Inferring Work Task Automatability from AI Expert Evidence

Paul Duckworth[1], **Logan Graham**[1], **Michael A. Osborne**[1,2]
(1) Machine Learning Research Group, (2) Oxford Martin School,
University of Oxford, UK

Abstract

Despite growing alarm about machine learning technologies automating jobs, there is little good evidence on what activities can be automated using such technologies. We contribute the first dataset of its kind by surveying over 150 top academics and industry experts in machine learning, robotics and AI, receiving over $4,500$ ratings of how automatable specific tasks are today. We present a probabilistic machine learning model to learn the patterns connecting expert estimates of task automatability and the skills, knowledge and abilities required to perform those tasks. Our model infers the automatability of over 2,000 work activities, and we show how automation differs across types of activities and types of occupations. Sensitivity analysis identifies the specific skills, knowledge and abilities of activities that drive higher or lower automatability. We provide quantitative evidence of what is perceived to be automatable using the state-of-the-art in machine learning technology. We consider the societal impacts of these results and of task-level approaches.

Introduction

Machine learning (ML), in combination with complementary technologies such as robotics and software-based standardization, have rapidly become real substitutes and complements to human labor. This work aims to better understand that automation, and its effects on work. One example is Amazon Go, a recently opened grocery store that uses computer vision to replace cashiers, of which over 3.5 million are employed in the United States (Grewal, Roggeveen, and Nordfaelt 2017; OES 2017). Further, the 500 000 designers in the US are beginning to use constraint-based generative design to automate creative designs of buildings, industrial components, and more (Autodesk 2017; OES 2017). As a result, we as researchers in these fields often confront examples of media and public concern about technologies we develop. What remains uncertain is the magnitude and direction of impact on employment of machine learning technologies. While recent advances in technology seem able to automate *intelligent* work, we lack good data on the scope of such automation.

We collected a detailed *task*-based survey of 150+ machine learning, robotics, and automation researchers. This is the first dataset of its kind with over 4,500 datapoints about what specific tasks are automatable according to *current* technology. In this "nowcasting" exercise, technologists provide knowledge of the extent to which a task can or cannot be automated with technology that exists *today*. We use a probabilistic model to infer the automatability of thousands of activities for which it would be prohibitively difficult to collect reliable data. By collecting task-specific data to model automatability, we believe we can develop richer, more accurate frameworks about what can be automated by the current state-of-the-art in intelligent technology. We believe this allows society to better understand and prepare for automation.

The contributions of this paper are a novel dataset of the automatability of workplace activities; a probabilistic method for inferring automatability of unmeasured activities; activity-level analysis of automatability and its drivers; and consequent patterns across worktypes, occupations, income, and education. We use more detailed numeric attributes and incorporate more expert knowledge than used in previous studies.

We demonstrate that we can accurately model expert opinions regarding the current state of automation, and introduce a methodology that goes beyond the limitations of the literature. We call for more and better measurement of automation at the activity level. We discuss how this is needed to better prepare governments, employees, and businesses for the effects of automation.

Related Work

Traditionally, approaches to predicting what is automatable have developed frameworks based on the "types" of occupations and the skills they require (Autor 2013; Acemoglu and Restrepo 2016; Frey and Osborne 2017). One popular framework places occupations on a manual-cognitive spectrum and a standardizable-dynamic spectrum. These frameworks are broad and don't best reflect that *tasks* (or groups of tasks), not *entire occupations*, are the unit of automation. As a result, public perception is conflicted about the true effect of automation on work (Smith 2016).

Recent work assumes that occupations are better analyzed as evolving combinations of detailed tasks, skills, and/or environments (Arntz, Gregory, and Zierahn 2016; Manyika et al. 2017a; Acemoglu and Restrepo 2016). With increasingly granular job data available from sources such as from O*NET (National Center for O*NET Development), that break down occupations into hundreds of continuously updated numerical components, we believe there is an opportunity to evaluate occupations by focusing on their tasks, skills, and environments. Two recent reports use a task-first approach. A recent report by McKinsey (Manyika et al. 2017a) uses an unclear approach to model the opinions of automation potential of an unknown number of industry-based experts unfamiliar with the frontier of technology today. Another recent analysis by the OECD (Arntz, Gregory, and Zierahn 2016) derives high-level task-level estimates from occupation-level estimates, and uses worker (not task) characteristics in their inference procedure.

We differ from previous approaches in three key aspects: first, we seek expert knowledge at the most granular task level, similar to (Manyika et al. 2017a) and (Grace et al. 2017). Second, we ask what is automatable *today* and do not make speculative assumptions about future developments or uptake of future technological advancements. Third, we present a robust and (soon) openly available probabilistic methodology and dataset.

Data Representation

Expert Survey We conducted an online survey of 156 academic and industry experts in machine learning, robotics and intelligent systems about how automatable specific tasks are using technology available today. Each expert was presented with 5 occupations and their 5 "most important" tasks, taken from the Occupational Network (O*NET) 2016 database (National Center for O*NET Development). The complete list of 70 occupations whose tasks are annotated is shown in Table 2 in Appendix A, with occupations chosen to be representative of the feature space, with an emphasis on high-employment and hence familiar occupations. Five sample occupations and their surveyed tasks are displayed in Table 3 Appendix A.

Each expert answered the following question: *"Do you believe that technology exists today that could automate these tasks?"*, then labeled each task as either: *Not automatable today* (score of 1.0), *Mostly not automatable today (human does most of it)* (2.0), *Could be mostly automated today (human still needed)* (3.0), *Completely automatable today.* (4.0), or *Unsure*. Respondents also reported overall confidence in their answers (distribution shown in Appendix B).

Our dataset contains 4 599 task level responses from 156 academic and industrial experts from around the world, and across various scientific and industrial fields. Due to the mix of fields, we broadly label these as experts in artificial intelligence and recognize experts offer varying perspectives.

We combine each task's multiple expert labels using Independent Bayesian Classifier Combination (IBCC), a principled Bayesian approach to combine multiple classifications (Kim and Ghahramani 2012; Simpson et al. 2013). IBCC creates a posterior over labels that reflects the individual labellers' tendencies to agree with other labellers over ultimately chosen label values. We averaged IBCC task scores into their task's work activity (described below). Labels concentrate around whole and half values and we round the final values to the closest 0.5 (a half-class).

We believe a survey of many experts, combined using IBCC, is a transparent and reliable method of obtaining data about the current state of task automation. It requires no forecasting or prediction by the participants. The distribution of task-level expert responses, and the IBCC combined task-label distribution are shown in Table 5 in Appendix B. The distributions of field-relevant academic experience, and the geographic location are shown in Figure 7 Appendix B. 97% of participants had field-relevant academic experience: most participants coming from computer science, ML, robotic or AI backgrounds, and 52% of responses from the US, UK and Germany, with the rest from 30 other countries.

