The Impact of Artificial Intelligence on the Labor Market

Michael Webb*

Stanford University

May 15, 2019

Abstract

The impact of a technology on the economy depends on what tasks the technology can perform. We develop a new measure of the exposure of occupations to any given technology by quantifying the overlap between the text of job task descriptions and the text of patents in that technology. We find that low-wage jobs are most exposed to robotics, while middle-wage jobs are most exposed to software. By contrast, high-wage "non-routine" jobs are most exposed to artificial intelligence. Artificial intelligence may therefore have different effects on jobs and inequality compared to previous technologies.

Keywords: artificial intelligence, robotics, technology, patents, occupations
JEL-codes: J23, J24, O33

*Email: mww@stanford.edu. I am grateful to Nick Bloom, Tim Bresnahan, Erik Brynjolfsson, Raj Chetty, Diane Coyle, Matthew Gentzkow, Caroline Hoxby, Xavier Jaravel, Chad Jones, Pete Klenow, Will Rafey, Isaac Sorkin, and John Van Reenen for helpful comments and suggestions. Greg Thornton and Arshia Hashemi provided excellent research assistance. I acknowledge financial support from the Future of Life Institute FLI-RFP-AI1 program, grant #2016-158713 (5388).

1 Introduction

Technologies like steam, electricity, and the computer have each transformed the economy (Bresnahan and Trajtenberg, 1995; Jovanovic and Rousseau, 2005). They have changed the work people do and how much they get paid to do it (Goldin and Katz, 2007; Michaels et al., 2013). Computer software, for example, has replaced humans in performing many routine activities, such as clerical work (Autor et al., 2002). This appears to have reduced employment and wages in affected occupations, even as it has increased them in complementary activities (Autor, 2015).

In recent years, concern has shifted to the impacts of artificial intelligence, especially on employment (United States, 2016; House of Commons, 2016). Artificial intelligence has different capabilities to traditional software, so is likely to affect different kinds of economic activity. Because artificial intelligence has not yet been widely adopted by firms, analyzing the relationship between technology adoption and changes in employment (as, for example, Graetz and Michaels (2015) and Acemoglu and Restrepo (2017) have done for robots) is not yet feasible. Prior research on the potential impacts of artificial intelligence has consisted of case studies (e.g., Remus and Levy, 2017) and subjective assessments by experts (Frey and Osborne, 2017; Arntz et al., 2017; Grace et al., 2017; Manyika et al., 2017; Felten et al., 2017). Assessing which parts of the economy are exposed to artificial intelligence is difficult because it requires combining knowledge of specific technological capabilities with detailed knowledge of economic activities, and doing so at the scale of the whole economy.

Rich data on technologies and on occupations have long been available. The key challenge has been to find a systematic way of linking a specific technology to the specific economic activities it might affect. We develop a new method to do this for arbitrary technologies and economic domains. The key idea is that verb-object pairs, such as "diagnose disease", capture both technological capabilities and economic activities in a transparent, parsimonious way. We call these verb-object pairs "capability pairs". Capability pairs can be extracted automatically from text corpora using natural language processing algorithms.

To measure the exposure of occupations to different technologies, our approach combines two inputs: (i) the text of patents in a given technological field, and (ii) a database of occupations, O*NET, which includes descriptions of the component tasks involved in every occupation in the economy. We apply our method separately to artificial intelligence, traditional software, and robotics. For each of these technological fields, we calculate the intensity of patenting in that field directed towards the capabilities involved in each occupation in O*NET. Specifically, we take each capability pair (like "diagnose disease") extracted from each task description in each occupation, and measure how often similar capability pairs occur in the patents of a given technology. This allows us to characterize the kinds of occupations likely to be affected by each technology.

To validate our method, we conduct out-of-sample tests of our measure by comparing our results to a seminal paper on the effect of software on jobs (Autor, Levy, and Murnane, 2003), and to the most careful study we know on the likely effects of artificial intelligence on the work of a particular occupation – namely, lawyers (Remus and Levy, 2017). Our method independently recovers the key results from those two papers, using only patent text and job descriptions as input.

Our results suggest that artificial intelligence has a set of "non-routine" capabilities that fundamentally distinguish it from software. This leads to high-wage occupations being more exposed to artificial intelligence than low- and medium-wage occupations. All three technologies appear unlikely to affect those parts of jobs requiring strong interpersonal skills. We also find that men tend to work in occupations with much higher exposure to all three technologies than women.

While we use our method to study occupations, there are many other potential uses for such a general-purpose measure that directly links specific technologies to specific economic activities in a consistent and automated fashion. Our measure is likely to be useful for any field of economics that would benefit from fine-grained measures of technological progress, such as macroeconomic research on economic fluctuations and growth, and microeconomic research on productivity, labor, and innovation. As Syverson (2011), for example, notes, "research that ties observable attributes of [an] industry's technology... to the quantitative importance of productivity-influencing factors would be an incredible advance in our ability to explain productivity growth." We hope that our measure marks an important step in that direction.

We stress that our measure only quantifies the *exposure* of occupations to new technologies. It does not measure actual adoption of those technologies. It is thus a leading indicator of "realized" technological change in the economy. It is also important to emphasize that our measure only captures possibilities for "direct" impacts of technologies on tasks, as when spreadsheet applications took over performing calculations from humans. It misses "indirect" impacts, as when the invention of the automobile reduced the demand for horse breeders. It also does not capture new tasks that arise as a result of new technologies.

The rest of the paper proceeds as follows. Section 2 provides a qualitative description of what artificial intelligence is and the features of the technology that affect its economic applicability, in order to ground the later empirical analysis. Section 3 describes the construction of our technological exposure measures. Section 4 presents the results. Section 5 concludes.

2 Conceptual framework

An influential framework (Autor, Levy, and Murnane, 2003) for assessing the exposure of economic activities to the previous wave of computer technologies distinguishes between "routine" and "non-routine" tasks, as reviewed below. No such framework exists for artificial intelligence. We therefore introduce one to ground our empirical analysis. Here, we use the term "artificial intelligence" to refer to machine learning algorithms. Specifically, we have in mind algorithms for supervised learning and reinforcement learning. We recognize that "artificial intelligence" has broader scope, and our use is not meant prescriptively. Our focus is motivated by the dramatic practical advances in supervised and reinforcement learning in recent years (Krizhevsky et al., 2012; Mnih et al., 2015) and their broad economic applicability.

A supervised learning algorithm learns functions from training data consisting of example input-output pairs. For instance, it may learn a function that maps an image (input) into a textual description of the image (output) from a bank of examples of image-description pairs. With appropriate training data, it could learn to map a customer's financial history into their likelihood of repaying a loan, or sentences in one language into another. The algorithm learns the statistical relationships present in the training data, so, in a loose sense, the data must "contain" enough information for it to learn the function. A reinforcement learning algorithm, by contrast, learns how to achieve an objective by acting in a dynamic environment, which provides it with feedback. This might be controlling the machinery in a factory to optimize energy efficiency, or exchanging messages with a human to troubleshoot a technical problem. These algorithms learn by a kind of trial-and-error, so they require experiments in the environment, or a good simulator of it, and a way of evaluating performance.

These requirements suggest three features of tasks that affect their amenability to artificial intelligence: the availability of data, the nature of the objective, and the "levers" available. In supervised learning, data availability is important. Certain kinds of information are regularly stored at scale, like system logs, sensor data, and records kept by humans. Some tasks, such as operating a power plant or auditing a company's accounts, involve processing such information and little else. Other tasks, like discussing a treatment plan with a patient or advising a student on their research, consist of interactions that are not usually recorded, perhaps because doing so would be unethical or illegal. This makes training data difficult to obtain. Data must also be sufficiently comprehensive. A software company may have enough recordings of customer support interactions to train an algorithm to answer queries about its current products. But as soon as it releases a new product, new training data will be required. In general, tasks with stable, systematic relationships between inputs and decisions, such as sorting agricultural products, translating languages, or interpreting medical images, are much easier to learn than tasks with frequent changes or one-off exceptions, such as those handled by customer service supervisors, flight attendants, or any kind of team leader.

In reinforcement learning, learning by trial-and-error takes the place of learning from examples. Experimentation by an algorithm in the real world can be costly or unethical, so learning in simulation is often preferred (Rusu et al., 2016). A simulator may itself be learned from observational data, again requiring that such data be available (Rezende et al., 2014). Alternatively, a human could build a simulator manually. This may be possible for human-designed systems, like robots or industrial processes, and for well-understood aspects of the natural world, like classical mechanics. But manually building a simulator of a human to interact with would be more difficult. This means that, in practice, robotic locomotion and dextrous manipulation are more amenable to reinforcement learning than, say, motivational speaking, product development, or public relations. For the same reason, physical activities in contexts like assembly lines that are strictly regimented, hence easy to simulate, are easier to learn than physical activities in more variable contexts like construction sites, homes, or offices.

Algorithms need to evaluate their performance in order to improve, and this represents a second significant constraint on the kinds of task they can learn. Optimizing for energy efficiency is possible because energy efficiency is easily measurable. But it is less easy to measure success in tasks like writing a corporate strategy, negotiating an agreement, leading a team, or even tidying a room, and optimizing a misspecified objective can be highly counterproductive (Amodei et al., 2016). Delayed feedback, which characterizes many kinds of project work, makes learning even harder (Sutton, 1984). An algorithm could ask for guidance from a human (Christiano et al., 2017), but a human may not have the patience to give it, especially if they are a paying customer. Most algorithms also need too many examples to be able to learn from a single human in real time. Tasks with objectives that can be evaluated immediately and automatically, such as parts of online marketing and design engineering, are thus easier for algorithms to learn than those with "fuzzier" or longer-term objectives, such as many managerial activities.

Finally, a constraint on any system is the action space, or "levers", available to it. Most humans can walk and speak, many can sign contracts, and some can get the president on the phone. Algorithms run on silicon chips. They interact with the world only by issuing electronic commands to other systems via human-constructed interfaces. These interfaces may be costly to create, such as when physical equipment or legacy software systems are involved (David, 1990). Prudence or regulation may limit interfaces that pose safety, operational, financial, or other kinds of risk. And many activities, such as those of investigative journalists, executive recruiters, or managers, depend on human relationships, which are hard for algorithms to acquire.

How does artificial intelligence differ from software or robotics? With artificial intelligence, the human programmer defines the learning algorithm, which then learns for itself from data or experimentation how to achieve an objective. With software, as we mean it here, the human programmer must code a sequence of steps that directly completes the task (Autor, Levy, and Murnane, 2002). Many applications, such as algorithms for self-driving cars, involve a mixture of both, so the distinction is not always clean. In the labor economics literature, tasks whose execution steps can be expressed as a codified sequence of instructions by a programmer, such as tracking inventory, are known as "routine" (Autor, Levy, and Murnane, 2003). Tasks for which it is impractical to write down the execution steps, such as describing images, are known as "non-routine". Software can therefore carry out routine tasks, in this sense, but not non-routine ones. In principle, artificial intelligence should be capable of some non-routine tasks, and this is exactly what we find using our method. Finally, we think of robots as any physical machine controlled by a computer. This control could be by software or artificial intelligence.

3 Methodology

3.1 Overview

To assess the exposure of occupations to artificial intelligence, software, and robots empirically, we use the text of patents to identify the capabilities of each technology, then quantify the extent to which each occupation in the economy involves these capabilities. Figure 1 provides an overview of our method. We first walk through the figure at a high level, before getting into the details.

We define capabilities as verb-object pairs, such as "diagnose disease", and call such pairs "capability pairs". On the patent side, we first choose the set of patents corresponding to a particular technology. For example, we define the set of artificial intelligence patents as those that use certain keywords, such as "neural network", in their titles or abstracts. We extract all the titles from this set of patents. Such patent titles might include "Method for diagnosing diseases" and "Method for recognizing aircraft". From this list of titles, we extract all capability pairs. This results in a long list of pairs, such as (diagnose, disease), (recognize, aircraft), and so on. For any given pair, we can calculate how often that pair, or ones similar to it, occurs in the list of all pairs. (We explain how we group "similar" pairs below.) For example, pairs similar to (diagnose, disease) might represent 0.1% of all pairs extracted from the titles of our set of artificial intelligence patents.

