The Missing Link? Using *LinkedIn* Data to Measure Race, Ethnic, and Gender Differences in Employment Outcomes at Individual Companies*

Alexander Berry
Econ One Research

Molly Maloney
UCI

David Neumark
UCI, NBER

June 2024

Abstract

Stronger enforcement of discrimination laws can help to reduce disparities in economic outcomes with respect to race, ethnicity, and gender in the United States. However, the data necessary to detect possible discrimination and to act to counter it is not publicly available – in particular, data on racial, ethnic, and gender disparities within specific companies. In this paper, we explore and develop methods to use information extracted from publicly available *LinkedIn* data to measure the racial, ethnic, and gender composition of company workforces. We use predictive tools based on both names and pictures to identify race, ethnicity, and gender. We show that one can use *LinkedIn* data to obtain reasonably reliable measures of workforce demographic composition by race, ethnicity, and gender, based on validation exercises comparing estimates from scraped *LinkedIn* data to two sources: ACS data, and company diversity or EEO-1 reports. We apply our methods to study the race, ethnic, and gender composition of workers who were hired and those who experienced mass layoffs at two large companies. Finally, we explore using LinkedIn data to measure race, ethnic, and gender differences in promotion. In our analyses of layoffs and promotions, we find suggestive evidence of discrimination at some of the companies we study, including evidence of "intersectional" discrimination against black and Hispanic women.

* This paper was prepared for the NBER-CRIW Conference: Race, Ethnicity, and Economic Statistics for the 21st Century. We have received helpful comments from Anjali Adukia, Randall Akee, Keith Finlay, Lawrence Katz, Mark Lowenstein, and conference participants. The data used in this paper are proprietary, and hence we cannot produce the full dataset for other researchers to use. However, the paper fully explains how the data can be accessed for purchase. We received modest financial support for this project from Econ One Research ("Econ One"). Alexander Berry is employed by Econ One and David Neumark does consulting work on discrimination cases with Econ One. Econ One had no right of review or comment on this research, and all conclusions are our own.

Introduction

Disparities in economic outcomes with respect to race, ethnicity, and gender are persistent in the United States. There is little doubt that labor market discrimination continues to contribute to these disparities, and that continued if not stronger enforcement of discrimination laws in the United States will help to reduce these disparities. However, the data necessary to detect possible discrimination and to act to counter it is not publicly available – in particular, data on racial, ethnic, and gender disparities within specific companies. Nor – not surprisingly – are such data readily provided by companies.

In this paper, we explore and develop methods to use information extracted from publicly available *LinkedIn* data to measure the racial, ethnic, and gender composition of company workforces. We use predictive tools based on both names and pictures to identify the race, ethnicity, and gender of employees. And we explore using this information, along with information from job histories on *LinkedIn*, to develop estimates of racial, ethnic, and gender differences in employment, layoffs, hiring, and promotions.

This paper builds on an emerging body of research that leverages data from private companies – and especially data on workers, firms, job openings, etc. – to better understand U.S. labor markets. It also dovetails with greater efforts, via legislation, to increase transparency about labor markets, in part to increase the information workers have about jobs, to reduce labor market frictions and increase labor market competition, as well as to reduce discrimination.[1]

A particularly valuable application of our research is that it can be used to strengthen enforcement of discrimination laws. The strength of discrimination laws in the United States rests on class action or

---

[1] For example, New York recently passed a law that requires firms to post pay ranges in advertisements for all job positions (https://www.littler.com/publication-press/publication/new-york-becomes-latest-state-require-salary-transparency-job-postings) and California's recently enacted pay transparency law requires posted pay ranges by demographic group (https://www.adp.com/spark/articles/2023/03/pay-transparency-laws-your-questions-answered.aspx). Further, a recent federal Executive Order (13665) prohibits federal contractors from retaliating against workers who disclose or discuss compensation information (Trotter et al., 2017). Despite expectations, some recent work suggests that pay transparency may reduce workers' bargaining power (Cullen and Pakzad-Hurson, 2021), because higher wage offers can lead to more renegotiation when pay is transparent. We regard this as a still-open question requiring more research.

other "pattern and practice" lawsuits on behalf of large numbers of a company's employees. The large potential penalties/awards in these lawsuits serve both to attract resources from attorneys to pursue discrimination claims and to incentivize firms to avoid these claims. Although federal and state agencies (like the EEOC, at the federal level) can file lawsuits against companies alleged to have discriminated, most enforcement – and enforcement against the largest companies – stems from private-sector attorneys.[2] But there are three problems, all of which our research can help address.

First, only the federal government (through EEO-1 reporting) and state governments (via similar authority) obtain data on, e.g., the racial or ethnic composition of firms' workforces, or of specific occupations within those firms. These data are confidential.[3]

Second, these data only measure employment (with limited information on occupational distributions). They do not measure hiring, retention, layoffs, or promotions.

Third, and most important, private attorneys – the key agents in the enforcement of discrimination laws – are severely hampered in trying to target the companies that potentially engage in discrimination. Complaints of discrimination are typically initiated by a small number of employees who may have personally experienced discrimination (or believe they have), but do not have information on statistical

[2] We are not aware of empirical comparisons of suits filed by the EEOC or state agencies vs. private attorneys. However, in our experience most discrimination lawsuits are filed by private attorneys. Moreover, the EEOC, for example, acknowledges that it narrowly targets cases for litigation: "The EEOC files employment discrimination lawsuits in select cases. When deciding whether to pursue litigation, the General Counsel will consider several factors, including the nature of the violation, the issues presented, and the wider impact the lawsuit could have on the EEOC's efforts to combat workplace discrimination. Because of limited resources, the EEOC cannot file a lawsuit in every case where the agency finds discrimination and the EEOC's efforts to secure voluntary compliance through the agency's conciliation process are unsuccessful" (https://www.eeoc.gov/eeoc-litigation).

[3] There is some potential movement towards the Office of Federal Contract Compliance Programs (OFCCP) releasing EEO-1 reports under FOIA requests. (See, e.g., https://content.govdelivery.com/accounts/USDOLOFCCP/bulletins/3495276.) It is at this time unclear how easy it will be for companies to stop release of these data by objecting on grounds of trade secrets, financial information, etc. Of course, OFCCP data only cover federal contractors. OFCCP is reluctant to release these data, so objectors may be able to block release easily (https://news.bloomberglaw.com/daily-labor-report/labor-department-reluctant-to-reveal-contractor-diversity-data). In addition, there is a recent agreement between the EEOC and the Census Bureau to provide EEO data that can be accessed and linked to other data sources at Census Research Data Centers (https://www.eeoc.gov/newsroom/eeoc-signs-agreement-census-bureau-provide-secure-access-agency-data-qualified-researchers), and it is also possible that one could import the *LinkedIn* data we describe and match to the EEO data. There would be no ability to identify individual companies, but it might be possible to satisfy Census reporting requirements while providing some descriptive information on the correspondence between EEO data and *LinkedIn* data. This is a potential project for future research using the full *LinkedIn* dataset, in contrast to the limited extract we study in this paper.

patterns at their employers. Plaintiffs' attorneys work on contingency fees, and hence have to decide whether to invest large sums in filing charges and commencing discovery before they can see any data on potentially discriminatory behavior. The uncertainty involved can deter them from taking on cases and reduce the efficiency of how their resources are targeted.[4]

The kinds of information we extract from *LinkedIn* data could potentially lead to more efficient targeting of anti-discrimination efforts. By helping attorneys identify where there is suggestive evidence of discrimination, these data could help make discrimination law more efficient, allowing attorneys to concentrate their efforts and resources on the companies where there is a higher probability that discrimination is occurring.[5] Moreover, the methods we develop could provide anti-discrimination enforcement agencies with additional tools to monitor companies and to target investigations of discriminatory behavior. While they typically obtain employment data (like EEO-1 reports to the EEOC), these reports do not cover other employment decisions – like hiring, layoffs, or promotions. These authorities can request additional data, but this process itself can be expensive and contested. Thus, the ability to better identify ex ante employers that might be discriminating can help in the targeting of public resources as well.

Our core research questions are: How can the *LinkedIn* data best be used to characterize companies' employment by race, ethnicity, and gender? How reliable are these data? And, can the *LinkedIn* data be used to study other company workforce decisions – in particular hiring, retention/layoffs, and promotions – and their relation to workforce demographics?

Our focus is to a large extent on the Professional, Scientific, and Technical Services Sector, owing in large part to strong representation on *LinkedIn*. That is not to say that our methods cannot be

[4] This oversimplifies slightly, as state anti-discrimination agencies sometimes partner with private attorneys. But in our experience, this is rare.

[5] Anecdotally, we have spoken with a handful of plaintiffs' attorneys who work on discrimination cases about the potential value of using the *LinkedIn* data in this way. We have confirmed that some indicated they would find it useful. One indicated they had already used the data in this way (although of course absent the kind of validation, etc., we explore in this paper). And below we cite use of *LinkedIn* data in a different manner in a discrimination case.

usefully applied to other industries, although some of our conclusions about the representativeness of the *LinkedIn* data might not apply as strongly and future work could assess this. At the same time, jobs in this sector are significant with regard to enforcement of discrimination laws, since this sector is an important source of high-paying jobs and upward mobility.

Social media data vs. scientific samples

The *LinkedIn* database is not the population of workers, and is not a scientific sample. At the same time, it covers a very large number of workers and might be expected to generate quite reliable measures.[6] The evidence we report in this paper indicates that the *LinkedIn* data correspond reasonably well, with some exceptions, to other probability-sample estimates of workforce demographics. However, the value of what we can do with the *LinkedIn* data does not hinge only on the representativeness of the data being so good as to be able to claim that the estimates (e.g., of the share black) are unbiased.

First, a growing body of research trying to study labor markets using social media data acknowledges the tradeoff between probability sampling and the ability to learn from social media data what we cannot learn from other data. As examples, Schneider and Harknett (2019a, 2019b) use targeted ads on Facebook to study work schedules, based on a 1.2% response rate. Similarly, a number of labor economists use data on job postings or job applications to study monopsony, discrimination, minimum wages, and other topics (e.g., Azar et al., 2022; Borup and Montes Schütte, 2022; Clemens et al., 2021; Marinescu and Wolthoff, 2020; and Neumark et al., 2019) – in our view learning a great deal more than we could otherwise, despite data sources being not fully representative.

Second, it is critical to emphasize that the core value of the *LinkedIn* data for enforcement of discrimination laws is that it provides some reliability in estimates of race, ethnic, or gender differences in the outcomes the data are used to measure, to improve targeting of what would generally be more thorough investigations. As explained earlier, absent this type of information, private attorneys considering class action discrimination claims have only anecdotal evidence to rely on, and hence can be

[6] *LinkedIn* claims 200 million active U.S. users.

deterred from initiating lawsuits; and enforcement agencies have limited data on potential discrimination. Thus, data and estimates from *LinkedIn* can provide a critical complement to this limited information, and hence improve targeting of anti-discrimination efforts, even if estimated differences by race, ethnicity, and gender have some biases due to imperfect representation.

Alternative data sources?

This project considers the development of a data source that can fill substantial gaps in labor market data available for the United States. We have rich household and worker data, but these data contain no firm identifiers and typically do not include information on the positions people hold within companies. In principle, the LEHD could be used to provide descriptive information similar to some things we can measure with the *LinkedIn* data. However, there are a number of limitations of the LEHD data.

First, a core limitation is that the LEHD could never be used as an enforcement tool the way we are envisioning, both because company names could never be revealed, and because those for whom this tool would be useful would never be able to access these data, to use them to study a single company. Second, the LEHD would pose severe challenges to doing this in a timely manner, both because the LEHD is updated slowly, and because securing access to the LEHD and then working with the data is a slow process.[7] Further, the LEHD data does not have any information on job titles at companies that could be used to study employees' positions within companies. In contrast, the *LinkedIn* data do not present these restrictions, and – importantly from both an enforcement and research perspective – are up-to-date and immediately accessible. Thus, although the LEHD is an extraordinarily valuable and powerful data set that can be leveraged for the analysis of discrimination,[8] it cannot be readily used for the specific questions we will be studying with the *LinkedIn* data. At the same time, it is possible that our work with the *LinkedIn* data will prompt work with the LEHD to provide a more extensive and higher quality

[7] As an example, one of the most recent LEHD publications we could find on *Google Scholar* is a 2022 publication using LEHD data through 2016 (McKinney and Abowd, 2022); and the earliest working paper version of this paper appears to be from 2020.
[8] See, e.g., Barth et al. (2021), Brick et al. (2023), and Hu (2019).

characterization of employment and other dynamics one can measure with the LEHD in relation to race, ethnicity, and gender. We regard these as complementary efforts, with different strengths, weaknesses, and purposes.

The National Establishment Time Series (NETS) is a proprietary data set in which company names are public and can be used in research (e.g., Burnes et al., 2014), but it contains no worker information aside from employment. (As indicated above, however, we do make some use of the NETS data in this project to select companies for our validation work.)

Aside from these datasets familiar to researchers, there is some information on the demographic composition of firms' workforces from companies that make public their "diversity reports." For example, *Google* releases an annual diversity report.[9] Its 2022 report provides the percentage of hires by race, ethnicity, and gender, as well as workforce representation and attrition (the only instance we have found of reporting on retention). But it says nothing about the positions different workers occupy in the company. *LinkedIn*'s 2021 report[10] provides the race, ethnic, and gender composition of its overall, tech, and non-tech workforces, and also the composition of a vaguely defined "leadership" category (a category required in confidential EEO-1 reports). It is not difficult to find other similar reports, providing the same limited information.[11] Conversely, there is ample information on companies' resistance to providing this information,[12] and we might expect that it is supplied selectively across companies. Also, note that neither these reports, nor the EEO-1 reports, typically provide information on retention or hiring. The methods we develop and describe in this paper can provide information on the race, ethnicity, and gender dimensions of all of these aspects of firms' workforces.

Finally, a very recent paper by Kline et al. (2024) provides company-level evidence on

[9] For 2022, see: https://about.google/belonging/diversity-annual-report/2022/.
[10] See https://news.*LinkedIn*.com/2021/october/2021-workforce-diversity-report.
[11] See, e.g., https://news.linkedin.com/en-us/2022/october/2022-workforce-diversity-report; https://www.apple.com/diversity/.
[12] See, e.g., https://www.proxypreview.org/2022/contributor-articles-blog/data-transparency-key-to-improving-diversity-equity-and-inclusion-in-the-workplace; https://circaworks.com/articles/eeo-1-report-and-voluntary-disclosure/.

discrimination. In particular, the authors conducted a correspondence study of discrimination based on race (distinctively black names) of 108 companies in the Fortune 500 and report the race gaps in callbacks by company.[13] This study provides another way to garner evidence on discrimination at the company level. It has the advantage, relative to what we do, of providing more rigorous experimental evidence on hiring discrimination (Neumark, 2018). It has three disadvantages, however. First, it is limited to hiring. Second, strong evidence of discrimination in callbacks does not necessarily point to companies where many workers (blacks, in this case) are being harmed, because the companies with low callback rates for blacks may, in reality, get few black applicants. Third, in this study companies received a large number of artificial job applications, with as many as 1,000 applications to some firms (Table F5 in the paper). One might expect companies to respond to this inundation by implementing procedures to detect such artificial applications, rendering the method less useful. Nonetheless, such evidence – like ours – may be able to assist enforcement agencies in targeting investigations.[14]

Approach and methods

We conduct our research using extracted, publicly available *LinkedIn* data offered by the company *Proxycurl*.[15] Given a company's *LinkedIn* profile, *Proxycurl* returns *LinkedIn* profile data for that company's employees with public profiles. The data reflect the publicly available *LinkedIn* profiles at the time they are scraped. We can query the current *LinkedIn* profile for each employee who has linked a company of interest as an employer, either past or current, depending on parameter selections. This means we can get information on past as well as current employment.[16] However, the data do not simply cover

---

[13] They sent up to eight applications for each vacancy, or about 84,000 applications for 11,000 jobs at the 108 companies covered. Much of the focus of the paper is on how to rate companies' discriminatory behavior, given uncertainty in the estimates.