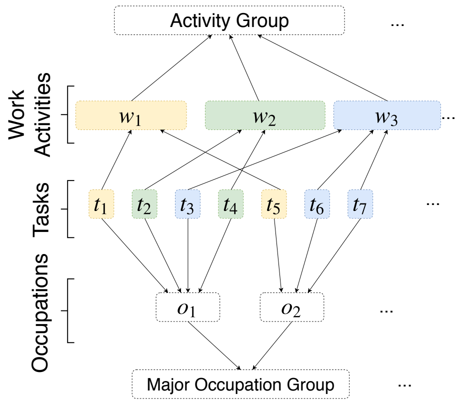

Figure 1: O*NET Database Taxonomy including occupations o, work activities w, tasks t, and high-level groupings.

Occupations, Activities and Tasks An occupation o is represented as a set of discrete *tasks* an employee may be required to perform, $o = \{t_1, t_2, \dots, t_n\}$. Each occupation may have a different number of tasks, and all tasks are detailed enough to be unique to one occupation.

Similar tasks are grouped into a *work activity* w such that $w = \{t_i, \dots, t_j\}$. While a task is occupation specific, work activities are generic activities performed across multiple occupations. We further aggregate results to major occupational groups and high-level activity groups. A diagram demonstrating this hierarchy is shown in Figure 1.

Automation by Work Activity To create activity-specific feature vectors, we aggregate the roughly $20,000$ tasks into their $2,067$ work activities (which are called Detailed Work Activities in O*NET) to create a feature vector $\boldsymbol{x}_w$ for specific work activity w. Each occupation is represented as a vector $\boldsymbol{x}_o$, comprising of the numerical ratings of its skills, $(\boldsymbol{x}_o^s)$, knowledge $(\boldsymbol{x}_o^k)$ and abilities $(\boldsymbol{x}_o^a)$, i.e. $\boldsymbol{x}_o = \left[\boldsymbol{x}_o^s, \boldsymbol{x}_o^k, \boldsymbol{x}_o^a\right]$. We represent each task t of occupation o with the feature vector of occupation o, because we assume that an occupation's skills, knowledge, and abilities informs those needed to perform its constituent tasks. These features are measured quantitatively on a 1 to 5 scale by dozens of employees and experts in the O*NET database.

The activity feature vector is a weighted average of its constituent task vectors: $\boldsymbol{x}_w = \sum_{t \in w} w_{(t,w)} \boldsymbol{x}_o$, where $w_{(t,w)}$ is a normalized weight of the task's relative importance to its occupation and its work activity, i.e. $w_{(t,w)} =$

$I_{(t,o)} I_{(t,w)} / \sum_{t \in w} I_{(t,o)} I_{(t,w)}$. The relative importance of the task to its occupation is calculated as $I_{(t,o)} = {}^{I_t}/\sum_{t \in o} I_t$, while the relative importance of a task to its work activity is $I_{(t,w)} = {}^{I_t}/\sum_{t \in w} I_t$. Task importance I_t is a numeric measure also supplied by O*NET.

Automation by Occupation We also explore what the automatability of activities implies about the automatability of the occupations that perform them. First, we infer automation scores $\hat{\boldsymbol{y}}_w$ for all work activities (including unlabeled ones), as will be described in the next section. We construct an occupation automation score $\hat{y}_o$ for occupation o using the importance-weighted average of its constituent work activities: $\hat{y}_o = \sum_{w \in \Omega_o} I_{(w,o)} \hat{y}_w$, where Ω_o, is the set of all work activities performed by occupation o, and $I_{(w,o)}$ is the importance score of work activity w normalized over this set.

We present automatability over *major occupation groups* which are the highest level occupation categorisation provided by the Bureau of Labor Statistics' SOC system. We represent a major occupation group G as a set of occupations $\{o_1, o_2, \dots\}$ and construct the automation score $\hat{y}_G$ by taking the employment-weighted average of the automatability scores of its constituent occupations, $\hat{\boldsymbol{y}}_o$ for all occupations in the group, i.e. $o \in G$.

On Using IBCC on Training Data We believe using IBCC to achieve a single rating from multiple experts is advantageous both because it is fully Bayesian and reflects a higher chance of accurately recovering the true automatability label of a task in an environment of uncertainty and subjectivity. The main idea is to use the agreement of responses to learn a belief over the correctness of each individual classifier (each human expert) in order to weight their responses. The alternative – averaging task scores – we believe misrepresents the ordinal classification task as a continuous regression task. A side-effect of this approach is that the ground truth labels are more polarised at the extreme values (one and four), when compared with simply (mean) aggregating the task-level responses together. A complication of this is that when comparing two models based upon their Root Mean Squared Error (RMSE), lower absolute error is achieved by not using the (more polarized) IBCC combined labels. However, we believe they are more representative when combining multiple (semi)-reliable expert sources. In future work, we would like to explore modeling at the individual user-task label level.

The interpretation of the learning task and the social scientific nature of the data influence the correct choice of metric and model. This is often found in computational social science. We save deeper rationalizations of our data, model, and task setup choices for further work.

Model Comparison and Validation

We seek a flexible function estimation capable of modeling complex, non-linear relationships between the features (skills, abilities, knowledge) and (perceived) automatability in high-dimensional space. Given the social scientific nature of the study, we also desire a measure of model uncertainty. We compare models based on their "tolerance accuracy" score – the percent of posterior prediction means, $\hat{\boldsymbol{y}}_w$, that are within 0.5 of the ground truth post-IBCC survey value $\boldsymbol{y}_w$. This is a sensible score for our task, and allows more flexibility in our multiclass ordinal setting than strict accuracy or average error. This score also takes into account that we compare models with heterogenous output types – some output discrete labels, and some output continuous values. We optimised the hyperparameters of all models using 10-fold cross-validation.

Our first candidate model class is that of Gaussian Processes (GPs) (Rasmussen and Williams 2006), which have previously been applied to *occupation*-based data in (Frey and Osborne 2017) and (Bakhshi et al. 2017). We specifically use the ordinal likelihood function introduced in (Chu and Ghahramani 2005) to reflect the nature of having discrete labels but with an ordinal interpretation (*not at all* to *completely* automatable). We use the squared exponential, or RBF, kernel, as it consistently performed well compared to other kernels and was less likely to overfit. We optimize the kernel hyperparameters by minimizing the negative marginal log likelihood $\log p(\boldsymbol{y}_w \mid \boldsymbol{x}_w)$ as described in (Rasmussen and Williams 2006), using the open source software GPFlow (Matthews et al. 2017).

Table 1: Model tolerance accuracy & negative log likelihood.