We now turn to occupations. In our database of occupations, any given occupation, such as "doctor", consists of a collection of tasks. Each task is described in free-form text, such as "Interpret

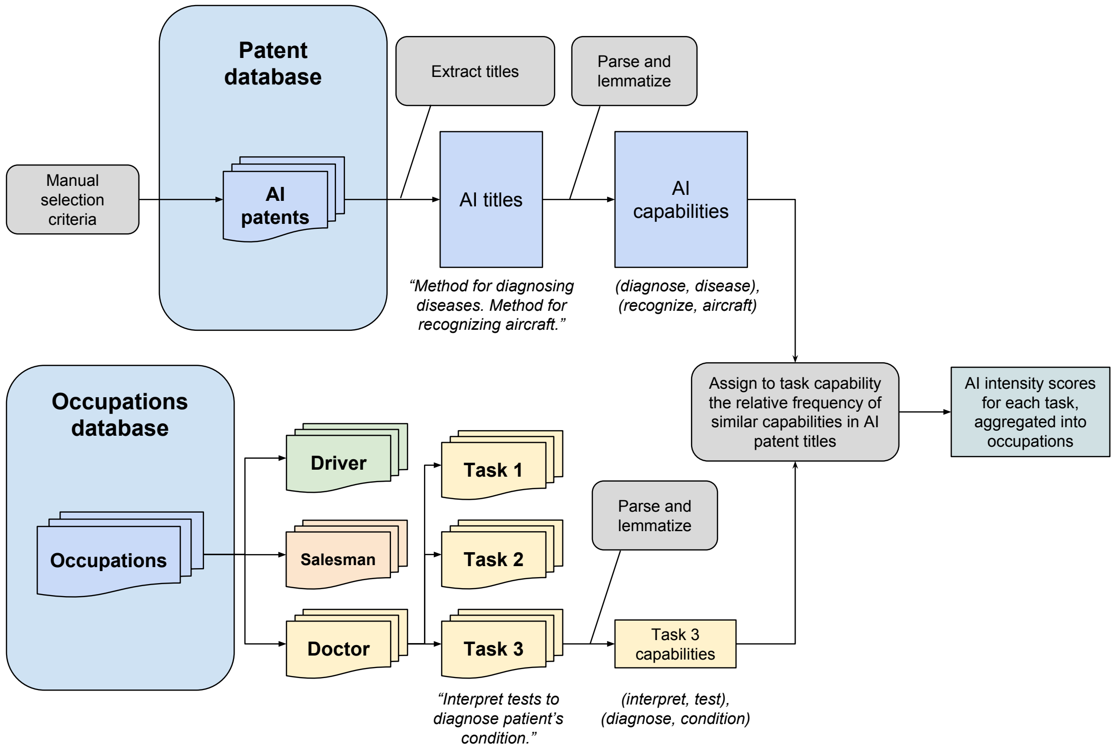

Figure 1: Illustration of process for constructing technology exposure measures

tests to diagnose patient's condition". From each task description, we extract all capability pairs. For the task just mentioned, the pairs would be (interpret, test) and (diagnose, condition). Most occupations have 20-40 extracted pairs in total. To each pair, we assign the relative frequency of similar pairs in our patent titles. For example, if pairs similar to (diagnose, condition) represented 0.1% of all pairs extracted from the titles of our set of artificial intelligence patents, we would assign a score of 0.001 to that capability pair. To get a single overall exposure score for the "doctor" occupation, we take an average of all the capability pairs mentioned in the task descriptions of that occupation, weighted by the "importance" of the task to the occupation (defined below).

3.2 Extracting capabilities from patents

We select patent publications corresponding to each technology from the Google Patents database using a keyword search of patent titles and abstracts as defined in Table A1 in the online appendix.[1] We use a keyword search because there is no patent classification code (or collection of codes) that directly corresponds to any of our three technologies. We follow Bessen and Hunt (2007) in our choice of keywords to identify software patents, and choose the keywords for artificial intelligence and robotics by hand. Manual checks suggest that our keywords work well for identifying applications

[1]Data accessed via Google BigQuery at https://bigquery.cloud.google.com/table/patents-public-data:patents.publications_201710.

Table 1: **Top extracted verbs and characteristic nouns.**

AI		Software		Robots	
Verb	**Example nouns**	**Verb**	**Example nouns**	**Verb**	**Example nouns**
recognize	pattern, image, speech, face, voice, automobile, emotion, gesture, disease	**record**	information, data, position, log, reservation, location, transaction	**clean**	surface, wafer, window, glass, floor, tool, casting, instrument
predict	quality, time, performance, fault, behavior, traffic, prognosis, treatment	**store**	program, data, information, image, instruction, value	**control**	robot, arm, motion, position, manipulator, motor, path, force
detect	signal, abnormality, defect, object, fraud, event, spammer, human, cancer	**control**	access, display, unit, image, device, power, motor	**weld**	wire, part, tong, electrode, sensor, component, nozzle
identify	object, type, damage, illegality, classification, relationship, importance	**reproduce**	data, picture, media, file, sequence, speech, item, document, selection	**move**	robot, body, object, arm, tool, part, substrate, workpiece
determine	state, similarity, relevance, importance, characteristic, strategy, risk	**detect**	defect, error, malware, fault, condition, movement	**walk**	robot, structure, base, stairs, circuit, trolley, platform, maze
control	process, emission, traffic, engine, robot, turbine, plant, discharging	**generate**	data, image, file, report, map, key, password, animation, diagram	**carry**	substrate, wafer, tray, vehicle, workpiece, tool, object, pallet
generate	image, rating, lexicon, warning, description, recommendation	**measure**	characteristic, rate, performance, time, distance, thickness	**detect**	position, state, collision, obstacle, force, angle, leak, load, landmine
classify	data, object, image, pattern, signal, text, electrogram, speech, motion	**receive**	signal, data, information, message, order, request, instruction, command	**drive**	unit, wheel, motor, belt, rotor, vehicle, automobile, actuator

Notes: For each of artificial intelligence (AI), software, and robots, this table lists the top eight verbs by pair frequency extracted from the title text of patents corresponding to the technology, together with characteristic direct objects for each verb chosen manually to illustrate a range of applications. Patents corresponding to each technology are selected using a keyword search. A dependency parsing algorithm is used to extract verbs and their direct objects from patent titles.

of these technologies. We restrict our search to the earliest-filed published patent document within each patent family. Descriptive statistics for our extracted patents are displayed in Table A2 in the online appendix.

We extract capability pairs only from patent titles. This is because titles have a much higher signal-to-noise ratio than the other patent text fields. Specifically, a patent's title tends to express the main application of the invention, whereas the abstract, description, and claims contain technical implementation details that are irrelevant for our purposes.

To extract capability pairs from any given sentence (such as a patent title), we perform the following sequence of steps. First, we use a dependency parsing algorithm (Honnibal and Johnson, 2015) to determine the syntactic relations of the words in the sentence. This algorithm attains 91.85% accuracy on the standard dependency parsing benchmark used in the natural language processing literature. Next, for each verb, we select its direct object as identified by the algorithm, if it exists, and store the resulting pair. We then lemmatize the verb, so that, say, "predicting" and "predicted" are both recorded as "predict"; and lemmatize the noun, so that, say, "person" and "people" are both recorded as "person". Stop words such as "use" and "have", which do not express capabilities, are dropped, as is common in the natural language processing literature (Jurafsky and Martin, 2014). Note that the entire process just described is fully automated. Table A3 in the online appendix displays example artificial intelligence patent titles and their extracted capability pairs.

Table 1 illustrates the most frequent verbs and characteristic nouns for each technology extracted using this method. In line with our conceptual framework, software is described as performing routine tasks such as recording, storing, and reproducing information. Artificial intelligence, by contrast, is described performing non-routine tasks such as recognizing speech and physical objects, and predicting faults and prognoses. Robots are described performing physical activities such as cleaning surfaces, welding parts, and moving workpieces.

3.3 Measuring the exposure of occupations

We use the O*NET database, produced by the US Department of Labor, as our source of information on occupations. In the database, each occupation consists of a collection of tasks described in natural language. Table 2 illustrates some of the component tasks of precision agriculture technicians, an occupation that we will find has high exposure to artificial intelligence. For each task, such as "analyze geospatial data to determine agricultural implications of [various factors]", we use the same dependency parsing algorithm as for patents to extract capability pairs. The table displays these extracted pairs for each task. Figure A1 in the online appendix plots the distribution of the number of capability pairs extracted across occupations. The vast majority of occupations (92%)

Table 2: **Tasks and exposure scores for precision agriculture technicians.**

Task	Weight in occupation	Extracted pairs	AI exposure score x100
Use geospatial technology to develop soil sampling grids or identify sampling sites for testing characteristics such as nitrogen, phosphorus, or potassium content, ph, or micronutrients.	0.050	(develop, grid)	0.050
		(identify, site)	0.234
		(test, characteristic)	0.084
Document and maintain records of precision agriculture information.	0.049	(maintain, record)	0.000
Analyze geospatial data to determine agricultural implications of factors such as soil quality, terrain, field productivity, fertilizers, or weather conditions.	0.048	(analyze, datum)	0.469
		(determine, implication)	0.837
Apply precision agriculture information to specifically reduce the negative environmental impacts of farming practices.	0.048	(apply, information)	0.000
		(reduce, impact)	0.151
Install, calibrate, or maintain sensors, mechanical controls, GPS-based vehicle guidance systems, or computer settings.	0.045	(maintain, sensor)	0.000
Identify areas in need of pesticide treatment by analyzing geospatial data to determine insect movement and damage patterns.	0.038	(identify, area)	0.234
		(analyze, datum)	0.469
		(determine, movement)	0.502

Notes: Table displays six of the twenty-two tasks recorded for precision agriculture technicians in the O*NET database. For each task, the weight is an average of the frequency, importance, and relevance of that task to the occupation, as specified in O*NET, with weights scaled to sum to one. The capability pairs in the third column are extracted from the task text by a dependency parsing algorithm. The AI exposure score for an extracted pair is equal to the relative frequency of similar pairs in the titles of AI patents. The score multiplied by 100 is thus a percentage; for example, pairs similar to "determine implications" represent 0.84% of pairs extracted from AI patents.

have more than 15 extracted pairs in total, and most (76%) have more than 20 extracted pairs.

Before calculating an exposure score for each capability pair, we first group the nouns in each pair into conceptual categories. We do this because the nouns used in O*NET task descriptions are quite general, whereas the nouns used in patents vary in their generality. For example, if an O*NET task capability pair refers to "animals", and a patent capability pair refers to "mammals", we want to account for the fact that "mammal" is an instance of "animal". This would not be possible using a thesaurus, since "mammal" and "animal" are not synonyms.

Instead, we use WordNet (Miller, 1995), a database developed at Princeton University that groups nouns into a hierarchy of concepts. For example, the ancestors of "economist" are "social scientist", "scientist", "person", "causal agent", "physical entity", and "entity". At each conceptual level, the conceptual categories are mutually exclusive. This allows us to assign each of the nouns occurring in our capability pairs to a single conceptual category, for a given conceptual level. We use "aggregated capability pair" to refer to a pair consisting of a verb and a noun conceptual category.

For the conceptual level that includes "person", for example, "recognize economist" would be part of the aggregated capability pair "recognize person".

In choosing the conceptual level at which to group the nouns, we face a trade-off between specificity and coverage. For example, if we group into categories at the conceptual level of "dog", we lose all words that exist only at a more general level, such as "mammal" and "animal". Figure A2 in the online appendix displays the share of capability pairs extracted from O*NET tasks that would be lost for this reason at each level of aggregation. Due to the level of generality at which O*NET tasks are expressed, we would lose more than a quarter of all capability pairs if we grouped at WordNet level 5, for example. (Levels with higher numbers are more specific.) We therefore use WordNet level 3 for our main results, and re-run our analyses at levels 2, 4, and 5 to check their sensitivity. While the level of aggregation does make some difference, the results for these other levels are qualitatively very similar to our baseline specification. These robustness checks are reported in Section F.2 in the online appendix.

We now describe how we go from the set of aggregated capability pairs extracted from an occupation's task descriptions to final occupation exposure scores. Denote the set of technologies T. For a given technology, $t \in T$, let f_c^t denote the raw count of occurrences of aggregated capability pair c extracted from technology t patent titles, and let C^t denote the full set of aggregated capability pairs for technology t. The relative frequency, rf_c^t, of aggregated capability pair c in technology t patent titles is

$$rf_c^t = \frac{f_c^t}{\sum_{c \in C^t} f_c^t}.$$

As an alternative to relative frequency, we can instead use the inverse rank as the exposure score. When we do this, our results are very similar. This robustness check is reported in the online appendix.

We assign to each of the *occupational* aggregated capability pairs that pair's relative frequency, rf_c^t, in technology t patent titles. These scores for artificial intelligence are displayed in the final column of Table 2.

