized scores for more information (here).
B. Method appendix

Let $L$ denote the number of layers (i.e., Transformer blocks), let $d$ denote the hidden size (i.e., the dimensions of the embeddings), and let $A$ denote the number of self-attention heads. The model size we consider uses $L = 24$, $d = 1024$, and $A = 16$. All hyperparameters can be found in Table B.1. For $m = 1, \ldots, N$, let $T_m$ be a piece of text (e.g., a sentence or a paragraph) from $T$, where we ignore the subscripts $i, k, j$. Each $T_m$ is a sequence of tokens (e.g., words or subwords), $t_{m,1}, \ldots, t_{m,N_m}$, generated by WordPiece tokenization (Wu et al., 2016). BERT takes as input a concatenation of two sequences, $t_{m,1}, \ldots, t_{m,N_m}$ and $t_{m',1}, \ldots, t_{m',N_{m'}}$ for $m \neq m'$. The two sequences are delimited by special tokens, namely

\[
\text{tok} (T_m, T_{m'}) = [\text{CLS}], t_{m,1}, \ldots, t_{m,N_m}, [\text{SEP}], t_{m',1}, \ldots, t_{m',N_{m'}}, [\text{EOS}],
\]  

where $\text{tok} (\cdot, \cdot)$ produces the special tokenization of two input texts, $\text{CLS}$ abbreviates classification, $\text{SEP}$ separation, and $\text{EOS}$ end-of-sentence. The lengths of $T_m$ and $T'_m$ are constrained such that $N_m + N_{m'} < \bar{N}$, where $\bar{N}$ is fixed beforehand. From B.1, BERT constructs three temporary input embeddings that, once summed, act as the input embedding to the Transformer. First, each token $t$ in $\text{tok} (T_m, T_{m'})$ is passed through three deterministic functions,

\[
l : (t, \mathcal{V}) \rightarrow [1, |\mathcal{V}|], \quad \text{(B.2)}
\]

\[
p : (t, \text{tok} (T_m, T_{m'})) \rightarrow [1, \bar{N}], \quad \text{(B.3)}
\]

\[
s : (t, \text{tok} (T_m, T_{m'})) \rightarrow [1, 2], \quad \text{(B.4)}
\]

where $l (t, \mathcal{V})$ is a lookup function that returns the index of token $t$ in vocabulary $\mathcal{V}$ of all tokens, $p (t, \text{tok} (T_m, T_{m'}))$ is a position function that returns the index of token $t$ in the concatenated sequences $\text{tok} (T_m, T_{m'})$, and $s (t, \text{tok} (T_m, T_{m'}))$ is a sequence function that returns the sequence index that token $t$ belongs to in $\text{tok} (T_m, T_{m'})$. Essentially, each token is represented as three integers that determines the both position of the token in the vocabulary and the concatenated sequences, respectively, and which of the individual sequences that the token belongs to. Last, each integer (i.e., token, position, and sequence) is passed through an embedding layer $\phi_h : \mathbb{R} \rightarrow \mathbb{R}^d$ for $h = l, p, s$ that returns a $d$-
We show the full process of turning tokens into input embeddings in Figure B.1. These input embeddings are then passed through the layers of the deep Transformer model. During training, BERT considers two primary objectives, i.e., the MLM objective and the NSP objective, as described in the main text.

Notes: This figure shows how BERT translates tokens into an input embedding by summing embeddings of the integers that describe both the token’s position in the vocabulary and the concatenated sequences and which of the sequences the token belongs to.

dimensional embedding. The three embeddings are summed to generate the final input embedding $e_t = e(t; V, T_m, T_{m'})$. That is,

$$ e(t; V, T_m, T_{m'}) = \phi_l(l(t, V)) + \phi_p(p(t, \text{tok}(T_m, T_{m'}))) + \phi_s(s(t, \text{tok}(T_m, T_{m'}))) $$

(B.5)

We show the full process of turning tokens into input embeddings in Figure B.1. These input embeddings are then passed through the layers of the deep Transformer model. During training, BERT considers two primary objectives, i.e., the MLM objective and the NSP objective, as described in the main text.