Model	Accuracy	(std)	−Log-likelihood
Ordinal RBF GP	**0.645**	(0.028)	**385.6**
Gaussian RBF GP	0.643	(0.102)	382.3
Ordinal DNN	0.604	(0.071)	–
Random Forest	0.517	(0.081)	–
Ordinal Regression	0.451	(0.036)	–
$\hat{\boldsymbol{y}}_w$ = avg($\boldsymbol{y}_w$)	0.575	(0.065)	–
Proportional Random	0.374	(0.049)	–

For other candidate models, we consider ordinal logistic regression (Pedregosa-Izquierdo 2015), a random forest, and a neural network with an ordinal loss function (Hart 2017), with a 4-layer (120-60-120-7) fully connected layer architecture and 10% layer-wise dropout. A proportional random assignment and constant mean predictor are compared as a lower baseline on predictive performance. Notably, just predicting with the value of the mean achieves fourth-best performance. This is due to the concentration of the values, as the mean-predictor offers no useful information. While we believe our metric is the best for assessing models, this highlights the challenges of modelling subjective data with the ordinal classification interpretation we've taken.

Results for each model are displayed in Table 1 (standard deviation in brackets). The ordinal GP model consistently outperforms comparative methods at prediction of posterior mean values of automatability over the space of work activities. While the non-ordinal GP model and the optimised deep neural network perform on average similarly to the ordinal GP model, they do so much less reliably.

Experiments and Results

Question 1: What is automatable?

Using the best performing GP model to infer the automatability score for all 2,067 work activities (including the train-

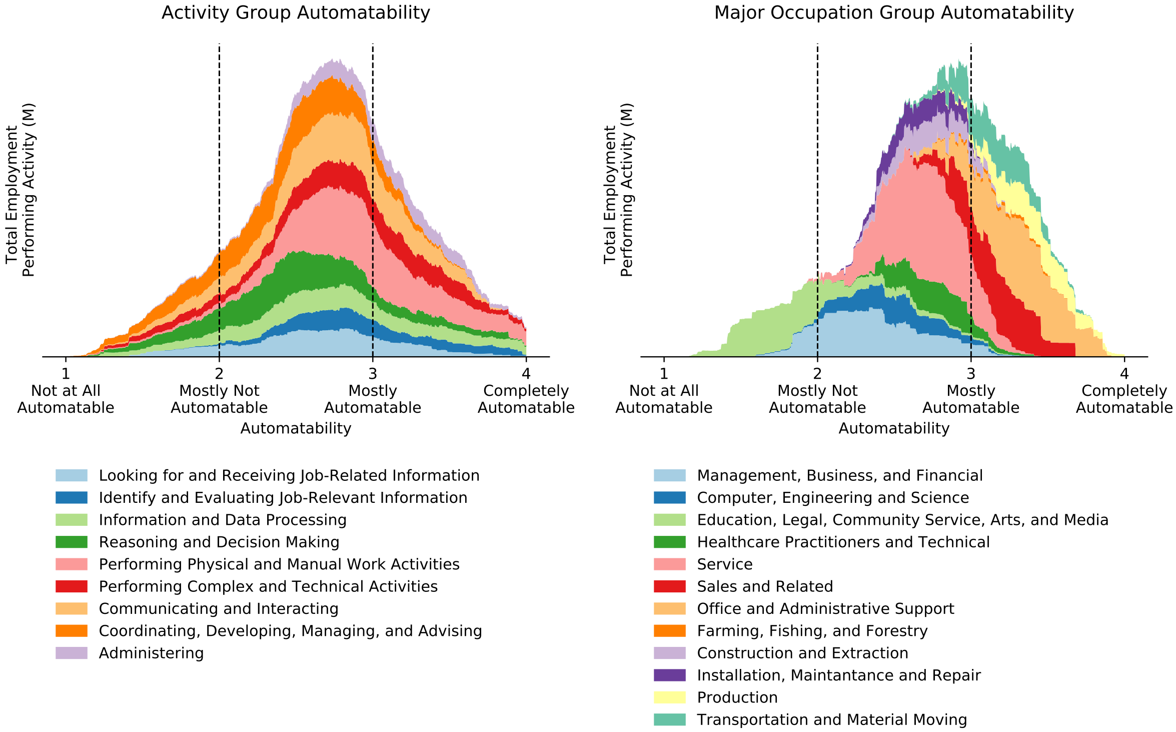

Figure 2: (left:) Amount of employment affected across automatability scores, by 9 high level activity groups. (right:) Employment affected across automatability scores, by 12 major occupation groups.

ing set), we examine examples of automatable and not-automatable activities. Table 4 in Appendix C presents a sample of work activities with the highest and lowest automatability from the unlabeled data (with uncertainties).

We observe that activities such as “route mail to correct destinations” (3.82) or “cut fabrics” (3.93) have a high automation potential (and indeed, are widely automated). However, we also notice that white collar activities are also highly automatable with current technology: “Operate digital imaging equipment” (3.63),“Send information, materials or documentation” (3.39), and “Advise others on ways to improve processes or products” (3.38). Insights such as these propose likely future automatable areas, where automation could be achieved in the real world with relatively little further attention to the underlying technology. The mean of automatability scores is 2.65, indicating that that the model, learned from expert estimates, believes that tasks are on average marginally more likely to be more automatable than not.

In lieu of profiling the long list of activities by their automatability here, we consider instead what *groups* of activities are automatable, and the implications for occupations when using an activity-first approach.

In Figure 2 (left) we plot the automatability of activities by the number of currently-employed individuals who perform them, and classify into nine high-level activity groupings. It becomes evident that while most activities are between mostly and mostly not automatable, work tends to lie closer to “mostly automatable”. Eight times as much work lies between “mostly” and “completely” automatable than between “mostly not” and “not at all” automatable, when weighted by employment. Activities classified as “reasoning and decision making” and “coordinating, developing, managing, and advising” are less likely than others to be automatable. However, “administering”, “information and data processing” and (perhaps surprisingly) “performing complex and technical activities” are more likely to be automatable.

Additionally, we average the automatability scores of an occupation’s activity automatabilities to create an occupation-level automatability score. (How to properly aggregate activities for an occupation-level score is a subject of further research, but we use this as preliminary exploration.) We classify occupations into 12 high-level “Major Occupation Groups”, as in Figure 2 (right). We see that the model predicts very high automation potential in office, administrative support (orange), and sales occupations (red), which together employ about 38 million people in the United States. This stands in contrast to the popular emphasis on the automation of physical processes such as production (yellow), farming, fishing and forestry (dark orange), and transportation and material moving (brown), which employ about 20 million people in total.

In contrast, two Major Occupation Groups appear very robust to automation: education, legal, community service, arts, and media occupations (light green), and to a lesser extent, management, business, and financial occupations (light blue).

Trends across Income and Education Using our occupation-level mappings, we present a preliminary analysis of how automatable activities cluster across income and education, to understand the likely impact on employees. We use median annual income data from the Occupational Employment Statistics from the Bureau of Labor Statistics (OES 2017), and the average expert estimate that one requires at least a bachelor's degree to perform an occupation from O*NET (National Center for O*NET Development).

The impact is perhaps expected. The highest paid, most educated occupations tend to be the least automatable. They tend to be much smaller occupations. However, it is worth noting that even being paid well and having a bachelor's degree does guarantee an occupation's activities are not automatable. "Air Traffic Controllers" make about $125,000 a year, yet are deemed mostly automatable (2.93). "Cytogenetic Technologists", for example, require a Bachelor's degree (with a 100% likelihood from O*NET), with an estimated occupation automatability of 3.03.