For each occupation i, we take a weighted average of these capability-level scores to produce an overall technology t exposure score for the occupation,

$$Exposure_{i,t} = \frac{\sum_{k \in K_i} \left[w_{k,i} \cdot \sum_{c \in S_k} rf_c^t \right]}{\sum_{k \in K_i} \left[w_{k,i} \cdot |\{c : c \in S_k\}| \right]}.$$

In this expression, K_i is the set of tasks in occupation i, and S_k is the set of capabilities extracted from task $k \in K_i$. Finally, $w_{k,i}$, the weight of task k within occupation i, is an average of the frequency, importance, and relevance of task k to occupation i, as specified in the O*NET database, with weights scaled to sum to one.

An occupation's exposure score for technology t thus expresses the intensity of patenting activity in technology t directed towards the tasks in that occupation. Figure A3 in the online appendix shows the distribution of AI scores across occupations.

As an out-of-sample validation, we compare our results for lawyers to those of Remus and Levy (2017), who conduct a careful manual exercise for that occupation similar in spirit to ours. The two sets of results conform closely. Using a binary classification metric, and weighting by amount of time spent on each task, the percentage of tasks for which our method correctly predicts their exposure classification is 95.4%. Further details are provided in Section D.1 in the online appendix.

4 Occupation-level results

In line with our conceptual framework, occupations with high artificial intelligence exposure scores include gas and power plant operators and power dispatchers; logistics managers and chemical engineers; and fraud and credit analysts. They also include political scientists and astronomers, disciplines already making significant use of machine learning. Occupations with low scores include chief executives, baristas, fitness trainers, account managers, flight attendants, university professors, and home health aides.

To study these scores more systematically, we regress our exposure scores for each technology on the five occupation routineness measures developed in Autor, Levy, and Murnane (2003) and Acemoglu and Autor (2011). These routineness measures are hand- crafted composites of job characteristic scores provided in O*NET; for example, the "routine cognitive" measure combines O*NET scores for how structured an occupation's work is and the importance of "repeating the same tasks". Further details on the construction of these measures are provided in Section E in the online appendix.

The results are plotted in Figure 2. Consistent with our conceptual framework, the coefficient on the non-routine cognitive analytic measure is exceptionally high for artificial intelligence and much lower for software, suggesting that artificial intelligence has a set of non-routine capabilities that fundamentally distinguish it from software. The coefficients on the interpersonal measure are low for all three technologies, suggesting that, for now, interpersonal tasks will be harder to automate than other activities. The high non-routine manual scores reflect the fact that this score captures controlling equipment, which is a capability to some extent of all three technologies. As a second validation exercise, we directly assess our software exposure scores against the findings in

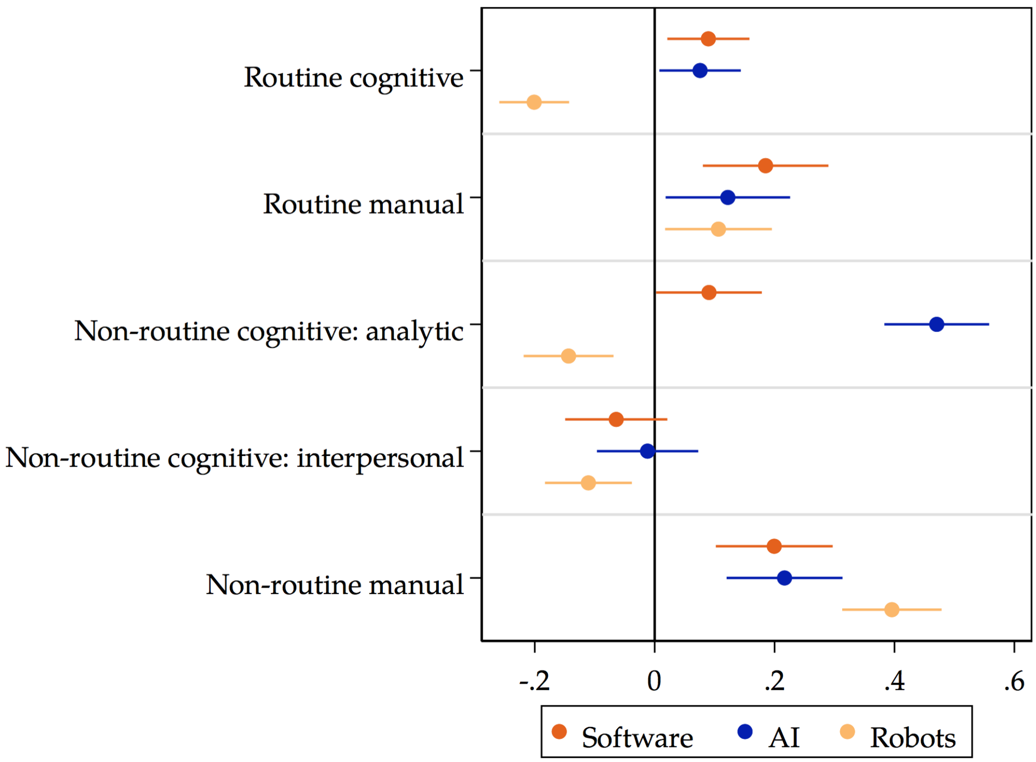

Figure 2: **Occupation-level regression results.**

Notes: Plot shows the coefficient estimates and 95% confidence intervals from each of three regressions, one per technology. The specification is $Exposure_{i,t} = RoutineCog_i + RoutineMan_i + NonRoutineCogA_i + NonRoutineCogI_i + NonRoutineMan_i + \epsilon_{i,t}$, where i indexes occupation, t indexes technology, and $Exposure_{i,t}$ measures the intensity of patenting in technology t directed towards the capabilities involved in occupation i. Exposure scores and routineness measures are each standardized to have a mean of zero and a cross-occupation standard deviation of one. For the standardizations and regressions, observations are weighted by employment reported in the May 2016 Occupational Employment Statistics. Each regression has $n = 961$ occupations, and the R^2 statistics are 14.1%, 14.7%, and 38.2% for software, AI, and robots respectively.

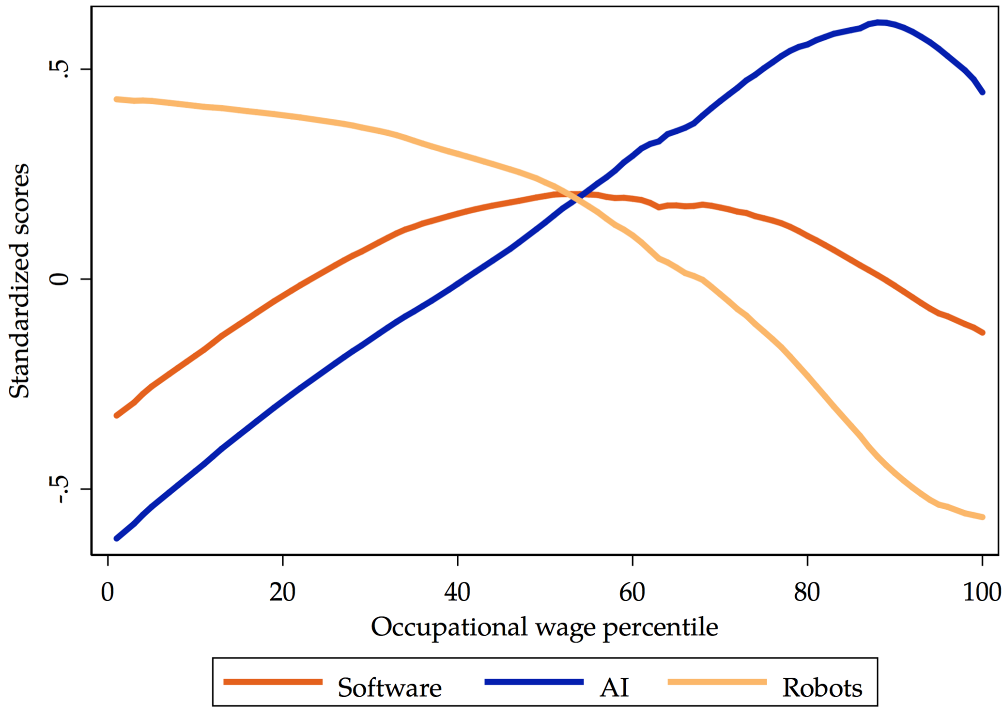

Figure 3: **Smoothed scores by occupational wage percentile.**

Notes: Plot shows the average of standardized occupation-level exposure scores for each technology by occupational wage percentile rank using a locally weighted smoothing regression (bandwidth 0.8 with 100 observations), following Acemoglu and Autor (2011). Exposure scores and their standardization are as described in Figure 2. Wage percentiles are measured as the employment-weighted percentile rank of an occupation's mean hourly wage in the May 2016 Occupational Employment Statistics.

Autor, Levy, and Murnane (2003), and find that our method independently recovers that paper's key qualitative results. Further details are provided in Section D.2 in the online appendix.

We match our occupation-level scores into the 2015 American Community Survey and May 2016 Occupational Employment Statistics, two datasets that provide wage and employment data linked to occupations. Figure 3 plots the average exposure score for each technology by occupational wage percentile. Robotics capabilities, as measured by patenting intensity, are most common in low-wage occupations, software capabilities in middle-wage occupations, and artificial intelligence capabilities in high-wage occupations. This result for software is consistent with the routinization hypothesis (Autor et al., 2006). The result for robots reflects the fact that manual work tends to be lower-paid. It is striking that the artificial intelligence scores peak around the 90th wage percentile, suggesting that artificial intelligence may have substantially different effects on jobs, wages, and inequality compared to these other technologies. We also find that, for all three technologies, on average, men work in occupations with much higher exposure scores than women. These supplementary results are described in Section C in the online appendix.

5 Conclusion

What do these findings suggest about the future economic impacts of these technologies? Our exposure measure identifies occupations that are likely to be affected by a technology, but not how they will be affected. The high artificial intelligence score for astronomers, for example, seems more likely to reflect the technology's usefulness as a tool for astronomers rather than its potential to replace them. More generally, an algorithm that diagnoses medical conditions could replace doctors; it could help them make better diagnoses; or it could lead them to specialize in different activities. Our results should thus not be taken as predicting automation or job losses, but as highlighting areas where some kind of impact, which could be positive or negative, should be expected.

The factors that affect whether a given technological capability ends up complementing or substituting for human labor are not yet well-understood. The economic environment, including factor prices and consumer preferences, aspects of the production process, such as the complementarities of tasks within occupations, and institutional features, such as regulation, are all likely to play a role (Lindbeck and Snower, 2000; Bessen, 2016).

A technology's effects are also not limited to existing tasks within occupations. The functionality of an occupation may be automated by redesigning the production process altogether (Bresnahan, 1999). For example, a railway ticket inspector may be replaced not by a humanoid robot capable of traversing a train carriage, but by a ticket barrier. Changes can also occur at the level of entire industries. The automobile reduced demand for horse breeders, but not by automating horse breeding. Technologies can perform activities that were never cost-effective for humans to do, such as label billions of photographs each day. (Our method could be inverted to find the capabilities of technologies that do *not* feature in existing occupations.) And they can create demand for new occupations, as the automobile did for car dealers, mechanics, and roadside restaurant workers (Gordon, 2016).

The roles of these factors are likely to vary over time and across capabilities, such that there is probably no simple mapping between the invention of a technological capability and equilibrium outcomes. The lack until now of a general-purpose measure that directly links specific technologies to specific economic activities in a consistent and automated fashion has hindered research on these questions. Our measure, combined with historical data, could be used as an input to a more structured modelling exercise to determine the relative importance of these different economic, technological, and institutional factors for employment and wages. This could shed light both on the past and on the likely effects of new technological capabilities. Our measure could also be used to investigate the potential for artificial intelligence to benefit other technological fields, themselves measured using patents; and could be extended to other sources of text, such as technical articles or company regulatory filings, to create measures fine-tuned for specific domains.

References

Acemoglu, Daron and David H Autor, "Skills, Tasks and Technologies: Implications for Employment and Earnings," in David Card and Orley Ashenfelter, ed., *Handbook of Labor Economics*, Vol. 4, Part B, Elsevier, 2011, pp. 1043–1171.

__ **and Pascual Restrepo**, "Robots and Jobs: Evidence from US Labor Markets," Technical Report w23285, NBER Working Paper Series 2017.

Amodei, Dario, Chris Olah, Jacob Steinhardt, Paul Christiano, John Schulman, and Dan Mané, "Concrete Problems in AI Safety," June 2016.