[14] Moreover, in the case of audit studies, individual testers have legal standing to sue (see EEOC guidance: https://www.eeoc.gov/laws/guidance/enforcement-guidance-whether-testers-can-file-charges-and-litigate-claims-employment). Interestingly, case law appears to go back to an early civil rights case in which black clergymen who were removed from segregated bus terminals were deemed to have been discriminated against even though their goal was to test the law, rather than to ride the bus. The standing of testers has been established in housing cases (*Havens Realty Corp. v. Coleman*, 455 U.S. 363, 374 (1982)). And the EEOC, in the document referenced above, argues the language of Title VII (covering employment discrimination) parallels that in Title VIII (covering housing discrimination).

[15] See https://nubela.co/proxycurl/linkdb for more details.

[16] See https://nubela.co/proxycurl/docs for more details.

all employees at a single point in time. Rather, the scraping captures all individuals who have ever worked at the company, as long as they have not deleted their data. We also typically know when they were employed at the company, so we can approximate employment by year, as well as hiring, exits, etc. We extract *LinkedIn* profile data on all current and former employees for seven companies, the selection of which we describe below.

The information returned often includes detailed information about previous work experience (including place of employment, title, description, start date, and end date) and education (including school, field of study, degree obtained, start date, and end date). All *LinkedIn* data are self-reported, and voluntarily reported. These profiles may also include information about a worker's skills, activities, volunteer work, languages spoken, certifications, and recommendations, among other topics. Critically, for our purposes, it also often includes their profile picture, along with their name.

We use the *DeepFace* package in Python to classify workers by race, ethnicity, and gender, based on their *LinkedIn* profile pictures and a common training dataset ("picture classification"). We also use R packages (*rethnicity*, *gender*), which use statistical data to classify based on names ("name classification").[17] The *DeepFace* package is trained on the *FairFace* dataset for race/ethnicity identification. The gender prediction model for picture classification is trained on Wikipedia data. These both return probabilities that the worker is in each group. For details on both types of classifications, see Serengil and Ozpinar (2020, 2021). We also supplement with name classification when picture is not available, and we combine information when both are available.[18]

[17] There is other information that, in principle, could be used to classify people, such as other information in the images or information in *LinkedIn* entries on schools, locations, interests, etc. As an example, the Consumer Financial Protection Bureau (CFPB) combines information on surnames and geography to predict race and national origin to monitor compliance with fair lending laws and discrimination in the consumer credit industry. CFBP (2014) indicates that a Bayesian Improved Surname Geocoding (BISG) method works better than relying on only name or geography. Implementing something like this with the *LinkedIn* data would require a training sample, and the available geographic information is not as useful because it is much less detailed. Developing new classification methods is beyond the scope of this paper.

[18] In our view, this kind of classification is consistent with U.S. discrimination laws. The EEOC notes that Title VII of the Civil Rights Act does not define "race." But its guidance is based on perception – i.e., whether one is perceived as belonging to a group, In particular, "Race discrimination includes discrimination on the basis of ancestry or physical or cultural characteristics associated with a certain race, such as skin color, hair texture or

We use binary classifications (black/non-black, Hispanic/non-Hispanic, and female/male), based on highest probabilities returned by these programs. We use names as the primary method for gender and ethnicity, and pictures as the primary method for race, based on evidence described below. Where the classification probability is missing for the primary method, the other method is used. We also change which method we use if the primary method gives a quite uncertain classification and the secondary method gives a far more certain classification, as explained in more detail below.[19]

One of our core goals in this paper is to validate our classifications against external data, to see how reliable the *LinkedIn* data are. For example, one might wonder whether particular demographic groups are under- or over-represented on *LinkedIn*. One approach to validating the *LinkedIn* data is to leverage corresponding information in two other data sources – the National Establishment Time Series (NETS), and the American Community Survey (ACS). For this validation exercise we proceed in two steps. First, we use data from the NETS, along with *LinkedIn* information from the *LinkedIn* website, to identify companies that are in a broad industry category that has good representation on *LinkedIn*. In particular, we focus on the Professional, Scientific, and Technical Services Sector. As documented in Table 1, which reports results for the top 10 *Fortune* companies, this sector (see the highlighted rows) has good representation on *LinkedIn* – in the sense that a large share of the companies' workers appears on the website, based on current employment (column (2)).[20]

---

styles, or certain facial features" (https://www.eeoc.gov/laws/guidance/questions-and-answers-about-race-and-color-discrimination-employment). We would interpret this to correspond to physical characteristics or distinctly racial names. The EEOC offers similar guidance with regard to discrimination based on national origin (ethnicity): "national origin discrimination means discrimination because an individual (or his or her ancestors) is from a certain place or has the physical, cultural, or linguistic characteristics of a particular national origin group" (https://www.eeoc.gov/laws/guidance/eeoc-enforcement-guidance-national-origin-discrimination). On the other hand, data used in discrimination lawsuits is often self-reported, and our validation exercises compare the *LinkedIn* data to self-reported classifications in the ACS and to EEO-1 reports or other information companies provide that is almost surely based on these reports.

[19] We considered using probabilities to construct estimates of demographic shares of the workforce, etc., based on weighted averages using these probabilities. However, we found that this was not as accurate for the race and ethnicity coding, because there are other minority groups (such as Asians) that can receive some probability weight, which results in lower estimated shares black or Hispanic than we get from using the highest probability.

[20] Column (2) captures those as reported on a company's *LinkedIn* page. Column (3) is current employment included in *Proxycurl*'s database (which is also restricted to public *LinkedIn* profiles). The low number for CVS Health in the *Proxycurl* database is because of search constraints imposed when using the *Proxycurl* database by buying

We then use the NETS data to select companies and areas to make the ACS and *LinkedIn* data comparable – i.e., so that when we extract ACS data on workers by Place-of-Work PUMA (POWPUMA) and industry, we should be sampling by and large from employees of these companies in the corresponding geographic area. In particular, we identify companies in this sector that meet three criteria: (1) fairly negligible employment at other firms in the same industry and POWPUMA; (2) most of the company's employment in the POWPUMA is in the industry; and (3) the company has strong representation on *LinkedIn*. The idea behind criterion (1) is that these firms constitute most of industry employment in the POWPUMA. Thus, ACS workers in the industry and POWPUMA are likely to work for these companies. The idea behind criterion (2) is that the company's employment in the POWPUMA is concentrated in one industry. This is critical because we have an industry identifier in the ACS but not in *LinkedIn*. Thus, if the company had POWPUMA employment in other industries, the ACS data for a single industry might not be representative of the company's POWPUMA employment. Because these companies are largely unique in their industry-location cells, if they also have good representation on *LinkedIn* (criterion 3), the measures of race, ethnic, and gender composition from the two data sources should correspond. Thus, we measure the race, ethnic, and gender composition of ACS employment in those industry-location cells, and then compare to our estimates based on the *LinkedIn* data.

These ACS restrictions greatly limit the number of company comparators that we are able to benchmark against. We therefore also use a second validation approach, comparing the *LinkedIn* results to companies' DEI reports, when available, or other sources of information on the demographic composition of their workforces. This approach avoids the constraints on companies dictated by the first approach. In particular, we were able to find information in diversity reports or other information companies provided, and we use these, when available, for both the companies selected to validate against the ACS data, and other companies we selected.[21]

---

tokens for a specific number of searches, which constrains the search to workers who include company urls in their *LinkedIn* profiles. Based on our investigations, non-professional workers may be less likely to do this.

[21] There is already some limited evidence of the reliability of the *LinkedIn* data. Specifically, in a large gender pay

Companies selected

We identified four companies that meet the criteria for the ACS validation discussed above, and that are not too large (allowing the required *LinkedIn* data extraction within our budget constraints). Table 2 reports the companies meeting the following constraints:

1. Company's NAICS industry employment in POWPUMA vs. all NAICS industry employment in POWPUMA > 70%
2. Company's NAICS industry employment in POWPUMA vs. all company employment in POWPUMA > 80%
3. Firm employment > 800.

The table also shows *LinkedIn* employment – in this case based on ever employed, since our validation with the ACS is not based on only one year of data. We had to constrain the choice among the companies meeting these criteria these based on number of employees, and we also constrained it based on *LinkedIn* data showing a large share of employment in a single nearby geographic area, since otherwise we would not expect the ACS to provide a very relevant comparison. The four companies we selected for this validation exercise are shown in the shaded rows of Table 2. The non-shaded rows are for companies that met our criteria with regard to POWPUMA, but were either very large, had low representation on *LinkedIn*, had low representation in the geographic area (which likely has to do with international employment not measured in the NETS), or were quite small so that we would not be able to learn that much from the data.

The additional companies selected were GlaxoSmithKline (GSK), SpaceX, and Meta. GSK made our initial list for the ACS validation; employment was not sufficiently geographically concentrated to be

---

discrimination lawsuit, *LinkedIn* data were extracted on jobs Oracle employees held prior to coming to Oracle, as well as on their education. (See *Expert Trial Report of David Neumark in the Matter of Jewett et al. v. Oracle America, Inc.*, December 2021, redacted.) It was possible to match about 55% of Oracle employees in the company's data to *LinkedIn* observations, and to establish that the matched data were representative in one dimension; in particular, in that case the estimated gender pay gap was similar in the full company data and the subsample matched on *LinkedIn*.

useful for the ACS validation, but we found other information on its workforce. Meta and SpaceX are two high-profile technology companies. And for one (Meta), we actually found EEO-1 data posted, whereas for the other (SpaceX), we found nothing reliable. Thus, we could validate the Meta data from *LinkedIn* against the EEO-1 data, and also illustrate the potential value of our approach for a company (SpaceX) for which neither EEO-1 nor diversity report information was available. We restricted our choices to other companies in the same sector, for comparability. Finally, we chose among these based on the ability to cover a number of companies while remaining within our data budget constraint. Clearly future work with greater funding could expand the scope of these types of analyses.

Data extraction and classification

For these companies, we extract publicly available data from *LinkedIn* from the *Proxycurl LinkedIn* database. To do this, we provide the company's *LinkedIn* url. We request current and past employees (who can be distinguished in the database). The application then returns all data from public profiles (employment history, education, skills, etc.).

Across the seven companies for which we extracted data, we obtain 112,280 worker profiles, of which 78,639 are in the United States. The numbers and distributions of these observations are displayed in Table 3. We get a sizable number of observations from all companies except Research Corporation of the University of Hawaii, and very large numbers of observations for GSK, Meta, and SpaceX. We retain only those working in the United States. We break the data for each person into separate entries for each job at each company at which they worked. Together, this results in 557,329 separate observations, although many of these are not at our companies of interest.

We then use the extracted *LinkedIn* data to classify workers, based on the *DeepFace* package in Python for picture classification, and R packages (*rethnicity*, *gender*) for name classification.[22] Before reporting our findings on demographic composition and more, there are some results about classification that are of interest, and which dictate how we use this information.

[22] We utilized *RetinaFace* for the face detection backend, and otherwise used a pre-trained neural network that comes with the package.

There are certainly caveats to using these methods. First, some pictures on *LinkedIn* make identification difficult. For example, some show multiple people, obscured images, or have bad lighting, as shown in the examples in Figure 1. In addition, sometimes there is no picture available. Overall, we run the *DeepFace* classification code for 96,651 profiles (which includes pictures from non-U.S. profiles). Of these, 22.95% were missing a picture, and 1.09% had an image in which a face could not be detected. The distributions of these cases were roughly stable across the companies, as shown in Table 4.

Second, we cannot always classify people by name. We run the name classification code for 75,393 names.[23] We are unable to classify gender for 9.89% of names, and unable to classify race/ethnicity for 4.55% of names. This could occur if the name recorded on an individual's *LinkedIn* profile does not include their first name (e.g., "Lt. Higgins") or if their name includes many non-Latin characters, for example. For predicting gender, if a name is sufficiently uncommon that it does not appear in the Social Security database used to predict gender, then no gender probability will be assigned to it. For race prediction, we do not use an individual's last name to predict race when they only provide a single initial in place of their last name (the program classifies all one-letter names as Asian).

Third, some names were problematic. For example, some first names – like "Alex" – are not strongly gendered. Some last names – like "Monte" – could be of Hispanic origin or another ethnicity. Some names are classified as more likely to be black or white, without providing a strong confirmation – e.g., "Steve Fulton," with a 72% probability of being black. And, of course, last name (or even first name) changes can obscure race or ethnicity. In these cases, pictures may provide more definitive information.

We use additional information on how the two programs classify people by race, ethnicity, and gender to settle on our classification "algorithm." First, as shown in Figure 2A, the distributions of probabilities that observations are female, whether using names or pictures, are bimodal, with probabilities clustered near zero or one. This reflects the fact that names are highly gendered, as is physical appearance. There is a little more mass at the endpoints using names (about 85%) than using

[23] Note that this differs from the number of profiles above because we do not run the algorithm separately for repeated names as the predicted gender/race will not differ for individuals with the same name.

pictures (about 80%), which is why we use names as the first source of classification by gender.

Second, the story for ethnicity and race classification is more complicated. The charts in Figure 2A for both ethnicity and race show large spikes at zero probability, but do not show much evidence of bimodality. Similarly, there is far less mass at the lower and upper ends of the range for Hispanic or black classifications than for gender classifications, and correspondingly more mass between these points for Hispanic or black classifications – and more so for Hispanic classifications. These findings reflect less definitive assignments of probabilities, potentially owing to less distinct physical differences than those by gender, or because shading of pictures can obscure race or ethnicity. And it is likely because these differences are less pronounced for Hispanics that there is, in the bottom panel of Figure 2A, a good deal more mass at lower values but above zero probability for Hispanic than black, and conversely much more mass at zero (more accurately, in the band 0-2) probability for blacks. We learn more about what is happening at the higher probabilities from Figure 2B, which shows more details at the higher probabilities by zooming in on the upper halves of the distributions.[24] We now see much more clearly that for black classifications the distribution of probabilities based on pictures is more bimodal, with a spike at 100. We thus rely on pictures as the first source of classification for blacks. For Hispanic classifications, there is a much more pronounced mass of probabilities at the top of the distribution using names than using pictures, so we use names as the first source of classification for Hispanics.

As a result of these considerations, as well as the inability sometimes to classify people by gender, race, or ethnicity based on either a picture or a name, we use the following algorithm to classify people.

<u>Race</u>

1. If the picture probability is non-missing, we classify people as black based on the picture if the probability black based on picture is the highest among all "race" categories.[25]

[24] Note that in Figure 2B the vertical scales are not the same in each graph, so that we could better highlight the details.

[25] The classification programs do not separately code race and ethnicity as commonly defined by, e.g., the U.S. Census, but rather include these in the same overall "race" classification.

2. If the picture probability is missing but the name probability is non-missing, we classify people as black based on the name if the probability black based on name is the highest among all "race" categories.
3. If both are non-missing, we rely on pictures, except when the picture classification is highly uncertain, but the name classification is not. Specifically, when no race probability based on pictures (among the 6 groups)[26] > .5, but the probability black based on name > .9, we classify people as black.[27]

Ethnicity

1. If the name probability is non-missing, we classify people as Hispanic based on the name if the probability Hispanic based on name is the highest among all "race" categories.
2. If the name probability is missing but the picture probability is non-missing, we classify people as Hispanic based on the picture if the probability Hispanic based on picture is the highest among all "race" categories.
3. If both are non-missing, we rely on names, except when the name classification is highly uncertain, but the picture classification is not. Specifically, when no race probability based on name (among the 4 groups) > .5, but the probability Hispanic based on picture > .9, we classify people as Hispanic.[28]

Gender

1. If the name probability is non-missing, we classify people as female based on the name if the probability female is higher.
2. If the name probability is missing but the picture probability is non-missing, we classify people as female based on the picture if the probability female based on picture is higher.

[26] For pictures, these are Asian, black, Indian, Latino/Hispanic, Middle Eastern, and white. For name, these are Asian, black, Hispanic, and white.
[27] This only results in a re-classification if the probability black based on the picture was not the highest.
[28] This only results in a re-classification if the probability Hispanic based on the name was not the highest.

We ran the classification code for all worker-company observations. After creating the race, ethnicity, and gender identifiers, we had 29,843 observations that were missing either race or gender, approximately 5.4% of the data. In terms of race, we have 97.44% coverage, and our gender variable covers 96.82% of the data.[29]

Table 5 reports on the probabilities of classification (reported as percentages) by each category based on pictures and names, including the initial classifications and the final classifications. We can see in this table some of the same results from Figures 2A and 2B, and also the consequences of our final rules for classification. For example, we noted that pictures are far more reliable for classifying black vs. non-black than names. This is reflected in the first row, columns (1)-(8), in the higher probabilities black for pictures at each of the percentiles reported. Note, though, that we obtain far more classifications based on name, so as a result the probabilities in columns (10)-(12), which are often based on name, are lower than in columns (2)-(4). In contrast, the probabilities Hispanic are far higher for names. For gender, the probabilities are very high in both cases, with the difference (lower probabilities based on pictures) only apparent at lower percentiles (column (2) vs. column (6)).

