Evaluating Algorithmic Systems for Privacy and Accountability

Ryan Steed

May 7, 2025

Submitted in partial fulfillment of the requirements for the degree of
Doctor of Philosophy in Machine Learning & Public Policy

Heinz College & Machine Learning Department
Carnegie Mellon University
Pittsburgh, PA

Dissertation Committee

Alessandro Acquisti *(Chair)*
Zhiwei Steven Wu
Rayid Ghani
Arvind Narayanan

Funding: This research was supported in parts by the Tata Consultancy Services (TCS) Presidential Fellowship (R.S.), the Meta Research Ph.D. Fellowship (R.S.), the National Science Foundation under Grant No. 2319919 (A.A., R.S.), the U.S. National Institute of Standards and Technology (NIST) under Grant No. 60NANB20D212 (A.C.), National Science Foundation (NSF) grant 1939606 (Z.S.W., T.L.), NSF grant 2120667 (Z.S.W., T.L.), Alfred P. Sloan Foundation grant G-2015-14111 (A.A.), MacArthur Foundation grant 22-2203-156318-TPI (A.A.), the Mozilla Foundation, Science Foundation Ireland via the ADAPT Centre of Digital Content Technology funded under the European Regional Development Fund (ERDF) through Grant #13/RC/2106_P2 (A.B.), and the University of Notre Dame IBM Tech-Ethics Lab (B.V.).

To Dana, my North Star

Acknowledgments

This thesis exists thanks to effort, insight, and wisdom from my esteemed collaborators: my co-authors, Steven Wu, Michael Wick, Diana Qing, Swetasudha Panda, Eduardo Abraham Schnadower Mustri, Terrance Liu, Ari Kobren, Aylin Caliskan, Alessandro Acquisti, and the star-studded OAT team, Briana Vecchione, Deborah Raji, Victor Ojewale, and Abeba Birhane; and the talented research assistants who contributed to work in this thesis, Donna Zhu, Annie Qian, and Xingyu Chen.

I am grateful to Steven Wu, Rayid Ghani, and Arvind Narayanan, for thoughtful questions and good advice, and to Aylin Caliskan and Rahul Simha, who first inspired me to pursue a Ph.D.

My deepest gratitude belongs to Alessandro Acquisti, my incredible research advisor and the best mentor a student could wish to have.

I have been lucky to meet many wonderful peers along the way: Zijun, Sachin, Logan, Jessica, Eduardo, and the rest of PeeX lab; Sarah, Roy, Priyanka, Miranda, Jeremy, Jayshree, danah, and my other fellow DP/census wonks; Steven and the Social AI lunch club; Zeid, Nupoor, Lingwei, Keegan, and my other classmates and comrades; and the countless colleagues and seminar attendees who have shaped my work with their feedback.

Thanks as always to the kind and supportive people at Carnegie Mellon University, especially Michelle Wirtz, Olivia Wells, James Trimbee, Tricia Straw, Diane Stidle, and Emily Marshall for answering all my emails; and thanks to Jessica Guo, Lingwei Cheng, Logan Crowl, and all Heinz Ph.D. reps for tirelessly working to support doctoral students.

I cannot express all my appreciation and love: for Dana, my designated understudy for poster sessions (and a genuine, "boot leather" inspector); for Mom, my chief editor and spelling coach; for Michael and James, my lifelong interlocutors and inner council; for Dad, my go-to consultant pro bono; and for Aunt Megan, Grandpa Tim, Aunt Tara, Uncle Todd, Kate, and the rest of my family, who dared to ask what I'm working on (or to even mention AI). This thesis was completed in spite of our cat, Hubble, who has been a great impediment to its progress.

Finally, to Mom: You will always be my biggest supporter and my first call. Thank you.

Abstract

Machine learning algorithms and other statistical techniques are widely used to make inferences with personal information, but systems built for this purpose may be detrimental to privacy and social equity. Recent research proposes techniques intended to make these inferences while preserving individuals' privacy. This thesis 1) develops approaches for evaluating the social impacts of machine learning systems and "privacy-preserving" approaches to analytics and 2) theorizes about the role of these evaluations in holding accountable the architects and operators of machine learning systems.

