Understanding How Startups Use AI-as-a-Service

November 20, 2024

Stephen Michael Impink, HEC Paris
Mari Sako, University of Oxford
Robert Seamans, New York University

This report provides the latest data from the 6th AI Startup Survey responses captured in 2024.

Access to high-quality data is a strategic advantage when developing AI products (Mayer-Schönberger & Ramge, 2022). Data is a key input used to train all types of algorithms, including the most recent large language models (LLMs) supporting generative AI. These latest technical advances that use neural network algorithms have unlocked anticipated potential in natural language comprehension capabilities, making chatbots and digital assistants “conversational.”[1] While open-source training datasets – such as the WikiQA Corpus, Yahoo Language Data, and Ubuntu Dialogue Corpus – are freely available to pre-train a model, startups need to collect their own data to fine-tune their performance. For competitive advantage, firms may need access to proprietary data from customers and suppliers (Bessen et al., 2022). Incumbents, especially the so-called Big Tech firms – Alphabet, Amazon, Apple, Meta, Microsoft – have vast piles of customer data collected about users via their platforms. Moreover, they have complementary cloud services businesses that enable easy access to data storage and computational ability. Their advantage dwarfs the data collection capacity of startups and smaller competitors. For this reason, startups may turn to the generation and use of synthetic data (Bolón-Canedo, Sánchez- Maroño, & Alonso-Betanzos, 2013; Lucini, 2021). Alternatively, they may be unable to move forward with developing their own AI products internally because they lack the necessary resources to compete in the market.

While all types of machine-learning-based models are "data hungry," the advent of LLMs is making access to data a more important source of competitive advantage for AI startups than before. One reason is that developing an LLM from scratch is very costly, so AI startups are more likely to use an existing LLM model pre-trained with publicly available data (Ray, 2023) to develop specific use cases. Adapting and refining a chosen LLM for an identified use case requires access to training data and an external database for retrieval augmented generation (RAG). Despite the strategic importance of data access, AI projects may not acknowledge this upfront in the AI project lifecycle (Vial, Jiang, Giannelia, & Cameron, 2021). From ideation, blueprint, proof-of-concept, minimal viable product, and production, data tends to be disconnected

[1] https://aichat.com/2019/06/27/data-is-the-key-to-develop-a-truly-conversational-chatbot/

from the organization undertaking the AI project until it hits the production phase, when live data integration has to occur without much preparation in earlier phases.

Evidence from the AI Startup Survey

To date, good evidence exists on specific aspects of this issue from the past rounds of the AI venture survey (Bessen, Impink, Reichensperger, & Seamans, 2023). First, access to data is important for startups, and there is self-awareness about this among many startups: about 37% of survey respondents agree that training data is most important to their firm's success. Second, data provided by their customers are the most dominant type of data for startups. Eighty-three percent of survey respondents reported using their customers' data to train their AI models, and more than half reported using data available from third parties, including publicly available data. These results provide evidence that more than 40% of startups use data from multiple sources. This is unsurprising as different data types are needed for pre-training, fine-tuning, testing, and RAG. Customers' proprietary data also correlates with increased future performance, including venture capital funding (Bessen et al., 2022a), suggesting that the inability to access data constitutes a barrier for startups to scale.

Third, survey evidence also exists on contracting for data. Among the startups that use customer data, 45% of respondents use legal contracts with customers to specify data use (Bessen et al., 2023)(p.29). More specifically, 52% of respondents retain data reuse rights. These rights promote data portability, creating value for startups by facilitating collaboration with other providers, leveraging data network effects, and accessing new markets. These pieces of evidence constitute a great starting point for digging deeper into the nature of contracts and what the contracts say about data.

The AI venture survey has led to addressing a variety of related research questions including:

a) Does access to data constitute barriers to entry for startups? For example, does it create anti-competitive trends in which Big Tech companies as incumbents have advantages over new entrants due to the large resource stocks they can command and their first-mover advantages (Mayer-Schönberger & Ramge, 2022; Sokol & Comerford, 2016)?

b) Do policy shifts such as the EU's GDPR impose undue costs on AI startups (Bessen, Impink, Reichensperger, & Seamans, 2020; Jia, Jin, & Wagman, 2018)? How does regulation of data and AI influence the nascent AI development industry?

c) How widespread is the adoption of ethical AI principles among AI startups, and do these principles translate into ethical AI use practices such as turning down business due to ethical principles (Bessen, Impink, & Seamans, 2024), and implementing robust corporate governance such as having audits and AI oversight boards (Bessen, Impink, & Seamans, 2023)?

Aligning incentives in incomplete contracts

Contracting for data to train AI algorithms poses challenges over and above the usual challenges arising from contracting for innovative activities whose outcomes are difficult to specify fully at the outset (Gilson, Sabel, & Scott, 2009). These challenges include difficulty in allocating liabilities when training AI systems go wrong, difficulty in allocating benefits from training due to the black box nature of learning algorithms[2], and lack of clarity around data ownership (Sako, 2023). Incomplete contracting is a good theoretical lens through which to interpret the empirical phenomenon of AI startups accessing client data by giving clients an equity stake in their venture. The lack of interpretability in machine learning (ML) algorithms (Lipton, 2018) due to the "black box" nature of ML processes and outputs has received much attention. Moreover, interpretability issues have led to recommendations for improving explainability (Wachter, Mittelstadt, & Russell, 2017) and reducing opacity of various sorts (Burrell, 2016).[3] However, no prior work has explored the implications of this problem for contracting between the vendors of AI systems and data.

[2] We use the terms "learning algorithm" and "AI system" interchangeably in this paper, although the former might be considered a subset of the latter.

[3] An important form of opacity centers on the mismatch between mathematical procedures of machine learning algorithms and human styles of semantic interpretation. Imposing human interpretative reasoning on a mathematical process of statistical optimization, e.g., to classify spam emails, does not eliminate ambiguities around whether an explanation provided is a correct one.

Efforts to improve accuracy are difficult to observe and, therefore, contract on. Humans have complete visibility over the data (as an input) and the results (the output) for AI systems. But what happens in between, especially with deep learning and neural networks, is a black box with little transparency about how or why the model produced a specific output. Lack of transparency can create complex liability issues. When AI tools and services that function without much human intervention fail, it becomes challenging to determine who bears responsibility. The failure might result from a technical malfunction or human error in how the model was designed and trained. Also, improving an algorithm involves trial and error, making it difficult to attribute the cumulative benefit of adding a new feature (i.e. a measurable characteristic of a phenomenon) to one party or the other, as the existing features may not be orthogonal.

The property rights approach to addressing incomplete contracts focuses our attention on how changes in the distribution of asset ownership affect the incentives of the parties who work, directly or indirectly, with those assets (Hart, 1989; Hart & Moore, 1990). It starts with an observation that real-world contracts are almost always "incomplete" in the sense that not all eventualities and contingencies are specified because they were either unforeseen or too expensive to enumerate in sufficient detail (Hart, 1995). For example, the "quality" of professional work like legal advice and the level of "care" in maintaining machinery are often too costly, if not impossible, to elaborate in an enforceable manner in a contract. Each party to a transaction will have certain rights under the contract, but its incompleteness means that some "residual rights" remain unspecified in the contract. When these rights pertain to the use of an asset, property ownership is the institution that allocates these residual rights of control. All rights to the asset not expressly assigned in the contract accrue to the "owner" of the asset. Specifically, the owner has the right to use the asset in any way within the bounds of a prior contract, custom, or existing law.

With this theoretical backdrop, whenever two parties to a transaction each have their asset, A and B, and the two are complementary, incentives are better aligned by having the control of both assets vested in the same party.[4] Separating control would lead to holdup problems (Brynjolfsson, 1994)(p.1653). While

[4] Complementarity is a matter of degree. In the strong form, assets A and B are complementary if A is

unified ownership (vertical integration) is a clean solution, incomplete contracts may be addressed by quasi-vertical integration, as well as relational contracting to enhance flexibility, adaptation, and renegotiation over time as new information becomes available or as circumstances change (Frydlinger, Hart, & Vitasek, 2019).

