Expanding AI Adoption is an Opportunity for Job Creation

Aniket Baksy and Avi Gupta

1 Key Takeaways

- AI adoption has been slow and concentrated in large firms, preventing it from reaching its full potential and potentially increasing economic inequality.
- High adoption costs are responsible for this slow adoption. We focus on two main drivers of these costs:
    - AI technologies are expensive to customize to specific business needs and require large intangible investments.
    - AI technologies require data, which is expensive to collect, securely store, and analyze.
- Expanding AI adoption by reducing adoption costs can create medium-skill jobs and unleash AI's productivity-enhancing effects.
- We propose a two-pronged policy approach:
    - Making AI easier to use via public support for flexible AI-enabled tools that emphasize usability.
    - Democratizing data access by:
        - Enabling easy access to government-held datasets.
        - Making AI deployment easier for business by creating a novel clearinghouse-like data licensing and computational resource infrastructure.
        - Creating a data curation workforce by retraining medium-skill workers.

2 AI Technologies: What they are, what they do, who uses them

2.1 Defining AI

We adopt the definition of AI proposed by the National Institute of Standards and Technology (NIST):

> AI technologies and systems are considered to comprise software and/or hardware that can learn to solve complex problems, make predictions or undertake tasks that require human-like sensing (such as vision, speech, and touch), perception, cognition, planning, learning, communication, or physical action.[1]

[1] National Institute of Standards and Technology. (2019) U.S. Leadership in AI: A Plan for Federal Engagement in Developing Technical Standards and Related Tools. https://www.nist.gov/system/files/documents/2019/08/10/ai_standards_fedengagement_plan_9aug2019.pdf

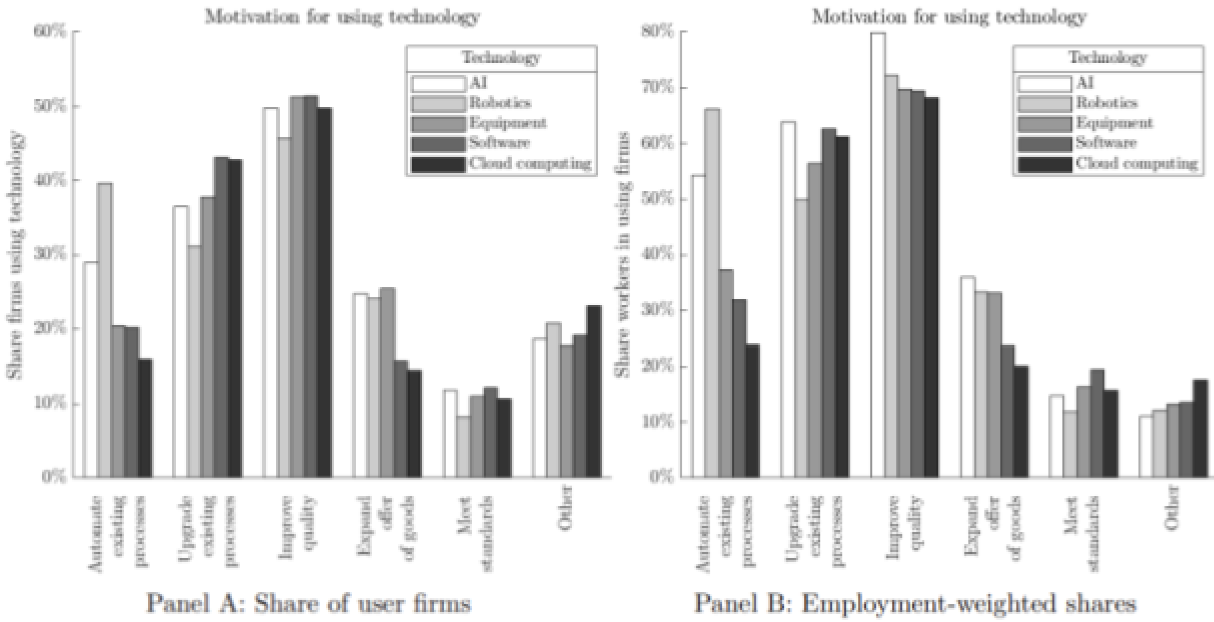


Figure 1: Motivations for technology adoption, ABS 2016-18, from Acemoglu et al. (2022)

2.2 The current abilities of AI technologies

Given the volumes of data available today and improvements in hardware, state-of-the-art AI systems today have achieved near superhuman levels of proficiency at various business-relevant prediction and related tasks, including image recognition[2], speech recognition[3], and text generation[4]. In the business context, AI excels at well-defined tasks with measurable outcomes which can be modeled as decisions or predictions. From Figure 1, the top business use cases for AI include improving existing product quality, and upgrading and automation of routine business processes[5].

Most AI systems learn from a set of labeled examples to predict outcomes for previously unseen data. The quality of AI predictions depends on the type of AI model chosen and, crucially, the amount of data the model is trained on. For example, an ice cream vendor could input data on past sales along with the day and time of the sale and the hourly temperature into an AI system and use it to predict sales for the coming week given the weather forecast.

---

[2] Russakovsky, O., Deng, J., Su, H. *et al.* 2015. ImageNet Large Scale Visual Recognition Challenge. *Int J Comput Vis* 115, 211–252. https://doi.org/10.1007/s11263-015-0816-y.

[3] Schmelzer, Ron. 2020. “Machines that Can Understand Human Speech: The Conversational Pattern of AI.” Forbes, June 28, 2020. https://www.forbes.com/sites/cognitiveworld/2020/06/28/machines-thatcan-understand-human-speech-the-conversational-pattern-of-ai/.

[4] Brown, Tom B., Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, et al. 2020. “Language Models Are Few-Shot Learners.” https://arxiv.org/abs/2005.14165.