It is worth mentioning here that we deliberately did not tune or adjust our methodology regarding classifications to try to better match the proportions we see in the ACS or the diversity/EEO-1 reports. We are trying to demonstrate the value of these data in doing these classification exercises for other companies. Since a potential user presumably would not have any data source other than the *LinkedIn* data, there would be no basis for adjustments to the classifications. As a result, potential users should be most interested in the accuracy of race, ethnic, and gender classifications that use out-of-the-box algorithms (as we do), without any further fine-tuning.[30]

Classification of companies' workforces and validation with ACS data

---

[29] Evaluations of these methods point to fairly reliable classification. For *DeepFace*, see Serengil and Ozpinar (2021). As discussed in Blevins and Mullen (2015), the *gender* package in R assigns a probability that an individual is female based on the historical frequency at which women are observed with that name using Social Security Administration data since 1930s.

[30] At the same time, we recognize that additional exploration could help identify other code that works better, refine the existing code, or find alternative ways to use the resulting probability estimates.

Based on race, ethnicity, and gender classifications, we first report our results for demographic classifications of companies' workforces for the four companies for which we can perform our validation exercise with ACS data. For these comparisons, we restrict the relevant areas in the *LinkedIn* data as follows: Virginia for BWXT; Hawaii for Research Corporation of the University of Hawaii; Virginia for Chesterfield County, and Illinois for Fermi Research Corporation. For the ACS comparison we limit observations to those observations with at least a portion of relevant experience at the company between 2012 and 2021 (a 10-year window). We use the *LinkedIn* data to construct each job spell of each person at the corresponding company, and compare the ACS data to the *LinkedIn* data where we measure the demographic composition of employment in each year for any spell of employment covering that year. It is the case that *LinkedIn* employment at each company generally trends upwards over time, which we imagine is due at least in part to increasing numbers of employees on *LinkedIn*. Regardless of the reason, we weight the ACS data by year to be proportional to the representation in the *LinkedIn* data.

These results are reported in Table 6. The ACS numbers are weighted annual averages, based on 2012-2021 data for the same POWPUMA and 4-digit NAICS code. We also show similar results for professional, technical, and managerial occupations, which are likely over-represented on *LinkedIn*.[31] The data are also displayed in more digestible form in Figure 3. The results for the percent female do not indicate tight concurrence of the estimates, but there is some correspondence. For example, looking at the overall ACS numbers, the rank order across the four companies is the same for the ACS and *LinkedIn* data, and the *LinkedIn* percentages are notably higher where the ACS estimates are (for females at Research Corporation of Hawaii and for blacks and females for Chesterfield County), and vice versa. The data for the Research Corporation of the University of Hawaii should probably be discounted, given the low representation in the *LinkedIn* data and the ACS data. The results for the percent Hispanic do not correspond very well. The estimates for the percent black, excluding Research Corporation of the

[31] That said, the sample sizes when we restrict to these occupations are only a bit smaller, consistent with (a) most workers at these companies being in these occupations, and (b) little apparent bias from the hypothesized over-representation of these occupations on *LinkedIn.* A potential caveat, however, is that differences between the occupations represented on *LinkedIn* and overall workforces may be more marked for other industries.

University of Hawaii, exhibit reasonable correspondence between the two measures, with the rank order the same in both data sets, and the values matching reasonably well. The results are very similar, generally, with the occupational restrictions, except for the decline in the percent black at BWXT.

Of course, one issue is that the ACS samples are not very large.[32] In addition, despite our best efforts, we pick up workers at other companies in the ACS data, and the geographic match is not precise.[33] We thus, in the next section, turn to comparisons between the *LinkedIn* data and other sources of direct measures of workforce composition at the companies in question.

Classification of other companies' workforces and other validation efforts

We looked for other sources of information on company demographics, including company diversity reports and other information posted on their websites. We also found that some companies post actual EEO-1 reports.[34] Table 7 indicates, for each company, what kind of information we could identify. The notes to the table, which also list the sources, provide details on any assumptions or computations we use to obtain comparable estimates.[35]

As the table indicates, we obtain numbers to compare, either for a single year or, in the case of Meta, many years, for three companies: Chesterfield County and Fermi Research Alliance, LLC, for which we also did the ACS validation exercise, and Meta, which fortuitously provides data from two sources, in one case (its diversity report) for many years.[36]

For the first two, the observation counts in *LinkedIn* are quite low, both because they are smaller companies and because we restrict to a single year. And as the table notes, for Fermi, in particular, there

---

[32] The ACS is a 1% random sample for the years we use (https://usa.ipums.org/usa/acs.shtml).

[33] We cannot map directly to POWPUMA in the *LinkedIn* data for two reasons. First, the geographic information is less specific in *LinkedIn* (we use state). Second, the geographic information in *LinkedIn* is current and may not correspond to when the person worked at the company.

[34] We explored with both the U.S. EEOC and the California equivalent – the Civil Rights Division – obtaining such data. But at the company level they are confidential and could not be shared.

[35] There are some SpaceX numbers available from "Zippia." It is unclear where these data come from. According to its website, Zippia gets company information from employee self-reporting, public and open data sources on the internet, and proprietary data licensed from other companies. Data sources include, but are not limited to, the BLS, company filings, H1B filings, public websites on the internet, and other public and private datasets. (See https://www.zippia.com/employer/zippia-faq/.) We do not regard these data as reliable and do not use them.

[36] For the overlapping data – the percentages black and Hispanic for 2021 – the data are very close but do not match exactly.

are so few observations that the comparison should probably be ignored. For Chesterfield County, however, there is some correspondence. In both data sources, the percent female is the highest, followed by the percent black, and then the percent Hispanic; and the numbers roughly correspond (e.g., the percent female is much higher in both data sources), although the percent Hispanic is a good deal higher in the *LinkedIn* data.

The data for Meta are perhaps the most interesting, because (i) we have far more observations, and (ii) we can check some data by year. The numbers for the percent black are close, and the pattern of increase in this percentage is similar in the two data sources, as is the amount of the increase (2.9 percentage points in the diversity report, and 3.1 percentage points in *LinkedIn*). The numbers for percent Hispanic are not as close, perhaps because of how Hispanics are defined (although we find no mention of this in the Meta documents). In particular, again the share Hispanic is higher in the *LinkedIn* data.[37] But, again, the pattern of increase is similar, as is the amount of the increase (2.7 percentage points in the diversity report, and 1.6 percentage points in *LinkedIn*). The comparison with the 2021 EEO-1 Report looks similar for the percentages black and Hispanic, which is not surprising. And the percentage female is reasonably close.

What do we conclude from the validation exercises? First, there is clearly some correspondence between measures of workforce demographic composition for the *LinkedIn* data and other sources. For our ACS comparisons, given that we do not have an exact match, and that there is sampling error in the ACS data, we would not expect exact matches, so this is encouraging. Put differently, there is no reason to assume the ACS data are more reliable. But the rough correspondence is encouraging for using the *LinkedIn* data, and of course the *LinkedIn* data can in principle be used to study any company, whereas the ability to use the ACS to learn something about a company's workforce is highly limited to companies

[37] There are a couple of reasons the Hispanic classifications could differ. First, EEO-1 data are based on self-reports whereas our *LinkedIn* classification is not, and there is research indicating potential subjectivity with which people classify themselves as Hispanic (e.g., Antman and Duncan, 2024; Loewenstein et al., 2024). Second, we rely on names, which, among other things, could inflate the share Hispanic when non-Hispanics marry and change their names. There may, of course, be other reasons as well.

with a sizable share of industry employment in a POWPUMA. Second, and reassuringly for the *LinkedIn* data, the correspondence appears to be much tighter for large companies – although admittedly this is based on data for one company (Meta).

Recall, though, our perspective on the utility of the *LinkedIn* data. The data do not provide scientifically valid estimates of workforce composition with known sampling properties. Rather, they are interesting as a guide for further exploration by government agencies or attorneys seeking to enforce discrimination laws. Our take is that these data may be useful for both larger and smaller companies, although of course more reliable – and hence more useful – for larger companies. That actually meshes well with the way data are likely to be used in enforcing discrimination laws, as the class action suits that rely on statistical evidence typically are against large companies.[38]

Overall demographic composition

Having established what we view as reasonable reliability of the *LinkedIn* data, in Table 8 we report the overall demographic composition for each company – the shares black, Hispanic, and female.[39] For comparability with the earlier analysis, we report this for data beginning in 2012, but we extend to the end of the data (2023). For the four companies with which we did the ACS validation, we have much larger numbers of observations for this and the following analyses. This is mainly because we no longer restrict to the geographic areas identified in *LinkedIn* to correspond to the ACS POWPUMA. In addition, we do not restrict the time period to the 10 years covered by the ACS analysis. The percent black varies substantially, from a low of 5.7% at Research Corporation of the University of Hawaii to 23.5% at Chesterfield County. The variation in the percent Hispanic is less pronounced, but ranges from 6.1% at

[38] As examples, there have been fairly recent class action discrimination lawsuits against IKEA (https://www.consolelaw.com/court-unseals-order-conditionally-certifying-age-discrimination-collective-action-suit-against-ikea-filed-by-console-mattiacci-law/), Google (https://www.nytimes.com/2022/06/12/business/google-discrimination-settlement-women.html), Walmart (https://www.cohenmilstein.com/case-study/wal-mart/, including a very large case two decades ago), and Twitter (https://www.reuters.com/legal/twitter-beats-disabled-workers-lawsuit-over-layoffs-now-2023-05-08/).

[39] Later, we present some evidence on the intersections of minority and gender categories, although we are not able to validate these intersectional categories against the sources explored in the prior sections. In principle, other ethnic groups could be considered, although in some cases the classification models would have to be trained to identify them.

BWXT to 17.9% at SpaceX. In contrast, the percent female is lowest at SpaceX (17.7%) and highest at Chesterfield County (53.0%). It is also substantially higher than SpaceX at some of the other private employers, like GSK (46.4%) and Meta (38.2%). We caution, again, that any inference of discrimination would have to consider the composition of potential workers, which can vary with other factors – probably most notably geography and educational levels and fields.

An application to mass layoffs

Based on what we have learned about the reliability of the *LinkedIn* data, we explore using these data to study potential discrimination in mass layoffs. Given that many separations in normal times are voluntary quits,[40] from the point of view of learning about discrimination in layoffs it is more informative to look at a period of mass layoffs. That is, we refer to those "laid off," but we mean those who separated in a period of mass layoffs – when the share of layoffs in separations is much higher.

In particular, we examine layoffs at Meta, which is reported to have laid off 11,000 employees (13% of its workforce) in 2022, and 10,000 in 2023,[41] and at SpaceX, which is reported to have laid off about 10% of its workforce in 2019.[42,43] We use the *LinkedIn* data to measure all separations, identifying the spell of time an employee spends at the same company and inferring that an employee has separated with a company if they either stop working for at least four months or if their next employment is with a different company.[44]

[40] See, e.g., https://www.bls.gov/news.release/pdf/jolts.pdf, indicating about a 2-to-1 ratio of quits vs. layoffs/discharges.

[41] See: https://www.forbes.com/sites/qai/2022/12/07/meta-layoffsfacebook-continues-to-cut-costs-by-cutting-headcount/?sh=5e36a1898456; https://www.washingtonpost.com/technology/2023/05/23/meta-layoffs-misinformation-facebook-instagram/.

[42] See: https://www.cnn.com/2019/01/11/tech/spacex-layoffs/index.html; https://www.bnnbloomberg.ca/spacex-layoffs-include-577-positions-at-california-headquarters-1.1197570.

[43] One can also confirm these layoffs and find information on layoffs at other companies from https://www.warntracker.com/, which appears to be an amalgamation of announcements of large layoffs from the Worker Adjustment Training Notification (WARN) system.

[44] If a spell of time at one company is entirely overlapped by a spell at another company, we drop it from our data set. This can happen, for example, if a Ph.D. student is employed as a teaching assistant with a university for five years, but also lists a summer internship on their resume for a summer during their Ph.D. We would then drop the spell of time at the internship because it is completely overlapped by the teaching assistant position, rather than considering the teaching assistant as having separated from the university during the tenure of their internship. We do not drop partially overlapping positions, however.

As shown in Table 9, consistent with the news stories, we see inordinately high numbers of separations from Meta in 2022 and 2023, as well as a somewhat higher share in 2021.[45] We suspect the 2023 numbers may be lower than reported layoffs because either the layoffs were not yet implemented when we extracted the data in fall of 2023, or some people do not update their profiles until they get a new job. Similarly, some of the 2022 layoffs may be reflected in the 2023 data. For SpaceX, we see the highest share in 2019, again consistent with news stories.

We compare the demographic composition of those laid off, vs. the workforce as a whole, in Table 10. To do this, we use all years of data, and estimate a linear probability model for separations for each individual in each year. We include year dummy variables, which will capture the aggregate differences in separation probabilities, including the spikes associated with mass layoffs documented in Table 9. And we include main effects of race, ethnicity, and gender to allow for group-specific separation probabilities. We then also add interactions between dummy variables for race, ethnicity, and gender with a dummy variable for the mass layoff years. Because a higher share of separations will be involuntary in mass layoff years, the overall year dummy variables for these years will capture the higher share of involuntary separations for non-black, non-Hispanic males. The key variables are the race/ethnicity/gender by mass layoff interactions, which will capture the race, ethnic, or gender differences in the probabilities of a separation in the mass layoff years. We interpret the latter as reflecting the differences in layoff rates for blacks, Hispanics, or women in the mass layoff years.

In the top panel, we use only these variables in the regression models. For Meta, the estimates in the first column indicate higher separation rates in general for blacks, Hispanics, and women, with statistically significant differences in the range of 1.3 to 1.6 percentage points (relative to an 11.8% rate for the non-mass layoff years for non-black, non-Hispanic males). These differences could reflect higher involuntary separations in non-mass layoff years, but they may reflect voluntary separations.[46] The second

[45] Table 9 also exhibits the rising employment over time at these companies, as noted earlier.
[46] Enforcement agencies or private attorneys might find these differences sufficiently suggestive to undertake further efforts to obtain company data, which often contain the reason for the separation.

column reports the interactions with the dummy variable for the mass layoff years. For all three groups the separation rate is significantly higher – a 1.40 percentage point higher rate for blacks, significant at the 5% level for blacks, and 2.18 and 2.57 percentage points higher rates for Hispanics and women, significant at the 5% level. (These can be compared to a 15.9% rate for the mass layoff years for non-black, non-Hispanic males.) Thus, there is reason to believe that blacks, Hispanics, and women all experienced higher mass layoff rates at Meta in 2021-2023.

For SpaceX, the analysis (in the third and fourth columns) focuses on 2019 only. Here, we find a lower separation rate for Hispanics and a higher rate for women, in general. During the mass layoff years, the separation rate is significantly higher only for Hispanics (by 3.2 percentage points, relative to a separation rate of 19.5% for non-black, non-Hispanic males in 2019). Thus, only for Hispanics is there evidence of a higher mass layoff rate.

To be clear, these analyses do not provide a rigorous "test" of discrimination because other factors could account for disproportionate layoffs among some groups. On the other hand, this kind of evidence could be far more helpful when trying to identify claims of systemic discrimination than anecdotal evidence one or a few plaintiffs present to a government agency or private attorney.[47] Moreover, the *LinkedIn* data can be used to construct some additional controls to augment the models. We have constructed controls for both highest educational degree, and company tenure.