In our framework, the object of interest is the sentence or paragraph embedding of $T_m$, and generally, three strategies exist to compute these embeddings once the final word embeddings have been estimated: \textit{mean-pooling} that uses the average of all the word embeddings from $T_m$, \textit{max-pooling} that uses the maximum of all the word embeddings from $T_m$, or \textit{first-pooling} that uses the first of all the word embeddings from $T_m$ (i.e., the [CLS] token). We use \textit{mean-pooling} to construct our sentence embeddings. Technically, we add a pooling operation to the

---

\[ \phi_l(l([CLS])) \phi_l(l(t_{m,1})) \ldots \phi_l(l(t_{m,N_m})) \phi_l([SEP]) \phi_l(l(t_{m',1})) \ldots \phi_l(l(t_{m',N_{m'}})) \phi_l([EOS]) \]

\[ \phi_p(p([CLS])) \phi_p(p(t_{m,1})) \ldots \phi_p(p(t_{m,N_m})) \phi_p([SEP]) \phi_p(p(t_{m',1})) \ldots \phi_p(p(t_{m',N_{m'}})) \phi_p([EOS]) \]

\[ \phi_s(s([CLS])) \phi_s(s(t_{m,1})) \ldots \phi_s(s(t_{m,N_m})) \phi_s([SEP]) \phi_s(s(t_{m',1})) \ldots \phi_s(s(t_{m',N_{m'}})) \phi_s([EOS]) \]

\[ e_{[CLS]} e_{t_{m,1}} \ldots e_{t_{m,N_m}} e_{[SEP]} e_{t_{m',1}} \ldots e_{t_{m',N_{m'}}} e_{[EOS]} \]
Table B.1: Hyperparameters for pre-training BERT and RoBERTa

| Hyperparameter                  | BERT  | RoBERTa |
|---------------------------------|-------|---------|
| Number of layers, $L$           | 24    | 24      |
| Hidden size, $d$                | 1024  | 1024    |
| Feed-forward/filter size        | 4096  | 4096    |
| Attention heads                 | 16    | 16      |
| Dropout                         | 0.1   | 0.1     |
| Warm-up steps                   | 10,000| 30,000  |
| Weight decay, $L_2$             | 0.01  | 0.01    |
| Peak learning rate              | 0.0001| 0.0004  |
| Learning rate decay             | Linear| Linear  |
| Gradient clipping               | 0.0   | 0.0     |
| Maximum tokens, $N_0$           | 512   | 512     |
| Minibatch size, $B$             | 256   | 8,000   |
| Max steps, $S$                  | 1,000,000| 500,000|
| Activation function             | GELU  | GELU    |
| Optimizer                       | Adam  | Adam    |
| Adam, $\beta_1$                | 0.9   | 0.9     |
| Adam, $\beta_2$                | 0.999 | 0.98    |
| Adam, $\epsilon$               | 0.000001| 0.000001|

*Notes:* This table shows the hyperparameters used to pre-train BERT and RoBERTa, respectively.

The output of BERT/RoBERTa to derive the sentence embeddings, which is proposed by Reimers and Gurevych (2019). This is essentially a fine-tuning layer such that BERT/RoBERTa generates better sentence embeddings. For more details, we refer interested readers to Reimers and Gurevych (2019). The hyperparameters used for pre-training follow from Table B.1.
C. Degree of task measures

Table C.1: Components of the task measures

| Scale             | Subscale                  | Element ID | Element name                                           |
|-------------------|---------------------------|------------|--------------------------------------------------------|
| Abstract          | Non-routine cognitive: Analytical | 4.A.2.a.4  | Analyzing data/information                             |
|                   |                            | 4.A.2.b.2  | Thinking creatively                                     |
|                   |                            | 4.A.4.a.1  | Interpreting information for others                    |
|                   | Non-routine cognitive: Interpersonal | 4.A.4.a.4  | Establishing and maintaining personal relationships    |
|                   |                            | 4.A.4.b.4  | Guiding, directing and motivating subordinates         |
|                   |                            | 4.A.4.b.5  | Coaching/developing others                             |
| Manual            | Non-routine manual         | 4.A.3.a.4  | Operating vehicles, mechanized devices, or equipment   |
|                   |                            | 4.C.2.d.1.g| Spend time using hands to handle, control or feel objects, tools or controls |
|                   |                            | 1.A.2.a.2  | Manual dexterity                                       |
|                   |                            | 1.A.1.f.1  | Spatial orientation                                    |
| Routine           | Routine cognitive          | 4.C.3.b.7  | Importance of repeating the same tasks                 |
|                   |                            | 4.C.3.b.4  | Importance of being exact or accurate                  |
|                   | Routine manual             | 4.C.3.d.3  | Pace determined by speed of equipment                  |
|                   |                            | 4.A.3.a.3  | Controlling machines and processes                      |
|                   |                            | 4.C.2.d.1.i| Spend time making repetitive motions                   |

Notes: This table shows the task measures and their components. Each measure is a standardized sum of its subscales, and equivalently, each subscale is a standardized sum of its individual elements. See Acemoglu and Autor (2011) for further details.