[5] Acemoglu et al. (2022). “Automation and the Workforce: A Firm-Level View from the 2019 Annual Business Survey.” Census Bureau working paper CES 22-12. https://www.census.gov/library/working-papers/2022/adrm/CES-WP-22-12.html. Deloitte’s AI dossier provides a wide range of applications for AI technologies in modern firms, including automation of routine tasks, making traditional data analytics more forward looking, and promoting cybersecurity through predictive modeling of attacks.

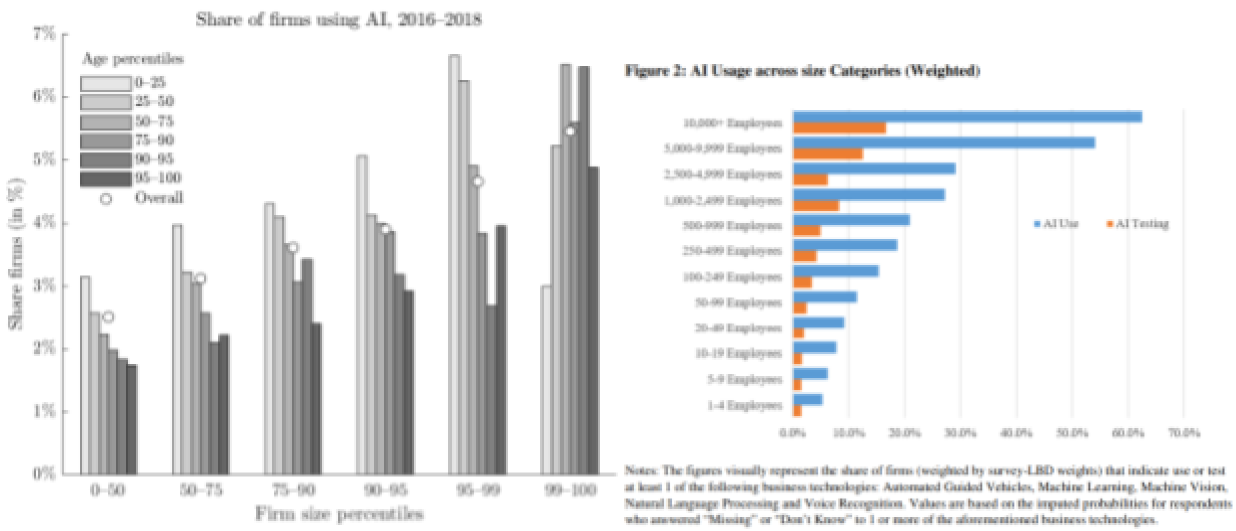


Figure 2: Adoption of AI technologies. LEFT: Acemoglu et al (2022), based on ABS (2019). RIGHT: McElheran et al (2021), based on ABS (2018).

The rapid advance in AI capabilities has raised concerns about job losses due to AI-enhanced automation. The relationship between AI adoption and the labor market is, however, nuanced[6]. Adoption can increase a firm's productivity, enabling its expansion and potentially raising its employment. On the other hand, productivity gains at AI-adopting firms can entrench their dominance, destroying jobs at other firms. Direct evidence for AI-driven job destruction remains weak[7]. We caution that given the current low levels of AI adoption and the dramatic advances in modern AI, AI's final impact on jobs is very uncertain. Irrespective of AI's ultimate impact on the labor market, our goal in this brief is to propose policies that create jobs by accelerating business adoption of AI technologies.

2.3 The skewed adoption of AI technologies

AI adoption has accelerated in the 21st century. However, Figure 2 shows that the adoption of AI technologies remains low and confined to relatively large firms. Evidence from the Annual Business Survey (ABS) (2019) suggests that 3.2% of all firms, accounting for 12.6% of employment, use AI, with large jumps in adoption in 2015[8] and in the aftermath of the COVID-19 pandemic[9]. Of the top 1% of all firms by employment, about 5.5% use AI, but only 2.5% of the bottom 50% do.

---

[6] Aghion et. al. (forth. 2023). "The Effects of Automation on Labor Demand: A Survey of the Recent Literature." in Ing and Grossman (eds.) *Robots and AI: A New Economic Era.* (forthcoming, 2023) https://library.oapen.org/bitstream/handle/20.500.12657/55791/9781000626483.pdf?sequence=1

[7] See Domini et. al. (2022). "For whom the bell tolls: The firm-level effects of automation on wage and gender inequality." *Research Policy.* 51(7) and references therein. https://www.sciencedirect.com/science/article/pii/S0048733322000592#b5

[8] Bloom et al (2022), "The Diffusion of Disruptive Technologies." NBER Working paper 28999. https://www.nber.org/system/files/working_papers/w28999/w28999.pdf. This jump in adoption may have been associated with the public release of the TensorFlow and PyTorch libraries. Globally, AI adoption followed similar trends (Squicciarini, M. and H. Nachtigall (2021), "Demand for AI skills in jobs: Evidence from online job postings", OECD Science, Technology and Industry Working Papers, No. 2021/03, OECD Publishing, Paris, https://doi.org/10.1787/3ed32d94-en.).

[9] McKendrick (2021), "AI Adoption Skyrocketed Over the Last 18 Months", *Harvard Business Review*. https://hbr.org/2021/09/ai-adoption-skyrocketed-over-the-last-18-months

	Artificial Intelligence (1)
No adverse factors	42%
Technology not applicable to this business	49%
Technology too expensive	7%
Technology not mature	2%
Lacked access to required data	1%
Required data not reliable	0%
Lacked access to human capital or talent	1%
Laws and regulations	1%
Concerns regarding safety and security	1%
Lacked access to capital	1%