This framework was developed initially in a context in which human capital (which is inalienable because it cannot be bought and sold) was seen to be complementary to physical assets (such as machinery, factories, and inventories). In the present context of contracting for AI systems and data, we can treat learning algorithms and data as complementary assets. Indeed, learning algorithms cannot learn without data, and data cannot be "monetized" without access to learning algorithms. Then, contractual difficulties arise from a situation in which the learning algorithm owner is different from the data owner, and these difficulties can be resolved by vertical integration, leading to both being owned by one entity. This eliminates the need to allocate liabilities to either party whenever training AI systems goes wrong. Moreover, all the training benefits can be kept within the firm. This leads to the following hypotheses.

Lemley and McCreary provide evidence that at the turn of the century, most successful Silicon Valley startups went for an initial public offering (IPO), whereas by the 2010s, most were bought up by incumbents. Exits by IPO declined from one in two exits during 1995-2000 to one in ten exits during 2010-2015, and exits by incumbent acquisition of VC-backed firms increased exponentially (Lemley & McCreary, 2021) (pp.17-18). The reasons for this trend are the monopoly power of Big Tech firms and the fact that venture capitalists need to cash out on exits. However, over and above these reasons, the acquisition of startups by their customers may result from attempts to resolve the incomplete contracting problem.

"indispensable" to B, and vice versa, so each asset is valueless without access to the other asset. In a weaker form, A and B are complementary in the sense that joint use leads to synergistic value creation.

Latest Survey Results

Below, we provide results on the new survey questions added to this round by region. In the appendix, we provide details on new responses to prior questions.

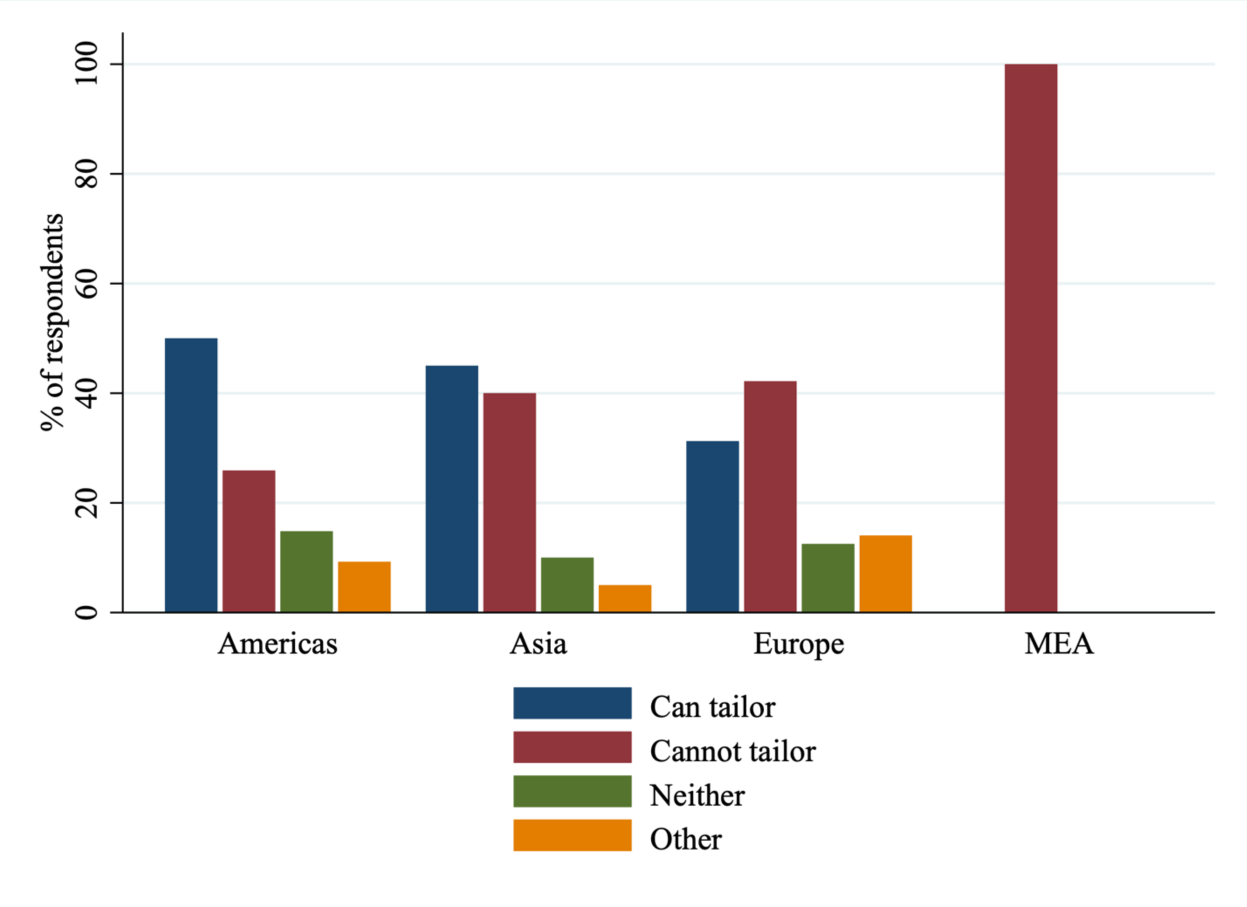
Question 1: Which statement about customizing AI products applies to your startup?

(a) Your startup licenses an AI product that customers can tailor themselves using their training data

(b) Your startup licenses an AI product that customers cannot tailor themselves using their training data

(c) Neither of these statements

(d) Other



Notes: There is a significant negative correlation -0.17 (SD 0.84) between “can tailor” and “can reuse.”

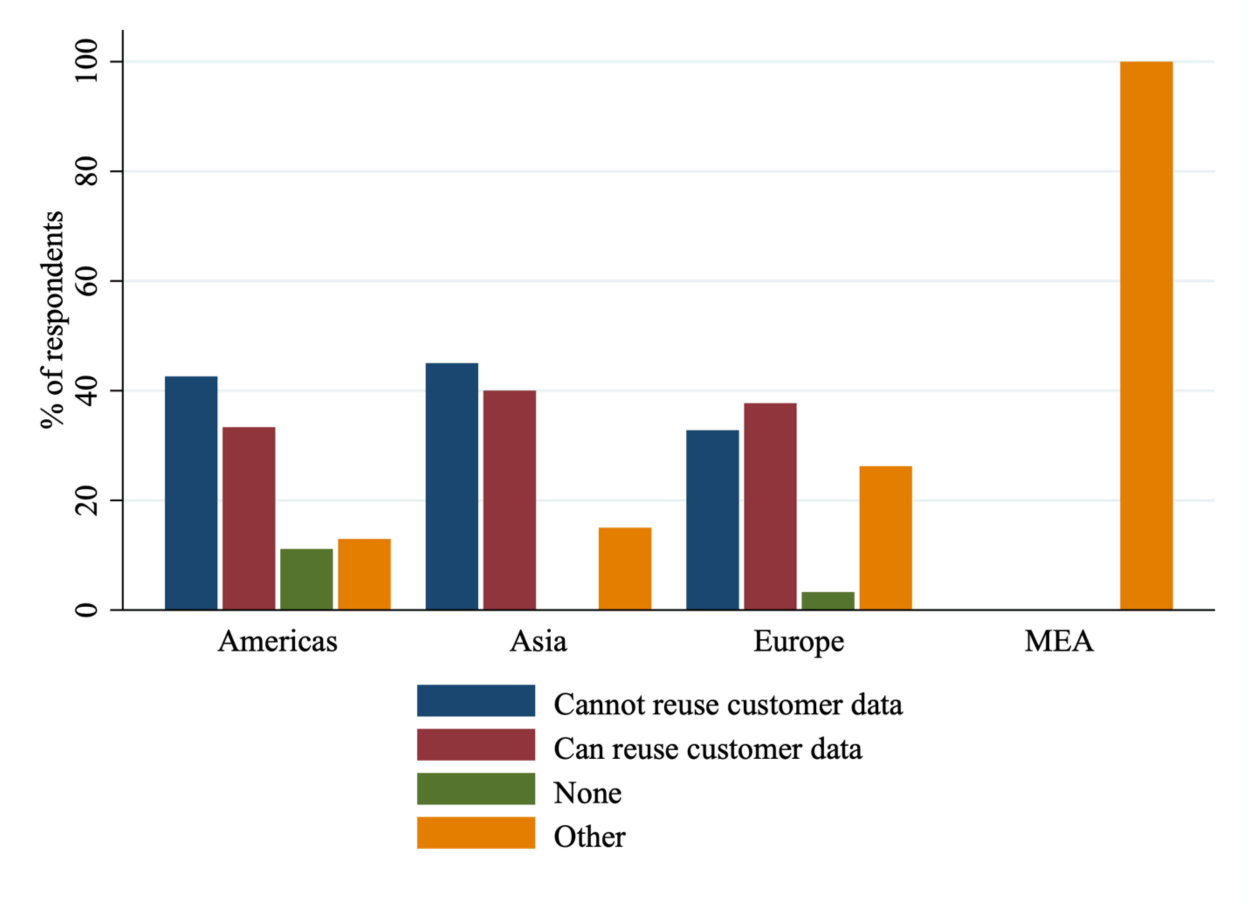
Question 2: Which statement about use and reuse of customer data applies to your startup? *(Please select all that apply.)*