LinkedIn profiles contain self-reported education data for individuals, with fields for start and end dates, degree names, fields of study, educational institutions, and grades. Individuals can add as many education entries as they would like (e.g., a bachelor's degree and a master's degree). Since these fields

---

[47] Moreover, establishing evidence of discrimination in court does not require the same level of rigor as the most cutting-edge economics research, for two reasons. First, a simply "conjecture" that unobservables could explain a difference otherwise attributable to discrimination does not constitute evidence; it has to be supported by data. Most directly, in *Bazemore v. Friday*, 478 U.S. 385, 399-400, 403 n.14 (1986), the court rejected defendant's argument that plaintiffs' analysis was unsound because certain factors were omitted, finding that defendant "made no attempt … to demonstrate that when these factors were properly organized and accounted for there was no significant disparity between the salaries of blacks and whites." Second, the standard of evidence in a civil case is a "preponderance of evidence," often interpreted as "more likely true than not." (See, e.g., https://www.law.cornell.edu/wex/preponderance_of_the_evidence.)

are self-reported, they are not guaranteed to be accurate, but our assumption is that given the importance of education in finding employment, many individuals will choose to self-report education accurately and few individuals will omit it. We cleaned the education degree name fields for individuals in our data (beginning with 18,432 unique degree names, and standardizing these).[48] We found that 86.9% of employees in our data reported some information on education that could be classified as a high school, associate's, bachelor's, master's, or doctoral degree.[49] We also constructed company tenure from start and stop dates (the latter if there was more than one employment spell) of employment at the companies.[50]

The estimates with these controls are reported in the lower panel of Table 10. Most of the estimates are very similar. The one exception is that there is now strong evidence that females at SpaceX experienced mass layoffs at a higher rate, with the probability higher by 4.2 percentage points.

The last analysis we did for layoffs considers the intersections between race and gender and ethnicity and gender. This evidence is reported in Table 11. The one striking finding is that, at SpaceX, the probability of a mass layoff was much higher for black females than for black males. The differential for black males is small and statistically insignificant, while the differential for black females relative to black males (which is close to the overall differential relative to white males) is 16.6 percentage points, significant at the 5% level.

Hires

We did similar calculations for hires. For Meta, the representation of blacks, Hispanics, and women among hires was generally very similar to that among the workforce, as shown in Table 12 (with some variation from year to year). For SpaceX, the comparisons suggest lower hiring of blacks and Hispanics relative to the workforce shares, especially in more recent years, but higher hiring of women. Of course, with regard to hiring the more relevant question is the comparison of the race, ethnic, or

[48] As corroborating evidence of the validity of our education coding, we found far higher percentages with doctoral degrees at Fermi, GSK, and Research Corporation of the University of Hawaii, and a far higher percentage with high school as the highest degree for Chesterfield County.
[49] This category includes Ph.D., M.D., J.D., Pharm.D., and anything else with the word "doctor" in it.
[50] One might also want to control for position (job). However, this is likely empirically infeasible and not very meaningful, given that there are very large numbers of unique/idiosyncratic job titles.

gender composition of hires relative to applicants or the potential hiring pool. Thus, we do not report any statistical tests. Nonetheless, this table suggests that one could use estimates like these, relative to "benchmark" estimates of the composition of the hiring pool sometimes used in discrimination cases, usually from the ACS, to obtain provisional evidence on discrimination in hiring.[51]

Exploring promotions

Finally, we consider the use of *LinkedIn* data to measure promotions. We expect, ex ante, that measuring promotions will be more problematic. A hire or separation is relatively straightforward to measure in the *LinkedIn* data from the work history. But even if people report every position they hold while at a company, it is not certain that every position change is a promotion; conversely, some promotions may occur without position changes.[52] Moreover, job titles are not reported uniformly on *LinkedIn*. For example, in our sample of 57,097 work experiences at Meta, we observe 24,720 unique job titles, 21,460 of which only appear once. Thus, we caution that any use of the *LinkedIn* data to measure promotion differences by race, ethnicity, and gender must be viewed as more provisional, and a user of the data for these purposes will likely have to make their own assessment, based on job title transitions in the data, as to the reliability of any estimated differences across groups.[53] Reflecting this, the analysis we describe should perhaps best be viewed as exploring how one might use and assess the *LinkedIn* data for this purpose.

We begin by presenting evidence on the measured job title changes we see in the *LinkedIn* data at the companies we studied. Table 13 reports the 20 most common job title changes at each company. One

---

[51] For an example comparing hires to such benchmarks in a discrimination case, see Expert *Report of David Neumark in the matter of Heldt et al. v. Tata Consultancy Services, Ltd.*, February 2017. For examples and discussion of EEOC guidance using benchmarks from the Decennial Census or the ACS, see https://www.eeoc.gov/federal-sector/management-directive/instructions-federaal-agencies-eeo-md-715-1 and Amano-Patiño et al. (2022). For an example at the state level see https://www.twc.texas.gov/sites/default/files/enterprise/docs/equal-employment-opportunity-minority-hiring-practices-report-2016-twc.pdf.

[52] Indeed, based on our experience in discrimination litigation, even with full access to a company's data the definition of promotions can sometimes be ambiguous.

[53] Note that, in practice, this might be done for a specific set of jobs, like those covered in a class action claim. The same could also be true of a hiring analysis.

can only assess subjectively whether most of these job title changes look like promotions or not. Our view is that most of them do for some companies, but not others. For BWXT, many transitions are not clearly promotions (e.g., Product Engineer to Design Engineer). For Chesterfield County, we think that most are (e.g., GIS Analyst to Senior GIS Analyst, and Planner to Senior Planner). For Fermi, GSK, Meta, and SpaceX, nearly all transitions look like promotions, with many including words like "senior" or "lead" or numbered levels indicating clear promotions. For Research Corporation of the University of Hawaii, some of the transitions look like promotions, but there are too few to reliably use the data anyway.[54]

Based on this assessment, it appears that we can obtain some reliable information on promotion differences by race, ethnicity, and gender for Fermi, GSK, Meta, and SpaceX, among the seven companies we considered, by considering job title changes as promotions. As an illustration, we present estimates of promotion models for each of these companies. The analysis is similar to our earlier layoff analysis. The results are reported in Table 14, however in this case we simply aggregate across all years and estimate promotion rate differentials by race, ethnicity, and gender, since we do not have any reference to particular years of interest as we did in the layoff analysis.

The top panel reports the more parsimonious models. At Meta, the promotion rate differentials for blacks, Hispanics, and women are positive, and the difference for women is statistically significant (1.98 percentage points, which can be compared to a 5.6% promotion rate for non-black, non-Hispanic males). The promotion rate for women is also higher at SpaceX, but the promotion rates for blacks and Hispanics are significantly lower (2.5 and 2.8 percentage points, respectively, compared to a 7.0% promotion rate for non-black, non-Hispanic males). At GSK and Fermi, the only significant differences are for blacks,

[54] To be clear, we are not arguing that all job changes are necessarily promotions. One could conceivably select a subset of job changes for a company and study them as promotions. We have identified two studies that ask how promotions (as identified by workers) differ from other job changes. McCue (1996) studies the differences between self-reported promotions and other job changes (at the same firm) in the PSID. In her data, these are roughly equal shares. For workers with 10 or fewer years of tenure, both changes are associated with wage growth relative to no job change, with the wage growth higher for promotions. At higher tenure levels, only promotions are associated with wage growth. In a study using employer-employee data from Portugal, van der Klauw and da Silva (2011) study two kinds of promotions – clear moves up a hierarchy, or promotions as reported by firms. They report that these two measures do not overlap well, with about 40% of employer-reported promotions not associated with a change in job descriptions, and only about 30% of hierarchical step-ups are also reported as promotions by employers. Both types of "promotions" are associated with substantial wage increases.

with significantly lower promotion rate differentials of 2.4 percentage points at GSK and 1.2 percentage points at Fermi. We reiterate that a particular reason the evidence on promotions should also be viewed as suggestive is because we do not necessarily measure promotions the way these companies do. Nonetheless, once again this kind of evidence could be far more helpful when trying to identify claims of systemic discrimination than anecdotal evidence one or a few plaintiffs present to a government agency or private attorney.

We did the same supplemental analyses as for layoffs. First, we add the education and tenure controls, in the lower panel of Table 14. These generally have little impact, except that the lower promotion rate for Hispanics at Fermi more than doubles in magnitude and becomes significant at the 5% level. The intersectional results, in Table 15, are striking. In every single case, blacks or Hispanics who are also female experience far lower promotion rates, with the differential statistically significant at the 5% level for both black women and Hispanic women, at each of the four companies.

Finally, we did one additional refinement, trying to identify job changes that were more clearly promotions. We matched between strings in pairs of job titles associated with positions held consecutively at a company to identify job changes that were very likely promotions, and redid the promotions analysis considering only these job changes as promotions. Specifically, we first did string matching between all pairs of job transitions at a company, using *matchit* in Stata. Based on this, we concluded that it was difficult to verify that pairs with a similarity score below 0.8 were promotions, while those above were more likely to be promotions. To illustrate, Table 16 shows 10 randomly selected pairs of job titles held consecutively in the 0.7-0.8 and the 0.8-0.9 ranges of the similarity score. While some of the pairs in the top panel may be promotions, at least six of the 10 listed in the bottom panel appear to be (e.g., the addition of "Lead" in the first pair, or "Senior" in the second).

Based on examining these consecutive jobs pairs, we decided to code as promotions job pairs where the two job titles have a similarity score above 0.8 and the second job title contains one of the following words and the first job title does not:

- SENIOR

- SR
- II[55]
- PRINCIPAL
- LEAD
- MANAGER.

In addition, we code them as promotions if the same semantic similarity criterion is met and the first job title contains one of the following and second (next) job title does not:

- ASSISTANT
- ASSOCIATE.

As reflected in Table 16, this method is likely conservative in terms of identifying promotions. Some of the transitions with similarity scores below 0.8 may also be promotions, and these rules may also fail to classify some promotions above the 0.8 threshold because they use other words to signify the promotion (e.g., the second job title has the word "Head" in it). However, all of the examples we classify as promotions do appear to be promotions.[56]

Table 17 reports estimates of the same specifications as in Table 14 (the lower panel, with the additional controls), but using this narrower definition of promotions. All other job changes not classified as promotions are treated as non-promotions. Note that, as reported in the table, this results in much lower promotion rates, and the reduction relative to Table 14 varies by company, which can be a consequence of job titles more easily identifying promotions at some companies than others.[57] The lower promotion rates are reflected in smaller estimated coefficients. The conclusions for Meta are different from Table 14, with the models now yielding evidence of lower promotions for blacks, and no longer indicating higher

---

[55] The addition of "II" to the job title and "I" in the previous job title captured nearly all job transitions with numbers/levels that appeared to correspond to promotions.

[56] As this discussion indicates, one could obviously choose somewhat different rules to classify promotions, so our analysis should be viewed as suggestive of the ability to do this and what one can learn from it, rather than as a definitive analysis of promotions.

[57] Indeed, the reduction is smaller for Meta and SpaceX, for which Table 13 suggests that job changes are more readily identifiable as promotions.

promotions for women. For SpaceX the lower promotions for blacks and Hispanics persist, but the higher promotion rate for women is no longer evident. For GSK and Fermi the results are not qualitatively different.

As this analysis indicates, the classification of promotions from the *LinkedIn* data is subjective and certainly not definitive. However, in a litigation context one could choose a subset of job title changes that, for a particular company, have been identified as corresponding to promotions, although the *LinkedIn* job titles would not map perfectly to the job titles used at a company.

Potential limitations

One inherent limitation of our approach is that it is more applicable (and reliable) for companies with large shares of employees on *LinkedIn*. We suspect that companies with large shares of lower-skilled and lower-paid jobs are less well represented on *LinkedIn*, and the same could be true of lower-skilled and lower-paid workers at the companies on *LinkedIn*. On the other hand, the composition of employment at higher-pay, higher-skilled firms, and advancement through the ranks at these companies is critically important because these companies, and the higher-level jobs within them, are among the best jobs in the U.S. economy. These are also the jobs at which minority groups (and women, in tech jobs) are under-represented.[58]

A second limitation is the possibility of fake profiles on *LinkedIn.* We have no information on how pervasive this is. We also do not know any algorithm to identify these profiles.[59] Still, this is another reason one should not interpret the type of analysis we do with the *LinkedIn* data – or that others could do – as definitive with respect to measuring discrimination, but rather as indicative of possible discrimination. Finally, there is some evidence that *LinkedIn* seems to be fairly successful at stopping fake accounts.[60]

---

[58] See National Center for Science and Engineering Statistics (2023).
[59] Rather, "advice" on spotting them is based on reading and assessing individual profiles and does not appear very systematic (e.g., https://www.forbes.com/sites/forbesbusinesscouncil/2022/11/17/how-to-identify-a-fake-linkedin-profile-in-five-minutes-or-less/?sh=1421f73f1d7c; https://www.linkedin.com/pulse/dangers-fake-profiles-how-spot-one-david-smith-cv-writer).
[60] See https://www.cnbc.com/2022/12/10/not-just-twitter-linkedin-has-fake-account-problem-its-trying-to-fix.html.

A third limitation is that we cannot study pay (and hence pay discrimination) in the *LinkedIn* data. At present, we are not aware of other social media data that could be used in the same manner and that also includes pay. It is not inconceivable that this could be done however, perhaps by collecting survey data via targeted advertising like in the Schneider and Harknett (2019b) Facebook data collection on work schedules.

A fourth limitation is that we have not explored developing even better methods for classification by race, ethnicity, and gender, but instead use, by and large, "off the shelf" methods. There may be potential room for improvement, including, for example, training models based on data sets that might have more overlap with the *LinkedIn* data.

It is also important to clarify that the kind of evidence that can be produced with the *LinkedIn* data is not intended to be rigorous evidence of discrimination. Our goal is to try to use the *LinkedIn* data to produce estimates of race, ethnic, and gender differences in employment, hiring, separations (especially layoffs), and promotions. Our ability to look within companies can provide new descriptive evidence that is currently not available to researchers. And this evidence and methods, as we have argued, will be useful in enforcing discrimination laws. We want to be clear, however, that we are not proposing our measures based on the *LinkedIn* data as evidence of discrimination per se. Other factors can explain sorting of workers across firms, as well as differences in retention and promotion.

The intention is not that the *LinkedIn* data would necessarily be the data actually used to establish the definitive evidence of discrimination for either legal proceedings or research. The strongest evidence would typically require richer company data, both for reliability and comprehensiveness, and to rule out other non-discriminatory explanations.[61] These data typically become available at later stages of litigation. But our findings suggest that the *LinkedIn* data can be used to increase the precision with which potentially discriminatory companies could be targeted for further legal exploration, including filing of

[61] On the latter point, for example, one might get information on promotions from job titles at a company and be able to test whether education and prior jobs explain any difference using *LinkedIn* data to supplement company data. But more rigorous evidence would likely require performance ratings, as well as perhaps a more definitive identifier of promotions.

discrimination claims and opening of legal discovery to access the richer data that companies have on both workforce outcomes and potential factors accounting for those differences. And, conversely, the use of these data might prevent spurious lawsuits against companies less likely to be discriminating.

Summary and conclusions

We have shown that one can use *LinkedIn* data to obtain reasonably reliable measures of workforce demographic composition by race, ethnicity, and gender, based on validation exercises comparing estimates from scraped *LinkedIn* data to two sources – ACS data, and company diversity or EEO-1 reports. To be clear though, we do this validation for a small number of companies, limited by a restriction to one industry (which we suspect is better represented on *LinkedIn*) and to a small number of companies dictated by our research budget. This validation is further restricted to companies that can be compared to ACS data because they represent a high share of industry employment in a POWPUMA and companies that make public diversity reports or EEO-1 reports. Our evidence cannot speak to the universe of industries or companies.

We emphasize that the research we present in this paper is to some extent a "proof of concept" (or more accurately an assessment of a proof of concept), exploring how well our ideas for measuring the demographic composition of companies' workforces, and the relation of other decisions of these companies to demographics, can be measured. The methods we develop and explore – which we anticipate will be improved upon by others – can be used in a number of ways, potentially.

First, and most directly related to this paper, the *LinkedIn* data can be used by plaintiffs' attorneys or agencies charged with enforcing discrimination laws. We illustrate how this might be done by studying mass layoffs (and hiring) at two companies for which we extracted *LinkedIn* data, and by exploring the measurement and analysis of promotions. We first show that the data are sufficiently comprehensive to detect mass layoffs. And we then illustrate using the data to compare the race, ethnic, and gender composition of laid off workers to the workforce as a whole. Again, we caution that the *LinkedIn* data are unlikely to provide the level of rigor in measuring discrimination that is required to establish a legal claim of discrimination, both because the *LinkedIn* data do not provide a scientific sample (indeed, in an actual discrimination case, data on all employees covered by the case would typically be available), and because of a lack of some control variables – although it is possible to recover education and prior job history information from *LinkedIn*. Still, this kind of evidence may be much more convincing to attorneys or government agencies than anecdotal evidence from a handful of laid-off workers. Measuring promotions is harder because they are less directly captured in the *LinkedIn* data. Nonetheless, we provide evidence suggesting that one may be able to learn something about race, ethnic, and gender differences in promotions from these data.