Arntz, Melanie, Terry Gregory, and Ulrich Zierahn, "Revisiting the risk of automation," *Economics Letters*, October 2017, *159*, 157–160.

Autor, David H, "Why Are There Still So Many Jobs? The History and Future of Workplace Automation," *The Journal of Economic Perspectives*, August 2015, *29* (3), 3–30.

__ **, Frank S Levy, and Richard J Murnane**, "Upstairs, Downstairs: Computers And Skills On Two Floors Of A Large Bank," *Industrial and Labor Relations Review*, April 2002, *55* (3), 432–447.

__ **,** __ **, and** __ , "The Skill Content of Recent Technological Change: An Empirical Exploration," *The Quarterly Journal of Economics*, 2003, *118* (4), 1279–1333.

__ **, Lawrence F Katz, and Melissa S Kearney**, "The Polarization of the U.S. Labor Market," *The American Economic Review*, 2006, *96* (2), 189–194.

Bessen, James and Robert M Hunt, "An Empirical Look at Software Patents," *Journal of Economics & Management Strategy*, March 2007, *16* (1), 157–189.

Bessen, James E, "How Computer Automation Affects Occupations: Technology, Jobs, and Skills," October 2016.

Bresnahan, Timothy F, "Computerisation and Wage Dispersion: An Analytical Reinterpretation," *The Economic Journal*, June 1999, *109* (456), 390–415.

__ **and Manuel Trajtenberg**, "General purpose technologies 'Engines of growth'?," *Journal of Econometrics*, 1995, *65* (1), 83–108.

Christiano, Paul, Jan Leike, Tom B Brown, Miljan Martic, Shane Legg, and Dario Amodei, "Deep reinforcement learning from human preferences," June 2017.

David, Paul A, "The Dynamo and the Computer: An Historical Perspective on the Modern Productivity Paradox," *The American Economic Review*, 1990, *80* (2), 355–361.

Felten, Edward W, Manav Raj, and Robert Seamans, "Linking Advances in Artificial Intelligence to Skills, Occupations, and Industries," December 2017.

Frey, Carl Benedikt and Michael A Osborne, "The future of employment: How susceptible are jobs to computerisation?," *Technological Forecasting and Social Change*, 2017, *114*, 254–280.

Goldin, Claudia and Lawrence F Katz, "The Race between Education and Technology: The Evolution of U.S. Educational Wage Differentials, 1890 to 2005," March 2007.

Gordon, Robert J, *The Rise and Fall of American Growth: The U.S. Standard of Living since the Civil War*, Princeton University Press, January 2016.

Grace, Katja, John Salvatier, Allan Dafoe, Baobao Zhang, and Owain Evans, "When Will AI Exceed Human Performance? Evidence from AI Experts," May 2017.

Graetz, Georg and Guy Michaels, "Robots at Work," Technical Report 1335, CEP Discussion Papers March 2015.

Honnibal, Matthew and Mark Johnson, "An Improved Non-monotonic Transition System for Dependency Parsing," *EMNLP*, 2015, pp. 1373–1378.

House of Commons, "Robotics and artificial intelligence," Technical Report, Science and Technology Committee 2016.

Jovanovic, Boyan and Peter L Rousseau, "General Purpose Technologies," in Philippe Aghion and Steven N. Durlauf, ed., *Handbook of Economic Growth*, Vol. 1, Part B, Elsevier, 2005, pp. 1181–1224.

Jurafsky, Dan and James H Martin, *Speech and language processing*, Vol. 3, Pearson London, 2014.

Krizhevsky, Alex, Ilya Sutskever, and Geoffrey E Hinton, "ImageNet classification with deep convolutional neural networks," *Advances in Neural Information Processing Systems*, 2012, pp. 1097–1105.

Lindbeck, Assar and Dennis J Snower, "Multitask Learning and the Reorganization of Work: From Tayloristic to Holistic Organization," *Journal of Labor Economics*, 2000.

Manyika, James, Jacques Bughin, Michael Chui, Mehdi Miremadi, Katy George, Paul Willmott, and Martin Dewhurst, "Harnessing automation for a future that works," Technical Report, McKinsey Global Institute 2017.

Michaels, Guy, Ashwini Natraj, and John Van Reenen, "Has ICT Polarized Skill Demand? Evidence from Eleven Countries over Twenty-Five Years," *The Review of Economics and Statistics*, April 2013, *96* (1), 60–77.

Miller, George A, "WordNet: A Lexical Database for English," *Communications of the ACM*, November 1995, *38* (11), 39–41.

Mnih, Volodymyr, Koray Kavukcuoglu, David Silver, Andrei A Rusu, Joel Veness, Marc G Bellemare, Alex Graves, Martin Riedmiller, Andreas K Fidjeland, Georg Ostrovski, Stig Petersen, Charles Beattie, Amir Sadik, Ioannis Antonoglou, Helen King, Dharshan Kumaran, Daan Wierstra, Shane Legg, and Demis Hassabis, "Human-level control through deep reinforcement learning," *Nature*, February 2015, *518* (7540), 529–533.

Remus, Dana and Frank S Levy, "Can Robots Be Lawyers: Computers, Lawyers, and the Practice of Law," *The Georgetown Journal of Legal Ethics*, 2017, *30*, 501–558.

Rezende, Danilo Jimenez, Shakir Mohamed, and Daan Wierstra, "Stochastic backpropagation and approximate inference in deep generative models," *arXiv preprint arXiv:1401. 4082*, 2014.

Rusu, Andrei A, Matej Vecerik, Thomas Rothörl, Nicolas Heess, Razvan Pascanu, and Raia Hadsell, "Sim-to-Real Robot Learning from Pixels with Progressive Nets," October 2016.

Sutton, Richard Stuart, "Temporal credit assignment in reinforcement learning." PhD dissertation, University of Massachusetts Amherst 1984.

Syverson, Chad, "What Determines Productivity?," *Journal of Economic Literature*, June 2011, *49* (2), 326–365.

United States, "Preparing for the future of Artificial Intelligence," Technical Report, Executive Office of the President, National Science and Technology Council, Washington D.C. 20502 2016.

Online Appendix to "The Impact of Artificial Intelligence on the Labor Market"

Michael Webb

May 15, 2019

A Additional methodological details

A.1 Capability pair score descriptives

Figure A4 plots our relative frequency scores by aggregated capability pair rank for the top 100 ranked pairs. The three technologies have relatively similar distributions of frequencies, although software has a longer tail (more than 30,000 aggregated capability pairs), reflecting the much larger number of software patents. This is illustrated in Figure A5, which plots the cumulative frequency by aggregated pair rank for the 5,000 most frequent pairs for each technology. For computational reasons, and the fact that scores in the tails are very close to zero, we use only the aggregated capability pairs that account for the top 80% most frequent aggregated capability pairs. Robustness checks show that the choice of cumulative frequency cutoff does not affect the results. These robustness checks are reported in Section F.3.

A.2 Inverse rank robustness check

As a robustness check, we can use the inverse rank instead of the relative frequency as our exposure score. Let R_c^t denote the rank of aggregated capability pair c in the set of aggregated capability pairs extracted from technology t patents. That is, the rank of the most frequently-occurring aggregated capability pair is 1, of the second-most is 2, and so on. The inverse rank, ir_c^t, of aggregated capability pair c in technology t patent titles is

$$ir_c^t = \frac{1}{R_c^t}.$$

When we use inverse rank rather than relative frequency, our results are qualitatively very similar. These results are reported in Section F.4.

B Supplementary tables and figures

Table A1: **Patent selection criteria.**

Technology	Definition
AI	Title/abstract include “neural network”, “deep learning”, “reinforcement learning”, “supervised learning”, “unsupervised learning”, or “generative model”
Software	Title/abstract include “software”, “computer”, or “program” AND title/abstract exclude “chip”, “semiconductor”, “bus”, “circuit”, or “circuitry”
Robots	Title/abstract include “robot”

Notes: Patents corresponding to each technology are selected using these keyword inclusion/exclusion criteria.

Table A2: **Patent descriptives.**

	AI	Software	Robots
# Patents	16,485	1,371,026	114,496
# Words in titles	156,255	12,448,437	875,696
# Words/patent	9.5	9.1	7.6
# Extracted pairs	8,264	711,214	54,238
# Extracted pairs/patent	0.50	0.52	0.47

Notes: Table displays descriptive statistics for the patents corresponding to each technology, selected as in Table A1. Verb-object pairs are extracted automatically using a dependency parsing algorithm, as described in the main text.

Table A3: **Extracting capabilities from patent titles.**

Text	**Extracted pairs**
Adaptive system and method for predicting response times in a service environment	(predict, time)
Method of and apparatus for determining optimum delivery route for articles	(determine, route)
Methods and apparatus for reinforcement learning	
Device for forecasting total power demand	(forecast, demand)
Method and device for classifying images on basis of convolutional neural network	(classify, image)
A method for diagnosing food allergy	(diagnose, allergy)
Neural network language model training method and device and voice recognition method	
Automatic butterfly species identification system and method, and portable terminal having automatic butterfly species identification function using the same	(have, function), (use, same)

Notes: Table displays a set of titles of artificial intelligence patents, selected as in Table A1, and corresponding verb-object pairs extracted automatically using a dependency parsing algorithm, as described in the main text. Stop words such as "use" and "have", which do not express capabilities, are dropped from subsequent analysis.

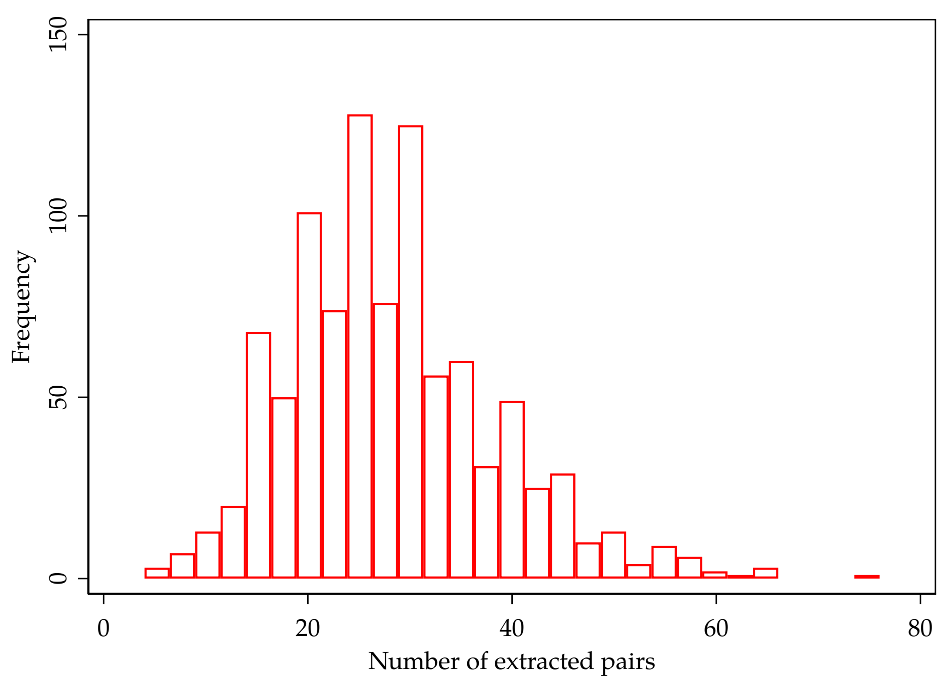

Figure A1: **Distribution of pair counts across occupations.**

Notes: Figure displays the distribution across occupations of the number of pairs extracted from an occupation's tasks in the O*NET database.