(a) Your startup uses the customer's data to train the AI product for the customer, but you cannot reuse customer data for any other purpose.

(b) Your startup uses the customer's data to train the AI product, and retain the right to pool customer data from different customers to improve the AI product.

(c) None or N/A

(d) Other



Notes: There is no significant correlation between the response to the firm outcome question (i.e., acquisition) and "can tailor" or "can reuse."

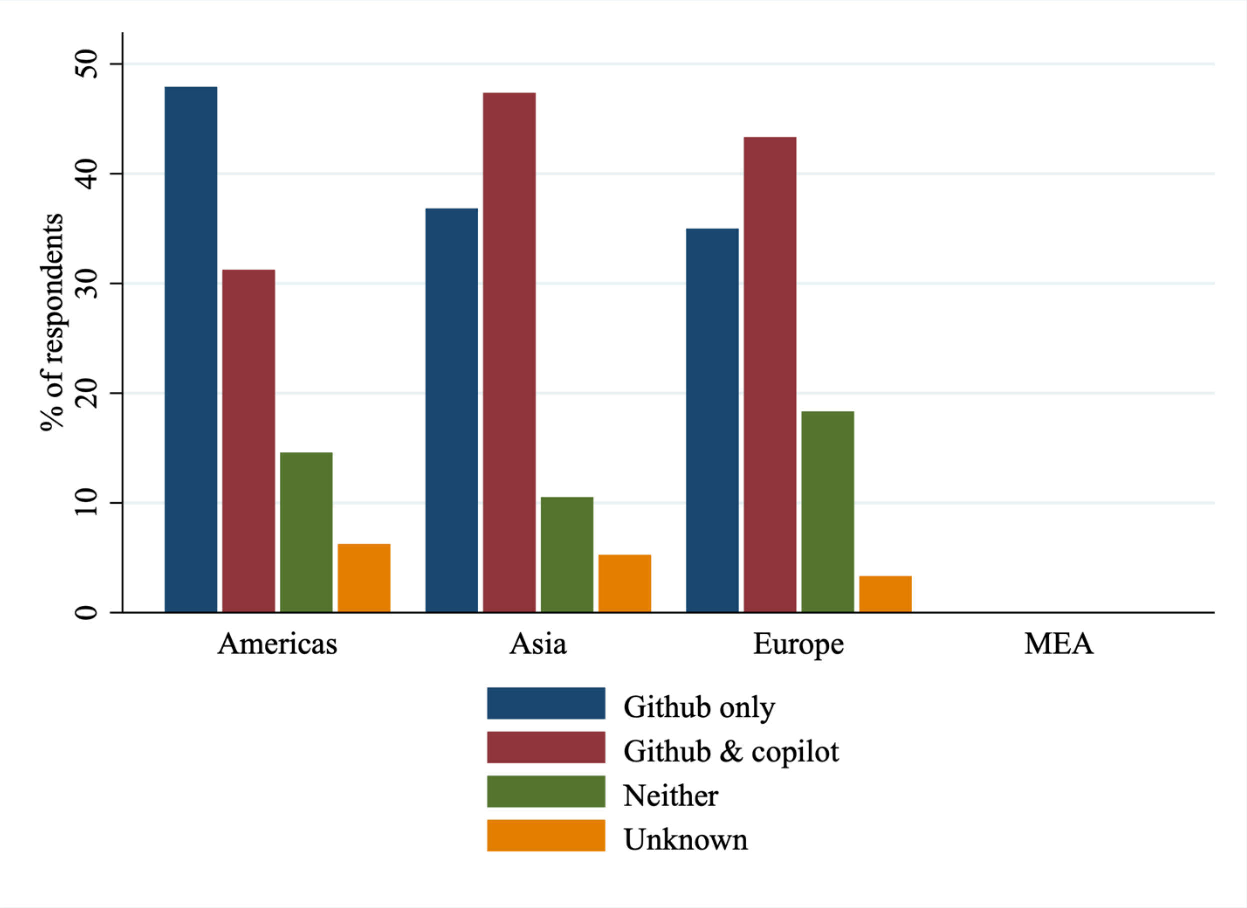
Question 3: Does your startup use GitHub (an opensource coding community) and GitHub's Copilot (a tool that suggests code as you type)?

(a) We use GitHub only

(b) We use GitHub and Copilot

(c) No, we use neither GitHub nor Copilot

(d) I don't know



References

Banko, M., & Brill, E. 2001. Scaling to very very large corpora for natural language disambiguation. Paper presented at the Proceedings of the 39th annual meeting of the Association for Computational Linguistics.

Bessen, J., Impink, S. M., Reichensperger, L., & Seamans, R. 2022a. The role of data for AI startup growth. *Research Policy*, 51(5): 104513.

Bessen, J., Impink, S. M., & Seamans, R. 2022b. Report: Self-governing ethical AI development in entrepreneurship.

Bessen, J. E., Impink, S. M., Reichensperger, L., & Seamans, R. 2020. GDPR and the Importance of Data to AI Startups. NYU Stern School of Business.

Bessen, J. E., Impink, S. M., Reichensperger, L., & Seamans, R. 2023. The business of AI startups. Boston Univ. School of Law, Law and Economics Research Paper (18-28).

Bessen, J. E., Impink, S. M., & Seamans, R. 2022c. Ethical AI development: Evidence from AI startups. Available at SSRN 3895939.

Bolón-Canedo, V., Sánchez-Maroño, N., & Alonso-Betanzos, A. 2013. A review of feature selection methods on synthetic data. Knowledge and information systems, 34: 483- 519.

Brynjolfsson, E. 1994. Information assets, technology and organization. *Management Science*, 40(12): 1645-1662.

Burrell, J. 2016. How the machine 'thinks': Understanding opacity in machine learning algorithms. *Big Data & Society*, 3(1): 2053951715622512.

Frydlinger, D., Hart, O., & Vitasek, K. 2019. A new approach to contracts: how to build better long-term strategic partnerships. *Harvard Business Review*, 97(5): 116-126.

Gilson, R. J., Sabel, C. F., & Scott, R. E. 2009. Contracting for innovation: vertical disintegration and interifrm collaboration. *Columbia Law Journal*, 109(3): 431-502.

Gregory, R. W., Henfridsson, O., Kaganer, E., & Kyriakou, H. 2021. The role of artificial intelligence and data network effects for creating user value. *Academy of Management Review*, 46(3): 534-551.