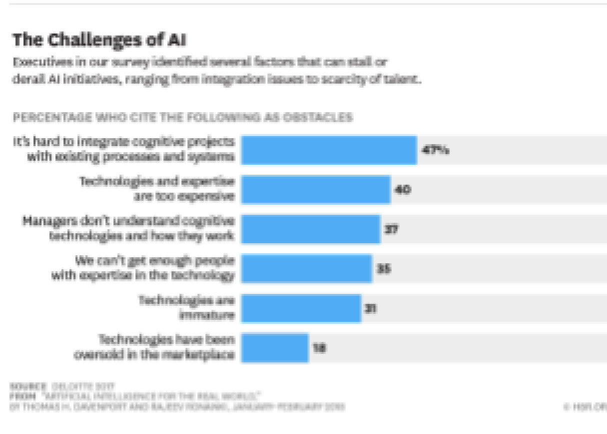


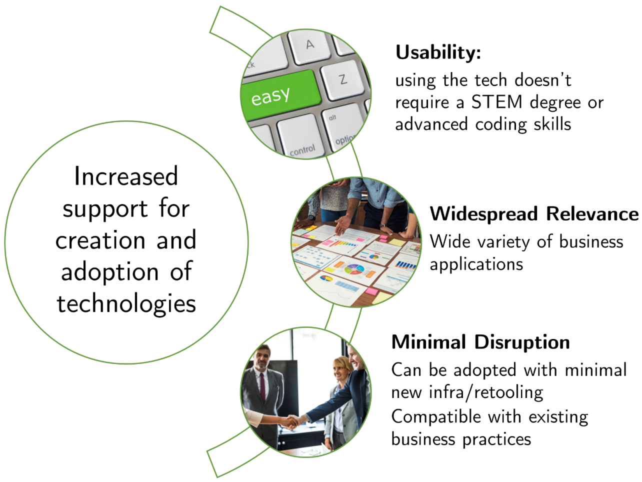
Figure 3: Challenges associated with AI adoption. LEFT: Acemoglu et al (2022) based on ABS 2016-18, RIGHT: Davenport and Ronanki (2018).

Low and concentrated AI adoption enhances inequality among firms and workers. Relatively low AI adoption by smaller firms may further entrench the dominance of larger "superstar" firms, which are usually more capital-intensive and technologically advanced. This would benefit relatively wealthy capital owners over workers. Moreover, employees at superstar firms are likely to disproportionately benefit compared to small business employees[10]. Finally, without widespread AI adoption, we are less likely to see follow-on AI-based innovations that could generate large welfare gains.

Why is adoption so skewed? In the 2019 ABS, 49% of firms that had not yet adopted AI claimed it was because AI was "not applicable to their business", which we interpret to mean that available AI technologies were incompatible with established business practices and infrastructure. Our interpretation is supported by surveys of AI adopters. Deloitte's 2020 State of AI in Enterprise Report[11] found that 39% of respondents thought implementation challenges, integration of AI into company functions, and data issues were the top 3 obstacles to AI adoption. Data issues faced by firms included combining data across disparate systems[12], incorporating data stored in formats unsuitable for analysis[13], and the costs associated with secure data collection and analysis.

These challenges are similar to those faced by firms at the onset of the information and communications technology (ICT) revolution of the 2000s. Successful ICT adopters underwent substantial restructuring of production processes and made large intangible investments[14]. Our goal in this brief is to propose policies to reduce AI adoption costs, which can expand adoption among smaller firms and simultaneously create jobs for workers displaced by technological change.

[10] Song et. al. (2019), "Firming Up Inequality", *The Quarterly Journal of Economics*, Volume 134, Issue 1, February 2019, Pages 1–50, https://doi.org/10.1093/qje/qjy025.
[11] J. Loucks, T. Davenport, and D. Schatsky. Deloitte state of ai in the enterprise, 2nd edition. Technical report, Deloitte Insights, 2018.
[12] E.g., customer assistants requiring customer information, financial data and virtual assistant configuration data, all of which resided in separate systems.
[13] E.g., recorded phone calls as an input for a virtual assistant system stored as audio files.
[14] Bresnahan, T., Brynjolfsson, E., and Hitt, Lorin M. "Information technology, workplace organization, and the demand for skilled labor: Firm-level evidence." *The Quarterly Journal of Economics*, 117(1):339–376, 2002. https://academic.oup.com/qje/article/117/1/339/1851770; Brynjolfsson E. and Hitt, Lorin M. "Beyond computation: Information technology, organizational transformation and business performance." *Journal of Economic perspectives*, 14(4):23–48, 2000. https://www.aeaweb.org/articles?id=10.1257/jep.14.4.23



3 Our Proposals

3.1 Supporting the Creation of Easy-to-use AI

Most AI research funding has historically supported specific applications[15] or basic research in core AI and related mathematics. Recently, the NSF's Institute for Human-AI Interaction and Collaboration has recognized the need to invest in research in AI usability and interaction.[16] Most proposals for new research on the NSF's website emphasize improving usability of AI tools or directly augmenting existing classes of workers, including farm workers[17], construction workers[18], and retail employees[19].

We propose targeted funding towards the development and commercialization of AI-enabled tools which can successfully demonstrate:

- **Usability:** the tool is suitable for users lacking advanced programming or STEM skills.
- **Widespread Relevance:** the tool simultaneously meets a wide range of business needs.
- **Minimal Disruption:** use of the tool does not involve creation of new IT infrastructure or significant changes to existing business practices.

An analogy for a usable and flexible technology with transformative business impact is the spreadsheet, whose diffusion also catalyzed widespread business adoption of computers.[20] The spreadsheet achieved this impact because users were empowered to create their own programs ("macros") and thus apply the data analysis power of the technologies to their own problems, without extensive training in coding[21].