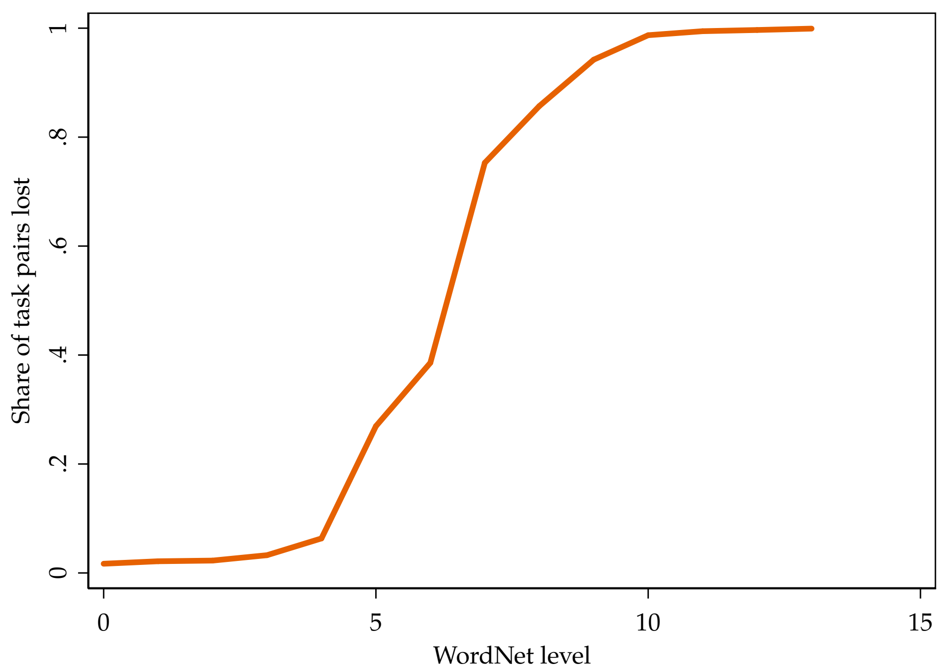

Figure A2: **Share of task pairs lost by WordNet aggregation level.**

Notes: WordNet assigns all nouns to positions in a conceptual hierarchy. For example, the ancestors of "dog" include "carnivore", "mammal", "vertebrate", "animal", and "physical entity". Our method groups together nouns in the same conceptual category, for a given WordNet level. If we aggregated at the level of "dog", we would lose all words that exist only at a more general level, such as "mammal" and "animal". This figure displays the share of capability pairs extracted from O*NET tasks that would be lost for this reason at each level of aggregation. (Levels with higher numbers are more specific.)

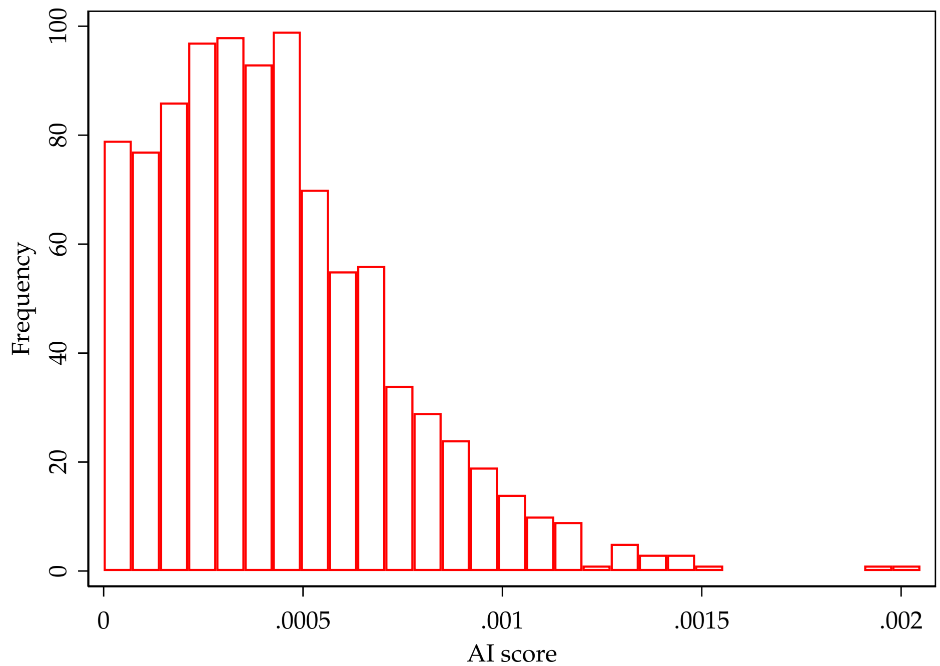

Figure A3: **Distribution of AI exposure scores across occupations.**

Notes: Figure displays the distribution across occupations of artificial intelligence exposure scores calculated using our method.

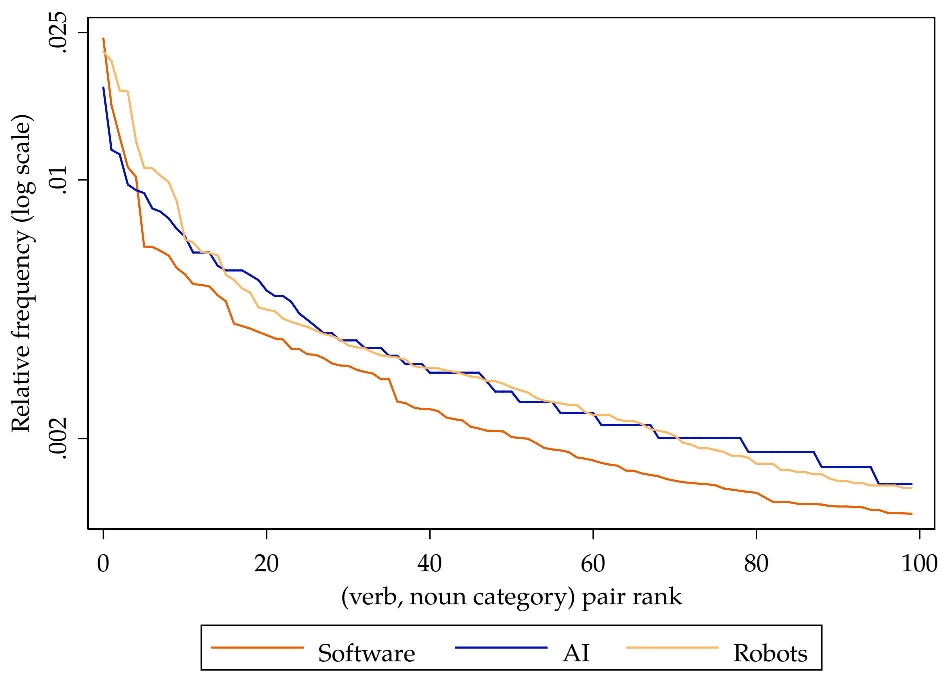

Figure A4: **Top 100 normalized aggregated capability pair scores by rank.**

Notes: Figure displays the relative frequency of the top 100 capability pairs, after grouping nouns into categories, sorted by rank for each technology.

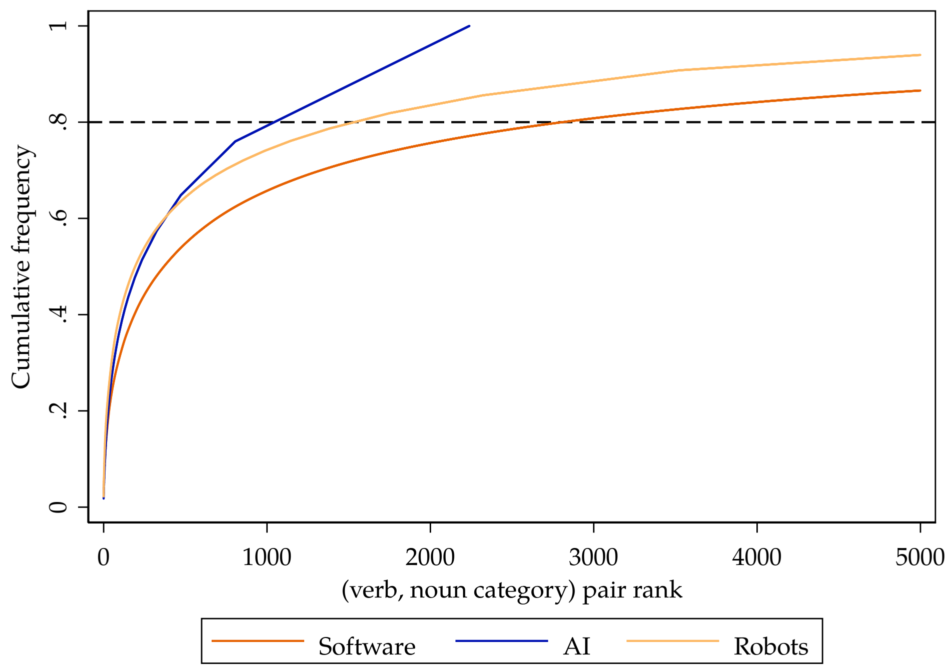

Figure A5: **Cumulative frequency of aggregated capability pair scores by rank.**

Notes: Figure displays the cumulative frequency of capability pairs, after grouping nouns into categories, sorted by rank, for the 5,000 most frequent pairs for each technology. Due to the different distributions of pair frequencies between technologies, only about 1,000 distinct pairs account for 80% of artificial intelligence capability pairs, whereas about 3,000 are needed to account for 80% of software capability pairs.

C Exposure scores by age, gender, and industry

To assess how people's propensity to work in higher- or lower-exposure occupations varies with demographic charateristics, we match our occupation-level exposure scores into the 2015 American Community Survey (ACS) 1% sample. We use the 2016 National Employment Matrix/SOC to ACS Crosswalk,[1] produced by the BLS, to link our exposure scores, which correspond to SOC 2010 occupation codes, to occupations in the ACS, which are reported using a slightly more aggregated classification. Where multiple SOC 2010 occupations correspond to a single ACS occupation, we assign to the ACS occupation the average of our exposure scores for the corresponding SOC 2010 occupations, weighted by employment as reported in the May 2016 Occupational Employment Statistics.

Figure A6 shows how average exposure varies with age for each technology. The interpretation of the figure is as follows. For each age, we take all individuals of that age with a reported occupation in the ACS and assign to them the exposure score of their occupation for a given technology. We take a simple average over individuals to produce the final score for each age. The figure plots these averages for each technology, smoothed using a locally weighted smoothing regression. The figure thus represents how the occupational mix of individuals, and hence their average exposure score, varies by age group. Whereas the propensity to work in occupations exposed to software and robots declines from the start of working life until about age 35, from which point the average exposure scores are relatively flat, the pattern is the opposite for artificial intelligence, for which the exposure score peaks around age 40. This partly reflects the greater education and experience requirements for many of the occupations most exposed to artificial intelligence.

[1] Downloaded from: https://www.bls.gov/emp/classifications-crosswalks/NEM_OccCode_ACS_Crosswalk.xlsx.

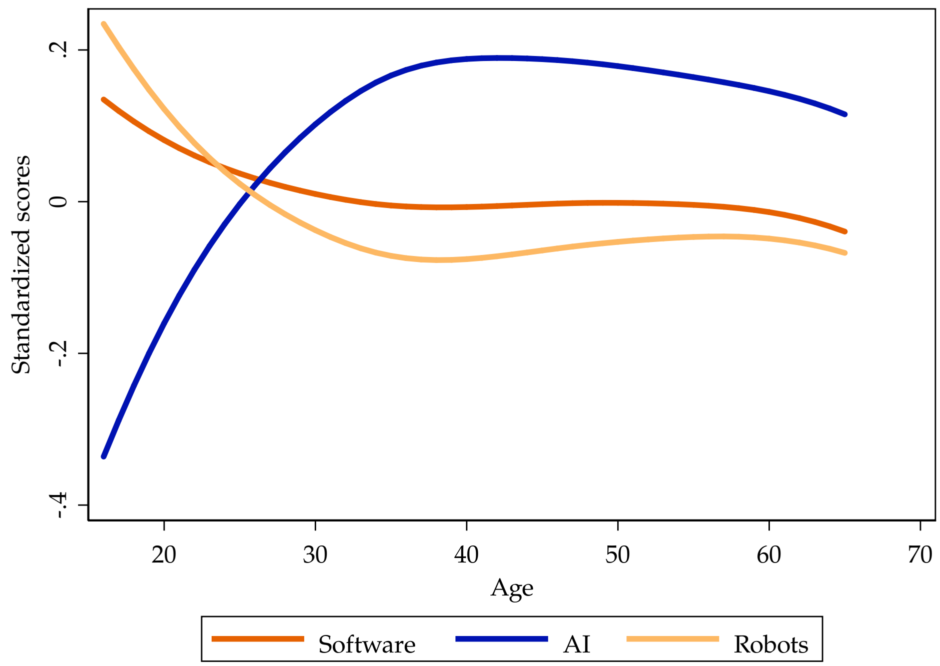

Figure A6: **Smoothed exposure scores by age.**

Notes: Plot shows the average of standardized occupation-level exposure scores for each technology by age of workers in the 2015 American Community Survey (ACS) 1% sample, smoothed using a locally weighted smoothing regression (bandwidth 0.8). Exposure scores and their standardization are as described in Figure 2 in the main text.