Gregory, R. W., Henfridsson, O., Kaganer, E., & Kyriakou, H. 2022. Data network effects: Key conditions, shared data, and the data value duality. *Academy of Management Review*, 47(1): 189-192.

Gurkan, H., & de Véricourt, F. 2022. Contracting, pricing, and data collection under the AI flywheel effect. *Management Science*, 68(12): 8791-8808.

Halevy, A., Norvig, P., & Pereira, F. 2009. The unreasonable effectiveness of data. IEEE intelligent systems, 24(2): 8-12.

Hart, O. 1989. Economist's Perspective on the Theory of the Firm, *Columbia Law Review*, 89: 1757.

Hart, O. 1995. Firms, contracts, and financial structure: Clarendon Press. London, UK.

Hart, O., & Moore, J. 1990. Property Rights and the Nature of the Firm. *Journal of Political Economy*, 98(6): 1119-1158.

Jia, J., Jin, G. Z., & Wagman, L. 2018. The short-run effects of GDPR on technology venture investment: National Bureau of Economic Research.

Lemley, M. A., & McCreary, A. 2021. Exit strategy. BUL Rev., 101: 1.

Lipton, Z. C. 2018. The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery. *Queue*, 16(3): 31-57.

Lucini, F. 2021. The real deal about synthetic data. *MIT Sloan Management Review*, 63(1): 1-4.

Mayer-Schönberger, V., & Ramge, T. 2022. The Data Boom Is Here-It's Just Not Evenly Distributed. *MIT Sloan Management Review*, 63(3): 7-9.

Ray, P. P. 2023. ChatGPT: A comprehensive review on background, applications, key challenges, bias, ethics, limitations and future scope. Internet of Things and Cyber- Physical Systems.

Sako, M. 2023. Contracting for Artificial Intelligence. *Communications of the ACM*, 66(4): 20-23.

Sokol, D. D., & Comerford, R. E. 2016. Does antitrust have a role to play in regulating big data? Cambridge Handbook of Antitrust, Intellectual Property and High Tech, Roger D. Blair & D. Daniel Sokol editors, Cambridge University Press, Forthcoming.

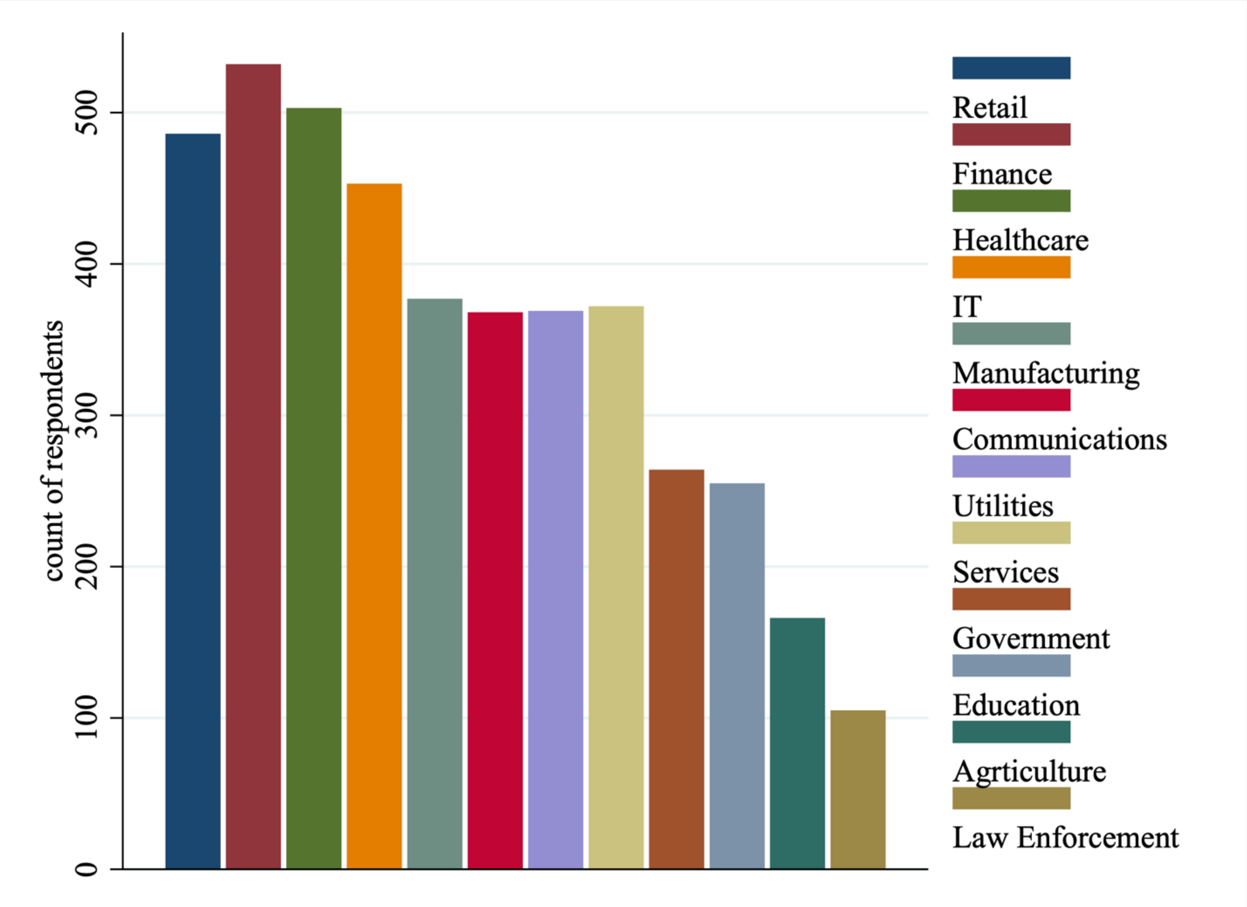
Vial, G., Jiang, J., Giannelia, T., & Cameron, A.-F. 2021. The data problem stalling AI. *MIT Sloan Management Review*, 62(2): 47-53.

Wachter, S., Mittelstadt, B., & Russell, C. 2017. Counterfactual explanations without opening the black box: Automated decisions and the GDPR. *Harvard Journal of Law & Technology*, 31: 841.