Part I focuses on the impacts of techniques intended to preserve *privacy* in machine learning and other analytic systems. Chapter 1 estimates the impact of differentially private public census statistics on evidence-based policy, finding that while statistical uncertainty creates inequalities in the distribution of education funding, noise injected for privacy likely has much less impact than existing data error (Steed, Liu, et al., 2022). Chapter 2 quantifies the impact of added noise on key findings from a large sample of social science studies. Chapter 3 develops a grounded theory of the adoption of privacy-preserving analytics from qualitative interviews, uncovering processes by which adopting organizations may decouple representations about privacy from the specifics of their implementation (Steed & Acquisti, 2025).

Part II explores approaches for evaluating social equity in machine learning systems and the use of evaluations as mechanisms for *accountability*. In Chapter 4, we develop a method for quantifying stereotypical associations in image embeddings and show that unsupervised image generation models automatically learn racial, gender, and intersectional biases (Steed & Caliskan, 2021). In Chapter 5, we taxonomize a dataset of artificial intelligence (AI) audit tools and interview 35 audit practitioners, finding that the tools practitioners need for AI accountability—including tools for harms discovery and advocacy—are comparatively under-resourced (Ojewale et al., 2025).

Contents

List of Figures

List of Tables

Introduction

Algorithmic systems, particularly machine learning (ML) systems, pose serious societal concerns related to privacy and social equity. They are used to make consequential decisions in finance, criminal justice, healthcare, and content moderation; but often, these systems do not work (Raji, Kumar, et al., 2022). They are used to make statistical inferences about people based on troves of personal data collected through a socioeconomic apparatus of mass surveillance (Zuboff, 2019; Cohen, 2019; FTC Staff, 2024); this arrangement perpetuates discrimination, inequality, and other social harms (Benjamin, 2020; Skinner-Thompson, 2020). Organizations have responded to privacy concerns by adopting further algorithmic techniques for protecting individual privacy while performing statistical analysis; while promising in theory, the practical impacts of these "privacy-preserving" techniques are uncertain.

Preventing and redressing the adverse impacts of algorithmic systems depends in part on ongoing, consequential, empirical evaluation. This thesis 1) develops approaches for evaluating the social impacts of machine learning systems and "privacy-preserving" approaches to analytics and 2) theorizes about the role of these evaluations in holding accountable the architects and operators of machine learning systems.

The results presented in this thesis weave together two key areas of contemporary technology policy: data privacy and "artificial intelligence" (AI) accountability. Well-established information privacy and data protection regimes—particularly in the European Union—focus on preserving the individual right to *privacy* by restricting the ways organizations can process personal data. Organizations across industry and government are pioneering deployments of differential privacy, federated learning, and other approaches to reconcile business models dependent on statistical inference with modern privacy regulations and consumer calls for privacy. But these techniques are complex in theory and implementation, and their role in privacy and data protection policy is still unsettled. Part I explores the adoption and impacts of these "privacy-preserving" systems.

Regulation of the societal impacts of AI systems specifically is less mature than regulations for data protection and information privacy. But policy attention has increased sharply, particularly in response to the popularity of products that use ML techniques to generate text and images. Recent policy proposals and enacted legislation in the United States and Europe place particular emphasis on independent *evaluation* of algorithmic systems as a mechanism for *accountability*: the ability to make consequential judgments about the performance of algorithmic systems relative to societal expectations (Birhane et al., 2024). Part II focuses on the practice of AI auditing.

Policy debates around algorithmic systems often center on hard trade-offs involved with their development and deployment—trade-offs between functionality and privacy, accuracy and fairness, innovation and safety. At a minimum, evaluations can help demarcate the frontier between these goals and provide an invaluable guide to considered, evidence-based policy-making. But a core insight of this thesis is that thorough evaluation can also bring to light deeper, systemic challenges in current practice—challenges that public policy must acknowledge and address. Debates over the use of noise infusion techniques in public statistics elide incredible certitude in data-driven policies and robustness shortcomings in scientific methods (Steed, Liu, et al., 2022; Manski, 2011). Attempts to simply delete social biases from AI text and image generators overlook stereotypes and inequalities endemic to large training datasets (Steed & Caliskan, 2021; Steed, Panda, et al., 2022). Deployments of "privacy-preserving" analytics in industry and government may fall short of advocates' hopes without careful oversight of technical details (Steed & Acquisti, 2025). Rigorous evaluations surface these deeper issues, clarify policy deliberation, and, in some cases, point the way to more robust reform.