---

[15] Key examples include computer vision and natural language processing. (Rahkovsky et al. (2021), "AI Research Funding Portfolios and Extreme Growth",Front. Res. Metr. Anal., April 2021., Sec. Research Assessment, doi: https://doi.org/10.3389/frma.2021.630124)
[16] https://blog.google/technology/ai/partnering-nsf-human-ai-collaboration/
[17] https://www.nsf.gov/awardsearch/showAward?AWD_ID=2202706&HistoricalAwards=false
[18] https://www.nsf.gov/awardsearch/showAward?AWD_ID=1928626&HistoricalAwards=false
[19] https://www.nsf.gov/awardsearch/showAward?AWD_ID=2231419&HistoricalAwards=false
[20] In a 1984 interview, Steve Jobs remarked that "VisiCalc … propelled the … success [of Apple] … more than any other single event."
[21] In an interview with Quartz, Dan Bricklin, the inventor of VisiCalc, the first commercial spreadsheet program, describes what people did with them: "... Early on, consultants told us they used it to help lay out slot machines on a casino floor. And doctors did calculations, too. *I didn't know about those applications, or that those people would have even thought to use it that way*. ... [spreadsheet users] were people who were able to figure out to use a tool for a specific problem, *even if it wasn't advertised to do it*. That made them innovators." (italics added). We describe the role played by spreadsheets in catalyzing PC adoption in more detail in the appendix.

One class of AI-enabled technologies satisfying the above requirement includes "low/no-code" platforms that make AI development accessible to non-coders.[22] This approach has already been adopted for automation of enterprise-level software workflows[23]. Although these platforms do not eliminate the need for AI engineers and data scientists, they allow non-engineers to contribute directly to AI development. Wider use of these platforms thus expands the set of people who can use AI effectively, accelerating AI adoption and contributing to follow-on innovation.

3.2 Democratizing Access to Data and Data Infrastructure

3.2.1 Publicly Accessible Data Can be used for Training

Today, massive amounts of data are publicly available both on private repositories and via the federal government[24]. However, making this data usable in AI contexts will require the development of data quality standards and privacy-preserving measures. We propose leveraging an expanded data curator workforce (see below) to release larger public datasets to private industry. This will reduce some of the high upfront costs associated with developing datasets for AI deployment.

3.2.2 The Clearinghouse: Publicly Accessible Models with Licensed Data

We propose creating an easy-to-use web 'clearinghouse' where companies can upload models (or choose from pre-trained models) to be trained on data licensed from other companies and/or publicly available data. The computation would be performed using cloud infrastructure available through the platform.[25] After training, the user would receive optimized coefficients for their model.

This approach enables firms of all sizes to unlock the benefits of large-scale data and computing infrastructure while minimizing risks to privacy and intellectual property. The clearinghouse platform would interface directly with low/no-code AI tools to simplify model development and incorporate cutting-edge publicly-available pre-trained AI models[26]. The clearinghouse would expose sufficient metadata[27] to allow users to optimize their model hyperparameters, but would not permit access to the underlying data[28]. The clearinghouse model can expand access to proprietary data because a trusted intermediary allows companies to train models on others' data without actually accessing it.

---

[22] Examples: DataRobot, Lobe.ai (acquired by Microsoft), Sway AI, Akkio, MonkeyLearn, and Levity.ai. Examples of DataRobot's interface can be found in our appendix.
[23] Pedram Ataee. "No-code ai platforms bring ai to everyone – here is how", TowardsDataScience, Mar 2022. https://towardsdatascience.com/no-code-ai-platforms-bring-ai-to-everyone-here-is-how-8f75b2f6ce9d
[24] Key examples include BEA/BLS for economic data, data.gov for government data, and NIST repositories for technical data.
[25] This could include private cloud computing providers such as AWS as well as government resources, such as supercomputers in use at national laboratories.
[26] Pre-trained AI models (e.g. DALL·E 2, GPT-3) can be fine-tuned using transfer learning to exhibit strong performance on new tasks using relatively lower amounts of data.
[27] Metadata could include feature descriptions, feature summary statistics (including feature coverage), and the quantity of data input.
[28] Large technology companies with massive datasets internally impose similar controls on their engineers – the interface provides enough information to work with the data, but not enough to copy or de-anonymize it.