Second, we imagine that researchers will develop other ways to use these data to measure and study the demographics of the workplace, such as the evolution of employment by race, ethnicity, and gender, overall (or at least in industries well-represented on *LinkedIn*), and extending to other questions like the progress women and minorities are making in reaching higher-level positions within companies. Indeed, while in this paper we only study data on a limited number of companies, the company from which we draw *LinkedIn* data does make available the entire public *LinkedIn* database. And researchers might be interested in more-refined evidence on discrimination in the form of unexplained demographic differences in hiring, layoffs, promotions, etc., constructing more-detailed controls from the *LinkedIn* data. Although this kind of evidence falls short of the level of rigor of experimental or quasi-experimental studies (Neumark, 2018), the ability to provide evidence on specific firms may offer unique insights.[62]

Third, researchers could potentially delve into other dimensions of discrimination. For example, some studies examine differences in patterns suggestive of discrimination between client-facing and other jobs, to assess the role of customer discrimination (e.g., Combes, et al., 2016; Neumark, 2024). Other research has considered the effects of skin tone on wage differentials and other outcomes among blacks, and relative to whites (e.g., Goldsmith et al., 2006; Monk, 2014). This kind of classification of blacks may be feasible using the *LinkedIn* pictures.[63] Another line of research conducts correspondence studies

[62] We remind the reader of the Kline et al. (2024) correspondence study approach to garnering company-specific evidence on race discrimination in hiring.
[63] This would be complicated by the fact that *LinkedIn* photos are taken in very different lighting situations as compared to DMV or passport photos, which are fairly standardized across individuals.

of differences in outcomes based on looks, even manipulating facial symmetry in photos (Bóo et al., 2013); again, this could be pursued based on *LinkedIn* pictures. To be clear, though, research along these lines might be more interesting to discrimination researchers than enforcement agencies, in which case the lower rigor attainable using the *LinkedIn* data might be more limiting.

Finally, we think our demonstration of how we use the *LinkedIn* data may prompt consideration of other potential research uses one could make of these data. One can, for example, construct work histories and educational histories, as well as job titles – although the job titles are somewhat idiosyncratic and company-specific, as our promotions analysis shows. Thus, for example, the *LinkedIn* data could in principle be used to study the impact of education on careers, and to study career trajectories including both changes within companies and mobility across companies. This is an interesting question in its own right, and also in relation to differences by race, ethnicity, and gender. As noted earlier, the LEHD, like other linked employer-employee databases, can be used to study mobility across firms, and mobility within firms, although in both cases based more on earnings than job titles.[64] One data source that can study specific job titles is the *ExecuComp* data studied by Bertrand and Hallock (2001). However, this data source covers only the upper echelons of companies (the top five executives), and hence is less suited to studying mobility. Another possible source of data is targeted surveys, like Bertrand et al.'s (2010) survey research on MBA, their careers, and the evolution of the earnings gap. A survey like this has the obvious advantage of the researchers being able to choose the questions, and the disadvantage of narrow coverage. Given these limitations, the *LinkedIn* data may be a unique source to study career trajectories and mobility, characterizing race, ethnic, and gender differences using the kinds of classifications we develop in this paper, providing new evidence that is complementary to what has been or can be learned from other data sources.

[64] We already noted Barth et al. (2021), which studies how the gender gap evolves as workers move across firms. For an example using Italian data, see Casarico and Lattanzio (2024).

References

Amano-Patiño, Noriko, Julian Aramburu, and Zara Contractor. 2022. "Is Affirmative Action in Employment Still Effective in the 21st Century?" U.S. Census Bureau, Center for Economics Studies Working Paper 22-54.

Antman, Francisca M., and Brian Duncan. 2024. "Ethnic Identity and Anti-Immigrant Sentiment: Evidence from Proposition 187." Unpublished paper, https://conference.nber.org/conf_papers/f182355.pdf.

Azar, José, Ioana Marinescu, and Marshall Steinbaum. 2022. "Labor Market Concentration." *Journal of Human Resources* 58(1): 167-199.

Barth, Erling, Sari Pekkala Kerr, and Claudia Olivetti. 2021. "The Dynamics of Gender Earnings Differentials: Evidence from Establishment Data." *European Economic Review* 134: 103713.

Bayard, Kimberly, Judith Hellerstein, David Neumark, and Kenneth Troske. 2003. "New Evidence on Sex Segregation and Sex Differences in Wages from Matched Employer-Employee Data." *Journal of Labor Economics* 21(4): 887-922.

Bertrand, Marianne, Claudia Goldin, and Lawrence F. Katz. 2010. "Dynamics of the Gender Gap for Young Professionals in the Financial and Corporate Sectors." *American Economic Journal: Applied Economics* 2(3): 228-55.

Bertrand, Marianne, and Kevin F. Hallock. 2001. "The Gender Gap in Top Corporate Jobs." *Industrial and Labor Relations Review* 55(1): 3-21.

Blevins, Cameron, and Lincoln Mullen. 2015. "Jane, John... Leslie? A Historical Method for Algorithmic Gender Prediction." *DHQ: Digital Humanities Quarterly* 9(3).

Bóo, Florencia López, Martín A. Rossi, and Sergio S. Urzúa. 2013. "The Labor Market Return to an Attractive Face: Evidence from a Field Experiment." *Economics Letters* 118(1): 170-2.

Borup, Daniel, and Erik Christian Montes Schütte. 2022. "In Search of a Job: Forecasting Employment Growth Using Google Trends." *Journal of Business and Economic Statistics* 40(1): 186-200.

Brick, Carmen, Daniel Schneider, and Kristen Harknett. 2023. "The Gender Wage Gap, Between-Firm Inequality, and Devaluation: Testing a New Hypothesis in the Service Sector." Forthcoming in *Work and Occupations*.

Burn, Ian, Patrick Button, David Neumark, & Luis Felipe Munguia Corella. 2022. "Does Ageist Language in Job Ads Predict Age Discrimination in Hiring?" *Journal of Labor Economics* 40(3): 613-67.

Burnes, Daria, David Neumark, and Michelle J. White. 2014. "Fiscal Zoning and Sales Taxes: Do Higher Sales Taxes Lead to More Retailing and Less Manufacturing?" *National Tax Journal* 67(1): 7-50.

Casarico, Alessandra, and Salvatore Lattanzio. 2024. "What Firms Do? Gender Inequality in Linked Employer-Employee Data." *Journal of Labor Economics* 42(2): 325-55.

Clemens, Jeffrey, Lisa B. Kahn, and Jonathan Meer. 2021. "Dropouts Need Not Apply? The Minimum Wage and Skill Upgrading." *Journal of Labor Economics* 39(S1): S107-49.

Combes, Pierre-Phillipe, Bruno Decreuse, Morgane Laouénan, and Alain Trannoy. 2016. "Customer

Discrimination and Employment Outcomes: Theory and Evidence from the French Labor Market." *Journal of Labor Economics* 34(1, Pt. 1): 107-60.

Consumer Financial Protection Bureau. 2014. *Using Publicly Available Information to Proxy for Unidentified Race and Ethnicity.* https://files.consumerfinance.gov/f/201409_cfpb_report_proxy-methodology.pdf.

Cullen, Zoe B., and Bobak Pakzad-Hurson. 2021. "Equilibrium Effects of Pay Transparency." NBER Working Paper No. 28903.

Deng, Jiankang, et al. 2019. "RetinaFace: Single-Stage Dense Face Localization in the Wild." arXiv: 1905:00641v2.

Goldsmith, Arthur H., Darrick Hamilton, and William Darity, Jr. 2006. "Shades of Discrimination: Skin Tone and Wages." *American Economic* Review 96(2): 242-5.

Google. 2022. "2022 Diversity Annual Report." https://about.google/belonging/diversity-annual-report/2022/.

Hellerstein, Judith K., Mark Kutzbach, & David Neumark. 2019. "Labor Market Networks and Recovery from Mass Layoffs: Evidence from the Great Recession Period." *Journal of Urban Economics* 113: 103192.

Hellerstein, Judith, David Neumark, and Melissa McInerney. 2008. "Spatial Mismatch or Racial Mismatch?" *Journal of Urban Economics* 64(2): 464-79.

Hellerstein, Judith, Melissa McInerney, and David Neumark. 2011. "Neighbors and Co-Workers: The Importance of Residential Labor Market Networks." *Journal of Labor Economics* 29(4): 659-95.

Hellerstein, Judith, and David Neumark. 2004. "Ethnicity, Language, and Workplace Segregation: Evidence from a New Matched Employer-Employee Data Set." *Annales d'Economie et de Statistique* 71/72: 19-78.

Hellerstein, Judith K., David Neumark, & Kenneth Troske. 1999. "Wages, Productivity, and Worker Characteristics: Evidence from Plant-Level Production Functions and Wage Equations." *Journal of Labor Economics* 17(3): 409-46.

Hu, Lingqian. 2019. "Racial/Ethnic Differences in Job Accessibility Effects: Explaining Employment and Commutes in the Los Angeles Region." *Transportation Research Part D: Transport and Environment* 76: 56-71.

Kline, Patrick M., Evan K. Rose, and Christopher R. Walters. 2024. "A Discrimination Report Card." NBER Working Paper No. 32313.

LinkedIn. 2021. "Our 2021 Workforce Diversity Report." https://news.*LinkedIn*.com/2021/october/2021-workforce-diversity-report.

Loewenstein, Mark, David Piccone, and Anne Polivka. 2024. "Measuring Potential Effects of the Proposed Race and Ethnicity Questions in the CPS." Unpublished paper, https://conference.nber.org/conf_papers/f180920.pdf.

Marinescu, Ioana, and Ronald Wolthoff. 2020. "Opening the Black Box of the Matching Function: The Power of Words." *Journal of Labor Economics* 38(2): 535-68.

McCue, Kristin. 1996. “Promotions and Wage Growth.” *Journal of Labor Economics* 14(2): 175-209.

McKinney, Kevin L., and John M. Abowd. 2022. “Males Earnings Volatility in LEHD Before, During, and After the Great Recession.” *Journal of Business and Economic Statistics* 41(1): 33-9.

Monk, Ellis P. 2014. “Skin Tone Stratification among Black Americans, 2001-2003.” *Social Forces* 92(2): 1313-37.

National Center for Science and Engineering Statistics (2023). *Diversity and STEM: Women, Minorities, and Persons with Disabilities*. National Science Foundation Directorate for Social, Behavioral and Economic Sciences, NSF 23-315.

Neumark, David. 2024. “Age Discrimination in Hiring: Evidence from Age-Blind vs. Non-Age-Blind Hiring Procedures.” *Journal of Human Resources* 59(2): 1-34.

Neumark, David. 2018. “Experimental Research on Labor Market Discrimination.” *Journal of Economic Literature* 56(3): 799-866.

Neumark, David. 2012. “Detecting Evidence of Discrimination in Audit and Correspondence Studies.” *Journal of Human Resources* 47(4): 1128-57.

Neumark, David, Ed. 2007. Improving School-to-Work Transitions. Russell Sage Foundation.

Neumark, David, Ian Burn, & Patrick Button. 2019. “Is It Harder for Older Workers to Find Jobs? New and Improved Evidence from a Field Experiment.” *Journal of Political Economy* 127(2): 922-70.

Neumark, David. 1999. “Labor Market Information and Wage Differentials by Race and Sex.” *Industrial Relations* 38(3): 414-45.

Schneider, Daniel, and Kristen Harknett. 2019a. “Consequences of Routine Work-Schedule Instability for Worker Health and Well-Being.” *American Sociological Review* 84(1): 82-114.

Schneider, Daniel, and Kristen Harknett. 2019b. “What’s to Like? Facebook as a Tool for Survey Data Collection.” *Sociological Methods & Research* 51(1): 108-40.

Serengil, Sefik Ilkin, and Alper Ozpinar. 2020. “LightFace: A Hybrid Deep Face Recognition Framework.” *2020 Innovations in Intelligent Systems and Applications Conference (ASYU)*, Istanbul, Turkey, 2020, pp. 1-5.

Serengil, Sefik Ilkin, and Alper Ozpinar. 2021. “HyperExtended LightFace: A Facial Attribute Analysis Framework.” *2021 International Conference on Engineering and Emerging Technologies (ICEET)*, Istanbul, Turkey, 2021, pp. 1-4.

Trotter, Richard G., Susan Rawson Zacur, and Lisa T. Stickey. 2017. “The New Age of Pay Transparency. *Business Horizons* 60(4): 529-39.

van der Klauw, Bas, and António Dias de Silva. 2011. “Wage Dynamics and Promotions Inside and Between Firms.” *Journal of Population Economics* 24: 1513-48.

Xie, Fangzhou. 2022. “Rethnicity: An R Package for Predicting Ethnicity from Names.” *Software X* 17: 100965.

Figure 1: Examples of *LinkedIn* Pictures

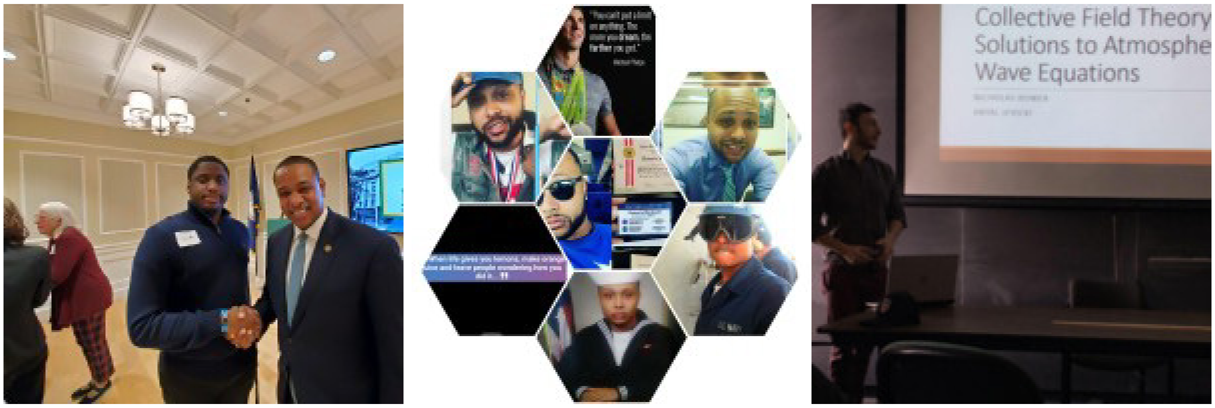


Figure 2A: Distributions of Probabilities of Gender, Ethnicity, and Race

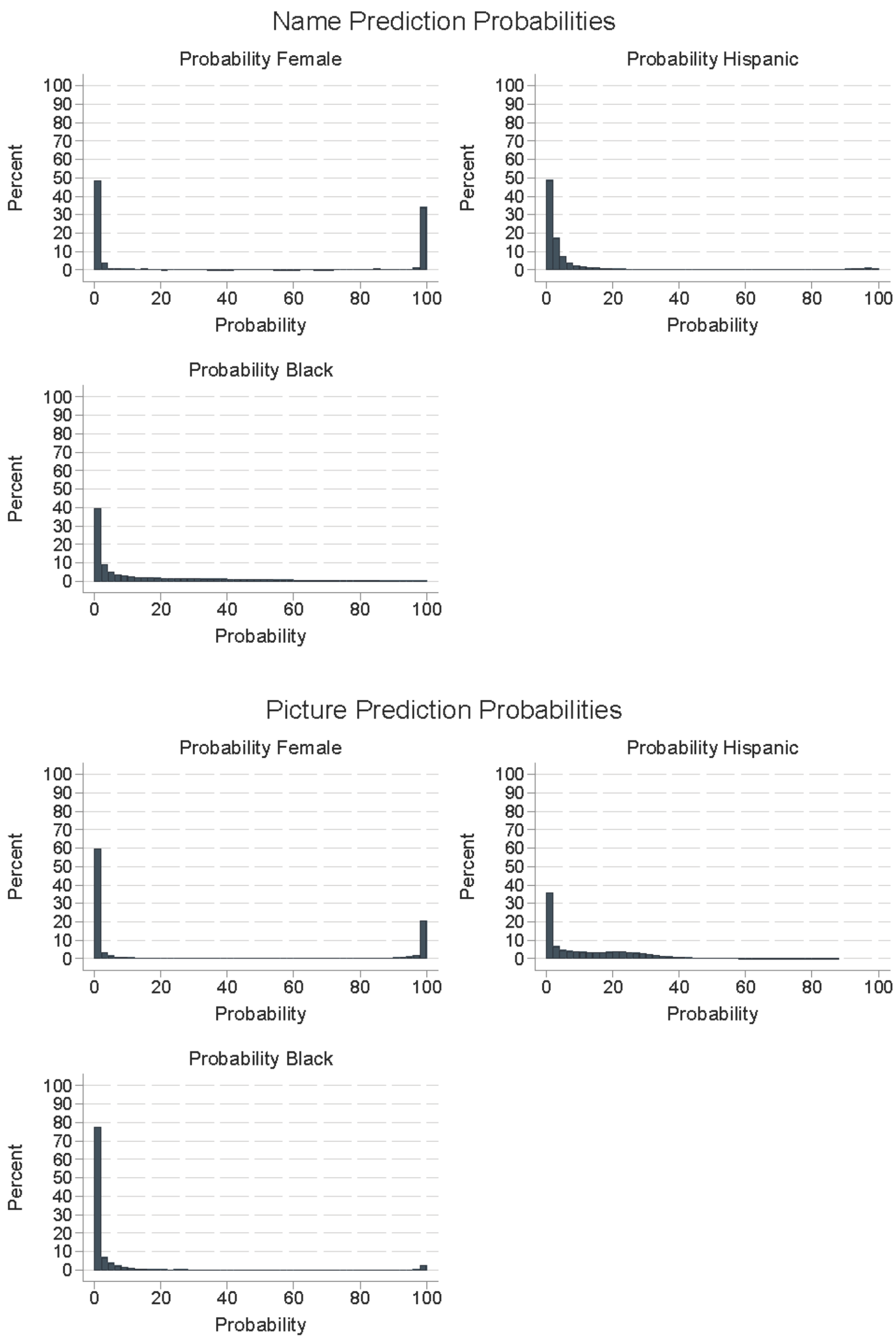


Figure 2B: Distributions of Probabilities of Gender, Ethnicity, and Race – Upper Tails