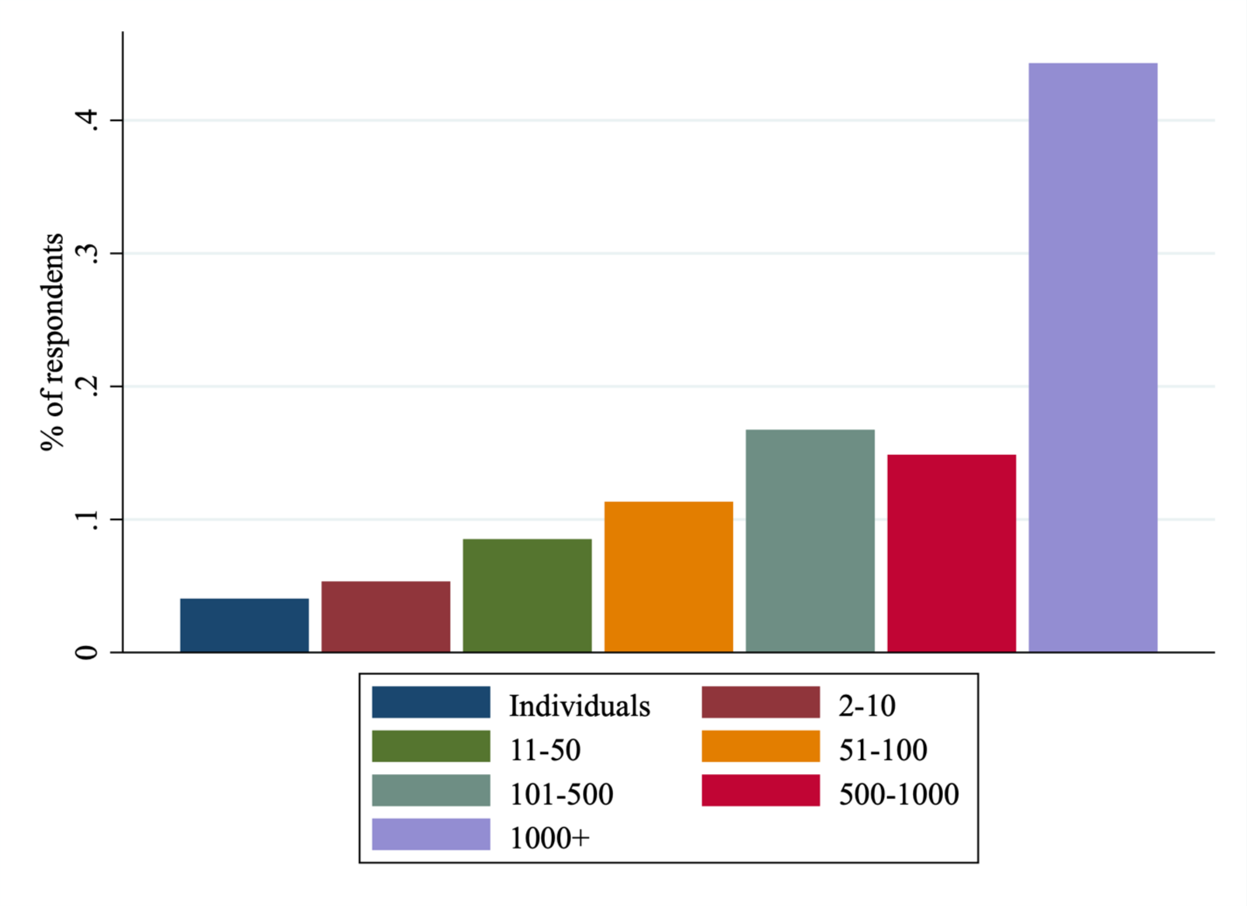
Appendix

Figures in this appendix provide results from the last 6 rounds of the survey.