The first part of this thesis is devoted to developing and applying methods to evaluate specific algorithmic systems.

Chapters 1 and 2 evaluate algorithmic techniques for differential privacy, a mathematical framework for limiting the sensitivity of ML and other data analysis methods to the presence or absence of any individual's information (Dinur & Nissim, 2003; Dwork et al., 2006). Chapter 1 examines the most well-known and controversial application of differential privacy to date: the Census Bureau's use of differential privacy to release statistics from the 2020 Decennial Census (Abowd et al., 2022). In 2017, federal agencies used census data to guide the distribution of over $1.5 trillion (Reamer, 2020). We evaluate how the Census Bureau's methods, if applied to other Census data products, could impact the allocation of federal education funds to school districts—over $16.5 billion in 2021. Differential privacy is typically achieved by adding statistical uncertainty, or noise, before sharing sensitive data. We find that misallocations due to noise from a differentially private mechanism occur on the margin of much larger misallocations due to existing data error, and we show that these misallocations particularly disadvantage marginalized groups (Steed, Liu, Wu, & Acquisti, 2022). We suggest policy reforms that could reduce the disparate impacts of both data error and privacy mechanisms.[1]

Chapter 2 further investigates the possible impacts of widespread use of differential privacy on the robustness of social science research, another key question of concern for researchers and policymakers (Hotz et al., 2022; Acquisti & Steed, 2023). To quantify and compare the impacts of noise injected for differential privacy, we replicated the key results of 50 empirical studies published in economics and social science journals, evaluating whether published findings replicate on noise-infused datasets. Under privacy budgets typical in industry, a non-negligible number of findings no longer support the original claims. In particular, the claims

[1] In other work, we evaluate a privacy concern deterring census respondents; we demonstrate a reconstruction attack that could identify subsidized households living in violation of occupancy guidelines, and show that differential privacy is more effective against this attack than previous disclosure avoidance methods (Steed, Qing, & Wu, 2024).

based on weaker original effect sizes are more likely to be nullified or even reversed. Again, we find that even modest amounts of *existing* data error can alter findings. Differentially private mechanisms may exacerbate these epistemic disparities—but the marginal impacts of privacy protection are smaller, especially when the magnitude of data error is large. We point towards robust estimation techniques that can better account for noise—from data privacy protections or from data error.

Chapter 4 evaluates two image generation models, SimCLR and iGPT—precursors to the popular products that are the subject of many contemporary AI policy discussions. We contribute a method for quantifying stereotypical associations in embeddings produced with image generation models and show that these embeddings encode racial, gender, and intersectional biases mirror empirical measurements of social stereotypes and statistical inequalities (Steed & Caliskan, 2021). We also demonstrate the tendency of these models to generate images that sexualize female-passing faces while placing male-passing faces in career-related attire—a tendency still exhibited by popular AI-powered photo editing apps (Heikkilä, 2022).[2]

The remainder of this thesis discusses the role of such evaluations in holding accountable the architects and operators of algorithmic systems.

Chapter 3 contributes a grounded theory of the growing adoption of privacy-preserving analytics (PPA) techniques, from differential privacy to federated learning, in industry and government. From interviews with PPA practitioners, we identify several mechanisms by which organizations—particularly organizations with large, existing surveillance businesses—may decouple representations about "privacy-preserving" algorithmic systems from the specifics of their implementation (Steed & Acquisti, 2025). Like other algorithmic systems, these techniques require continuous, consequential evaluation oversight to avoid "privacy theater"—adoption for show.