Figure A7 represents the same information as Figure A6, but now additionally splitting by gender. The most striking aspect of these results is the fact that for every age and every technology, men work in occupations with higher exposure scores on average than women. Indeed, for all three technologies, men spend most of their working life in occupations with above-average exposure scores, whereas women spend most or all of their working life in occupations with below-average exposure scores. Moreover, the declining slope for software observed in Figure A6 appears to be driven entirely by women. Investigating the drivers of these differences is a priority for future research.

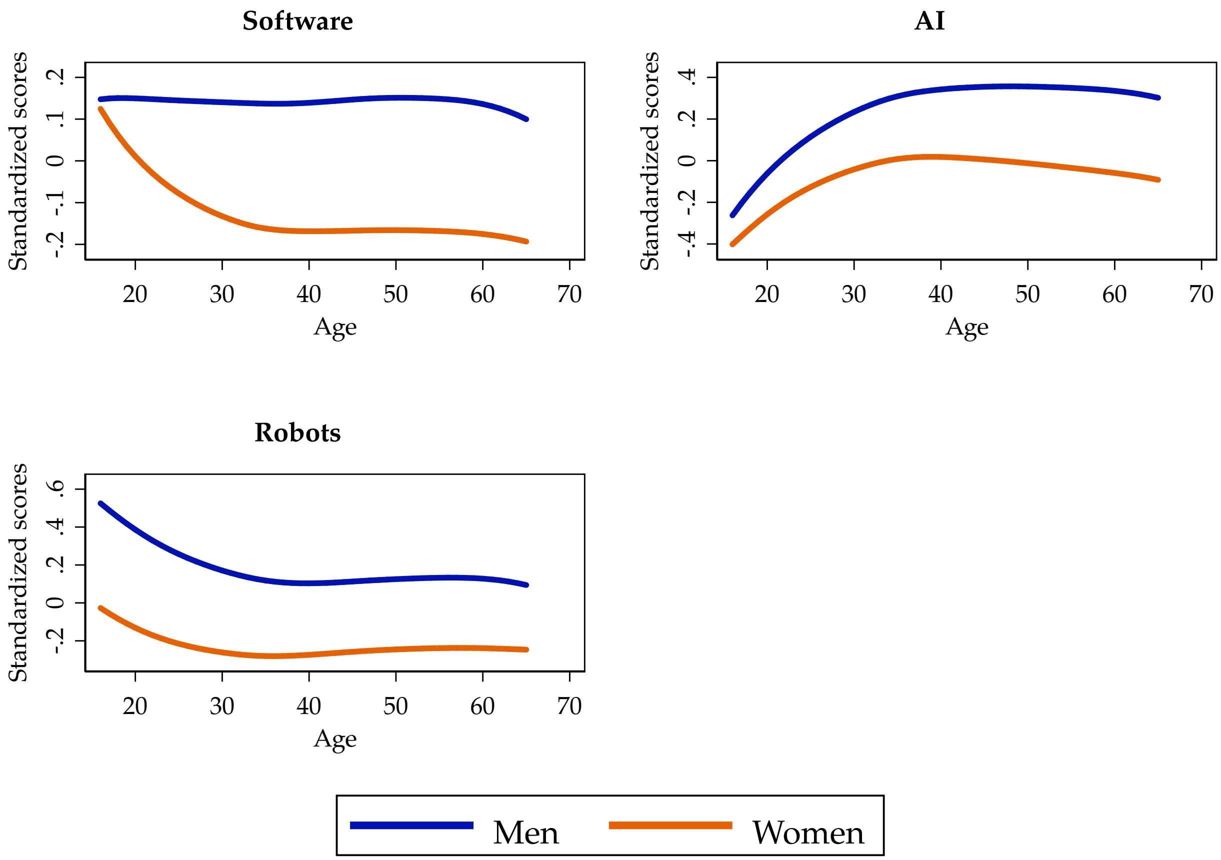

Figure A7: **Smoothed exposure scores by age and gender.**

Notes: Plot shows the average of standardized occupation-level exposure scores for each technology by age and gender of workers in the 2015 American Community Survey (ACS) 1% sample, smoothed using a locally weighted smoothing regression (bandwidth 0.8). Exposure scores and their standardization are as described in Figure 2 in the main text.

Finally, Figure A8 displays average exposure scores for artificial intelligence by industry. Industry scores are calculated by taking a simple average of the scores of individuals reported as working in each industry, with individuals assigned the exposure score of their occupation. The figure thus depicts the average exposure of the occupational mix of each industry. The high scores for utilities, agriculture, and manufacturing reflect the fact that many of the occupations in those industries involve inspection, such as for classification or anomaly detection (i.e., testing and quality assurance), or are technical or engineering roles that involve control of industrial systems, such as gas or power plant operators, or various kinds of analysis and optimization, such as the precision agriculture technicians occupation reported in Table 2 in the main text.

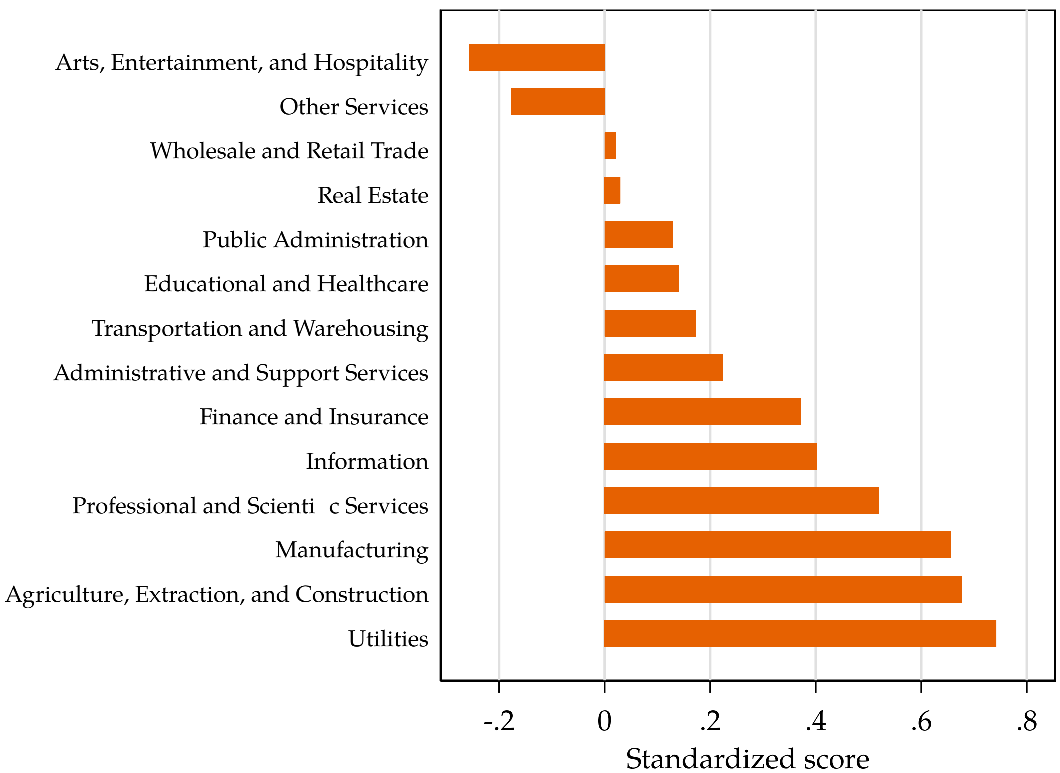

Figure A8: **Artificial intelligence exposure scores by industry.**

Notes: Plot shows the average of standardized occupation-level exposure scores for each technology by industry of workers in the 2015 American Community Survey (ACS) 1% sample. Exposure scores and their standardization are as described in Figure 2 in the main text.

D Out-of-sample validations

D.1 Comparison to Remus and Levy (2017)

Table A4: **Comparison to Remus and Levy (2017).**

Task (RL categorization)	**% invoiced hours**	**Corresponding O*NET extracted capability pairs**	**Potential impact (RL)**	**Mean AI score x100**
Negotiation	5.0%	negotiate settlements	Light	0.00
Other communications/interactions	5.0%	interview clients	Light	0.00
Fact investigation	9.6%	ascertain facts, establish ownership, examine records, gather evidence	Light	0.00
Advising clients	3.2%	advise clients, interpret laws [for clients], advise executors	Light	0.00
Legal writing	17.7%	prepare [legal] briefs, write opinions	Light	0.00
Court appearances and preparation	14.5%	argue motions, defend clients, defend lawsuits, formulate defense, present evidence, prosecute defendants, prosecute lawsuit, represent clients [in court], summarize cases [to judges and juries]	Light	0.00
Document drafting	4.0%	prepare draft, review documents	Moderate	0.00
Legal analysis and strategy	27.4%	analyze outcomes [of cases], determine advisability [of lawsuit], determine ramifications, develop strategies, evaluate findings, examine data, study constitution, verify bases [for legal proceedings]	Moderate	0.12

Notes: For each lawyer task category in Remus and Levy (2017), this table displays the percentage of hours invoiced by law firms for that activity, using data from Sky Analytics; the capability pairs extracted from the O*NET occupation 'Lawyers' using our method, matched manually to the Remus and Levy (2017) categorization; the Remus and Levy (2017) estimated potential impact of automation; and the mean of the artificial intelligence exposure scores calculated using our method for the corresponding capability pairs, multiplied by 100. The invoiced hours refer to hours of tier two-five firms; those of tier one firms are almost identical. The invoiced hours percentages sum to 86.4% because some infrequent task categories in the Remus and Levy (2017) categorization do not have any corresponding capability pairs in the O*NET data, as described in the text. The legal analysis and strategy category incorporates legal research (0.4% of invoiced hours), as described in the text. Converting an exposure score of 0 to "Light" and a score greater than 0 to "Moderate", the percent correctly predicted is 87.5%, and the percent correctly predicted weighted by share of invoiced hours is 95.4%. Any attempt to quantify the similarity between a quantitative and qualitative set of results is somewhat arbitrary, so these numbers should be taken merely as indicative.

As one validation exercise, we compare the output of our method with the findings of Remus and Levy (2017). Those authors conduct a very careful manual study of the work activities of lawyers, and assess which lawyering tasks are likely to be automated by machine learning applications. They conducted a series of interviews with computer scientists, legal technology developers, and practicing lawyers. They also obtained lawyer time usage data from Sky Analytics, a consulting firm that aggregates invoices billed by law firms for corporate clients. Each invoice specifies the task the lawyer performed for each increment of time billed, with tasks classified according to the American Bar Association's Uniform Task-Based Management System (UTBMS). The authors aggregate the 114 task codes of the UTBMS into thirteen categories, and manually classify each category as likely to face light, moderate, or strong automation impacts. Their main result (Table 2 in their paper) consists of these thirteen categories, and, for each, the percent of invoiced hours spent on the category, and the estimated potential impact of automation.

Their exercise is thus very similar in spirit to ours, with the difference that theirs is manual and specialized to lawyers, whereas ours is automated and covers the universe of occupations. We use O*NET as our source of data on the tasks involved in all occupations; they use Sky Analytics, which covers only legal tasks. They use interviews to assess the potential impact of automation on lawyers in particular; we use the text of patents to assess its potential impact on all occupations. The strength of their procedure is its accuracy: by gaining expert knowledge of technologies specific to the legal profession, reasoning carefully about the details of those technologies as they relate to the practice of law, and obtaining specialized data on lawyer time use, they are able to provide a highly nuanced assessment of the likely impact of automation on lawyers. By contrast, the strength of our procedure is its coverage: we are able to assess the tasks involved in every occupation in the economy in an automated and replicable manner. Naturally, there is a trade-off between accuracy and coverage. It is therefore instructive to consider how our analysis of lawyers' tasks compares to theirs.

To make the comparison, we manually match each of the tasks in the O*NET "Lawyers" occupation to one of the Remus and Levy (2017) categories, as in Table A4. Remus and Levy (2017) describe each of their categories in detail, making the matching a straightforward exercise. For a small number of categories, there are no corresponding O*NET tasks with extracted capability pairs. These are case administration and management (5.6% of invoiced hours), document review (3.6%), due diligence (3.4%), and document management (0.7%). Some tasks are related to both legal analysis and strategy (27.0%) and legal research (0.4%). These categories are both assigned 'Moderate' potential impacts, so we combine them into a single category.