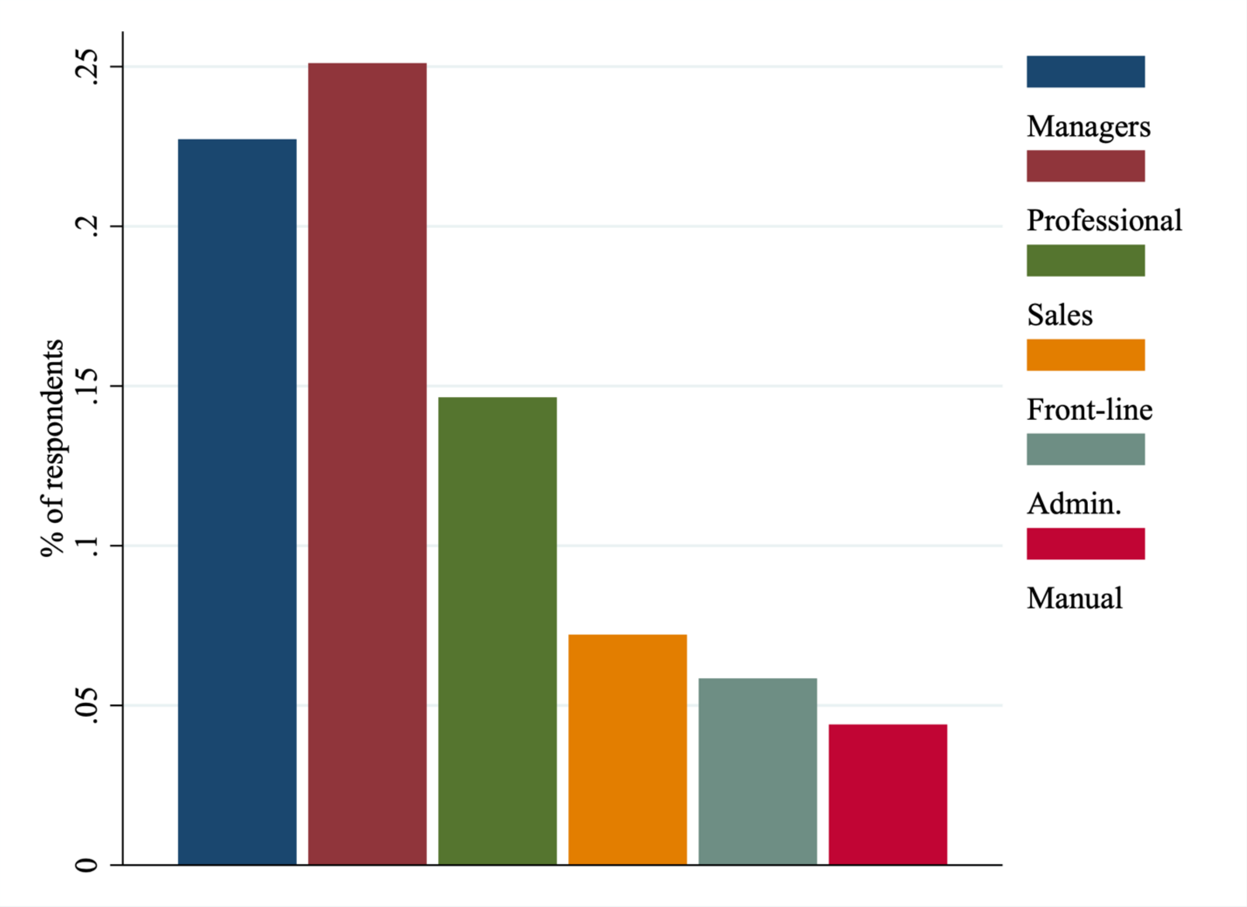
Figure A.1 – Count by Industry



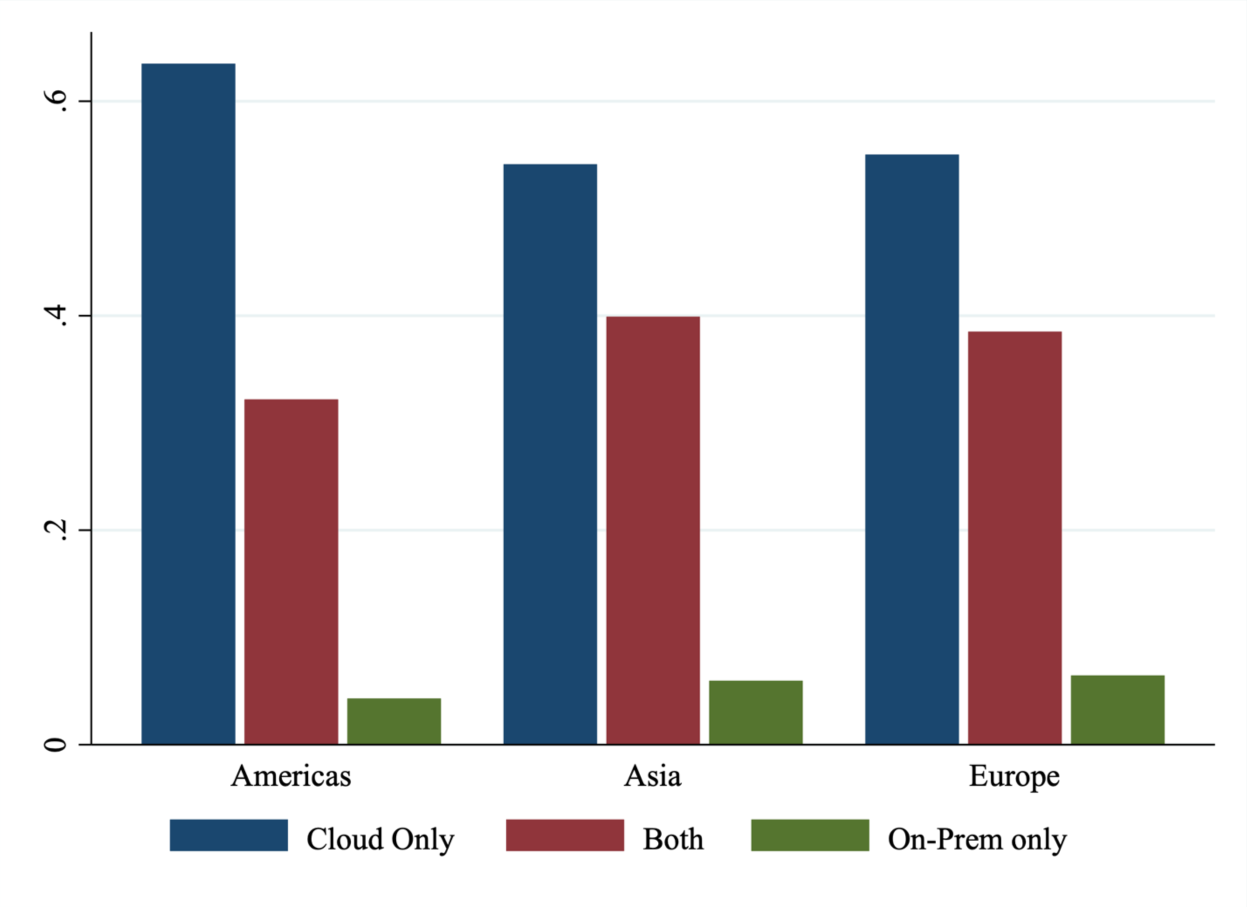
A.2 – Targeted Customer Size



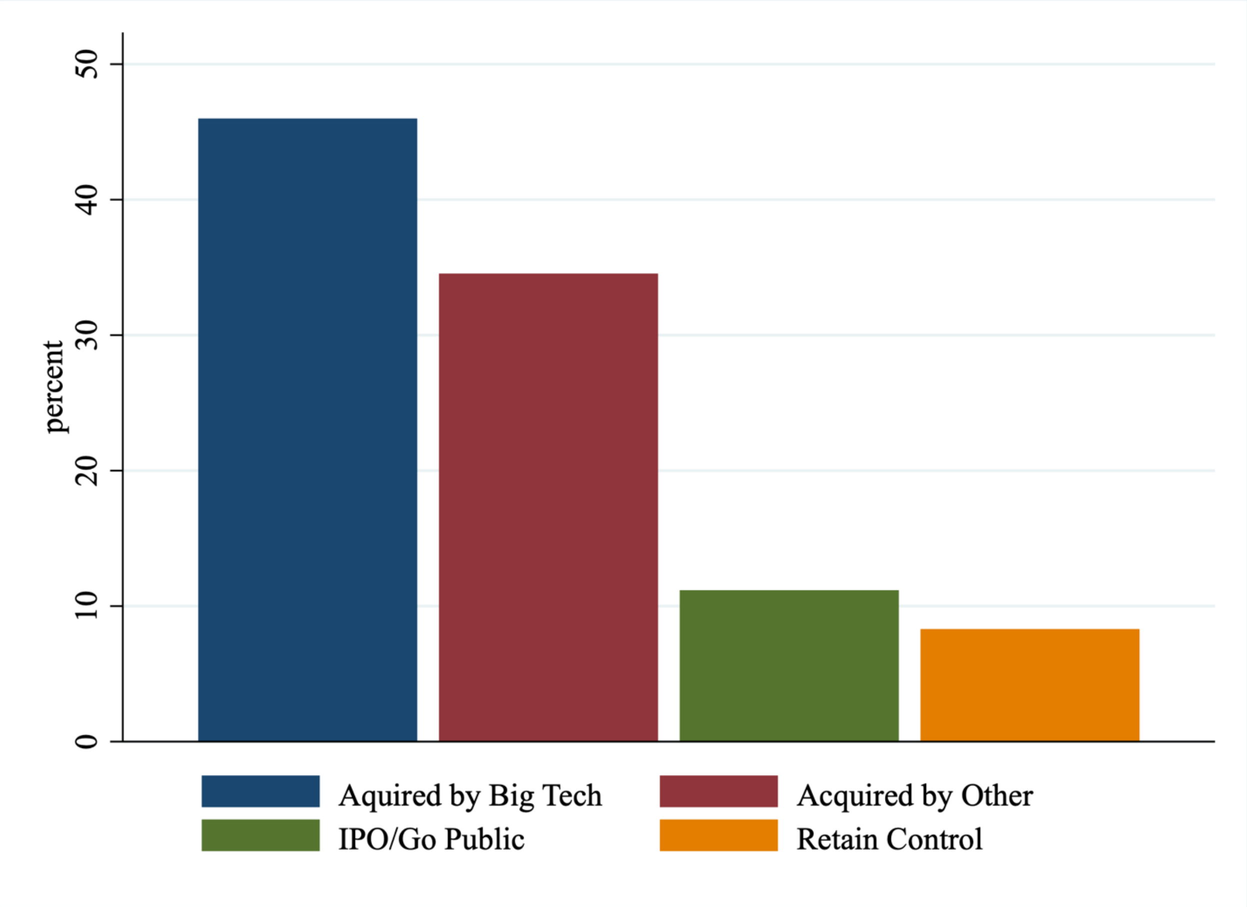
A.3 – Target for Respondents Product