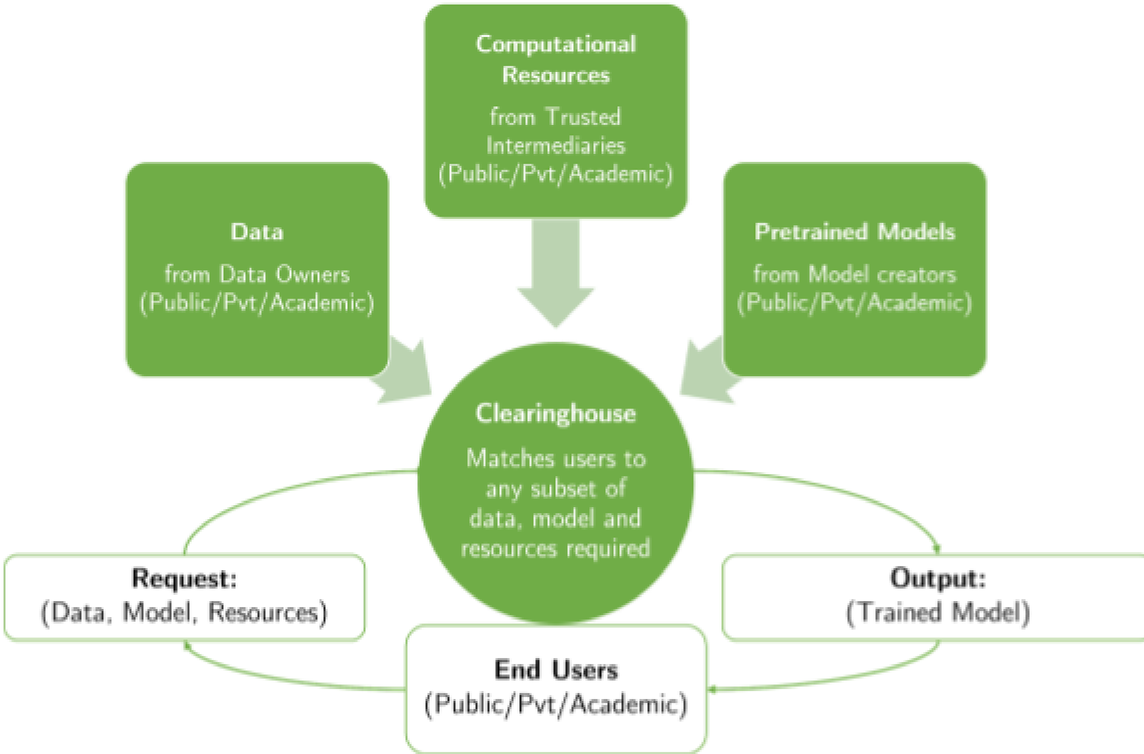


Figure 4: The Clearinghouse model for AI model deployment. End users, typically businesses or academics, submit a request consisting of any combination of a dataset, a model design, and computational resources. The clearinghouse matches the end user to data owners, model creators willing to license their pretrained models and trusted intermediaries willing to rent their computational resources. Model coefficients are returned to the end user on verification of all licensing arrangements. End users are free to upload their own data or their own model architecture to the platform.

Privacy is enhanced because datasets are not shared with end users, and therefore cannot be cross-referenced to expose personally identifiable information.

3.3 The Data Curator Workforce

Our final proposal is for public-private partnership in retraining workers to develop a *medium*-skilled workforce, who we will call data curators, engaged primarily in data management. Data curators would serve as a complementary workforce to data scientists.

3.3.1 What Data Curation would Involve

Data curation would involve data ingestion, exploratory data analysis, and data quality control. Data ingestion includes harmonizing data collected from separate sources and constructing raw datasets. Exploratory data analysis includes documenting and dealing with outliers and missing data, ensuring consistency across data features, and preliminary data visualization. Data quality control includes ensuring that data complies with fairness, legal, licensing, privacy and security standards. More details on the specific tasks associated with data curation are available in Appendix 5.2.

Working in data curation requires the ability to understand and vigilantly apply rules for identifying errors and ensuring compliance. Most of these rules are well-defined[29]. However, other rules may not be so, for instance when ensuring that data meets equity standards for representation of all groups being studied[30].

3.3.2 Why Data Curation?

Job displacement driven by trends in trade and technology has disproportionately affected workers in routine manual tasks. In addition, automation has disproportionately affected growth in a class of jobs providing non-college educated workers an initial entry to higher paying and fulfilling careers[31]. Commentators responding to this often emphasize the need for everyone to acquire "AI literacy" and for a growing fraction of workers to be able to work with today's AI tools[32].

Scaling up policies that seek to directly expand coding and AI skills is unrealistic. First, coding skills are already nontrivial to acquire and teach at the college level[33], and are likely even harder to teach to to mature students[34]. Second, even among programmers, AI skills are particularly challenging to acquire[35]. Third, since program development is a fundamentally collaborative process, adding less skilled coders to programming teams can adversely affect the productivity of existing coders[36].

Our proposed data curator position provides workers previously employed at medium-skill routine task-intensive jobs with a bridge to the non-routine quantitative AI work. The work context gap between traditional manufacturing jobs and data curation is smaller than the gap between such jobs and traditional data science occupations, making data curation more suitable for retraining. The diversity in tasks associated with data curation ensures that both mature and entry-level non-college

---

[29] For instance, a rule to identify outliers can be defined by the traditional data science team such that observations outside the 1st and 99th percentiles of the data distribution are outliers. Once the rule is defined, ensuring compliance with such rules is a routine task.

[30] For instance, see examples at Google PAIR (2021). "Data Collection: Section 1" in *People + AI Guidebook*. https://pair.withgoogle.com/chapter/data-collection/#section1.

[31] This is visible in the declining prime age labor force participation rates among lower educated workers. See, eg. Binder and Bound (2019) "The Declining Labor Market Prospects of Less-Educated Men", *J Econ Perspect.* 33(2): 163–190, Spring 2019. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7745920/

[32] A typical viewpoint: Biden telling coal miners to learn to code. Such views are also perhaps behind ideas such as Amazon's push to train its employees and The Last Mile's program for teaching inmates coding skills.

[33] Sobral, Sonia Rolland (2021). "Strategies on teaching introducing to programming in higher education". In Alvaro Rocha, Hojjat Adeli, Gintautas Dzemyda, Fernando Moreira, and Ana Maria Ramalho Correia (eds.) "*Trends and Applications in Information Systems and Technologies*", pp 133–150, Cham, 2021. Springer International Publishing.

[34] Retraining programs in the US have a history of failure, and one reason is because they place mature learners in college contexts, which are unsuitable learning environments for them (Fadulu, L. "Why is the US so bad at worker retraining?", *The Atlantic*, January 2018).

[35] Carrie J. Cai and Philip J. Guo. Software developers learning machine learning: Motivations, hurdles, and desires. 2019 IEEE Symposium on Visual Languages and Human-Centric Computing (VL/HCC), pages 25–34, 2019.

[36] An apocryphal tale: every mediocre coder creates two coding jobs. Atwood, J. (2009). Nobody hates software more than software developers, July 2009. https://blog.codinghorror.com/nobody-hates-software-more-than-software-developers/

educated workers can be readily trained for different aspects of the data curation pipeline, while also ensuring that the job itself does not consist solely of repetitively applying algorithms[37].

3.3.2 The demand for data curators

Expanded adoption of low/no-code AI, the creation of public data repositories and the clearinghouse model we propose above will all lead to large increases in demand for data curation services that will exceed even the current demand for data skills[38]. This high demand, along with the smaller gap between data curation and existing jobs, would establish both the demand for such positions and the visible viability of data curation as a career, generating appropriate engagement from displaced workers with training programs. Crucial to the success of such programs will be the involvement of local authorities for coordination and local firms to establish demand for these workers and to outline specific training requirements.