In our final matched set, six categories are rated "Light" and two as "Moderate". Our method assigns a score of 0 to all "Light" categories, and assigns a positive score to one of the two "Moderate" categories. Overall, then, the two sets of results line up well. Converting an exposure score of 0 to "Light" and a score greater than 0 to "Moderate", the percent correctly predicted is 87.5%, and the percent correctly predicted weighted by share of invoiced hours is 95.4%. Of course, any attempt to quantify the similarity between a quantitative and qualitative set of results is somewhat arbitrary, so these numbers should be taken merely as indicative.

We conclude with some additional details and comments. The O*NET task from which the capabilities "prepare draft" and "review documents" are extracted refers to documents "such as wills, deeds, patent applications, mortgages, leases, and contracts". Remus and Levy (2017) define document drafting as "the development of legal documents such as deeds, contracts, wills, and trusts", and document review as "reviewing documents for purposes of discovery in litigation or government investigations." We therefore assign "prepare draft" and "review documents" to the document drafting category rather than the document review category. Remus and Levy (2017) assign document review "Strong" potential impacts (the only category to receive this rating). Since "review documents" has an artificial intelligence score of 0, assigning this pair to the document review category (3.6% of invoiced hours) would lower the percent correctly predicted. Manual inspection of our patent data reveals that there are indeed artificial intelligence patents for document discovery; however, their titles describe, for example, "providing electronic discovery" rather than "reviewing documents", and our automated procedure would not match "provide discovery" to "review documents". It is possible that more sophisticated algorithms could address this issue, using word embeddings, for example, but likely at the cost of transparency. This is an interesting avenue for future work.

D.2 Comparison to Autor, Levy, and Murnane (2003)

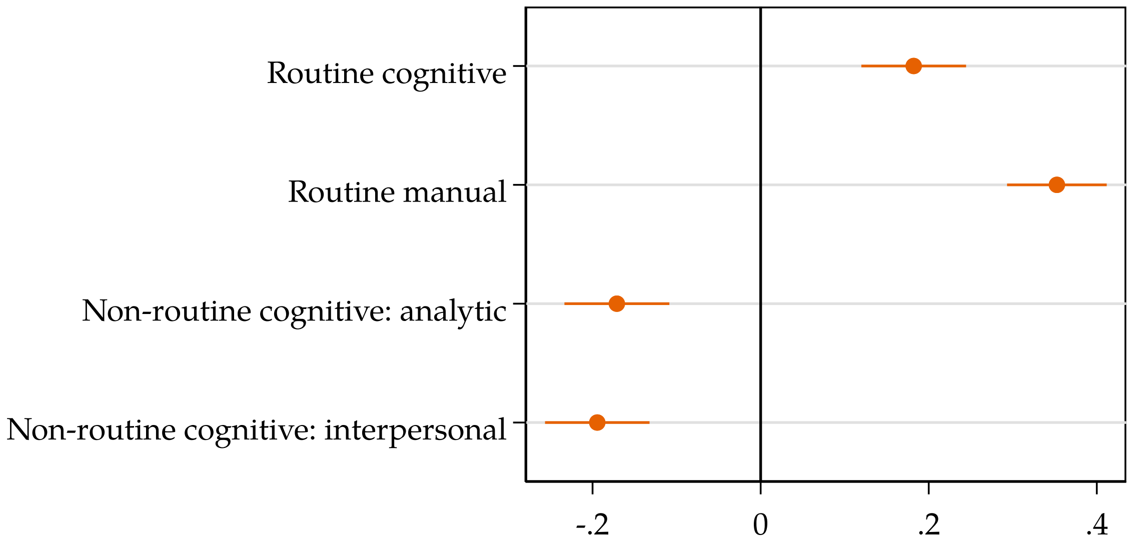

Figure A9: **Simple linear regression results for software.**

Notes: Plot shows the coefficient estimates and 95% confidence intervals from four simple linear regressions. Occupation-level routineness scores defined in Acemoglu and Autor (2011), which are updated versions of those in Autor, Levy, and Murnane (2003), are regressed on standardized occupation-level exposure scores calculated from software patent text and O*NET data. The routineness scores are functions of quantitative occupational characteristic scores provided in O*NET, as described in Section E. The specifications are $RoutineCognitive_i = Exposure_{i,Software} + \epsilon_i$, $RoutineManual_i = Exposure_{i,Software} + \epsilon_i$, and so on, where i indexes occupation. Observations are weighted by employment reported in the May 2016 Occupational Employment Statistics. Each regression has 961 observations, and the R^2 statistics are 3.3%, 12.4%, 2.9%, and 3.8% for the routine cognitive, routine manual, non-routine cognitive analytic, and non-routine cognitive interpersonal regressions respectively.

As another validation exercise, we compare the relationship between our software exposure scores and the O*NET routine/non-routine measures to the framework and results presented in Autor, Levy, and Murnane (2003). In that paper, the authors argue that software should substitute for routine cognitive and routine manual tasks, and complement non-routine cognitive tasks.

To test these theoretical predictions empirically, the authors study how the change in computer use in an industry is related to the change in the mix of jobs in that industry over various time periods. Specifically, in Table III they regress the average routine cognitive score of jobs in an industry, for various time periods, on change in computer use in that industry. They do the same for routine manual scores, non-routine cognitive analytic scores, and non-routine cognitive interpersonal scores. They exclude non-routine manual scores from their analysis.

They find that increased computer use is associated with rises in an industry's use of non-routine cognitive analytic occupations, larger rises in use of non-routine cognitive interactive occupations, falls in use of routine cognitive occupations, and even larger falls in use of routine manual occupations. They argue that this supports their theoretical predictions regarding the substitutability of software for routine occupations and its complementarity with non-routine cognitive ones.

Figure A9 shows the results of four regressions based on the measures used in this paper. In each case, observations are occupations, and the independent variable is our software exposure score for each occupation. We regress each of the four occupation-level routine/non-routine scores used in Autor, Levy, and Murnane (2003) separately on our software exposure score. We find that routine manual scores have the strongest positive association with our software exposure score, followed by routine cognitive scores, while the association between the non-routine cognitive scores and our software exposure score is negative. In other words, occupations with high routine manual and routine cognitive scores make intensive use of capabilities that feature in software patents, while occupations with high non-routine cognitive scores do not make intensive use of such capabilities. Our measure thus independently recovers the key qualitative results of the Autor, Levy, and Murnane (2003) framework, i.e., the substitutability of software for routine cognitive and routine manual occupations, using only patent text and job descriptions as input.

E Construction of routine/non-routine occupation measures

The creators of the O*NET database score each occupation along many characteristics, such as the extent to which it involves "being exact or accurate" or "coaching/developing others". These measures were hand-combined by Acemoglu and Autor (2011), updating Autor, Levy, and Murnane (2003), into composite "routine" (i.e., software-amenable) and "non-routine" scores for each occupation. We use these measures to characterize our exposure scores in Figure 2 in the main text. The component O*NET scales used to produce each score are as follows:

Routine cognitive	4.C.3.b.7 Importance of repeating the same tasks
	4.C.3.b.4 Importance of being exact or accurate
	4.C.3.b.8 Structured v. Unstructured work (reverse)
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
Non-routine cognitive: analytic	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
Non-routine manual	4.A.3.a.4 Operating vehicles, mechanized devices, or equipment
	1.A.2.a.2 Manual dexterity
	1.A.1.f.1 Spatial orientation
	4.C.2.d.1.g Spend time using hands to handle, control or feel objects, tools or controls

We use the replication code provided for Acemoglu and Autor (2011) to generate these measures for the latest version of O*NET, v22.0.

F Robustness checks

The following section reproduces Figures 2 and 3 in the main text, followed by the corresponding figures obtained by varying different elements of our baseline specification. The WordNet level checks are described in more detail in the main text, while the cumulative frequency cutoff and inverse rank checks are described in Section A.

F.1 Baseline results

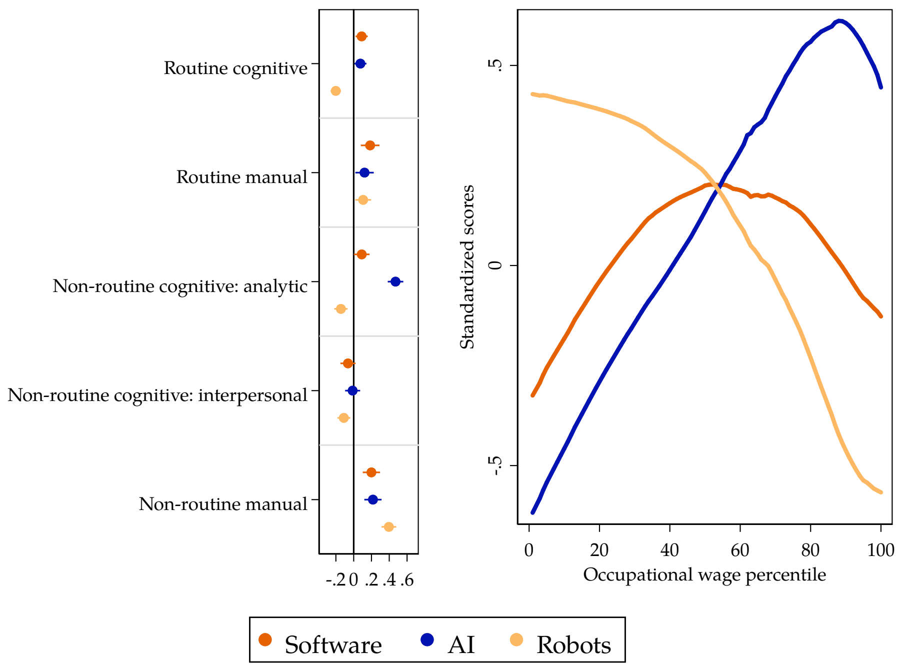

Figure A10: **Baseline results.**

Notes: The left panel shows coefficient estimates and 95% confidence intervals from each of three regressions, one per technology. The specification is $Exposure_{i,t} = RoutineCog_i + RoutineMan_i + NonRoutineCogA_i + NonRoutineCogI_i + NonRoutineMan_i + \varepsilon_{i,t}$, where i indexes occupation, t indexes technology, and $Exposure_{i,t}$ measures the intensity of patenting in technology t directed towards the capabilities involved in occupation i. For more details, see Figure 2 in the main text. The right panel shows the average of standardized occupation-level exposure scores for each technology by occupational wage percentile rank using a locally weighted smoothing regression (bandwidth 0.8 with 100 observations). For more details, see Figure 3 in the main text. In this baseline specification, nouns are grouped at WordNet level 3; the cumulative frequency cutoff for calculating capability pair scores is 80%; and capability pair scores correspond to the relative frequency of similar pairs in the titles of patents corresponding to a given technology.

F.2 WordNet levels

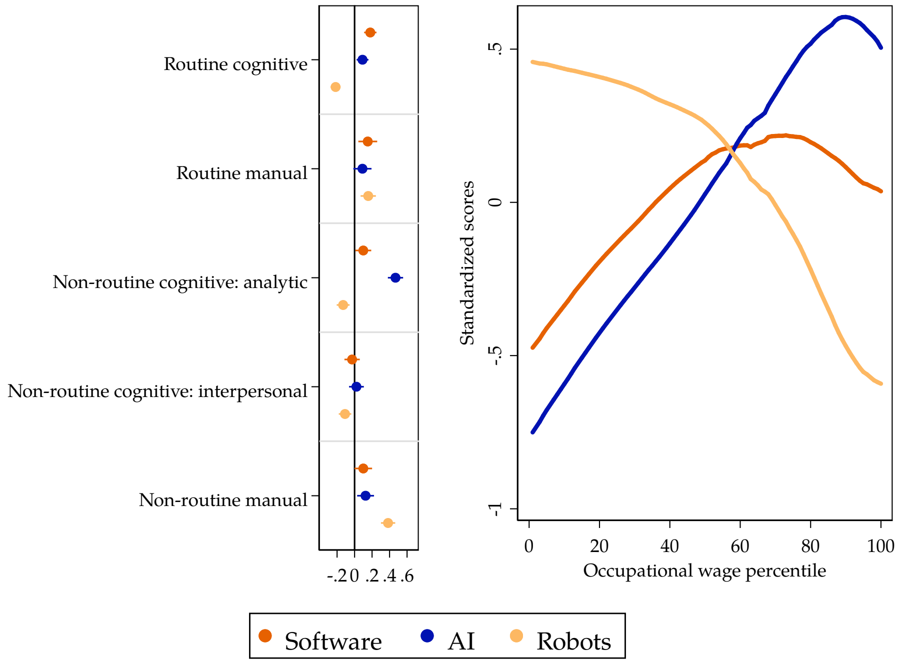

Figure A11: **WordNet level 2.**

Notes: This figure displays our results when nouns are grouped at WordNet level 2. This is a more general level of aggregation than level 3, used in our baseline specification. There are some noticeable changes in the wage percentile figure, particularly for software, but the overall patterns are very similar.