Nor is the opposite always true. "Preschool Teachers" and "Teacher Assistants" make just under $30,000 a year, yet are mostly non-automatable (1.71 and 1.87, respectively). O*NET experts estimate a 5% chance of needing at least a bachelor's degree to perform successfully as a "Heating and Air Conditioning Mechanics and Installer", an occupation which is also predicted mostly not automatable (2.38).

Indeed, how occupations evolve their set of activities in response to automation is an activate research question. We explore what *drives* automatability in Question 2, so that one might better predict how employers, employees, policymakers, and other stakeholders might respond.

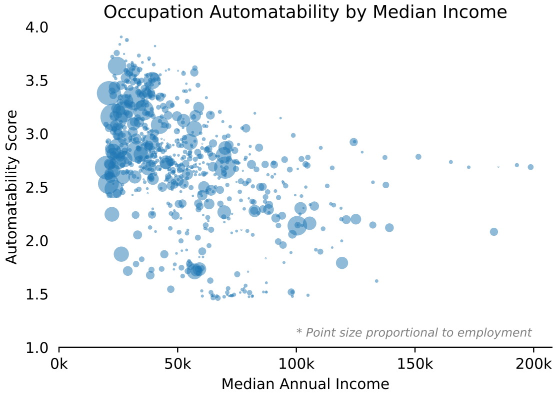

Figure 3: Occupation-level automatability scores by annual median income.

Model Disagreements with Ground Truth It is useful to consider where our model disagrees most with our ground truth labels. While, as discussed, it's difficult to confirm which estimate is true, when the model disagrees with the ground truth, it is using all information learned from all other ratings. Table 7 in Appendix C lists the 50 tasks with most disagreement between ground truth and model.

In the cases where the model *overpredicts* relative to the ground truth, we see common themes. Some activities, such as "connect electrical components or equipment" ($\boldsymbol{y}_w$: 1.0; $\hat{\boldsymbol{y}}_w$: 2.69) are characterised by dextrous physical action (of which the difficulty of automating is famously speculated as "Moravec's paradox" in (Moravec 1990)). Others involve information exchange in a situation with a clear goal, such as "communicate with customers to resolve complaints or ensure satisfaction" ($\boldsymbol{y}_w$: 1.0; $\hat{\boldsymbol{y}}_w$: 2.31). There are also cases involving complex process monitoring. Across these activities, it seems conceivable that while intelligent technology may not automate the *entire* activity, some combination of activity simplification, custom data gathering, and intelligent technology could automate a non-trivial amount of the activity.

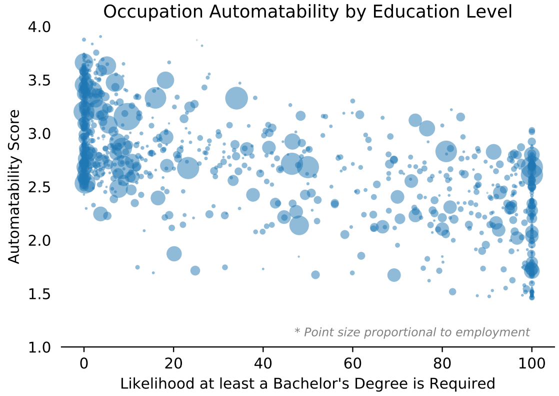

Figure 4: Occupation-level automatability scores by education attainment.

The cases where the model underpredicts are more varied in interpretation. Some model underpredictions could be a case of downweighting unreasonable expert beliefs. For example, "position construction forms or molds" ($\boldsymbol{y}_w$: 4.0; $\hat{\boldsymbol{y}}_w$: 2.80) – simple in theory, but influenced by many dynamic environmental factors to render it complex in practice. Other cases of underprediction might also be a reflection of a lack of data around the activity in feature space, leading to limited model learning. For example, there are some activities which are clearly automatable (and already automated) for which the model predicts low automatability (e.g. "Process customer bills or payments" and "Create electronic data backup to prevent loss of information"). Alternatively, the more questionable predictions may be artifacts of the O*NET taxonomy, in which the work activity title does not accurately reflect a critical nuance of its constituent tasks.

Question 2: What makes work automatable?

We now consider what increases or decreases the automatability of some activities. We compute the average derivative of automatability with respect to each numeric feature as described in (Baehrens et al. 2010) over the space of work activities. For the nth feature, this is computed as $AG(n) := \mathbb{E}(\partial m(\boldsymbol{x}) / \partial \boldsymbol{x}_n)$, where $m(x)$ is the posterior mean distribution. This measures the expected increase in automatability for a unit increase in the feature. Table 8 in Appendix C presents a sample of the highest and lowest average derivatives of the posterior mean function per feature.

These gradients seem to reflect what intelligent technology increasingly offers: work that is clerical, repetitive, precise, and perceptual can increasingly be automated. Increases in the features *Clerical, Number Facility, Depth Perception, Control Precision* and *Production and Processing* tend to increase an activity's automatability. Perhaps surprisingly, increases in *Economics and Accounting* and *Sales and Marketing* knowledge and ability also increase automatability, which reflects that we saw business-oriented sales & administrative work types having higher than average automatability.

On the other hand, work that is more creative, dynamic, and human oriented tends to be less automatable. While variable, the three strongest features driving decreased activity automatability are *Installation, Programming* and *Technology Design*. That is to say, the experts who answered our survey are relatively safe, or misperceive themselves to be.

The gradients might, for example, be used by employers/employees and policy makers to skill themselves differently, proactively change the characteristics of work activities, or set policy to incentivize the development of particular skills, knowledge, abilities, or occupations.

Uncovering the Drivers of Automation In our own analysis, we found that using activity-level ratings allows us to hypothesize richer frameworks about the drivers of automation than the previous occupation-level frameworks. For example, one particularly useful analysis is to examine clusters of similar activities, which score significantly high in a subset of a features but vary substantially in inferred automatability. Some of these activity clusters across features such as *Dynamic Strength, Persuasion, Critical Thinking,* and *Production and Processing* seemed to imply the predictive importance of (a) task-standardization versus dynamism; (b) a well-specified performance metric versus open-ended thinking; (c) single-party versus multi-party goal satisfaction; and (d) active thinking versus active physical interaction. We reserve final conclusions from this approach for further expanded and validated research. In summary, we believe this suggests that more activity-level data likely allows us to find richer frameworks for automatability prediction.