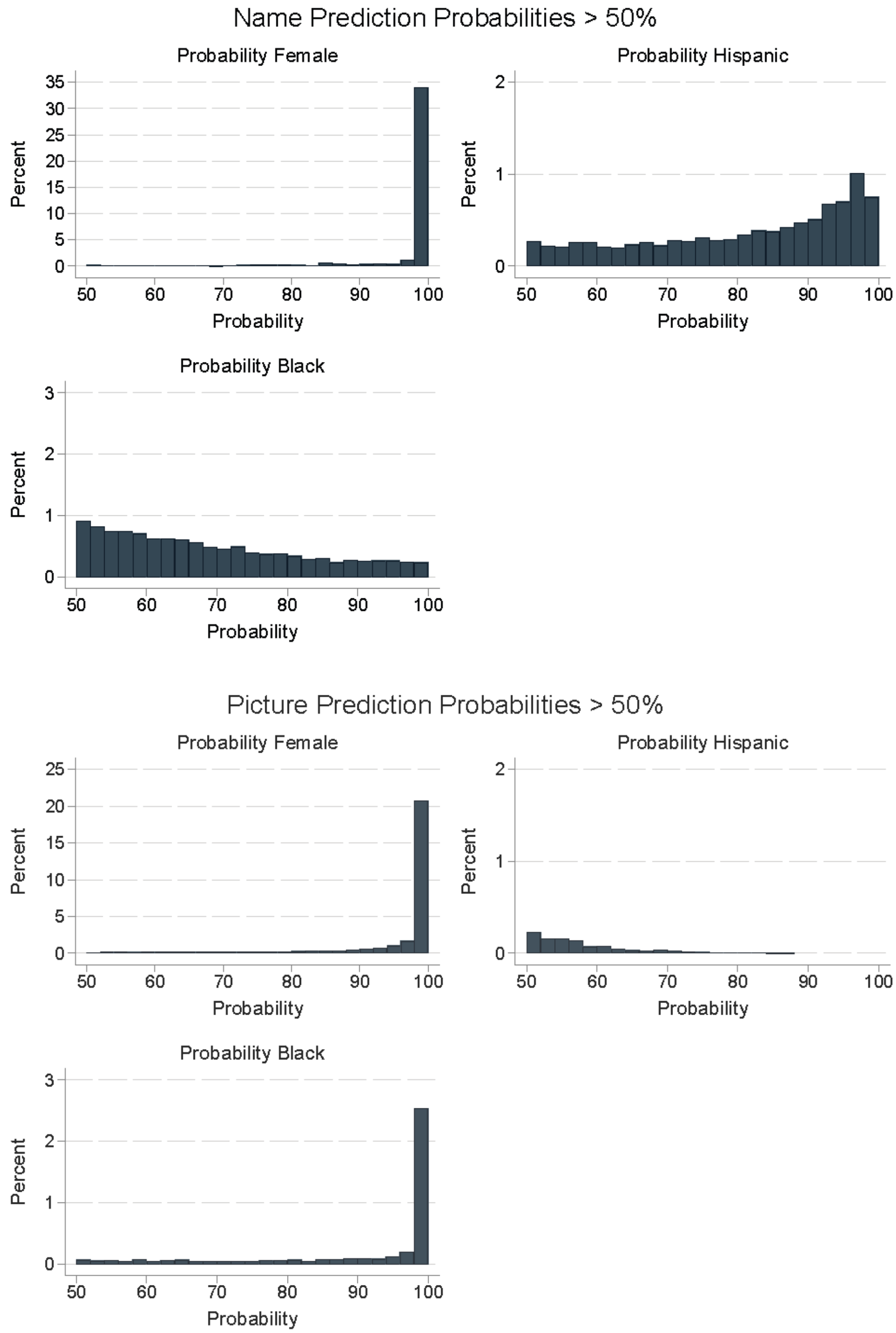
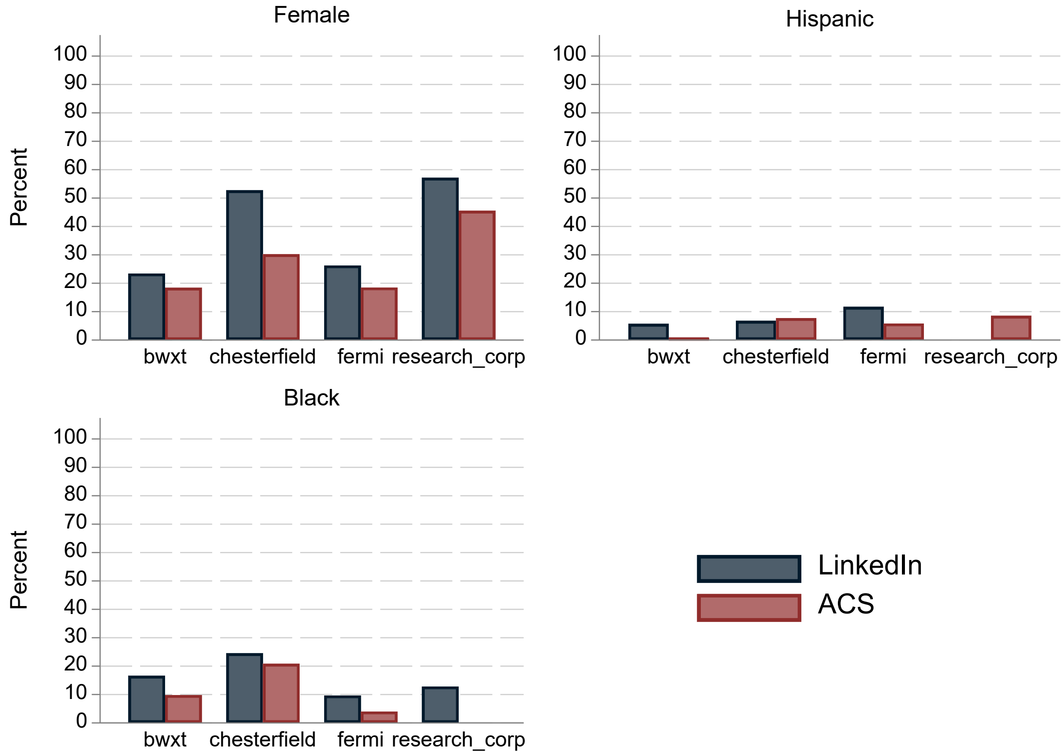
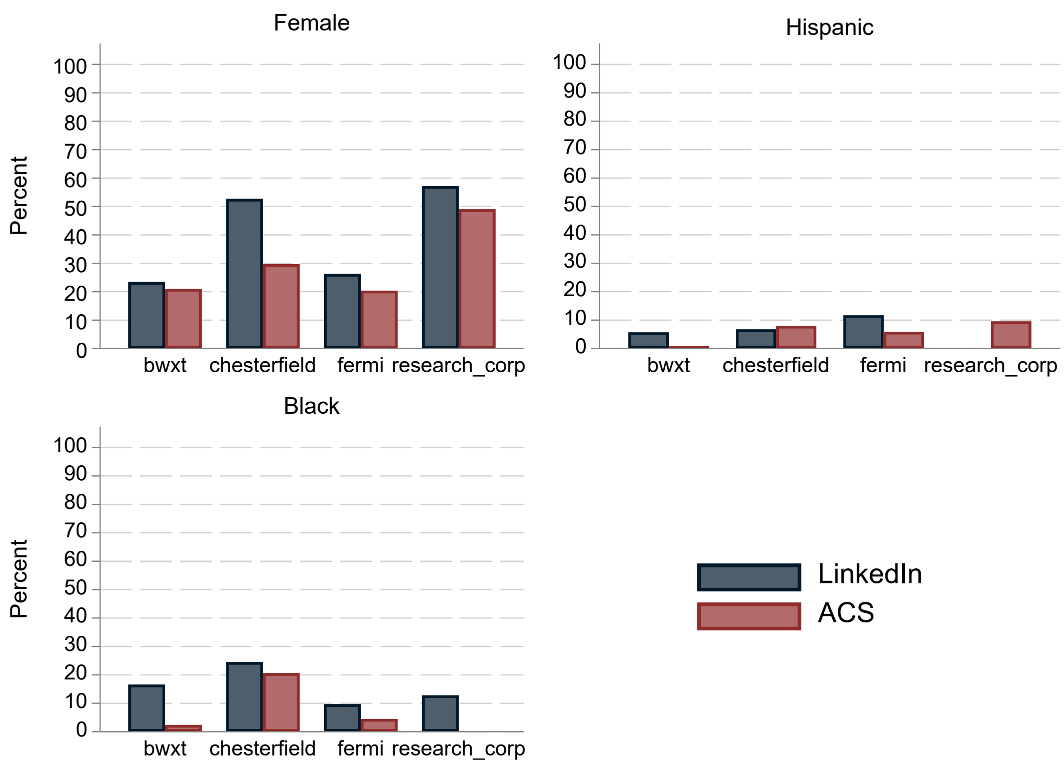


Figure 3: Validation Estimates for Percentages Black, Hispanic, and Female in *LinkedIn* and American Community Survey Data

A. All workers in industry



B. Professional/technical/managerial workers in industry



Notes: See notes to Table 6.

Table 1: Fortune 1-10 Companies (by Revenue) Comparing 10k and *LinkedIn* Employment

Company	**10k employment**	***LinkedIn* webpage employment**	***LinkedIn* database *(Proxycurl)* employment (current)**
	(1)	(2)	(3)
Walmart	2.2 million	389,386	94,192
Amazon	1.54 million	841,260	182,960
Apple	132,000	289,924	97,927
CVS Health	300,000	115,472	175
UnitedHealth Group	400,000	167,345	39,871
Exxon Mobil	63,000	57,735	20,939
Berkshire Hathaway	372,000 (many subsidiaries)	8,198	990
Alphabet	190,000	280,107 (Google)	124,743
McKesson	68,000	21,260	10,370
AmerisourceBergen	44,000	19,566	4,556

Source: https://www.zyxware.com/articles/4344/list-of-fortune-500-companies-and-their-websites and 10k reports. *LinkedIn* data are from July 2023.

Note: *Proxycurl* database employment is for those who link to company URL in their LinkedIn profile.

Table 2: Companies Identified using National Establishment Time Series (NETS) Data as having Large Share of Total Industry Employment in Place-of-Work PUMA, Large Share of Total Company's Employment in Place-of-Work PUMA, Good Representation on *LinkedIn*, and of Moderate Size for Initial Data Extraction

NAICS	Company	*LinkedIn* webpage empl.	*LinkedIn* webpage employment in nearby geographic "area"	% of POWPUMA industry empl. at the firm	% of firm empl. in industry
(1)	(2)	(3)	(4)	(5)	(6)
5417	Research Corporation of the University of Hawaii	286	229	82	100
5413	BWXT Technical Svcs Group Inc	1928	722	81	100
5415	Chesterfield County	2296	2022	73	99
5417	Fermi Research Alliance LLC	2069	1771	96	100
5415	Cognizant Tech Sltions US Corp	317047	34429 (US)	91	100
5417	Corning Research & Dev Corp	19106	5008	92	100
5417	Charles River Labs Intl Inc	14288	1619	86	100
5413	Tungland Corporation	215	172	76	100
5417	GlaxoSmithKline LLC	97000	5000	70	100

Note: *LinkedIn* data are from August 2023. NETS data are from 2019 (covering 2018).

Table 3: *LinkedIn* Worker Profiles Extracted

Company	**Total**	**U.S.**
BWXT Technical Svcs Group Inc	2,187	1,705
Chesterfield County	2,104	2,093
Fermi Research Alliance LLC	4,215	3,738
GlaxoSmithKline LLC	26,636	10,761
Meta	61,265	45,274
Research Corporation of the University of Hawaii	66	65
SpaceX	15,807	15,003
Total	112,280	78,639

Table 4: Missing Pictures or No Face Detected

Company	Total	Classified	Missing Picture (% of Classified)	No Face Detected (% of Classified)
BWXT Technical Svcs Group Inc	1753	1319	29.4	11.9
Chesterfield County	1249	952	27.8	12.1
Fermi Research Alliance LLC	3198	2486	26.2	9.2
GlaxoSmithKline LLC	22985	17106	33.2	3.4
Meta	54256	41855	28.7	3.1
Research Corporation of the University of Hawaii	60	48	22.9	9.1
SpaceX	13150	9652	32.6	11.0
Total	96651	73418	30.2	4.7

Table 5: Classification Methods Used and Resulting Probabilities Black, Hispanic, or Female

	Picture				Name				Final			
	N	25th percentile	Median	75th percentile	N	25th percentile	Median	75th percentile	N	25th percentile	Median	75th percentile
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
Black vs. non-black	2,648	62.7%	97.9%	100.0%	11,117	51.0%	60.8%	74.7%	6,289	53.8%	71.3%	95.8%
Hispanic vs. non-Hispanic	4,966	31.0%	36.2%	43.0%	7,964	61.5%	81.7%	93.4%	9,979	49.3%	72.9%	91.2%
Female vs. male	15,017	96.9%	99.8%	100.0%	24,851	99.5%	99.7%	99.9%	26,098	99.4%	99.7%	99.9%

Note: “final” column is based on the algorithm for classification described in the text. Probabilities are expressed in percentage terms.

Table 6: Validation Estimates for Percentages Black, Hispanic, and Female in *LinkedIn* and American Community Survey Data

Company	Data source	*LinkedIn* employment in area (N)	*LinkedIn* database public profiles (N)	ACS (N)	% black	% Hispanic	% female
Research Corporation of the University of Hawaii	*LinkedIn*	229	15		12.8	0	57.1
	ACS			67	0	8.5	45.5
	ACS, Prof/Tech Occ's			58	0	9.5	49.0
BWXT Technical Svcs Group Inc	*LinkedIn*	722	390		16.6	5.6	23.4
	ACS			116	9.8	0.7	18.4
	ACS, Prof/Tech Occ's			95	2.4	0.9	21.0
Chesterfield County	*LinkedIn*	2,022	796		24.5	6.7	52.7
	ACS			227	20.8	7.7	30.2
	ACS, Prof/Tech/Mgr Occ's			217	20.6	8.0	29.7
Fermi Research Alliance LLC	*LinkedIn*	1,771	684		9.7	11.6	26.3
	ACS			174	4.0	5.8	18.5
	ACS, Prof/Tech/Mgr Occ's			151	4.5	5.9	20.4

Sources/notes: *LinkedIn* and ACS data, 2012-2021, same POWPUMA and 4-digit NAICS. The *LinkedIn* data in this table are restricted to the relevant areas to correspond to the ACS POWPUMA, as described in the text. We use ACS weights, and for each company also reweight by year corresponding to the representation by year in the *LinkedIn* data. Professional, technical, and managerial occupations excluded, based on 2018 occupation codes: Community and Social Services Occupations 2001-2060; Service Occupations 3601-4655; all occupation 6005 and higher (Farming, Fishing, and Forestry; Construction Trades; Installation, Maintenance, and Repair; Production; Transportation and Material Moving; Military). See: https://www2.census.gov/programs-surveys/cps/methodology/Occupation%20Codes.pdf.