Chapter 5 explores the practice of evaluating algorithmic systems—specifically ML systems—and the landscape of tools available to its practitioners. With a taxonomy of hundreds of AI audit tools and interviews with 35 AI audit practitioners, we find that while many (often flawed) tools exist to aid with evaluation, the tools necessary to extend those evaluations into accountability—through harms discovery and advocacy, for example—are comparatively under-resourced (Ojewale, Steed, Vecchione, Birhane, & Raji, 2025). We contribute a set of extensive policy recommendations for building more robust AI accountability infrastructure (Raji, Vecchione, Birhane, Steed, & Ojewale, 2023b, 2023a, 2024).[3]

[2] In other work, we show that similar biases encoded in pre-trained large language models may persist after fine-tuning, even after bias corrections to the fine-tuning dataset (Steed, Panda, Kobren, & Wick, 2022).

[3] In other work, we map out the stakeholders and institutions involved in AI auditing across government, consulting, civil society, journalism, & academia to identify institutional antecedents of effective accountability (Birhane, Steed, Ojewale, Vecchione, & Raji, 2024).

Part I

Evaluating systems for privacy-preserving analytics

Chapter 1

Estimating Policy Impacts of Statistical Uncertainty and Privacy

This chapter is reproduced from:

Steed, R., Liu, T., Wu, Z. S., & Acquisti, A. (2022). Policy impacts of statistical uncertainty and privacy. *Science*, *377*(6609), 928–931. https://doi.org/10.1126/science.abq4481

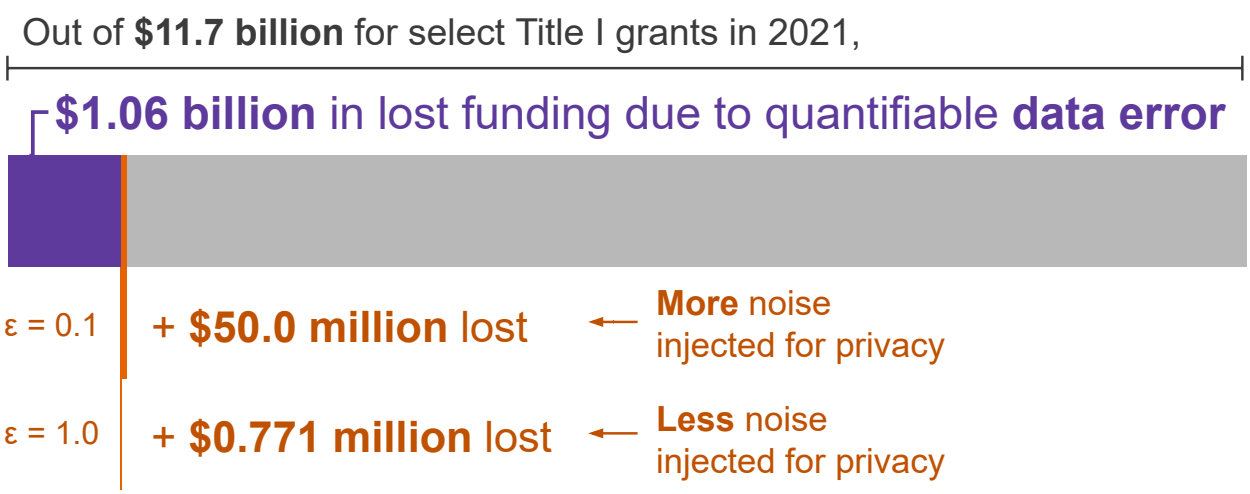


Figure 1.1: Expected lost entitlements due to data error and privacy protections. Out of a total of $11.7 billion in Title I basic, concentration, and targeted grants in 2021, we show expected sum of lost entitlements over 1000 trials due to quantifiable data error alone ("data deviations"; blue), and with the addition of noise injected for privacy ("privacy deviations"; blue plus red). Noise is injected with the ϵ-differentially private Laplace mechanism. The margins of error at 99% confidence are too small to be depicted—less than $4 million for all three bars. Note that for $\epsilon = 1.0$, the additional funding loss due to privacy deviations falls within the 90% margin of error for the impact of data deviations alone.