Societal Impact

Automation is a notably data-sparse yet opinion-heavy area of study (National Academies of Sciences, Engineering and Medicine 2017; Mitchell and Brynjolfsson 2017). We need more clarity about what can be automated, at the *actual* level of automation (tasks), and more granular frameworks based on more more granular data. Policymakers would be able to design better, more targeted policy responses, such as incentives to preemptively modify occupations or programs to reskill workers. Workers would be able to upskill or retrain in a more targeted way (towards certain tasks or away from certain tasks) to be robust to automation, instead of abandoning entire occupations. Further, we note the large *psychological* burden of fear, uncertainty, and doubt that comes with uncertain predictions; we hope to replace that with the optimism, clarity, and confidence that comes from every worker having better predictions to enable more effective responses. Last, we would be able to better spot ethically-challenging cases of activity automation, especially in healthcare and social services, *before* they happen, so that we can hold preparatory, informed ethical discussion.

We believe this necessitates collecting more data on automation. Governments, researchers, businesses, and employees would *all* benefit from uniting to do so. Despite automation being one of the dominant themes of work in the coming decades – a perhaps irreversible shift in how work is done in the future – we are only just taking our first steps towards granular, activity-level measurement. In this paper we offer our dataset, consider the value of activity-level data, and present preliminary results that reinforce previous ones and provide more fidelity and opportunities for deeper research.

We also offer our approach as one example of how to generate insights from limited data on automation. We will need to expand on limited data due to the nature of the challenge. First, collecting data is generally difficult and resource intensive. Further, automation is still poorly-defined at the micro-level. Is automation different if it replaces a modular activity, or an entire chain of activities? Is it intelligent automation if it solves the task by breaking it into unintelligent components? To supplement more data, we also need better and clearer definitions.

Conclusion & Future Work

By using a more granular approach to "now-casting" task-level automation, we can unlock more nuanced frameworks about what *actually* can be automated. Using task and activity-level data, we can likely better understand the drivers of automatability. However, our approach is a first step. We propose to the community six important research gaps that our dataset and approach should be useful for answering:

Real-world validation: Does our model accurately identify activities that are already automated? **Surprising automation**: Which activities are more, or less, automatable than previous models would have predicted? **Automatable vs. automated**: *Why* are some automatable activities not automated while others are? What mechanisms (like unfavourable economics, limited data, or no performance metric) prevent automation for predictively automatable activities? **Multivariate interaction patterns**: How does one feature modify a different feature's effect on a task's automatability? **Economic value**: What is the monetary value of automation potential for highly automatable activities? (See (Manyika et al. 2017b).) **Employee characteristics**: What are the patterns of demographics, industry, technology use, and other characteristics of employees performing activities with low- and high-automatability activities?

Acknowledgements

We are grateful for the support and comments of Jonathan Downing, Katja Grace, Allan Dafoe, John Salvatier, Matthew Willis, Nesta, MIRI, The Health Foundation (award #7559), the Oxford Martin Programme on Technology and Employment, and the Rhodes Trust.

References

Acemoglu, D., and Restrepo, P. 2016. Artificial Intelligence, Automation and Work.

Arntz, M.; Gregory, T.; and Zierahn, U. 2016. The Risk of Automation for Jobs in OECD Countries: A Comparative Analysis. *OECD Social, Employment and Migration Working Papers* 2(189):47–54.

Autodesk. 2017. What Is Generative Design.

Autor, D. H. 2013. The "task approach" to labor markets: an overview. *Journal for Labour Market Research* 46(3):185–199.

Baehrens, D.; Schroeter, T.; Harmeling, S.; Kawanabe, M.; Hansen, K.; and Muller, K.-R. 2010. How to Explain Individual Classification Decisions. *Journal of Machine Learning Research* 11:1803–1831.

Bakhshi, H.; Downing, J. M.; Osborne, M. A.; and Schneider, P. 2017. The future of skills employment in 2030.

Chu, W., and Ghahramani, Z. 2005. Gaussian processes for ordinal regression. *Journal of machine learning research* 6(Jul):1019–1041.

Frey, C. B., and Osborne, M. A. 2017. The future of employment: how susceptible are jobs to computerisation? *Technological Forecasting and Social Change* 114:254–280.

Grace, K.; Salvatier, J.; Dafoe, A.; Zhang, B.; and Evans, O. 2017. When Will AI Exceed Human Performance? Evidence from AI Experts.

Grewal, D.; Roggeveen, A. L.; and Nordfaelt, J. 2017. The Future of Retailing. *Journal of Retailing* 93(1):1–6.

Hart, J. 2017. Keras ordinal categorical crossentropy. `https://github.com/JHart96/keras_ordinal_categorical_crossentropy`.

Kim, H.-C., and Ghahramani, Z. 2012. Bayesian classifier combination. In *Artificial Intelligence and Statistics*, 619–627.

Manyika, J.; Chui, M.; Miremadi, M.; Bughin, J.; George, K.; Willmott, P.; and Dewhurst, M. 2017a. A Future that Works: Automation, Employment, and Productivity. *McKinsey Global Institute*.

Manyika, J.; Chui, M.; Miremadi, M.; Bughin, J.; George, K.; Willmott, P.; and Dewhurst, M. 2017b. Harnessing Automation for a Future that Works. *McKinsey Global Institute*.

Matthews, A. G. d. G.; van der Wilk, M.; Nickson, T.; Fujii, K.; Boukouvalas, A.; León-Villagrá, P.; Ghahramani, Z.; and Hensman, J. 2017. GPflow: A Gaussian process library using TensorFlow. *Journal of Machine Learning Research* 18(40):1–6.

Mitchell, T., and Brynjolfsson, E. 2017. Track how technology is transforming work. *Nature* 544(7650):290–292.

Moravec, H. 1990. *Mind Children: the Future of Robot and Human Intelligence*. Harvard University Press.

National Academies of Sciences, Engineering, and Medicine. 2017. Washington, D.C.: National Academies Press.

National Center for O*NET Development. O*NET OnLine.

2017. Occupational Employment Statistics.

Pedregosa-Izquierdo, F. 2015. *Feature extraction and supervised learning on fMRI : from practice to theory*. Theses, Université Pierre et Marie Curie - Paris VI.

Rasmussen, C. E., and Williams, C. K. 2006. *Gaussian Processes for Machine Learning*, volume 1.

Simpson, E.; Roberts, S.; Psorakis, I.; and Smith, A. 2013. Dynamic bayesian combination of multiple imperfect classifiers. In *Decision making and imperfection*. Springer. 1–35.

Smith, A. 2016. Public Predictions for the Future of Workforce Automation: Full Report. Technical report, Pew Research Center.

Supplemental Material

Appendix A: Expert Survey Description

Table 2: 70 occupations with tasks labeled to construct the training set.