Abstract tasks

Definition 4 (Abstract tasks). “Abstract tasks are problem-solving and managerial tasks, which require intuition, persuasion, and creativity and are unstructured and nonroutine. These abstract tasks are performed by workers who have high levels of education and analytical capability, and they benefit from computers that facilitate the transmission, organization, and processing of information. Abstract tasks are characteristic of professional, managerial, technical, and creative occupations, such as law, medicine, science, engineering, marketing, and design.”
Table C.2: Top and bottom 10 occupations on degree of abstract tasks

| Top 10 occupations                  | Score | Bottom 10 occupations                      | Score |
|-------------------------------------|-------|--------------------------------------------|-------|
| Mathematicians                      | 3.751 | Hosts and Hostesses, Restaurant, Lounge, and Coffee Shop | -2.967 |
| Operations Research Analysts        | 3.219 | Lockers, Room, Counter, and Dressing Room Attendants | -2.972 |
| Computer and Information Research Scientists | 3.215 | Tire Repairers and Changers                 | -2.872 |
| Graphic Designers                   | 3.144 | Dancers                                    | -2.744 |
| Electrical and Electronics Drafters | 2.967 | Counter and Rental Clerks                   | -2.199 |
| Architectural and Civil Drafters    | 2.942 | Passenger Attendants                        | -2.194 |
| Computer Programmers                | 2.702 | Food Servers, Nonrestaurant                 | -2.201 |
| Computer Numerically Controlled Tool Programmers | 2.672 | Insurance Sales Agents                      | -2.146 |
| Computer Occupations, All Other     | 2.652 | Parts Salespersons                          | -2.139 |
| Fine Artists, Including Painters, Sculptors, and Illustrators | 2.607 | Lifeguards, Ski Patrol, and Other Recreational Protective Service Workers | -2.103 |

Notes: This table shows the top and bottom 10 occupations on degree of abstract tasks estimated by the occupation2vec framework as the standardized correlation coefficient between occupation embeddings and the embedding of abstract tasks.

Figure C.1: Abstract tasks by major occupational group

Notes: This figure shows a boxplot of the occupational degree of abstract tasks by major occupational groups.
**Figure C.2:** Abstract tasks by education

*Notes:* This figure shows a boxplot of the occupational degree of abstract tasks by educational requirement.

**Figure C.3:** Comparing estimates of abstract tasks on wage and growth

*Notes:* This figure shows a smoothed polynomial regression of both standardized measures of abstract tasks in each 6-digit SOC occupation against its rank in the wage distribution (C.3a) and the employment growth distribution (C.3b).
### Table C.3: Top and bottom 10 occupations on degree of manual tasks

| Top 10 occupations                                      | Score | Bottom 10 occupations                  | Score |
|---------------------------------------------------------|-------|----------------------------------------|-------|
| Farmworkers and Laborers, Crop, Nursery, and Greenhouse| 2.346 | Political Science Teachers, Postsecondary | -2.339|
| Grading, Lapping, Polishing, and Buffing Machine Tool Setters, Operators, and Tenders, Metal and Plastic | 2.162 | Communications Teachers, Postsecondary | -2.209|
| Excavating and Loading Machine and Dragline Operators, Surface Mining | 2.171 | Social Work Teachers, Postsecondary | -2.169|
| Cleaners of Vehicles and Equipment                       | 2.034 | Sociology Teachers, Postsecondary | -2.146|
| Machine Feeders and Fitters                             | 2.007 | Philosophy and Religion Teachers, Postsecondary | -2.140|
| Agricultural Equipment Operators                        | 2.000 | Forensic Science Teachers, Postsecondary | -2.136|
| Laborers and Freight, Stock, and Material Movers, Hand  | 2.000 | Sociologists                          | -2.098|
| Cleaning, Washing, and Metal Polishing Equipment Operators and Tenders | 2.076 | Arbitrators, Mediators, and Conciliators | -2.073|
| Industrial Truck and Tractor Operators                  | 2.015 | Psychology Teachers, Postsecondary | -2.026|
| Crushing, Grinding, and Polishing Machine Setters, Operators, and Tenders | 2.134 | Law Teachers, Postsecondary | -2.059|

**Notes:** This table shows the top and bottom 10 occupations on degree of manual tasks estimated by the occupation2vec framework as the standardized correlation coefficient between occupation embeddings and the embedding of manual tasks.