3.4 Complementarities across our proposed Policies

Maximizing productivity benefits from our proposed policies will require seamless interoperability. Initial development of low/no-code AI platforms and training of data curators should be tailored to ensure clearinghouse compatibility. The complementaries between our proposed policies are illustrated by a hypothetical example: a small ice cream store could affordably hire a data curator to convert their daily sales into a dataset. They could then use a no-code AI platform via the clearinghouse to train an AI model on a dataset combining their ice cream sales with historical government weather data to accurately forecast product demand. The synergy between our proposals will enable increased AI adoption across the economy and create opportunities for medium-skill job creation. Moreover, establishing a transparent AI ecosystem will generate network effects that encourage accessible and equitable AI development[39].

4 Challenges

4.1 Data Privacy, Security and Storage

Limiting the transfer of sensitive data is essential to protecting privacy and IP rights. Our clearinghouse model avoids sharing data with end users while enabling smaller companies to achieve the economies of scale associated with training AI models on large datasets. The computational infrastructure platforms employed in the clearinghouse (e.g. AWS) already perform secure computations on sensitive data for end users, mitigating back-end security concerns.

---

[37] Certain kinds of data validation tasks may actually require the specific kinds of expertise that mature employees in routine tasks, who are disproportionately likely to be displaced by widespread adoption of automation technologies, possess. For instance, the input to a quality control AI system coupled with CNC machinery in a factory is data on attributes of past defective products, which an assembly line worker might be better qualified to validate than the typical white-collar data scientist.

[38] Employment of data scientists is projected to grow 36 percent from 2021 to 2031, much faster than the average for all occupations. Companies such as Collibra already perform services very similar to data curation, and their existence and high valuations suggests that the demand for their services will remain robust .

[39] For example, even privately developed low/no-code AI platforms will be strongly incentivized to ensure compatibility with the clearinghouse API.

4.2 Concerns about the effectiveness of retraining programs

Retraining programs in the US have a history of failure, primarily due to their improper design[40]. Appropriately designing retraining programs for the requirements of a displaced populace is a general challenge across all domains. One key obstacle to many retraining programs is the large gap between the skills most displaced workers possess and the skills they are being trained for, which may require fundamentally different activities and aptitudes.[41] We believe that the skills associated with a typical manufacturing job (e.g. attention to detail, ability to maintain focus across repetitive tasks, etc.) translate much more seamlessly to data curator roles than to computer programming, and that there will be a substantial demand for data curators in an accessible AI ecosystem. Therefore, we anticipate that our proposal will avoid at least some challenges associated with retraining programs.

4.3 Automation or Outsourcing of data curator tasks to low-wage countries

Since data curator tasks are likely to be relatively routine tasks, they are targets for automation and outsourcing to low-wage countries, especially as these countries raise their education and digital competencies[42]. We believe these concerns is unlikely to be relevant for three key reasons:

- **Privacy Regulation:** As policymakers enact increasingly restrictive data localization requirements, most data curator tasks which require direct data inspection will become increasingly difficult to outsource[43].
- **Supervision:** As data quality becomes ever more mission critical, direct supervision of the data pipeline will become more important for firms.
- **Tacitness:** A number of tasks in the data curator framework involve judgment calls which are difficult to codify, which makes them difficult to either automate or outsource.

[40] Key pitfalls include the reluctance of displaced workers to return to college settings in which most programs are conducted and concerns about the lack of practical applications of skills acquired by retraining. See Selingo, J. “The False Promises of Worker Retraining”, The Atlantic, January 2018.
[41] For instance, manufacturing jobs tend to be classified as routine manual jobs, while jobs in ICT industries tend to require cognitive skills and creativity.
[42] For instance, up to 46% of data entry work is expected to be outsourced by 2025. (ARDEM. “Why remote work is accelerating outsourcing of data entry tasks”, May 2022).
[43] For example, the EU’s GDPR requires that to transfer data collected within the EU outside the country, the receiving country’s data protection norms be in line with the GDPR level of protection.

5 References and works consulted

Stuart Russell and Peter Norvig. Artificial Intelligence: A Modern Approach. Prentice Hall, 4 edition, 2021.

National Institute of Standards and Technology. U.S. Leadership in AI: A Plan for Federal Engagement in Developing Technical Standards and Related Tools, 2019.

Ajay Agrawal, Joshua S. Gans, and Avi Goldfarb. Artificial intelligence: The ambiguous labor market impact of automating prediction. Journal of Economic Perspectives, 33(2):31– 50, May 2019.

Andrew A. Toole, Nicholas A. Pairolero, Alexander V. Giczy, James Q. Forman, Christyann Pulliam, Matthew Such, Kakali Chaki, David B. Orange, Anne Thomas Homescu, Jesse Frumkin, Ying Yu Chen, Vincent M. Gonzales, Christian Hannon, Steve Melnick, Eric Nilsson, and Ben M. Rifkin. Inventing AI: Tracing the diffusion of artificial intelligence with U.S. patents. Technical report, U.S. Patent and Trademark Office – Office of the Chief Economist, October 2020.

Daron Acemoglu, Gary Anderson, David Beede, Catherine Buffington, Eric Childress, Emin Dinlersoz, Lucia Foster, Nathan Goldschlag, John Haltiwanger, Zachary Kroff, Pascual Restrepo, and Nikolas Zolas. Automation and the workforce: A firm-level view from the 2019 annual business survey. Working Paper CES-22-12, United States Census Bureau, April 2022.

K. McElheran, J. F. Li, E. Brynjolfsson, Z. Kroff, E. Dinlersoz, L. Foster, and N. Zolas. Ai adoption in america: Who, what and where. Center for economic studies working paper, US Census Bureau, 2022.

Michael Chui, Bryce Hall, Alex Singla, and Alex Sukharevsky. The state of AI in 2021. Technical report, December 2021.

Prasanna Tambe, Lorin Hitt, Daniel Rock, and Erik Brynjolfsson. Digital capital and superstar firms. Working Paper 28285, National Bureau of Economic Research, December 2020.