1.1 Introduction

Differential privacy (Dwork et al., 2006) is an increasingly popular tool for preserving individuals' privacy by adding statistical uncertainty when sharing sensitive data. Its introduction into US Census Bureau operations (Abowd et al., 2022), however, has been controversial. Scholars, politicians, and activists have raised concerns about the integrity of census-guided democratic processes, from redistricting to voting rights. The debate raises important issues, yet most analyses of trade-offs around differential privacy overlook deeper uncertainties in census data (boyd & Sarathy, 2022). To illustrate, we examine how education policies that leverage census data misallocate funding because of statistical uncertainty, comparing the impacts of quantified data error and of a possible differentially private mechanism. We find that misallocations due to our differentially private mechanism occur on the margin of much larger misallocations due to existing data error that particularly disadvantage marginalized groups. But, we also find that policy reforms can reduce the disparate impacts of both data error and privacy mechanisms.

Differential privacy is the cornerstone of the Census Bureau's updated disclosure avoidance system (DAS) (Abowd et al., 2022). Designed to rigorously prevent reconstruction, reidentification, and other attacks on personal data, differential privacy formally guarantees that published statistics are not sensitive to the presence or absence of any individual's data by injecting transparently structured statistical uncertainty (noise) (Dwork et al., 2006). But even before differential privacy is applied, estimates from the decennial census, surveys such as the American Community Survey (ACS), and other Census Bureau data products used

for critical policy decisions already contain many kinds of statistical uncertainty, including sampling, measurement, and other kinds of nonsampling error (US Census Bureau, 2020). Some amount of those errors is quantified, but numerous forms of error are not (Groves & Lyberg, 2010), including some nonresponses, misreporting, collection errors, and even hidden distortions introduced by previous disclosure avoidance measures such as data swapping (Christ et al., 2022). If quantified and unquantified errors alike are not acknowledged and accounted for, policies that rely on census data sources may not distribute the impacts of uncertainty equally.

In 2021, the US federal government appropriated over $16.5 billion in Title I funds (including several special grants not analyzed here) to distribute to over 13,000 local education agencies (LEAs)—typically school districts—using a formula that takes as input census estimates of the number of children and children in poverty. School districts qualify for Title I grants on the basis of the number or share of children in poverty (Snyder et al., 2019). However, the formula does not account for deviations in the poverty estimates that could cause misallocations—cases where the funding amount allocated to a school district differs from its entitlement in an imaginary (boyd & Sarathy, 2022), noise-free world.

Researchers have recognized Title I as an important case study of policy-relevant privacy-utility trade-offs (Abowd et al., 2019), including misallocation after noise injection for differential privacy (Pujol et al., 2020). We extend this work by comparing the policy impacts of noise injected for privacy to the impacts of existing statistical uncertainty, contextualizing preliminary error analyses by Census Bureau scientists (Abowd et al., 2022). Our results empirically investigate analytical predictions and proposals from previous work on statistical estimation and federal funding formulas (Zaslavsky & Schirm, 2002; National Research Council, 2000, 2003).

We focus specifically on the way Title I implicitly concentrates the negative impacts of statistical uncertainty on marginalized groups. Weakening privacy protection will do little to help the most vulnerable—for these communities, participating in a census survey can be especially risky, despite the benefits of voting rights protection and school funding. Historically, abuse of census data facilitated internment of Japanese Americans and other injustices (boyd & Sarathy, 2022). Today, a parent with a restrictive lease may not mention their children to a census worker because they fear being kicked out by their landlord if their responses are reidentified (Cork et al., 2020).

1.2 Simulating Noise in Title I Allocations

Prior work on differential privacy in the context of Title I is purely analytical, analyzes abstracted components of funding formulas, or focuses only on basic grants (Abowd et al., 2019; Pujol et al., 2020). By contrast, we fully replicate the Title I provisions for allocating more than $11.6 billion in basic, targeted, and concentration grants using the same data sources and procedures as the Department of Education, which is responsible for calculating the official Title I grant amounts each year (Snyder et al., 2019). We measure the impact of data and privacy deviations on the 2021 allocations to 13,190 LEAs across the United States. The primary data input is the Census Small Area Income and Poverty Estimates (SAIPE)

from 2019—a table of counts of total population, children, and children in poverty in school districts from all 50 states (excluding Puerto Rico and other territories) that incorporates weighted survey estimates from the ACS (see Appendix A.2 for details).