### Figure C.4: Manual tasks by major occupational group

**Notes:** This figure shows a boxplot of the occupational degree of manual tasks by major occupational groups.

#### Manual tasks

**Definition 5 (Manual tasks).** “Manual tasks are innate tasks like dexterity, sightedness, and visual and language recognition, which require adaptability, flexibility, and in-person interactions. These tasks are typically unpredictable but straightforward. Manual tasks are characteristic of construction and service occupations, such as truck drivers, janitors, and house-cleaners.”
Figure C.5: Manual tasks by education

Notes: This figure shows a boxplot of the occupational degree of manual tasks by educational requirement.

Figure C.6: Comparing estimates of manual tasks on wage and growth

Notes: This figure shows a smoothed polynomial regression of both standardized measures of manual tasks in each 6-digit SOC occupation against its rank in the wage distribution (C.6a) and the employment growth distribution (C.6b).
D. Degree of artificial intelligence

Definition of artificial intelligence The following definitions jointly define artificial intelligence. The definitions are from IEEE Guide for Terms and Concepts in Intelligent Process Automation, IEEE Std 2755TM-2017 (link). Starting from Artificial Intelligence in Definition 7, we include the definitions of all related terms repeatedly until we have these nine definitions. By using our NLP algorithm, we get one embeddings for each definition, which we finally combine into one embedding by a simple average.

Definition 6 (Artificial General Intelligence). “Complex, computational artificial intelligence capable of providing descriptive, discovery, predictive, prescriptive, and deductive analytics with relevance and accuracy equal to or exceeding human experts in multiple general knowledge domains. Artificial general intelligence includes artificial intelligence systems capable of interacting naturally with humans and machines in a way undetectable to expert observers and consistently passing the Turing Test for artificial intelligence.”

Definition 7 (Artificial Intelligence). “The combination of cognitive automation, machine learning, reasoning, hypothesis generation and analysis, natural language processing, and intentional algorithm mutation producing insights and analytics at or above human capability.”

Definition 8 (Machine Learning). “The combination of cognitive automation, machine learning, reasoning, hypothesis generation and analysis, natural language processing, and intentional algorithm mutation producing insights and analytics at or above human capability.”

Definition 9 (Cognitive Automation). “The identification, assessment, and application of available machine learning algorithms for the purpose of leveraging domain knowledge and reasoning to further automate the machine learning already present in a manner that may be thought of as cognitive. With cognitive automation, the system performs corrective actions driven by knowledge of the underlying analytics tool itself, iterates its own automation approaches and algorithms for more expansive or more thorough analysis, and is thereby able to fulfill its purpose. The automation of the cognitive process refines itself and dynamically generates
novel hypotheses that it can likewise assess against its existing corpus and other information resources.”

**Definition 10** (Descriptive Analytics). “Insights, reporting, and information answering the question, “Why did something happen?” Descriptive analytics determines information useful to understanding the cause(s) of an event(s).”

**Definition 11** (Deductive Analytics). “Insights, reporting, and information answering the question, “What would likely happen IF...?” Deductive analytics evaluates causes and outcomes of possible future events.”

**Definition 12** (Diagnostic Analytics). “Insights, reporting, and information answering the question, “Why did something happen?” Diagnostic analytics determines information useful to understanding the cause(s) of an event(s).”

**Definition 13** (Predictive Analytics). “Insights, reporting, and information answering the question, “What is likely to happen?” Predictive analytics support high confidence foretelling of future event(s).”

**Definition 14** (Prescriptive Analytics). “Insights, reporting, and information answering the question, “What should I do about it?” Prescriptive analytics determines information that provides high confidence actions necessary to recover from an event or fulfill a need.”