Daniel Zhang, Nestor Maslej, Erik Brynjolfsson, John Etchemendy, Terah Lyons, James Manyika, Helen Ngo, Juan Carlos Niebles, Michael Sellitto, Ellie Sakhaee, Yoav Shoham, Jack Clark, and Raymond Perrault. ”The AI Index 2022 Annual Report”. Technical report, ”AI Index Steering Committee, Stanford Institute for Human-Centered AI, Stanford University”, March 2022.

J. Loucks, T. Davenport, and D. Schatsky. Deloitte state of ai in the enterprise, 2nd edition. Technical report, Deloitte Insights, 2018.

Timothy Bresnahan, Erik Brynjolfsson, and Lorin M. Hitt. Information technology, work- place organization, and the demand for skilled labor: Firm-level evidence. The Quarterly Journal of Economics, 117(1):339–376, 2002.

Erik Brynjolfsson and Lorin M Hitt. Beyond computation: Information technology, or- ganizational transformation and business performance. Journal of Economic perspectives, 14(4):23–48, 2000.

Michael Webb. The impact of artificial intelligence on the labor market. mimeo, 2020.

Leonid Kogan, Dimitris Papanikolaou, Lawrence D. W. Schmidt, and Bryan Seegmiller. Technology-Skill Complementarity and Labor Displacement: Evidence from Linking Two Centuries of Patents with Occupations. NBER working Papers 29552, National Bureau of Economic Research, Inc, December 2021.

Marguerita Lane and Anne Saint-Martin. The impact of artificial intelligence on the labour market. 2021.

Jae Song, David J Price, Fatih Guvenen, Nicholas Bloom, and Till von Wachter. Firming Up Inequality. The Quarterly Journal of Economics, 134(1):1–50, 2019.

Jaana Holvikivi. Conditions for successful learning of programming skills. In Key Competencies in the Knowledge Society, 2010.

Slintel. Market share of pytorch, August 2022.

u/fereal fire. Why is tensorflow so hard, July 2020.

TF Certification. How to learn tensorflow the painless way: 7 tips, March 2021.

Nico Jimenez. Tensorflow sucks, October 2017.

David H. Autor. Why are there still so many jobs? the history and future of workplace automation. Journal of Economic Perspectives, 29(3):3–30, September 2015.

Shankar Vedantam. Why more men don’t get into the field of nursing, October 2018.

Larysa Visengeriyeva, Anja Kammer, Isabel Ba ̈r, Alexander Kniesz, and Michael Plo ̈d. An overview of the end-to-end machine learning workflow, July 2022.

Erik Brynjolfsson. The turing trap: The promise & peril of human-like artificial intelligence. Daedalus, 151(2), 2022.

Pedram Ataee. No-code ai platforms bring ai to everyone – here is how, Mar 2022.

Gero Keil. Mapping the no-code ai landscape, July 2022.

Luciano Floridi and Massimo Chiriatti. Gpt-3: Its nature, scope, limits, and consequences. Minds and Machines, 30(4):681–694, 2020.

Karl Weiss, Taghi M Khoshgoftaar, and DingDing Wang. A survey of transfer learning. Journal of Big data, 3(1):1–40, 2016.

Google People + AI Research team. Data collection + evaluation.

Nithya Sambasivan, Shivani Kapania, Hannah Highfill, Diana Akrong, Praveen Kumar Paritosh, and Lora Mois Aroyo. ”everyone wants to do the model work, not the data work”: Data cascades in high-stakes ai. 2021.

Anaconda. The state of data science 2020, 2020.

Nanna Sandberg, Tom Hoy, and Martin Ortlieb. My ai versus the company ai: How knowledge workers conceptualize forms of ai assistance in the workplace. Ethnographic Praxis in Industry Conference Proceedings, 2019(1):125–143, 2019.

Larysa Visengeriyeva, Anja Kammer, Isabel Bär, Alexander Kniesz, and Michael Plod. An overview of the end-to-end machine learning workflow, July 2022.

Sonia Rolland Sobral. Strategies on teaching introducing to programming in higher education. In Alvaro Rocha, Hojjat Adeli, Gintautas Dzemyda, Fernando Moreira, and Ana Maria Ramalho Correia, editors, Trends and Applications in Information Systems and Technologies, pages 133–150, Cham, 2021. Springer International Publishing.

World Bank. Coding bootcamps for youth employment : Evidence from colombia, lebanon, and kenya. Technical report, World Bank, 2018.

Suzanne Straw, Susie Bamford, and Ben Styles. Randomised controlled trial and process evalution of code clubs. Technical report, National Foundation for Educational Research and the Raspberry Pi Foundation, March 2017.

Lola Fadulu. Why is the US so bad at worker retraining?, January 2018.

Peter Wayner. Coding bootcamps won’t make you a developer: Here’s what will.

Carrie J. Cai and Philip J. Guo. Software developers learning machine learning: Motivations, hurdles, and desires. 2019 IEEE Symposium on Visual Languages and Human-Centric Computing (VL/HCC), pages 25–34, 2019.

Jeff Atwood. Nobody hates software more than software developers, July 2009.

Mallory Dreyer and Kaleb Nygaard. Governments encourage smes to adopt new technology, June 2020.

Nicholas Carlini, Florian Tramer, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song, Ulfar Erlingsson, Alina Oprea, and Colin Raffel. Extracting training data from large language models, 2020.

AWS. Aws cloud security, 2022.

Jeffrey Selingo. The false promises of worker retraining. January 2018.

ARDEM. Why remote work is accelerating outsourcing of data entry tasks, May 2022.

Your Europe. Data protection under GDPR, June 2022.

6 Appendix

6.1 The Spreadsheet

The PC-bound spreadsheet application, first commercialized by VisiCalc and today exemplified by the likes of Microsoft's Excel and Google's Sheets applications, was a key driver of personal computer adoption in businesses. As late as 1983, the personal computer – then exemplified by the IBM PC, running an operating system called PC DOS licensed from the then fledgling company Microsoft – was widely dismissed as an expensive tool for hobbyists, and most computer users associated real work with mainframe computers. Analysts seeking specific information on a company would send a request to the mainframe terminal, receive a preformatted report within days or weeks depending on their priority, and then search through this report for the specific information they sought[44]. Much of company accounting was performed by hand or by inputting data afresh each time in a predetermined sequence as required by the programs on the mainframe[45].