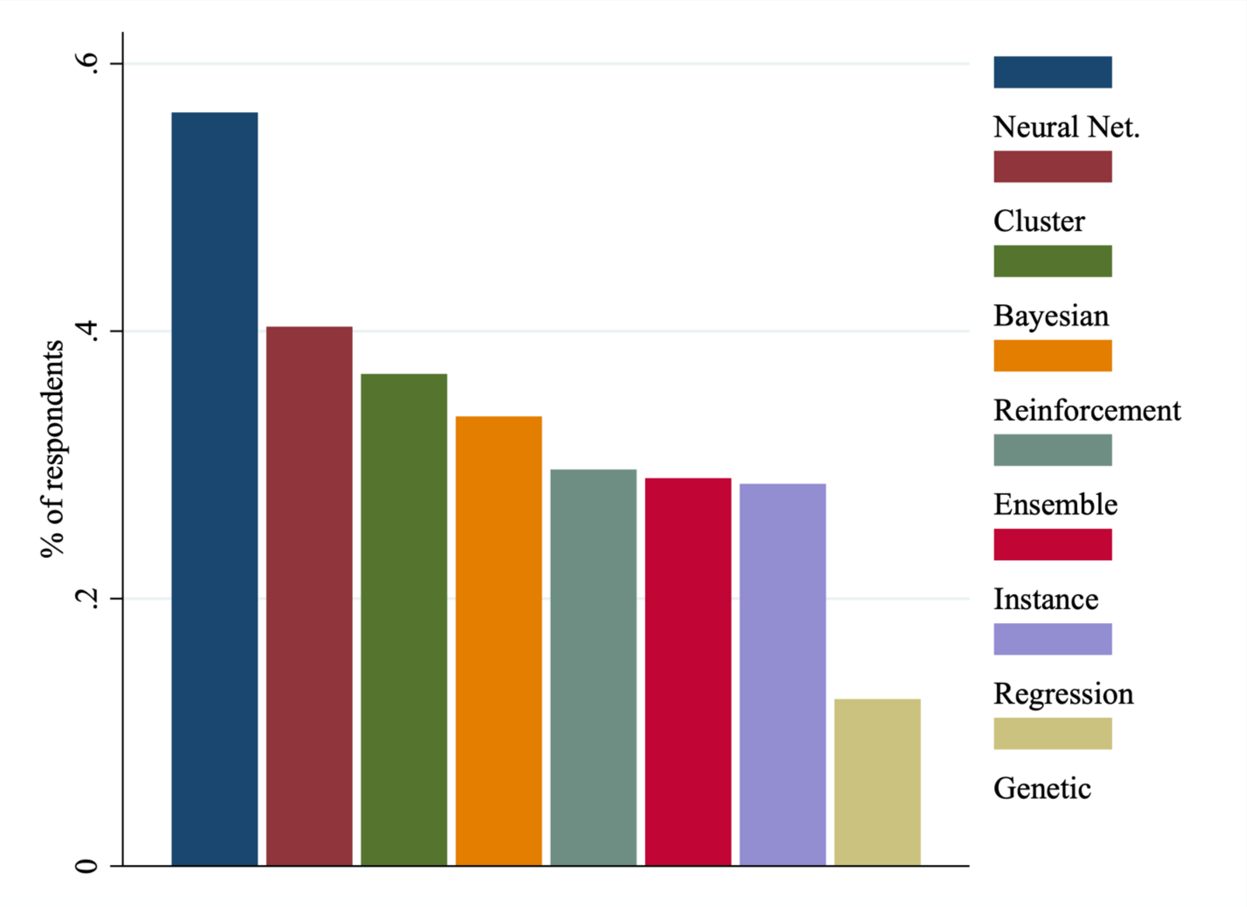
A.4 – Cloud Usage by Region



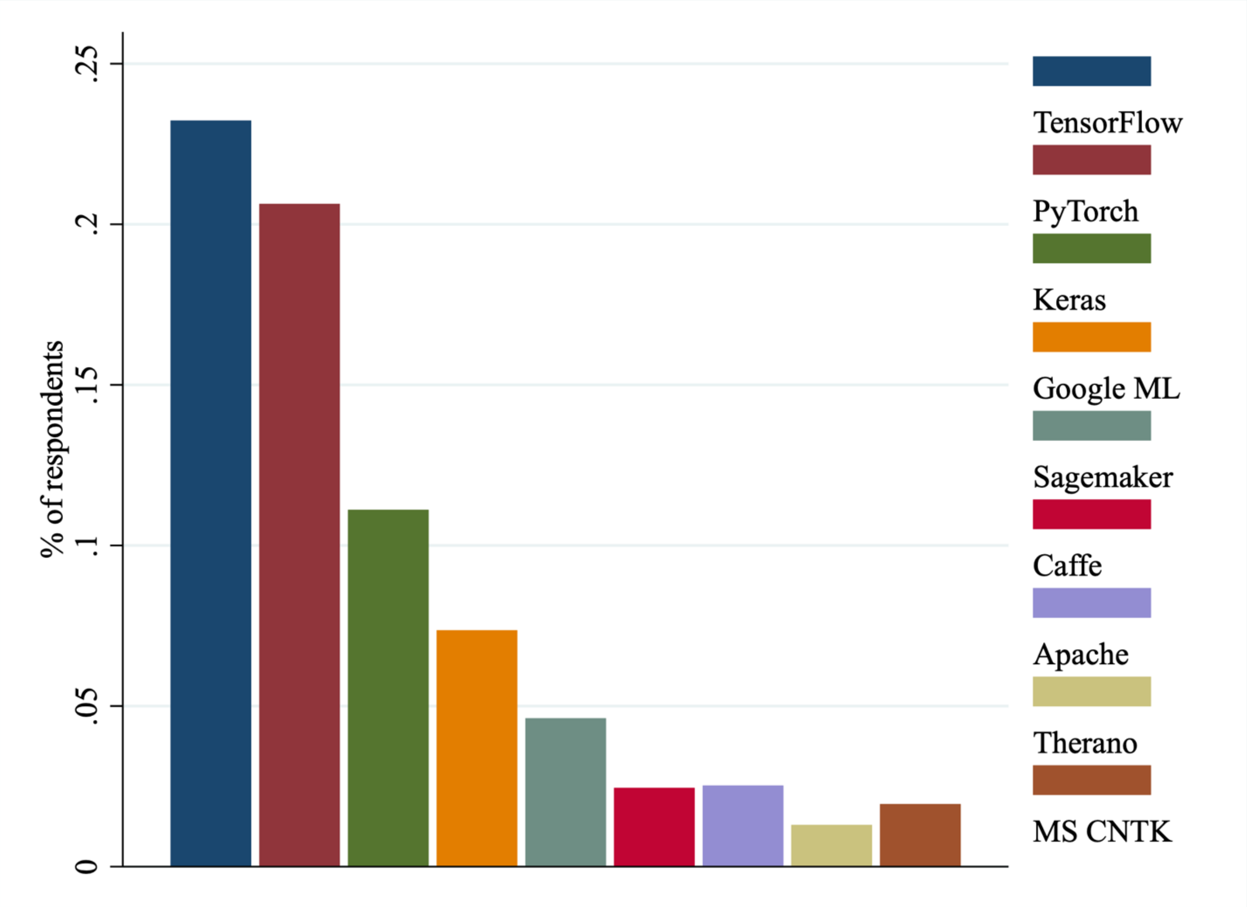
A.5 – Planned Exit Outcome over the Next Two Years



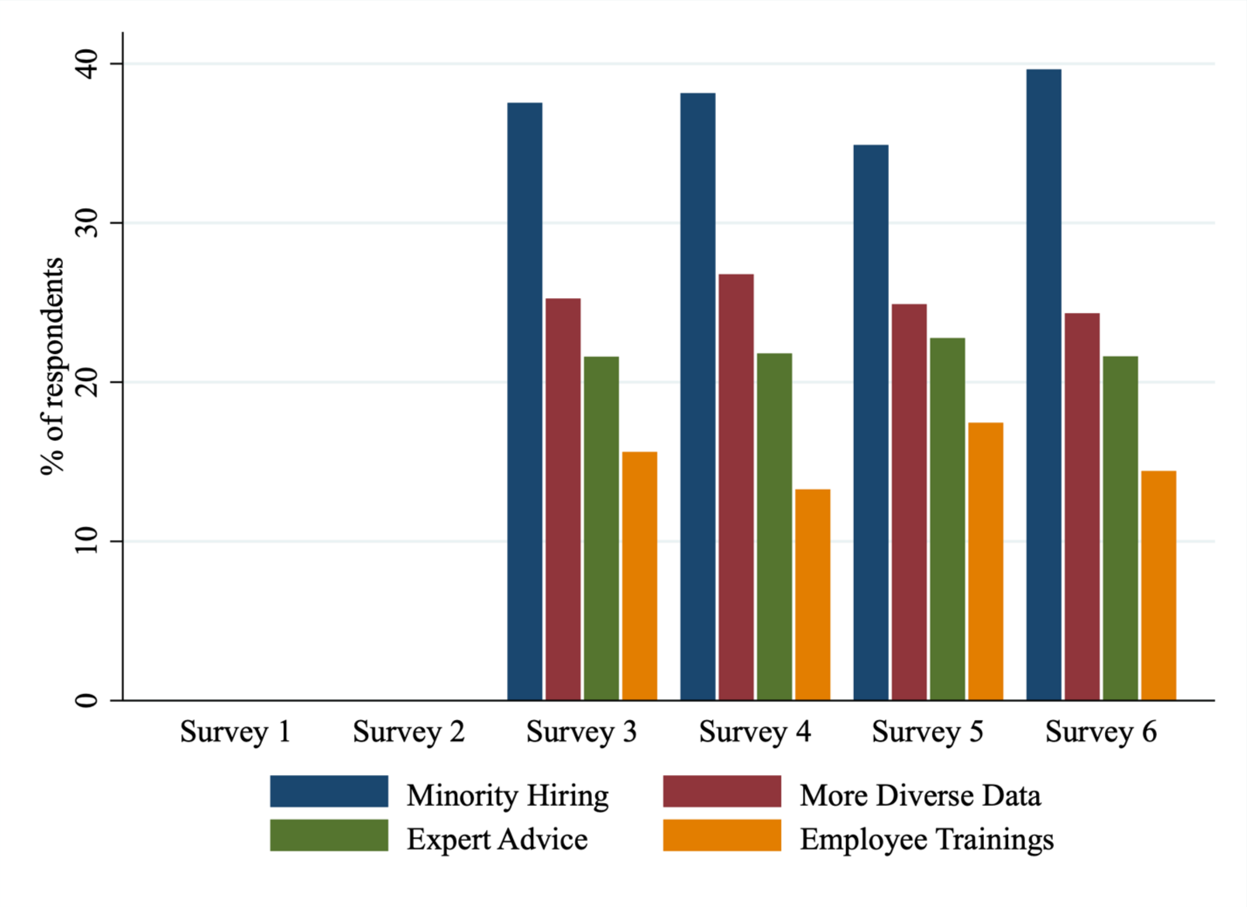
A.6 – Algorithm Types Used by Respondents