**Table D.1:** Top and bottom 10 occupations on degree of artificial intelligence

| Top 10 occupations                          | Score  | Bottom 10 occupations                          | Score  |
|--------------------------------------------|--------|-----------------------------------------------|--------|
| Data Scientists                            | 4.289  | Helpers–Brickmasons, Blockmasons, Stonemasons, and Tile and Marble Setters | -2.308 |
| Statisticians                              | 3.953  | Helpers–Painters, Paperhangers, Plasterers, and Stucco Masons | -2.357 |
| Operations Research Analysts               | 3.757  | Construction Laborers                          | -1.999 |
| Market Research Analysts and Marketing Specialists | 3.444  | Manicurists and Pedicurists                     | -1.942 |
| Mathematical Science Occupations, All Other | 3.419  | Stonemasons                                    | -1.856 |
| Database Architects                        | 3.251  | Hairdressers, Hairstylists, and Cosmetologists  | -1.853 |
| Financial Specialists, All Other           | 3.164  | Helpers–Pipelayers, Plumbers, Pipefitters, and Steamfitters | -1.849 |
| Actuaries                                  | 3.142  | Funeral Attendants                             | -1.837 |
| Credit Analysts                            | 3.062  | Plasterers and Stucco Masons                   | -1.825 |
| Statistical Assistants                     | 2.997  | Landscaping and Groundskeeping Workers         | -1.805 |

**Notes:** This table shows the top and bottom 10 occupations on degree of artificial intelligence estimated by the **occupation2vec** framework.

**Top and bottom 10 occupations**

**Major occupational groups and educational requirements**
Figure D.1: Artificial intelligence by major occupational group

Notes: This figure shows a boxplot of the occupational degree of artificial intelligence by major occupational groups.

Figure D.2: Artificial intelligence by education

Notes: This figure shows a boxplot of the occupational degree of artificial intelligence by educational requirement.

Wage and employment growth profiles

Correlation coefficients
Artificial intelligence by occupation wages 

(b) Artificial intelligence by occupation growth

Figure D.3: Comparing estimates of artificial intelligence on wage and growth

Notes: This figure shows a smoothed polynomial regression of both standardized measures of artificial intelligence in each 6-digit SOC occupation against its rank in the wage distribution (D.3a) and the employment growth distribution (D.3b).

|                  | Pearson’s correlation | Kendall’s rank correlation | Spearman’s rank correlation |
|------------------|-----------------------|---------------------------|-----------------------------|
| Felten et al. (2018) | 0.3819               | 0.2657                    | 0.3879                      |
| Brynjolfsson et al. (2018) | 0.1876             | 0.1377                    | 0.2069                      |
| Webb (2020)          | 0.2703               | 0.1548                    | 0.2300                      |

Notes: This table shows various correlation coefficients from comparing our occupational estimates of artificial intelligence to the ones of Felten et al. (2018), Brynjolfsson et al. (2018), and Webb (2019), respectively. The correlation coefficients include Pearson’s correlation, Kendall’s rank correlation, and Spearman’s rank correlation.
E. Degree of charisma

**Definition of charisma**  The following definitions jointly define charisma. The definitions are from Psychology Today, Cambridge University Press, Oxford University Press, and Wikipedia, respectively. In addition, we use American Psychological Association’s Dictionary of Psychology (see Definition 3). By using our NLP algorithm, we get one embedding for each definition, which we finally combine into one embedding by a simple average.

**Definition 15** (Charisma). "Charisma is an individual’s ability to attract and influence other people. While it is often described as a mysterious quality that one either has or doesn’t have, some experts argue that the skills of charismatic people can be learned and cultivated." (Psychology Today, 2021a)

**Definition 16** (Charisma). "A special power that some people have naturally that makes them able to influence other people and attract their attention and admiration." (Cambridge University Press, 2021a)

**Definition 17** (Charisma). "A quality possessed by some individuals that encourages others to listen and follow. Charismatic leaders tend to be self-confident, visionary, and change oriented, often with eccentric or unusual behavior." (Oxford University Press, 2021a)

**Definition 18** (Charisma). "Charisma is compelling attractiveness or charm that can inspire devotion in others. Scholars in sociology, political science, psychology, and management reserve the term for a type of leadership seen as extraordinary; in these fields, the term charisma is used to describe a particular type of leader who uses values-based, symbolic, and emotion-laden leader signaling." (Wikipedia, 2021a)
F. Degree of emotional intelligence

Definition of emotional intelligence  The following definitions jointly define emotional intelligence. The definitions are from American Psychological Association, Psychology Today, Cambridge University Press, Oxford University Press, and Wikipedia, respectively. By using our NLP algorithm, we get one embedding for each definition, which we finally combine into one embedding by a simple average.