O*NET-SOC Code	Title
11-1011.00	Chief Executives
11-3071.01	Transportation Managers
11-9033.00	Education Administrators, Postsecondary
11-9199.01	Regulatory Affairs Managers
13-1022.00	Wholesale And Retail Buyers, Except Farm Products
13-1075.00	Labor Relations Specialists
13-2053.00	Insurance Underwriters
15-1134.00	Web Developers
15-1143.01	Telecommunications Engineering Specialists
17-1011.00	Architects, Except Landscape And Naval
17-3022.00	Civil Engineering Technicians
21-1011.00	Substance Abuse And Behavioral Disorder Counselors
21-1023.00	Mental Health And Substance Abuse Social Workers
21-1093.00	Social And Human Service Assistants
23-1011.00	Lawyers
25-1011.00	Business Teachers, Postsecondary
25-1071.00	Health Specialties Teachers, Postsecondary
25-1194.00	Vocational Education Teachers, Postsecondary
25-2032.00	Career/Technical Education Teachers, Secondary School
25-2053.00	Special Education Teachers, Middle School
25-9041.00	Teacher Assistants
27-1011.00	Art Directors
27-1026.00	Merchandise Displayers And Window Trimmers
27-2011.00	Actors
27-2022.00	Coaches And Scouts
27-2042.01	Singers
29-1063.00	Internists, General
29-1199.01	Acupuncturists
29-2032.00	Diagnostic Medical Sonographers
29-2052.00	Pharmacy Technicians
29-9011.00	Occupational Health And Safety Specialists
31-9091.00	Dental Assistants
33-1021.01	Municipal Fire Fighting And Prevention Supervisors
33-3012.00	Correctional Officers And Jailers
33-9091.00	Crossing Guards
35-1011.00	Chefs And Head Cooks
35-2012.00	Cooks, Institution And Cafeteria
35-3011.00	Bartenders
35-9011.00	Dining Room And Cafeteria Attendants And Bartender Helpers
35-9021.00	Dishwashers
39-9011.00	Childcare Workers
41-2022.00	Parts Salespersons
41-4012.00	Sales Representatives, Wholesale And Manufacturing, Except Technical And Scientific Products
41-9021.00	Real Estate Brokers
43-3021.01	Statement Clerks
43-4121.00	Library Assistants, Clerical
43-4141.00	New Accounts Clerks
43-4181.00	Reservation And Transportation Ticket Agents And Travel Clerks
43-5021.00	Couriers And Messengers
45-2093.00	Farmworkers, Farm, Ranch, And Aquacultural Animals
47-1011.00	First-Line Supervisors Of Construction Trades And Extraction Workers

47-2021.00	Brickmasons And Blockmasons
47-2051.00	Cement Masons And Concrete Finishers
47-2181.00	Roofers
49-2022.00	Telecommunications Equipment Installers And Repairers, Except Line Installers
49-3021.00	Automotive Body And Related Repairers
49-9052.00	Telecommunications Line Installers And Repairers
51-1011.00	First-Line Supervisors Of Production And Operating Workers
51-2022.00	Electrical And Electronic Equipment Assemblers
51-4021.00	Extruding And Drawing Machine Setters, Operators, And Tenders, Metal And Plastic
51-4072.00	Molding, Coremaking, And Casting Machine Setters, Operators, And Tenders, Metal And Plastic
51-4121.06	Welders, Cutters, And Welder Fitters
51-6021.00	Pressers, Textile, Garment, And Related Materials
51-6031.00	Sewing Machine Operators
51-9111.00	Packaging And Filling Machine Operators And Tenders
51-9198.00	Helpers–Production Workers
53-1021.00	First-Line Supervisors Of Helpers, Laborers, And Material Movers, Hand
53-1031.00	First-Line Supervisors Of Transportation And Material-Moving Machine And Vehicle Operators
53-3022.00	Bus Drivers, School Or Special Client
53-7062.00	Laborers And Freight, Stock, And Material Movers, Hand

Table 3: Five randomly selected occupations and their surveyed tasks.

Title	Task	Importance
Chief Executives	Direct or coordinate an organization's financial or budget activities to fund operations, maximize investments, or increase efficiency.	4.54
	Appoint department heads or managers and assign or delegate responsibilities to them.	4.48
	Analyze operations to evaluate performance of a company or its staff in meeting objectives or to determine areas of potential cost reduction, program improvement, or policy change.	4.40
	Direct, plan, or implement policies, objectives, or activities of organizations or businesses to ensure continuing operations, to maximize returns on investments, or to increase productivity.	4.39
	Prepare budgets for approval, including those for funding or implementation of programs.	4.17
Lawyers	Represent clients in court or before government agencies.	4.59
	Present evidence to defend clients or prosecute defendants in criminal or civil litigation.	4.50
	Select jurors, argue motions, meet with judges, and question witnesses during the course of a trial.	4.50
	Study Constitution, statutes, decisions, regulations, and ordinances of quasi-judicial bodies to determine ramifications for cases.	4.47
	Interpret laws, rulings and regulations for individuals and businesses.	4.47
Diagnostic Medical Sonographers	Observe screen during scan to ensure that image produced is satisfactory for diagnostic purposes, making adjustments to equipment as required.	4.87
	Observe and care for patients throughout examinations to ensure their safety and comfort.	4.85
	Provide sonogram and oral or written summary of technical findings to physician for use in medical diagnosis.	4.84
	Select appropriate equipment settings and adjust patient positions to obtain the best sites and angles.	4.83
	Operate ultrasound equipment to produce and record images of the motion, shape, and composition of blood, organs, tissues, or bodily masses, such as fluid accumulations.	4.83
Cooks, Institution And Cafeteria	Clean, cut, and cook meat, fish, or poultry.	4.64
	Cook foodstuffs according to menus, special dietary or nutritional restrictions, or numbers of portions to be served.	4.61
	Clean and inspect galley equipment, kitchen appliances, and work areas to ensure cleanliness and functional operation.	4.61
	Apportion and serve food to facility residents, employees, or patrons.	4.58
	Direct activities of one or more workers who assist in preparing and serving meals.	4.27
Brickmasons And Blockmasons	Remove excess mortar with trowels and hand tools, and finish mortar joints with jointing tools, for a sealed, uniform appearance.	4.63
	Construct corners by fastening in plumb position a corner pole or building a corner pyramid of bricks, and filling in between the corners using a line from corner to corner to guide each course, or layer, of brick.	4.60
	Measure distance from reference points and mark guidelines to lay out work, using plumb bobs and levels.	4.47
	Break or cut bricks, tiles, or blocks to size, using trowel edge, hammer, or power saw.	4.39
	Interpret blueprints and drawings to determine specifications and to calculate the materials required.	4.31

Appendix B: Expert Survey Responses

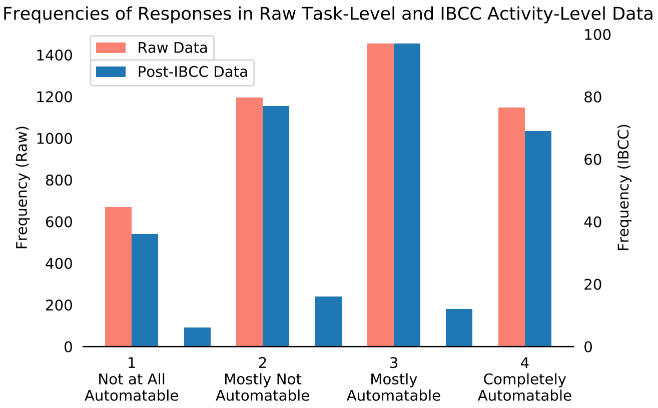

Figure 5: Distribution of expert task-level responses, and the IBCC combined activity labels.