Table 7: Validation Against Other Sources of Data on Workforce Composition

Company	Source	Statistics reported	Comparable *LinkedIn* data, for corresponding years
Research Corporation of the University of Hawaii	Nothing from company		
BWX Technologies, Inc.	Nothing from company		
Chesterfield County of Virginia	EEO Utilization Report, 2020	% black: 18.9 % Hispanic: 3.7 % female: 47.7	% black: 17.2 % Hispanic: 11.2 % female: 60.3
Fermi Research Alliance, LLC	Fermilab webpage, 2023	% black: 5.6 % Hispanic: 9.3 % female: 28.1	% black: 0 % Hispanic: 0 % female: 33 Note: based on only 4 *LinkedIn* observations in 2023.
GlaxoSmithKline (GSK)	GSK website	40% of senior roles were held by women, up from 38% in 2020; 50% of manager roles held by women 27.1% of senior leaders in the U.S. were "ethnically diverse" in 2021, up from 23.2% in 2020	Cannot compare because no way to reliably match roles, and "ethnically diverse" is vague.
SpaceX	Nothing from company		
Meta	2022 Annual Diversity Report	% black 2014: 2 2015: 2 2016: 2 2017: 3 2018: 3.5 2019: 3.8 2020: 3.9 2021: 4.4 2022: 4.9 Hispanic 2014: 4 2015: 4 2016: 4 2017: 5 2018: 4.9 2019: 5.2 2020: 6.3 2021: 6.5 2022: 6.7	% black 2014: 4.0 2015: 4.8 2016: 5.2 2017: 5.7 2018: 6.2 2019: 6.3 2020: 6.5 2021: 7.1 2022: 7.1 Hispanic 2014: 12.6 2015: 13.3 2016: 13.4 2017: 13.2 2018: 13.3 2019: 13.4 2020: 13.8 2021: 14.4 2022: 14.2
	2021 EEO-1 Report	Total/Professionals % black: 4.6/4.4 % Hispanic: 6.7/6.4 % female: 36.2/34.6	Total % black: 7.1 % Hispanic: 14.4 % female: 41.9

Sources/notes: **Chesterfield**. https://www.chesterfield.gov/DocumentCenter/View/446/EEOP-DOJ-Utilization-Report-PDF. Numbers reported for six categories of jobs: Officials/Administrators; Professionals; Technicians; Protective Services (Sworn); Protective Services (Unsworn); and Administrative Support. We have assumed these include the entire workforce. **Fermi**. https://www.fnal.gov/pub/about/demographics/. Numbers reported for six categories of jobs: Technical; Scientists; Postdocs; Mission Support; Engineers, Computing. We have assumed these include the entire workforce. Some numbers are reported as < 5 but < 0. We assume values of 2. **GSK**: https://us.gsk.com/en-us/responsibility/diversity-equity-and-inclusion/#inside-gsk. **Meta**: https://about.fb.com/wp-content/uploads/2022/07/Meta_Diversity-Data-Summary-Report_2022.pdf. % female is also reported by year, but only globally. % by ethnic group is also reported for Tech, Non-Tech, and Leadership. 2021 EEO-1 Report also reports numbers for: Executive/Senior Officials & Managers; First/Mid Officials & Managers; Professionals; Technicians; Sales Workers; Administrative Support; Craft Workers; Operatives; Laborers & Helpers; and Service Workers. Professionals are the vast majority.

Table 8: *LinkedIn* Demographic Composition Estimates for Each Company

Company	N	% Black	% Hispanic	% Female
Research Corporation of the University of Hawaii	54	5.7%	7.5%	47.1%
BWXT Technical Svcs Group Inc.	1,325	11.1%	6.1%	22.8%
Chesterfield County	1,050	23.5%	7.5%	53.0%
Fermi Research Alliance LLC	2,016	8.4%	15.4%	31.7%
GlaxoSmithKline (GSK)	5,886	10.1%	10.6%	46.4%
SpaceX	13,315	7.6%	17.9%	17.7%
Meta	44,034	6.5%	12.5%	38.2%

Notes: This table reports the demographic composition for each company, for the years 2012-2023. Observations are at the employee-firm level, and proportions are calculated based on the number of classified individuals in that company. In contrast to Table 6, this table does not restrict to the relevant *LinkedIn* area to correspond to the ACS POWPUMA.

Table 9: Separations at Meta and SpaceX in *LinkedIn* Data

	Workforce	Separations	Separations/Workforce
Meta			
2012	393	11	2.8
2013	666	20	3.0
2014	1,129	36	3.2
2015	1,776	56	3.2
2016	2,902	106	3.7
2017	5,096	189	3.7
2018	8,570	358	4.2
2019	12,837	611	4.8
2020	18,806	737	3.9
2021	29,703	2,510	8.5
2022	39,632	9,316	23.5
2023	30,930	5,096	16.5
Total	152,440	19,046	12.5
SpaceX			
2012	2,018	212	10.5
2013	2,791	393	14.1
2014	3,510	526	15.0
2015	4,358	668	15.3
2016	5,004	818	16.3
2017	5,691	796	14.0
2018	6,096	943	15.5
2019	5,959	1,164	19.5
2020	6,201	815	13.1
2021	6,856	1043	15.2
2022	6,347	984	15.5
2023	5,558	196	3.5
Total	60,389	8,558	14.2

Table 10: Meta and SpaceX Mass Layoff Differentials by Race, Ethnicity, and Gender

	Meta	**Meta**	**SpaceX**	**SpaceX**
Without education and tenure controls				
Black	0.0133	0.0037	0.0039	0.0047
	(0.0035)	(0.0038)	(0.0066)	(0.0069)
Hispanic	0.0159	0.0016	-0.0292	-0.0325
	(0.0025)	(0.0026)	(0.0033)	(0.0035)
Female	0.0134	-0.0035	0.0159	0.0138
	(0.0017)	(0.0017)	(0.0036)	(0.0039)
Black x mass layoff years		0.0140		-0.0091
		(0.0064)		(0.0184)
Hispanic x mass layoff years		0.0218		0.0322
		(0.0047)		(0.0132)
Female x mass layoff years		0.0257		0.0222
		(0.0031)		(0.0154)
Year dummy variables included	Yes	Yes	Yes	Yes
Mean separation rate for non-black/non-Hispanic males, all years	0.1177		0.1464	
Mean separation rate for non-black/non-Hispanic males, mass layoff years		0.1587		0.1946
Mean separation rate for non-black/non-Hispanic males, non-mass layoff years		0.0397		0.1411
N	141,045	141,045	58,947	58,947
With education and tenure controls				
Black	0.0146	0.0016	0.0012	-0.0002
	(0.0038)	(0.0041)	(0.0069)	(0.0075)
Hispanic	0.0147	0.0008	-0.0169	-0.0204
	(0.0028)	(0.0029)	(0.0042)	(0.0046)
Female	0.0125	-0.0050	0.0088	0.0049
	(0.0018)	(0.0018)	(0.0043)	(0.0048)
Black x mass layoff years		0.0188		0.0146
		(0.0069)		(0.0264)
Hispanic x mass layoff years		0.0209		0.0352
		(0.0051)		(0.0181)
Female x mass layoff years		0.0266		0.0412
		(0.0033)		(0.0185)
Year dummy variables included	Yes	Yes	Yes	Yes
Mean separation rate for non-black/non-Hispanic males, all years	0.1185		0.1656	
Mean separation rate for non-black/non-Hispanic males, mass layoff years		0.1586		0.2082
Mean separation rate for non-black/non-Hispanic males, non-mass layoff years		0.0402		0.1609
N	126,460	126,460	40,902	40,902

Notes: Observations are employee-years, and the outcome is whether a separation occurred. Estimates are from linear probability model. Standard errors clustered at the individual level are in parentheses. "Mass layoff years" are 2021-2023 for Meta and 2019 for SpaceX. Education controls in lower panel include dummy variables for high school degree, associate's degree (omitted), bachelor's degree, master's degree, or doctoral degree.

Table 11: Meta and SpaceX Mass Layoff Differentials by Race, Ethnicity, and Gender, Intersectional Effects

	Meta	SpaceX
Black	0.0033	-0.0044
	(0.0054)	(0.0074)
Hispanic	0.0066	-0.0183
	(0.0040)	(0.0050)
Female	-0.0031	0.0051
	(0.0020)	(0.0052)
Female x black	-0.0046	0.0261
	(0.0082)	(0.0258)
Female x Hispanic	-0.0133	-0.0114
	(0.0058)	(0.0118)
Black x mass layoff years	0.0256	-0.0081
	(0.0090)	(0.0277)
Hispanic x mass layoff years	0.0144	0.0215
	(0.0067)	(0.0195)
Female x mass layoff years	0.0258	0.0200
	(0.0037)	(0.0207)
Female x black x mass layoff years	-0.0168	0.1662
	(0.0141)	(0.0814)
Female x Hispanic x mass layoff years	0.0144	0.0750
	(0.0102)	(0.0510)
Year dummy variables included	Yes	Yes
Tenure and education controls included	Yes	Yes
Mean separation rate for non-black/non-Hispanic males, all years	0.1185	0.1656
Mean separation rate for non-black/non-Hispanic males, mass layoff years	0.1586	0.2082
Mean separation rate for non-black/non-Hispanic males, non-mass layoff years	0.0402	0.1609
N	126460	40902

Notes: Observations are employee-years, and the outcome is whether a separation occurred. Estimates are from linear probability model. Standard errors clustered at the individual level are in parentheses. “Mass layoff years” are 2021-2023 for Meta and 2019 for SpaceX. Education controls in lower panel include dummy variables for high school degree, associate’s degree (omitted), bachelor’s degree, master’s degree, or doctoral degree. Hispanic females make up 5.5 percent of observations (N=6,998) in the regression sample for Meta and 2.9 percent of observations (N=1,173) in the regression sample for SpaceX. Black females make up 2.4 percent of observations (N=3,045) in the regression sample for Meta and 1.08 percent of observations (N=442) in the regression sample for SpaceX.

Table 12: Meta and SpaceX Workforce and Hires Demographic Composition

	% black in workforce	% black in hires	% Hispanic in workforce	% Hispanic in hires	% female in workforce	% female in hires
Meta						
2012	4.8	4.6	7.9	8.6	35.6	33.9
2013	4.7	4.2	8.7	10.2	35.1	34.2
2014	4.7	4.6	9.7	11.4	33.2	30.6
2015	4.4	4.2	11.7	14.6	33.9	34.7
2016	4.7	5.1	12.4	13.5	34.8	35.3
2017	5.2	5.8	12.3	12.4	36.6	38.3
2018	5.6	6.2	12.2	12.1	36.9	37.3
2019	5.7	5.8	12.2	12.4	36.8	36.3
2020	6.0	6.6	12.2	12.1	36.2	35.6
2021	6.5	7.3	12.5	12.9	37.2	38.7
2022	6.3	6.0	11.9	11.0	36.1	34.6
2023	6.2	6.8	11.5	12.3	35.3	31.9
SpaceX						
2012	10.0	9.5	18.4	19.4	15.6	14.9
2013	9.4	8.0	19.2	18.6	16.3	17.5
2014	9.1	8.6	19.4	17.4	16.2	17.6
2015	8.9	8.0	18.8	15.8	16.0	16.6
2016	8.2	6.8	18.6	15.8	15.9	16.2
2017	8.5	9.2	19.2	18.1	15.6	16.5
2018	8.0	6.3	19.1	16.0	15.5	17.3
2019	7.8	6.8	19.8	18.9	15.1	14.6
2020	7.5	5.5	19.8	19.7	15.6	19.6
2021	7.5	6.4	19.4	16.0	16.2	20.6
2022	7.5	4.1	19.8	18.6	16.1	23.0
2023	7.5	3.1	20.8	16.5	15.8	19.1

Table 13: Most Common Job Title Changes at Each Company

BWXT Technical Svcs Group Inc.		Chesterfield County		Fermi Research Alliance LLC		GlaxoSmithKline LLC	
Job title change	#	Job title change	#	Job title change	#	Job title change	#
PRODUCT ENGINEER ---> DESIGN ENGINEER	2	BUSINESS TAX ASSESSMENT SPECIALIST ---> TAX COMPLIANCE AUDITOR	3	ASSOCIATE SCIENTIST ---> SCIENTIST	11	PRINCIPAL SCIENTIST ---> INVESTIGATOR	20
VICE PRESIDENT AND CHIEF INVESTOR RELATIONS OFFICER ---> VICE PRESIDENT, CONTRACTS & PROCUREMENT	2	BATTALION CHIEF ---> DIRECTOR OF COMMUNITY RELATIONS	2	SCIENTIST ---> SENIOR SCIENTIST	7	SCIENTIST ---> SENIOR SCIENTIST	17
QUALITY ASSURANCE ENGINEER ---> QUALITY CONTROL ENGINEER	2	PLANNING TECHNICIAN ---> PLANNER	2	RESEARCH ASSOCIATE ---> ASSOCIATE SCIENTIST	6	ASSOCIATE SCIENTIST ---> SENIOR SCIENTIST	14
SOFTWARE PROGRAMMER ---> COGNIZANT AREA ENGINEER	2	GIS ANALYST ---> SENIOR GIS ANALYST	2	WILSON FELLOW ---> SCIENTIST	4	ASSOCIATE SCIENTIST ---> SCIENTIST	9
LEAD AUDITOR ---> SOFTWARE PROGRAMMER	2	LIEUTENANT ---> CAPTAIN	2	ASSOCIATE SCIENTIST ---> SCIENTIST I	4	INVESTIGATOR ---> SCIENTIFIC LEADER	9
DESIGN ENGINEER ---> PROJECT ENGINEER	2	CHIEF DEPUTY TREASURER ---> TREASURER	2	ACCELERATOR OPERATOR I ---> ACCELERATOR OPERATOR II	4	SENIOR SCIENTIST ---> INVESTIGATOR	8
VICE PRESIDENT, INVESTOR RELATIONS AND CORPORATE PROCUREMENT ---> VICE PRESIDENT AND CHIEF INVESTOR RELATIONS OFFICER	2	SENIOR CONTRACT OFFICER ---> PROCUREMENT MANAGER II	2	ACCELERATOR OPERATOR ---> ENGINEERING PHYSICIST	3	SENIOR SALES REPRESENTATIVE ---> EXECUTIVE SALES REPRESENTATIVE	7
POLITICAL AFFAIRS MANAGER ---> SENIOR MANAGER, GOVERNMENT RELATIONS	2	ADMINISTRATIVE MANAGER - PURCHASING ---> SENIOR CONTRACT OFFICER	2	CHIEF INFORMATION OFFICER (CIO) ---> PROJECT SCIENTIST FOR LONG BASELINE NEUTRINO EXPERIMENT	2	PHARMACEUTICAL SALES REPRESENTATIVE ---> SENIOR PHARMACEUTICAL SALES REPRESENTATIVE	7
COGNIZANT AREA ENGINEER ---> SMARTFORMS SME	2	EMERGENCY COMMUNICATIONS OFFICER ---> ASSISTANT SHIFT SUPERVISOR	2	DISTRIBUTED COMPUTING SERVICES OPERATIONS DEPUTY DEPARTMENT HEAD ---> HPC/LQCD SITE CO-MANAGER	2	SENIOR COUNSEL ---> ASSISTANT GENERAL COUNSEL	7
NETWORK ENGINEER ---> SR. VOIP ENGINEER	2	PLANNER ---> SENIOR PLANNER	2	SCIENTIST I ---> SENIOR SCIENTIST	2	ASSOCIATE BRAND MANAGER ---> BRAND MANAGER	7
OPERATIONS ENGINEER ---> DEVELOPMENT ENGINEER	2	MICROCOMPUTER ANALYST ---> SENIOR MICROCOMPUTER ANALYST	2	JEFF METCALF INTERN ---> VOLUNTEER RESEARCH ASSISTANT	2	COUNSEL ---> SENIOR COUNSEL	5
QUALITY ASSURANCE ---> SHIFT SUPERVISOR	1	ASSISTANT DIRECTOR, HUMAN RESOURCE MANAGEMENT ---> DIRECTOR AND CHIEF LEARNING OFFICER, LEARNING & PERFORMANCE CENTER	1	CDF SPOKESPERSON ---> DISTINGUISHED SCIENTIST	2	INVESTIGATOR ---> SENIOR SCIENTIFIC INVESTIGATOR	4