Definition 19 (Emotional intelligence). “A type of intelligence that involves the ability to process emotional information and use it in reasoning and other cognitive activities, proposed by U.S. psychologists Peter Salovey and John D. Mayer. According to Mayer and Salovey’s 1997 model, it comprises four abilities: to perceive and appraise emotions accurately; to access and evoke emotions when they facilitate cognition; to comprehend emotional language and make use of emotional information; and to regulate one’s own and others’ emotions to promote growth and well-being. Their ideas were popularized in a best-selling book by U.S. psychologist and science journalist Daniel J. Goleman who also altered the definition to include many personality variables.” (American Psychological Association, 2021b)

Definition 20 (Emotional intelligence). “Emotional intelligence refers to the ability to identify and manage one’s own emotions, as well as the emotions of others. Emotional intelligence is generally said to include a few skills: namely emotional awareness, or the ability to identify and name one’s own emotions; the ability to harness those emotions and apply them to tasks like thinking and problem solving; and the ability to manage emotions, which includes both regulating one’s own emotions when necessary and helping others to do the same.” (Psychology Today, 2021b)

Definition 21 (Emotional intelligence). “The ability to understand the way people feel and react and to use this skill to make good judgments and to avoid or solve problems.” (Cambridge University Press, 2021b)

Definition 22 (Emotional intelligence). “Ability to monitor one’s own and other people’s emotions, to discriminate between different emotions and label them appropriately, and to use emotional information to guide thinking and behavior. It encompasses four competencies: the ability to perceive, appraise, and express ...
Table F.1: Top and bottom 10 occupations on degree of emotional intelligence

| Top 10 occupations                          | Score | Bottom 10 occupations                          | Score |
|---------------------------------------------|-------|-----------------------------------------------|-------|
| Marriage and Family Therapists              | 3.879 | Helpers–Brickmasons, Blockmasons, Stonemasons, and Tile and Marble Setters | -2.669 |
| Mental Health Counselors                    | 3.580 | Helpers–Pipelayers, Plumbers, Pipefitters, and Steamfitters | -2.246 |
| Clinical and Counseling Psychologists       | 3.363 | Roof Roofers, Mining                          | -2.029 |
| Psychiatrists                               | 3.128 | Helpers–Painters, Paperhangers, Plasterers, and Stucco Masons | -2.020 |
| Industrial-Organizational Psychologists     | 2.858 | Maintenance Workers, Machinery                 | -1.837 |
| School Psychologists                        | 2.845 | Shoe Machine Operators and Tenders             | -1.875 |
| Therapists, All Other                       | 2.679 | Woodworking Machine Setters, Operators, and Tenders, Except Sawing | -1.874 |
| Elementary School Teachers, Except Special Education | 2.636 | Helpers–Roofers                              | -1.840 |
| Kindergarten Teachers, Except Special Education | 2.438 | Maids and Housekeeping Cleaners               | -1.826 |
| Psychology Teachers, Postsecondary         | 2.392 | Cabinetmakers and Bench Carpenters            | -1.807 |

Notes: This table shows the top and bottom 10 occupations on degree of emotional intelligence estimated by the occupation2vec framework.

Emotions are: the ability to access and evoke emotions when they facilitate cognition; the ability to comprehend emotional messages and to make use of emotional information; and the ability to regulate one’s own emotions to promote growth and well-being. (…) Popularized interpretations of emotional intelligence include various other factors such as interpersonal skills and adaptability.” (Oxford University Press, 2021b)

Definition 23 (Emotional intelligence). “Emotional intelligence is most often defined as the ability to perceive, use, understand, manage, and handle emotions. People with high emotional intelligence can recognize their own emotions and those of others, use emotional information to guide thinking and behavior, discern between different feelings and label them appropriately, and adjust emotions to adapt to environments.” (Wikipedia, 2021b)

Top and bottom 10 occupations

Major occupational groups and educational requirements

Wage and employment growth profiles
Figure F.1: Emotional intelligence by major occupational group

Notes: This figure shows a boxplot of the occupational degree of emotional intelligence by major occupational groups.

Figure F.2: Emotional intelligence by education

Notes: This figure shows a boxplot of the occupational degree of emotional intelligence by educational requirement.
Figure F.3: Estimates of emotional intelligence on wage and growth

Notes: This figure shows a smoothed polynomial regression of our standardized measure of emotional intelligence in each 6-digit SOC occupation against its rank in the wage distribution (F.3a) and the employment growth distribution (F.3b).