In a given year, the SAIPE may vary due to several sources of error, including relative error in the county-level estimate, error from other data sources used (e.g., tax data), and errors from raking and recombination methods used to convert county estimates to school district estimates (US Census Bureau, 2020). To simulate the effects of these "data deviations"—quantified data errors (Spencer, 1985)—we generate alternative poverty estimates for each school district from a normal distribution around the published estimate of children in poverty in that district from the 2019 SAIPE, following prior work and Census Bureau guidance (US Census Bureau, 2020) (Appendix A.2).

We then add "privacy deviations"—noise deliberately injected to achieve differential privacy. The Census Bureau has not yet announced any concrete plans for updated disclosure avoidance in the ACS, and the SAIPE currently does not inject noise for privacy on top of its inputs. To illustrate how privacy deviations might affect these and similar products, and to guide policy-makers as the Census Bureau develops new disclosure avoidance measures, we follow prior work (Abowd et al., 2019; Pujol et al., 2020) in applying the Laplace mechanism, a commonly used noise-injection procedure that is provably differentially private (Dwork et al., 2006). Our hypothetical mechanism does not include the complex postprocessing applied to the discrete Gaussian mechanism used in the decennial census; we only round negative numbers to zero (Abowd et al., 2022).

The strength of differential privacy (described by the parameter ϵ) determines the magnitude of privacy deviations (lower ϵ implies stronger privacy and generally more noise). ϵ measures how much an individual's decision to respond to a census survey increases their risk of unwanted disclosure. It is not yet clear whether or how privacy deviations would be added to a statistical product like the SAIPE in practice, and because the SAIPE incorporates weighted survey estimates from the ACS, its sensitivity to changes in an individual's response is unclear. Instead, we try several reasonable privacy settings to provide an upper bound on the magnitude of privacy deviations that might be added in practice (Abowd et al., 2019) (Appendix A.2). We focus on $\epsilon = 0.1$ and $\epsilon = 1$ (Appendix A.7 additionally varies ϵ from 0.001 to 10). Previous work on Title I (Abowd et al., 2019) suggests $\epsilon \geq 2.52$; many applications use similarly high values (Abowd et al., 2022), whereas differential privacy advocates often prefer $\epsilon < 1$.

The Title I legislation includes two post-formula provisions to achieve secondary policy goals. The "hold harmless" provision (20 U.S.C. §6332) limits funding losses to between 5 and 15% per year and the "state minimum" provision (20 U.S.C. §6333) sets a formulaic floor on the total amount received by each state. We treat the allocations generated without these provisions as the official formula-based "entitlements" for each district. Later, we compare these entitlements and the real allocations produced with these provisions. For each privacy setting, we compute the misallocation due to deviations by comparing the simulated allocations after deviations to the official entitlements. We repeat this procedure 1000 times, drawing new data and privacy deviations in each trial. Our metric of group-weighted misallocation describes the expected misallocation borne by the average formula-eligible child

in a given group nationwide, assuming that misallocation to a district is borne equally by all its eligible students.

Of the roughly $11.7 billion distributed nation-wide in 2021, districts in our simulation expect to lose a total of $1.06 billion (summing all losses in each simulation, then averaging summed losses across 1000 simulations; SD = $0.04 billion) in entitlements to other districts due to the Title I formula's handling of existing (before differential privacy) data deviations alone (see the first figure). The standard deviation in misallocation (computed by averaging over 1000 trials) is about $835,000 (the average district receives around $880,000)—$237 per student. When we add privacy deviations (for a relatively strong privacy setting $\epsilon = 0.1$), the expected total entitlement loss only increases by $50 million (4.7%; marginal SD = $2.9 million). For a less strong privacy setting (smaller privacy deviations; $\epsilon = 1$), the increase is negligible. The marginal impact is small because—as in the 2020 Decennial Census (Abowd et al., 2022)—the magnitude of privacy deviations is comparable to the magnitude of data deviations only in the least populous districts, even at a relatively strong privacy setting ($\epsilon = 0.1$) (Appendix A.7).