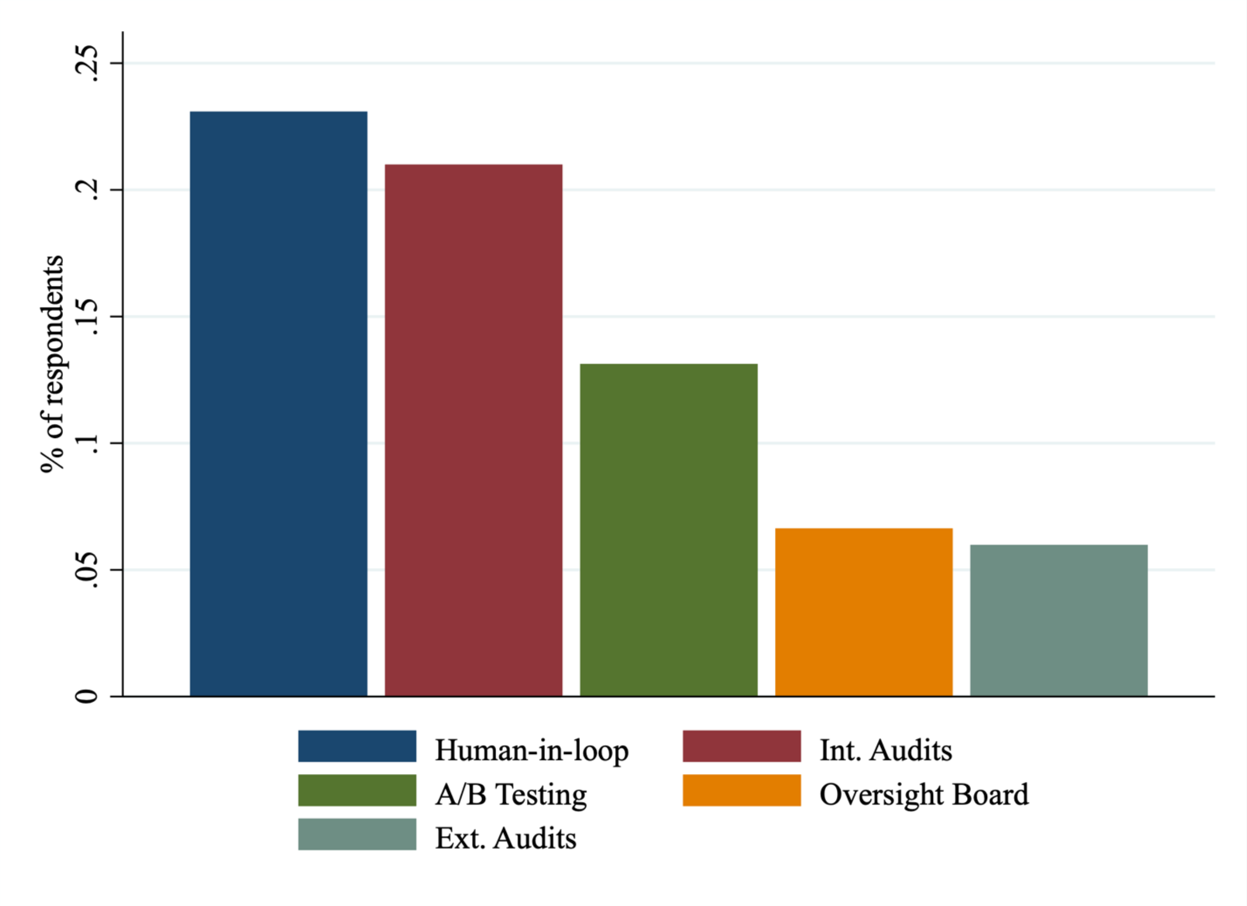
A.7 – AI Frameworks Used by Respondents



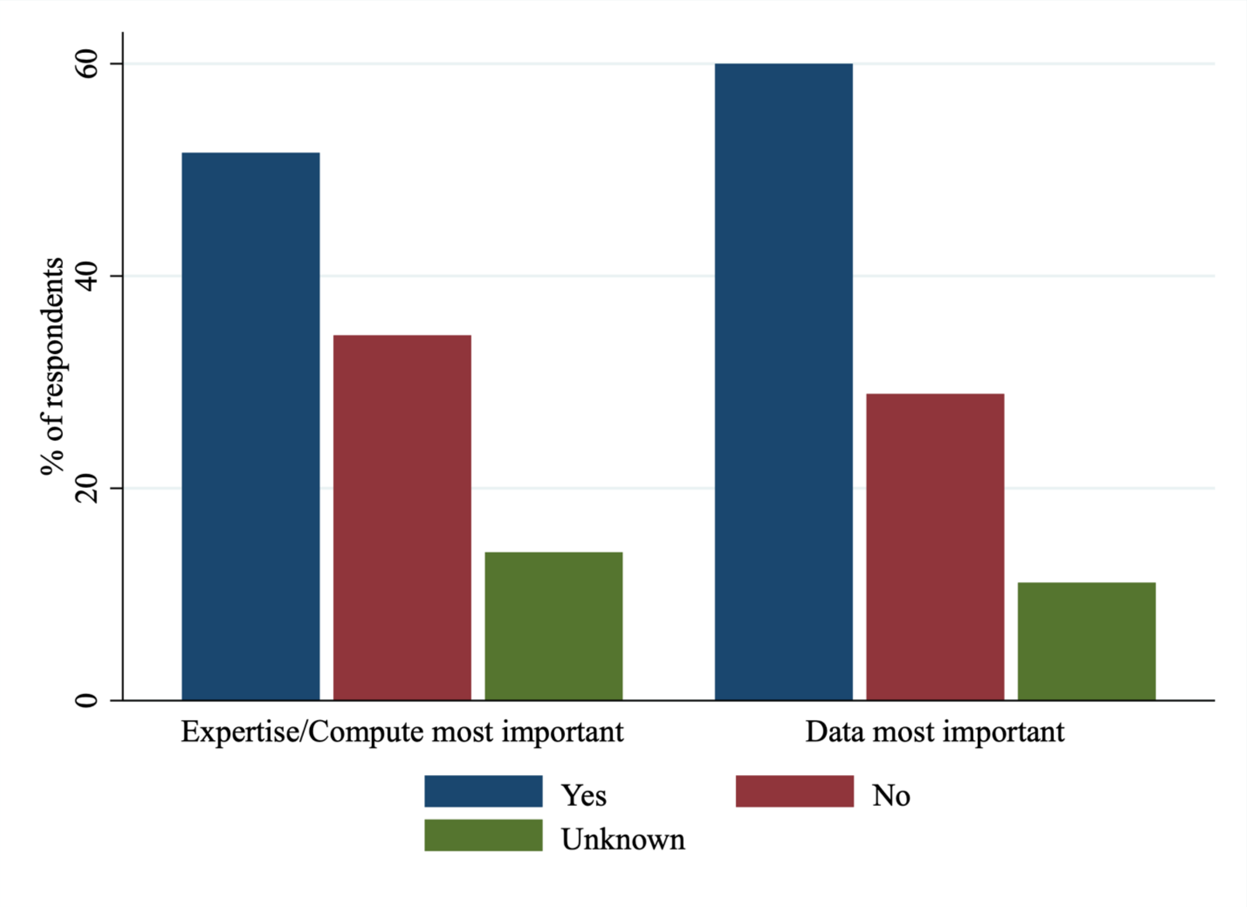
A.8 – Pro-ethics Action by Survey Round



A.9 – Governance of AI and Data



A.10 – Most Important Resource for Startups Developing AI Products


% of respondents
Expertise/Compute most important
Data most important
Yes
No
Unknown