The spreadsheet enabled both the storage of different kinds of data in one location and the integration of this data, and performed all necessary calculations on the personal computer that the analyst was using. Further, spreadsheet users could develop their own programs, called macros, which allowed them to adapt the data presentation and analysis power of spreadsheet programs for their own purposes[46]. The key difference between spreadsheet software like VisiCalc and existing business analytics software was just how easy it was for novices to use them. Ben Rosen, who would later found Lotus, raved on the launch of VisiCalc that "... VisiCalc comes alive visually. *In minutes, people who have never used a computer are writing and using programs.* ..." (italics added)

The success of VisiCalc led to the curious case of individual business owners beginning to buy PCs for the sole purpose of running VisiCalc. This success also led to the development of multiple competitors. The most successful of these was Lotus 1-2-3, a program which almost single-handedly drove the success of the IBM PC in modern businesses. Lotus would retain its dominance until the 1990s, when Microsoft's Excel, a spreadsheet program with a graphical user interface, began gaining popularity[47]. Today, spreadsheets in general are the second most used class of applications on computers, second only to email, and are a required job skill in up to 1 in 3 jobs[48]!

6.2 Tasks associated with Data Curation

We envision data curation as an occupation dedicated to performing the first step of a standard MLOps framework for AI use in business. This includes the following tasks.

---

[44] Palatto (2013), "30 Years Ago: PC Spreadsheets Bring Number Crunching to the Masses", EWeek.com, https://www.eweek.com/enterprise-apps/30-years-ago-pc-spreadsheets-bring-number-crunching-to-the-masses/
[45] https://www.quora.com/What-was-accounting-like-before-spreadsheets
[46] https://medium.grid.is/3-things-you-dont-understand-about-spreadsheets-part-2-43a4fb7b6a45
[47] Curiously, the first versions of MS Excel released in 1985 were written for use on Apple computers, and only in 1987 did Windows 2.x versions of Excel appear.
[48] https://medium.grid.is/3-things-you-dont-understand-about-spreadsheets-part-2-43a4fb7b6a45

- **Data Ingestion:** Collect data from different systems/verticals within the organization and combine them into complete datasets. May also include data enrichment or synthetic data generation.

- **Data Validation and Exploration:**
  - Data profiling to construct metadata about features.
  - Data validation to ensure that the dataset is error free, usually implemented by running user-defined error detection functions.

- **Exploratory Data Analysis:**
  - Dealing with outliers and erroneous observations.
  - Imputation or elimination of observations with missing values.
  - Ensuring consistency across labels for observations (e.g., ensuring that observations associated with a given zipcode are labeled with the correct county associated with that zipcode), in general ensuring that data is error-free for the purposes of the project.
  - Constructing data weights to ensure representativeness of sample data.
  - Data visualization to inform model construction.
  - Data splitting to create appropriate training, validation and test datasets.

- **Ensuring data compliance:** ensuring fairness, legal, privacy, licensing and security issues are dealt with.
  - Ensuring data is representative of the population, and applying weighting or resampling techniques if not
  - Ensuring that all permissions associated with data usage have been obtained, especially for licensed data stored locally
  - Ensuring that data storage complies with safeguards and standards associated with data privacy and protection (for instance, that Personally Identifiable Information is not inadvertently disclosed)

Importantly, note that these tasks involve a mix of routine and non-routine tasks, as well as requiring different levels of cognitive ability. This mix ensures that the data curation job strikes a balance between being a job for which workers hitherto engaged in relatively routine manual tasks can be retrained relatively easily, and yet being a job that provides a pathway to the more mathematics and coding intensive components of the MLOps framework.

6.3 Our proposals and status quo

Policy	Status Quo and Existing Examples to build on	Our Proposals
Promote Easy-to-Use AI Technologies	NSF's Future of Work at the Human-Technology Frontier supports research in creation of human-AI interfaces for specific occupational classes	Raise priority and support for technologies that augment business practices, codify precisely which technologies qualify based on the usability, widespread relevance and minimal disruption criteria. Disbursement either via business development grants as part of the Small Business Innovation Research Program (SBIR) or creation of a new program under the National AI Initiative Office.
Publicly Available Training Data	NIST and BLS/BEA/Census data repositories, existing data collection and management infrastructure associated with these datasets. Eg, calculation of the CPI requires the collection of 94,000 prices across the US each month[49].	Harmonize existing sources, apply protocols for data security and anonymization, make data available to the private sector. Possible models include direct management by govt (Research Data Center model as in the Census) or licensing data storage and delivery to trusted private parties (eg AWS) associated with clearinghouse model
Resource Clearinghouse Model	US FGDC's GeoPlatform.gov, National Research Cloud, Private Cloud Infra. (e.g., AWS), Pre-trained models licensed for general use (e.g. OpenAI's GPT-3 API). Wide range of standards, data formats, not much interoperability or scope for data/model reuse, limited B2B data sharing and expensive proprietary datasets.	Significant centralization of storage and standardization of data, raise the scope of what data is available as far as possible, public-private partnership in management, focus on business needs and for training AI models when it comes to determining access priority. Creation of novel IP regimes for resource, model and data licensing on a B2B basis.
Data Curator Workforce	Public training programs encouraging development of coding skills, private initiatives for upskilling own workforces, piecemeal support for introducing coding/AI skills in lower education levels	Targeted training for data curation tasks targeted at populations most likely to be displaced by AI deployment (lower skilled individuals in routine cognitive tasks, eg manufacturing sector or lower-end white collar jobs like secretaries)

[49] https://www.bls.gov/opub/hom/cpi/data.htm