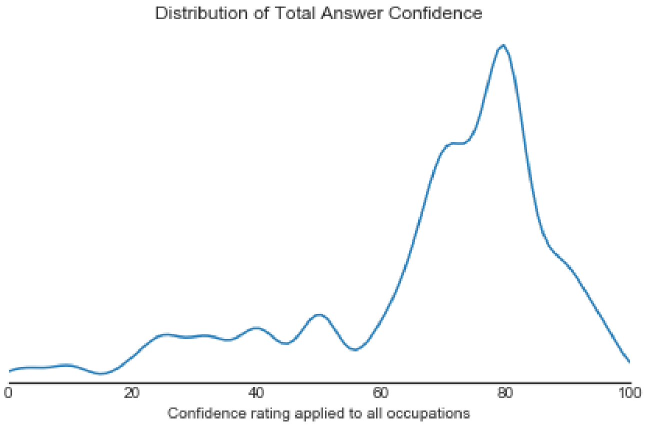

Figure 6: The distribution of respondents confidences they assigned to their answers (in total). ($\mu = 67.9$, $\sigma = 20.7$)

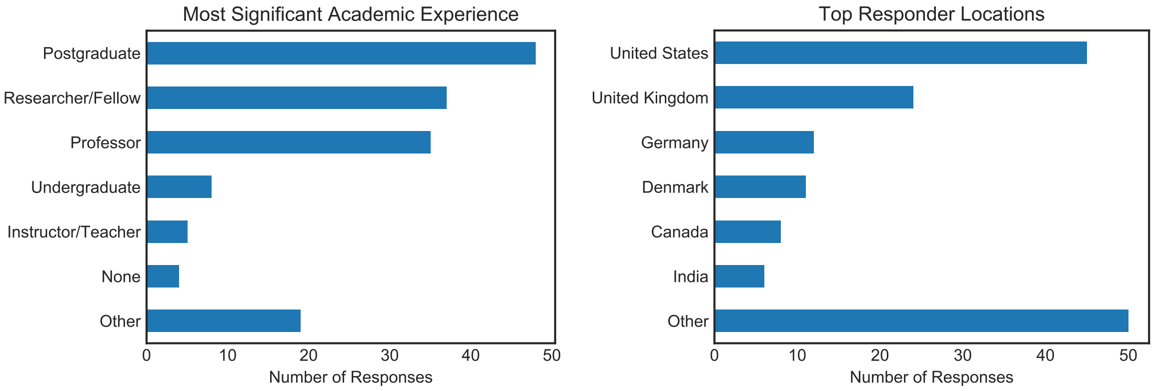

Figure 7: Expert survey response statistics. Responses by: (left:) academic experience. (right:) geographic location.

Appendix C: Inferred Work Activity Automatability

Table 4: The 25 most and least automatable work activities.

Activity	**Automatability Score (var)**
Examine physical characteristics of gemstones or precious metals.	4.00 (0.86)
Adjust fabrics or other materials during garment production.	4.00 (0.67)
Sew materials.	4.00 (0.91)
Assemble garments or textile products.	4.00 (0.68)
Sew clothing or other articles.	4.00 (0.72)
Repair textiles or apparel.	4.00 (0.71)
Attach decorative or functional accessories to products.	3.96 (0.62)
Operate sewing equipment.	3.95 (0.68)
Design templates or patterns.	3.95 (0.68)
Prepare fabrics or materials for processing or production.	3.95 (0.67)
Evaluate log quality.	3.93 (0.71)
Cut fabrics.	3.93 (0.64)
Estimate costs of products, services, or materials.	3.92 (0.67)
Store records or related materials.	3.92 (0.66)
Position patterns on equipment, materials, or workpieces.	3.87 (0.62)
Shape metal workpieces with hammers or other small hand tools.	3.85 (0.65)
Measure physical characteristics of forestry or agricultural products.	3.84 (0.65)
Maneuver workpieces in equipment during production.	3.83 (0.62)
Operate office equipment.	3.83 (0.62)
Route mail to correct destinations.	3.82 (0.64)
Select production input materials.	3.81 (0.61)
Polish materials, workpieces, or finished products.	3.81 (0.64)
Design jewelry or decorative objects.	3.80 (0.78)
Record shipping information.	3.80 (0.63)
Confer with customers or designers to determine order specifications.	3.80 (0.63)
Teach humanities courses at the college level.	1.06 (0.65)
Teach online courses.	1.07 (0.63)
Teach social science courses at the college level.	1.15 (0.61)
Coordinate training activities.	1.16 (0.69)
Conduct scientific research of organizational behavior or processes.	1.19 (0.70)
Choreograph dances.	1.19 (0.86)
Entertain public with comedic or dramatic performances.	1.21 (0.68)
Design video game features or details.	1.22 (0.79)
Advise others on educational matters.	1.32 (0.70)
Evaluate training programs, instructors, or materials.	1.33 (0.65)
Draft legislation or regulations.	1.35 (0.63)
Support the professional development of others.	1.36 (0.74)
Counsel clients on mental health or personal achievement.	1.38 (0.70)
Design psychological or educational treatment procedures or programs.	1.40 (0.65)
Guide class discussions.	1.40 (0.59)
Conduct research on social issues.	1.41 (0.71)
Lead classes or community events.	1.42 (0.66)
Counsel clients or patients regarding personal issues.	1.42 (0.61)
Display student work.	1.43 (0.58)
Develop methods of social or economic research.	1.43 (0.69)
Manage organizational or program finances.	1.44 (0.68)
Evaluate scholarly materials.	1.44 (0.66)
Evaluate effectiveness of educational programs.	1.44 (0.58)
Develop promotional strategies for religious organizations.	1.44 (0.77)
Stay informed about current developments in field of specialization.	1.44 (0.59)

Table 5: Average automatability scores of each of the nine high level work activity groups.

Activity Group	**Automatability Score (std)**
Performing Physical and Manual Work Activities	2.96 (0.45)
Identify and Evaluating Job-Relevant Information	2.88 (0.48)
Administering	2.79 (0.55)
Performing Complex and Technical Activities	2.70 (0.52)
Information and Data Processing	2.58 (0.56)
Communicating and Interacting	2.58 (0.47)
Looking for and Receiving Job-Related Information	2.52 (0.48)
Reasoning and Decision Making	2.44 (0.50)
Coordinating, Developing, Managing, and Advising	2.29 (0.49)

Table 6: Automatability scores of each of the 22 major occupation groups.