MANAGER, SUPPLIER OVERSIGHT & SECURITY ---> CORPORATE SECURITY PROGRAM MANAGER	1	SENIOR FAMILY SERVICES SPECIALIST ---> LEAD FAMILY SERVICES SPECIALIST	1	SYSTEM SERVICES SPECIALST III ---> DEPUTY GROUP LEADER (SYSTEM SERVICES SPECIALIST III)	2	SENIOR SCIENTIST ---> PRINCIPAL SCIENTIST	4
DIMENSIONAL INSPECTION MANAGER ---> ASSEMBLY ENGINEERING MANAGER	1	STAFF AUDITOR ---> SENIOR AUDITOR	1	OPERATIONS SPECIALIST ---> ENGINEERING PHYSICIST	2	SALES REPRESENTATIVE ---> SENIOR SALES REPRESENTATIVE	4
MATERIALS ENGINEERING TEAM LEAD ---> MANAGER OF MATERIALS MANAGEMENT	1	IT SPECIALIST III ---> SOLUTIONS ARCHITECT	1	RESEARCH ASSOCIATE ---> COMPUTER PROFESSIONAL	2	PATENT COUNSEL ---> SENIOR PATENT COUNSEL	4
INTERNAL AUDITOR ---> BUSINESS ANALYST	1	CHIEF OFFICER OF ELECTIONS ---> EASTERN VIRGINIA MARKET UNIT FLEET ADMINISTRATIVE ASSISTANT	1	ACCELERATOR OPERATOR II ---> SENIOR ACCELERATOR OPERATOR	2	INVESTIGATOR ---> PRINCIPAL INVESTIGATOR	4
MANAGER, CORPORATE DEVELOPMENT ---> DIRECTOR, CORPORATE DEVELOPMENT	1	SENIOR COMPUTER ANALYST ---> LEAD COMPUTER ANALYST	1	DEPUTY DIRECTOR FOR RESEARCH ---> HEAD OF FERMILAB QUANTUM INSTITUTE	2	ASSISTANT SCIENTIST ---> ASSOCIATE SCIENTIST	3
DIMENSIONAL INSPECTOR ---> TECHNOLOGIST	1	CODE COMPLIANCE SPECIALIST ---> PLANNER	1	VOLUNTEER RESEARCH ASSISTANT ---> ON-CALL EMPLOYEE / RESEARCH ASSISTANT	2	SCIENTIST ---> PRINCIPAL SCIENTIST	3
IT PROJECT MANAGER ---> CORPORATE PROCUREMENT ANALYST	1	SR. AUTOMATION TECHNICIAN ---> AUTOMATION SPECIALIST	1	DEPUTY GROUP LEADER (SYSTEM SERVICES SPECIALIST III) ---> DISTRIBUTED COMPUTING SERVICES OPERATIONS DEPUTY DEPARTMENT HEAD	2	BRAND MANAGER ---> SENIOR BRAND MANAGER	3
QUALITY ASSURANCE ENGINEERING SUPERVISOR ---> QUALITY ASSURANCE MANAGER	1	CAPTAIN ---> BATTALION CHIEF, EMERGENCY OPERATIONS DIVISION	1	PHYSICIST ---> DEPUTY DIRECTOR FOR RESEARCH	2	SENIOR PHARMACEUTICAL SALES REPRESENTATIVE ---> EXECUTIVE PHARMACEUTICAL SALES REPRESENTATIVE	3

Meta		**Research Corporation of the University of Hawaii**		**SpaceX**	
Job title change	#	Job title change	#	Job title change	#
SOFTWARE ENGINEER ---> SENIOR SOFTWARE ENGINEER	210	HUMAN RESOURCES ASSOCIATE I/II ---> HUMAN RESOURCES INFORMATION SYSTEMS SPECIALIST I/II/III	1	SOFTWARE ENGINEER ---> SOFTWARE ENGINEER II	50
SENIOR SOFTWARE ENGINEER ---> STAFF SOFTWARE ENGINEER	88	HUMAN RESOURCES INFORMATION SYSTEMS SPECIALIST I/II/III ---> HRIS/IT SYSTEMS ADMINISTRATOR	1	MANUFACTURING ENGINEER ---> MANUFACTURING ENGINEER II	26
SOFTWARE ENGINEER ---> ENGINEERING MANAGER	84	PROGRAM MANAGER @ THE VANGUARD MAUI HIGH PERFORMANCE COMPUTING CENTER	1	MANUFACTURING ENGINEER ---> LEAD MANUFACTURING ENGINEER	20

		(MHPCC) ---> EXECUTIVE DIRECTOR FOR THE VANGUARD MAUI HIGH PERFORMANCE COMPUTING CENTER (MHPCC)			
SOFTWARE ENGINEER ---> SOFTWARE ENGINEERING MANAGER	68	CRAZY ANT STRIKE TEAM CREW LEAD ---> CRAZY ANT STRIKE TEAM FIELD OPERATIONS COORDINATOR	1	SOFTWARE ENGINEER II ---> SENIOR SOFTWARE ENGINEER	15
DATA SCIENTIST ---> DATA SCIENCE MANAGER	53	HUMAN RESOURCES SUPPORT ASSOCIATE I ---> HUMAN RESOURCES RECRUITMENT ASSOCIATE I	1	SOFTWARE ENGINEER ---> LEAD SOFTWARE ENGINEER	14
RESEARCH SCIENTIST ---> SENIOR RESEARCH SCIENTIST	49	3D PLANETARIUM OPERATOR ---> 3D PLANETARIUM OPERATING INTERN	1	SENIOR SOFTWARE ENGINEER ---> LEAD SOFTWARE ENGINEER	13
DATA SCIENTIST ---> SENIOR DATA SCIENTIST	43			PROPULSION ENGINEER ---> PROPULSION ENGINEER II	11
PRODUCT DESIGNER ---> PRODUCT DESIGN MANAGER	31			STRUCTURES ENGINEER ---> STRUCTURES ENGINEER II	11
CLIENT SOLUTIONS MANAGER ---> CLIENT PARTNER	25			INTEGRATION AND TEST ENGINEER ---> INTEGRATION AND TEST ENGINEER II	11
ROTATIONAL SOFTWARE ENGINEER ---> SOFTWARE ENGINEER	24			MECHANICAL ENGINEER ---> SENIOR MECHANICAL ENGINEER	11
STAFF SOFTWARE ENGINEER ---> ENGINEERING MANAGER	20			SENIOR MECHANICAL ENGINEER ---> MANAGER-DETECTION INFRASTRUCTURE ENGINEERING	11
TECHNICAL SOURCER ---> TECHNICAL RECRUITER	17			SOFTWARE ENGINEER II ---> LEAD SOFTWARE ENGINEER	10
SOFTWARE ENGINEER ---> STAFF SOFTWARE ENGINEER	16			SOFTWARE ENGINEER ---> SENIOR SOFTWARE ENGINEER	9
SENIOR DATA SCIENTIST ---> STAFF DATA SCIENTIST	16			MANAGER-DETECTION INFRASTRUCTURE ENGINEERING ---> SR.MANAGER	9
DATA ENGINEER ---> DATA ENGINEERING MANAGER	16			BUILD RELIABILITY ENGINEER ---> BUILD RELIABILITY ENGINEER II	8
CONTENT DESIGNER ---> CONTENT DESIGN MANAGER	14			ASSOCIATE BUYER ---> BUYER	8
DATA ENGINEER ---> SENIOR DATA ENGINEER	12			STRUCTURES ENGINEER ---> LEAD STRUCTURES ENGINEER	8
UX RESEARCHER ---> SENIOR UX RESEARCHER	12			PROPULSION TECHNICIAN ---> LEAD PROPULSION TECHNICIAN	8
ENGINEERING MANAGER ---> SENIOR ENGINEERING MANAGER	12			MATERIAL PLANNER ---> MATERIAL PLANNER II	7
MACHINE LEARNING ENGINEER ---> SENIOR MACHINE LEARNING ENGINEER	12			LEAD PROPULSION ENGINEER ---> MANAGER, PROPULSION ENGINEERING	6

Table 14: Meta, SpaceX, GSK, and Fermi Promotion Rate Differentials by Race, Ethnicity, and Gender, 2012-2023

	Meta	**SpaceX**	**GSK**	**Fermi**
Without education and tenure controls				
Black	0.0052 (0.0032)	-0.0254 (0.0047)	-0.0242 (0.0089)	-0.0115 (0.0055)
Hispanic	0.0090 (0.0025)	-0.0278 (0.0032)	0.0180 (0.0103)	-0.0090 (0.0063)
Female	0.0198 (0.0017)	0.0195 (0.0043)	0.0067 (0.0057)	0.0057 (0.0056)
Year dummy variables included	Yes	Yes	Yes	Yes
Mean promotion rate for non-black/non-Hispanic males	0.0563	0.0704	0.1286	0.0296
N	141,045	58,947	24,612	8,984
With education and tenure controls				
Black	0.0054	-0.0270	-0.0224	-0.0082
	(0.0035)	(0.0066)	(0.0096)	(0.0083)
Hispanic	0.0103	-0.0152	0.0232	-0.0169
	(0.0027)	(0.0050)	(0.0108)	(0.0064)
Female	0.0194	0.0171	0.0077	0.0082
	(0.0018)	(0.0053)	(0.0060)	(0.0070)
Year dummy variables included	Yes	Yes	Yes	Yes
Mean promotion rate for non-black/non-Hispanic males	0.0569	0.0848	0.1324	0.0344
N	126,460	40,902	2,1907	6,855

Notes: Observations are employee-years, and the outcome is whether a promotion (job title change) occurred in the year. Estimates are from linear probability model. Standard errors clustered at the individual level are in parentheses. Education controls in lower panel include dummy variables for high school degree, associate's degree (omitted), bachelor's degree, master's degree, or doctoral degree.

Table 15: Meta and SpaceX Promotion Differentials by Race, Ethnicity, and Gender, Intersectional Effects

	Meta	SpaceX	GSK	Fermi
Black	0.0053	-0.0233	-0.0149	-0.0016
	(0.0042)	(0.0073)	(0.0139)	(0.0102)
Hispanic	0.0102	-0.0163	0.0284	-0.0116
	(0.0034)	(0.0054)	(0.0157)	(0.0075)
Female	0.0194	0.0178	0.0102	0.0126
	(0.0020)	(0.0063)	(0.0066)	(0.0086)
Female x black	-0.0791	-0.0817	-0.1264	-0.0389
	(0.0045)	(0.0096)	(0.0149)	(0.0133)
Female x Hispanic	-0.0860	-0.0851	-0.1671	-0.0429
	(0.0038)	(0.0080)	(0.0165)	(0.0110)
Year dummy variables included	Yes	Yes	Yes	Yes
Tenure and education controls included	Yes	Yes	Yes	Yes
Mean promotion rate for non-black/non-Hispanic males	0.0569	0.0848	0.1324	0.0344
N	126,460	40,902	21,907	6,855

Notes: Observations are employee-years, and the outcome is whether a promotion (job title change) occurred in the year. Estimates are from linear probability model. Standard errors clustered at the individual level are in parentheses. Education controls include dummy variables for high school degree, associate's degree (omitted), bachelor's degree, master's degree, or doctoral degree. Hispanic females make up 5.5 percent of observations (N=6,998) in the regression sample for Meta, 2.9 percent (N=1,173) for SpaceX, 5.7 percent (N=1,254) for GSK, and 4.05 percent (N=279) for Fermi. Black females make up 2.4 percent of observations (N=3,045) in the regression sample for Meta, 1.08 percent (N=442) for SpaceX, 5.4 percent (N=1,191) for GSK, and 2.5 percent (N=172) for Fermi.

Table 16: Pairs of Consecutive Job Titles (10 Randomly Selected in Each Range)

Similarity score 0.7-0.8		*Promotion based on our "rules"*
DEPUTY SHERIFF RECRUIT	DEPUTY SHERIFF	No
INTERNATIONAL MARKETING MANAGER	INTEGRATED MARKETING MANAGER | WHATSAPP	No
MANUFACTURING ENGINEER II, DRAGON PROPULSION	PROPULSION ENGINEER II	No
MEDICAL CENTER REPRESENTATIVE/EXECUTIVE SALES REP	SR EXECUTIVE REPRESENTATIVE/INTERIM SALES MANAGER	No
PHARMACEUTICAL SALES REPRESENTATIVE, RESPIRATORY	PHARMACEUTICAL SALES SPECIALIST, RESPIRATORY	No
PRODUCT DEVELOPMENT SPECIALIST	PRODUCT SPECIALIST	No
PROGRAM MANAGER, BRAND	PROGRAM MANAGER, DIGITAL	No
PROGRAM MANAGER, WHATSAPP	PROGRAM MANAGER	No
SENIOR SOFTWARE ENGINEER - BUSINESS SOLUTIONS	ENGINEERING MANAGER - BUSINESS SOLUTIONS	No
TOXICOLOGIST/MANAGER, SAFETY ASSESSMENT PROJECTS	SENIOR MANAGER, SAFETY ASSESSMENT PROJECT DEVELOPMENT	No
Similarity score 0.8-0.9		
RECOVERY ENGINEER	LEAD RECOVERY ENGINEER	Yes
AI RESEARCH SCIENTIST	SENIOR MEDICAL DIRECTOR	Yes
PRODUCT MANAGEMENT LEADERSHIP RECRUITER	SENIOR MATERIAL PLANNER	Yes
CLIENT SOLUTIONS MANAGER, RETAIL & B2B	LEAD DESIGNER & SENIOR SOFTWARE ENGINEER	Yes
LEAD SPACE LASER ENGINEER	SR MANAGER, BUSINESS PRODUCT OPERATIONS	Yes
MATERIAL PLANNER II	RECRUITING MANAGER | PRODUCT DESIGN	Yes
ADMINISTRATIVE SPECIALIST - GLOBAL BUSINESS MARKETING	EXPENSE PAY FULL STACK ENGINEER (VIA K2 PARTNERING SOLUTIONS)	No
SPECIALTY ACCOUNT DIRECTOR, REGIONAL ACCOUNTS, PAYER, POLICY AND VACCINE DIVISION	GLOBAL HEAD OF SCALED REGULATORY OPERATIONS	No
MANAGER, TECHNICAL PROGRAM MANAGEMENT, LEGAL	HEAD OF LATIN AMERICA, STRATEGIC RESPONSE POLICY	No
PAYMENTS & COMMERCE PROGRAM MANAGER, PARTNERSHIP OPERATIONS	DIRECTOR SALES AND MARKETING, DERMATOLOGY (2ND LINE LEADER)	No

Table 17: Meta, SpaceX, GSK, and Fermi Promotion Rate Differentials by Race, Ethnicity, and Gender, 2012-2023, Narrower Promotion Definitions

	Meta	SpaceX	GSK	Fermi
Black	-0.0020	-0.0075	-0.0021	-0.0032
	(0.0009)	(0.0028)	(0.0012)	(0.0009)
Hispanic	-0.0015	-0.0068	0.0018	-0.0025
	(0.0007)	(0.0020)	(0.0018)	(0.0013)
Female	-0.0000	-0.0007	-0.0004	-0.0018
	(0.0005)	(0.0021)	(0.0010)	(0.0014)
Year dummy variables included	Yes	Yes	Yes	Yes
Tenure and education controls included	Yes	Yes	Yes	Yes
Mean promotion rate for non-black/non-Hispanic males	0.0085	0.0223	0.0052	0.0039
N	126,460	40,902	21,907	6,855

Notes: Observations are employee-years, and the outcome is whether a promotion (restricted job title change) occurred in the year. Estimates are from linear probability model. Standard errors clustered at the individual level are in parentheses. Education controls include dummy variables for high school degree, associate's degree (omitted), bachelor's degree, master's degree, or doctoral degree.