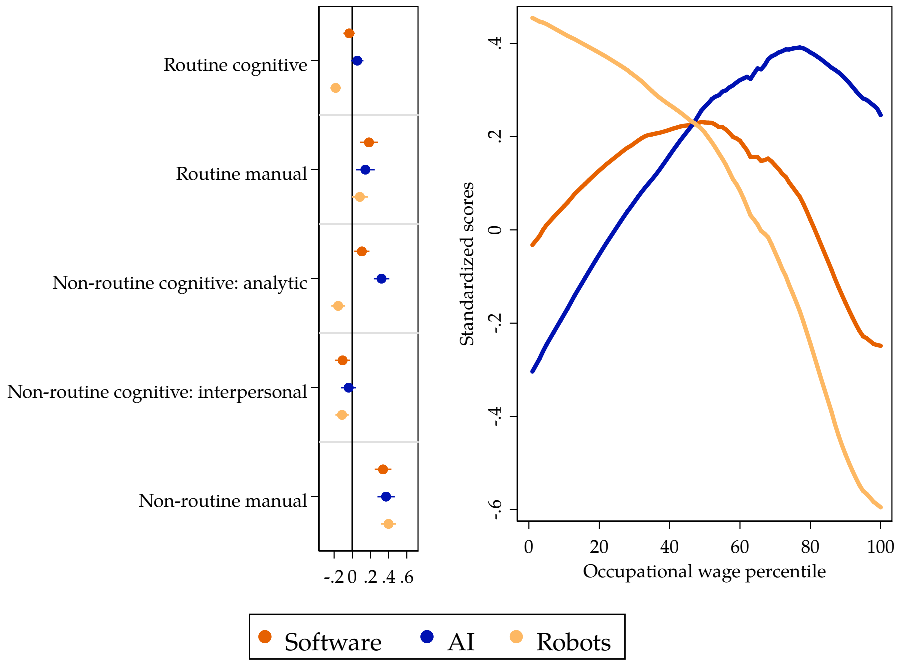

Figure A12: **WordNet level 4.**

Notes: This figure displays our results when nouns are grouped at WordNet level 4. This is a more specific level of aggregation than level 3, used in our baseline specification. Again, the results are qualitatively very similar to our baseline specification, although the slope of the robots line is a little more negative in the bottom half of the wage distribution.

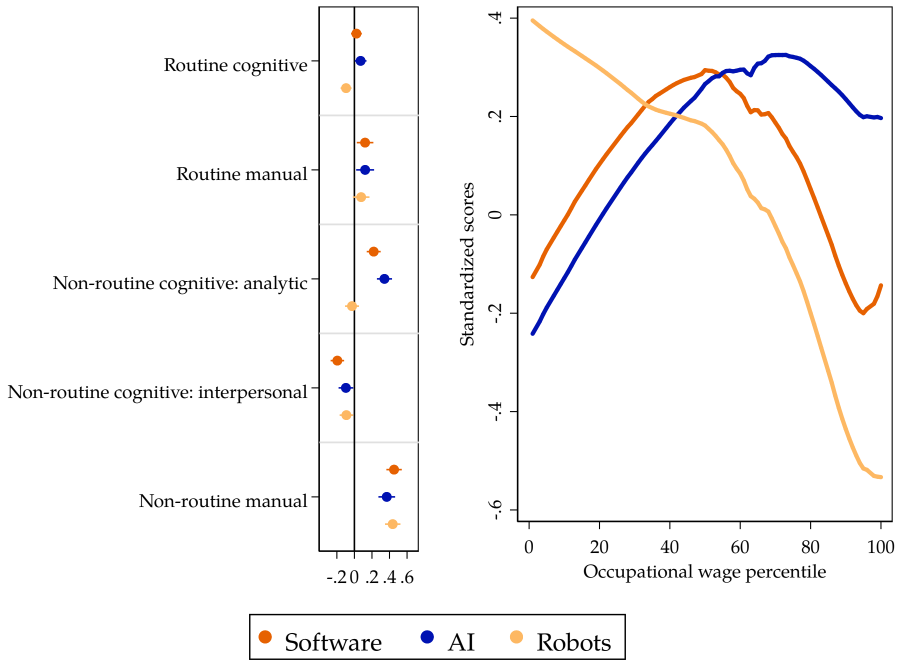

Figure A13: **WordNet level 5.**

Notes: This figure displays our results when nouns are grouped at WordNet level 5. This is a more specific level of aggregation than level 3, used in our baseline specification. There are more obvious differences to our baseline specification here; for example, the line for artificial intelligence is somewhat attenuated. Nevertheless, the overall patterns remain.

F.3 Cumulative frequency cutoff

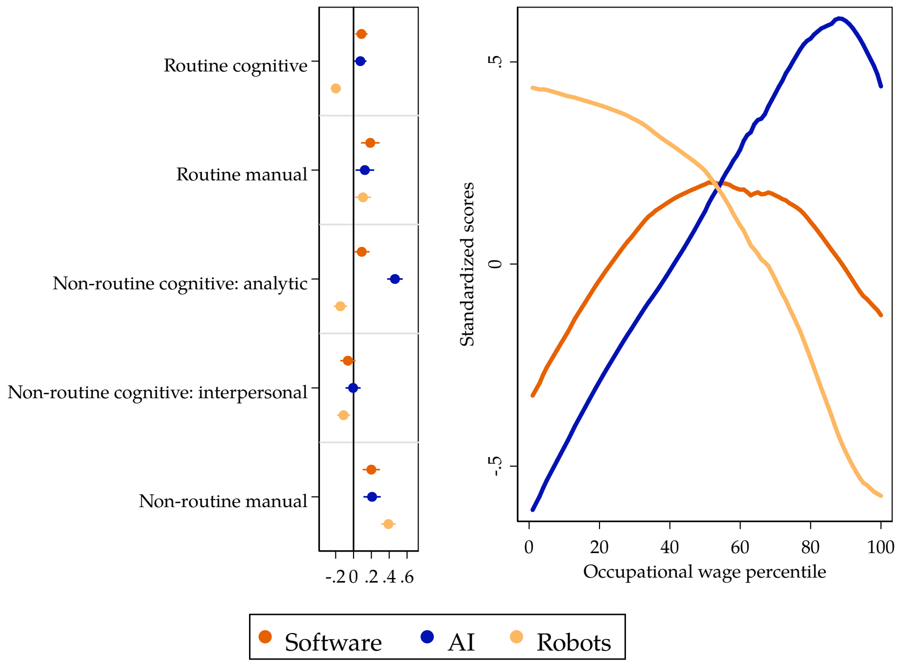

Figure A14: **Cumulative frequency cutoff = 70%.**

Notes: For computational reasons, and the fact that scores in the tails are very close to zero, we use only the aggregated capability pairs that account for the top 80% of aggregated capability pair occurrences in our baseline specification, as described in Section A.1. This figure displays our results when we instead set the cutoff to 70%. As expected, the results are essentially identical.

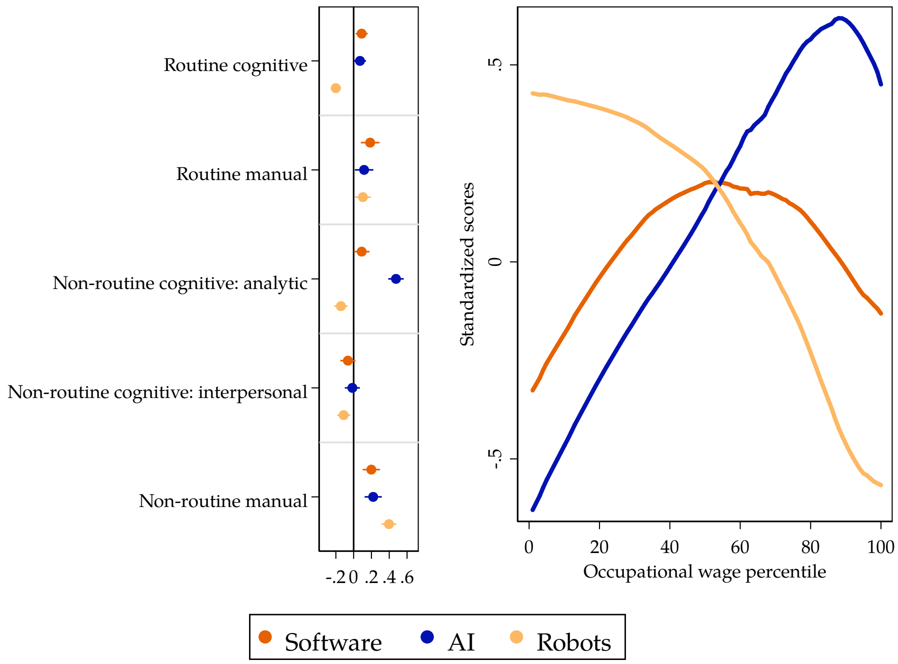

Figure A15: **Cumulative frequency cutoff = 90%.**

Notes: For computational reasons, and the fact that scores in the tails are very close to zero, we use only the aggregated capability pairs that account for the top 80% of aggregated capability pair occurrences in our baseline specification, as described in Section A.1. This figure displays our results when we instead set the cutoff to 90%. As expected, the results are essentially identical.

F.4 Inverse rank

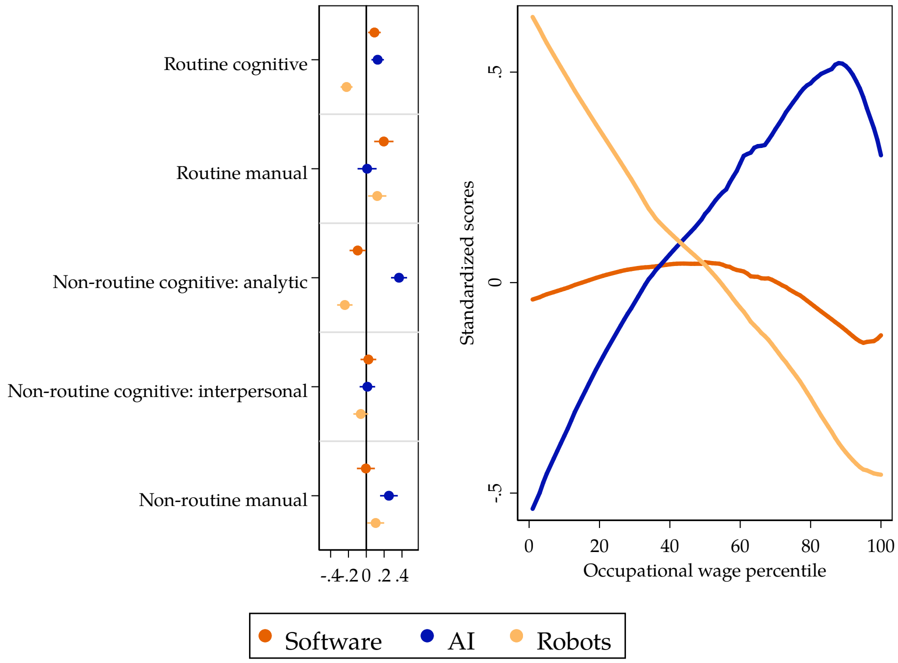

Figure A16: **Capability pair scores calculated by inverse rank.**

Notes: In our baseline specification, we set each aggregated capability pair score equal to the *relative frequency* of that aggregated capability pair in the titles of patents corresponding to a given technology. This figure displays our results when we instead set each aggregated capability pair score equal to the *inverse rank* of that aggregated capability pair, as described in Section A.2. The pattern for software is attenuated; the results for artificial intelligence are essentially identical; and the slope of the robots line is a little more negative in the bottom half of the wage distribution.

References

Acemoglu, Daron and David H Autor, "Skills, Tasks and Technologies: Implications for Employment and Earnings," in David Card and Orley Ashenfelter, ed., *Handbook of Labor Economics*, Vol. 4, Part B, Elsevier, 2011, pp. 1043–1171.

Autor, David H, Frank S Levy, and Richard J Murnane, "The Skill Content of Recent Technological Change: An Empirical Exploration," *The Quarterly Journal of Economics*, 2003, *118* (4), 1279–1333.

Remus, Dana and Frank S Levy, "Can Robots Be Lawyers: Computers, Lawyers, and the Practice of Law," *The Georgetown Journal of Legal Ethics*, 2017, *30*, 501–558.