Major Occupation Group	**Employment Weighted Automatability Score (std)**
Production	3.40 (0.19)
Office and Administrative Support	3.30 (0.18)
Farming, Fishing, and Forestry	3.16 (0.28)
Sales and Related	3.16 (0.20)
Transportation and Material Moving	3.12 (0.17)
Building and Grounds Cleaning and Maintenance	2.87 (0.10)
Healthcare Support	2.79 (0.15)
Healthcare Practitioners and Technical	2.75 (0.14)
Construction and Extraction	2.74 (0.15)
Food Preparation and Serving Related	2.66 (0.10)
Architecture and Engineering	2.66 (0.23)
Installation, Maintenance, and Repair	2.65 (0.19)
Business and Financial Operations	2.60 (0.31)
Personal Care and Service	2.59 (0.23)
Protective Service	2.53 (0.13)
Arts, Design, Entertainment, Sports, and Media	2.44 (0.35)
Computer and Mathematical	2.42 (0.18)
Life, Physical, and Social Science	2.37 (0.31)
Management	2.17 (0.13)
Legal	1.98 (0.58)
Community and Social Service	1.83 (0.16)
Education, Training, and Library	1.72 (0.21)

Table 7: The 25 work activities where our model disagrees positively and negatively with the ground truth label.

Activity	Ground Truth	Predicted	Disagreement
Connect electrical components or equipment.	1.0	2.69	1.69
Travel to work sites to perform installation, repair or maintenance work.	1.0	2.61	1.61
Clean food service areas.	1.0	2.53	1.53
Locate suspicious objects or vehicles.	1.0	2.52	1.52
Collect dirty dishes or other tableware.	1.0	2.49	1.49
Update knowledge about emerging industry or technology trends.	1.0	2.48	1.48
Arrange tables or dining areas.	1.0	2.47	1.47
Search individuals for illegal or dangerous items.	1.0	2.43	1.43
Collaborate with others to resolve information technology issues.	1.0	2.39	1.39
Operate vehicles or material-moving equipment.	2.0	3.31	1.31
Communicate with customers to resolve complaints or ensure satisfaction.	1.0	2.31	1.31
Exchange information with colleagues.	2.0	3.30	1.30
Direct operational or production activities.	2.0	3.16	1.16
Evaluate employee performance.	1.0	2.15	1.15
Collaborate with others to determine design specifications or details.	1.0	2.14	1.14
Examine animals to detect illness, injury or other problems.	2.0	3.07	1.07
Meet with individuals involved in legal processes to provide information and clarify issues.	1.0	2.04	1.04
Direct material handling or moving activities.	2.0	3.03	1.03
Advise customers on the use of products or services.	2.0	3.02	1.02
Test materials, solutions, or samples.	2.0	3.00	1.00
Monitor loading processes to ensure they are performed properly.	2.0	3.00	1.00
Clean medical equipment.	2.0	2.98	0.98
Assist practitioners to perform medical procedures.	2.0	2.98	0.98
Hire personnel.	1.0	1.98	0.98
Conduct employee training programs.	1.0	1.95	0.95
Maintain student records.	3.5	1.55	−1.95
Count prison inmates or personnel.	4.0	2.32	−1.68
Estimate supplies, ingredients, or staff requirements for food preparation activities.	4.0	2.38	−1.62
Advise others on career or personal development.	3.0	1.45	−1.55
Administer tests to assess educational needs or progress.	3.0	1.55	−1.45
Process customer bills or payments.	4.0	2.55	−1.45
Measure equipment outputs.	4.0	2.63	−1.37
Implement security measures for computer or information systems.	4.0	2.65	−1.35
Conduct research to gain information about products or processes.	4.0	2.73	−1.27
Record patient medical histories.	4.0	2.77	−1.23
Analyze test or performance data to assess equipment operation.	4.0	2.77	−1.23
Position construction forms or molds.	4.0	2.80	−1.20
Refer clients to community or social service programs.	3.0	1.82	−1.18
Maintain client records.	3.0	1.82	−1.18
Prepare reports detailing student activities or performance.	3.0	1.83	−1.17
Create graphical representations of structures or landscapes.	4.0	2.84	−1.16
Plan work operations.	4.0	2.85	−1.15
Create electronic data backup to prevent loss of information.	4.0	2.86	−1.14
Measure materials or objects for installation or assembly.	4.0	2.86	−1.14
Maintain inventory of medical supplies or equipment.	4.0	2.88	−1.12
Manage control system activities in organizations.	3.0	1.92	−1.08
Balance receipts.	4.0	2.92	−1.08

Refer customers to appropriate personnel.	4.0	2.94	−1.06
Care for animals.	4.0	2.94	−1.06
Maintain inventories of materials, equipment, or products.	4.0	2.98	−1.02

Appendix D: Sensitivity Analysis

Table 8: The 25 most automatability-increasing and decreasing features across the activity space.

Feature	Average Gradient (std)
Telecommunications	0.16 (0.03)
Clerical	0.14 (0.03)
Wrist-Finger Speed	0.13 (0.02)
Number Facility	0.11 (0.02)
Mathematics	0.09 (0.02)
Depth Perception	0.08 (0.01)
Mathematical Reasoning	0.08 (0.02)
Economics and Accounting	0.07 (0.02)
Response Orientation	0.07 (0.02)
Building and Construction	0.07 (0.04)
Control Precision	0.07 (0.02)
Arm-Hand Steadiness	0.06 (0.02)
Equipment Selection	0.06 (0.02)
Finger Dexterity	0.06 (0.01)
Perceptual Speed	0.06 (0.01)
Visual Color Discrimination	0.06 (0.01)
Static Strength	0.05 (0.01)
Sales and Marketing	0.05 (0.06)
Far Vision	0.04 (0.01)
Spatial Orientation	0.04 (0.02)
Flexibility of Closure	0.04 (0.01)
Night Vision	0.04 (0.02)
Manual Dexterity	0.03 (0.01)
Multilimb Coordination	0.03 (0.03)
Production and Processing	0.03 (0.02)
Installation	−0.18 (0.08)
Programming	−0.14 (0.04)
Technology Design	−0.14 (0.03)
Fine Arts	−0.11 (0.05)
Gross Body Equilibrium	−0.10 (0.07)
Dynamic Flexibility	−0.10 (0.03)
Speed of Limb Movement	−0.10 (0.02)
Psychology	−0.10 (0.02)
Personnel and Human Resources	−0.09 (0.02)
Sociology and Anthropology	−0.09 (0.03)
History and Archeology	−0.09 (0.03)
Science	−0.09 (0.04)
Food Production	−0.08 (0.07)
Management of Personnel Resources	−0.07 (0.02)
Glare Sensitivity	−0.07 (0.03)
Troubleshooting	−0.07 (0.02)
Gross Body Coordination	−0.06 (0.03)
Coordination	−0.06 (0.01)
Learning Strategies	−0.06 (0.02)
Law and Government	−0.06 (0.02)
Negotiation	−0.06 (0.01)
Management of Financial Resources	−0.06 (0.02)
Social Perceptiveness	−0.06 (0.01)
Chemistry	−0.06 (0.02)
Explosive Strength	−0.06 (0.04)

Interpretation: On average, an increase of an activity's *Clerical* score by one point (1 to 5 scale), tends to to *increase* its automatability by 0.14.