These costs are geographically asymmetrical. Certain population-sparse school districts, especially in the Northwest, benefit greatly on average from data deviations (Appendix Figure A1a)—their small sample sizes induce proportionally larger data deviations, and, because of their low absolute numbers of children in poverty (though poverty rates may still be high), they have more room to gain funding than to lose funding. Then, because the federal appropriation is fixed and allocations are zero sum, more populous districts, especially in the Southeast, pay for that proportional increase in funding with a small "tax" (Pujol et al., 2020). Less populous districts gain even more as they qualify for new grants (Zaslavsky & Schirm, 2002) (Appendix Figure A2). Notably, although less populous, usually rural districts gain funding on average from data deviations, their allocations are more volatile (Pujol et al., 2020) (Appendix Figure A5).

When we add privacy deviations (for relatively strong privacy, $\epsilon = 0.1$), gains by small districts are even more exaggerated (Appendix Figure A1b). Unlike data deviations, where the absolute variance increases with population size, our privacy deviations have the same variance in every district, exceeding data deviations in magnitude only in the least populous districts. Still, the marginal increase in cost to districts due to privacy deviations is much less than the base-level misallocations resulting from data deviations, and the marginal change reduces total misallocation about half the time.

1.3 Diversion from Marginalized Groups

Owing to Title I's distribution of quantified data deviations alone, Black students and Asian students can expect to lose around $5 and $8 per eligible student, respectively, whereas white students gain over $2 per eligible child on average (see the second figure). (The average district receives $1120 per eligible student.) Likewise, school districts with large Cuban, Puerto Rican, and other Hispanic communities expect to lose funding (between $3 and $14 per eligible student), whereas non-Hispanic districts gain (Appendix Figure A7). For a child in a particular district in an unlucky year, the disparity may be worse. Whether a demographic

group loses funding depends on whether its members tend to live in high- or low- poverty districts. Often, this happens because the poverty rate in the group itself is high. Groups that tend to live in denser, usually urban districts with more children in poverty lose out, whereas groups that live in sparse, often rural districts with fewer children in poverty (though the rate of poverty may be higher) gain. Geographically concentrated groups—such as tribal nations or racial subgroups (Appendix A.4)—experience more volatility in outcomes across trials, which depend on the population density and poverty rates where they live.

In a relatively strong privacy setting ($\epsilon = 0.1$), our differential privacy mechanism aggravates these disparities, especially for Black students, who lose more than twice as much funding on average after noise is injected—possibly because Black students are more likely to attend populous school districts where the costs of privacy deviations accumulate. But in less strong privacy settings ($\epsilon \geq 1$), disparities change very little from the status quo when privacy deviations are added (Appendix A.7).

To assess the impacts on noncategorical demographics, we also fit a generalized additive model (GAM) to the school district-level combined misallocations ($\epsilon = 0.1$) using district population characteristics: population density, median household income, proportion white, proportion Hispanic, proportion renter-occupied housing, and racial homogeneity (the Herfindahl-Hirschman index). Fitting the GAM on a sample of 100 trials, we find that districts with a median income between approximately $25,000 and $75,000 (about 56% of districts) can expect to lose out because of deviations, whereas most other districts gain (Appendix Figure A4). The 40% most population-dense districts can also expect to lose funding. Conversely, districts that are less than 5% Hispanic tend to benefit from data and privacy deviations.

1.4 Simple Reforms

Simple changes to the formula—including additional provisions currently required by law—can alleviate or aggravate disparities. For example, adding the hold harmless provision reduces the standard deviation in misallocation (relative to the formula entitlement) but drastically increases disparities in outcomes for racial minorities (see the second figure). Hold harmless prevents small districts from losing funding to data or privacy deviations, thereby increasing the tax on more populous districts and their non-white residents. The state minimum provision has a similar but smaller effect. Typically received by low population states, the state minimum slightly increases the amount of grants to low population districts, exacerbating disparities. This result illustrates a tension in evidence-based formula funding: Because estimates for less populous geographies have higher variance in both privacy and data deviations relative to their populations and entitlements, measures that overwhelmingly benefit those small areas burden larger areas.

We tested proposed policy changes that could alleviate this tension (Appendix A.6). We find that using multiyear averages with windows of increasing size decreases both overall misallocation and outcome disparities compared to when we use the averaged poverty estimates as a baseline (figs. S14 and S15). In general, using an average diminishes both data deviations and the privacy deviations required to achieve differential privacy, limiting both increases in

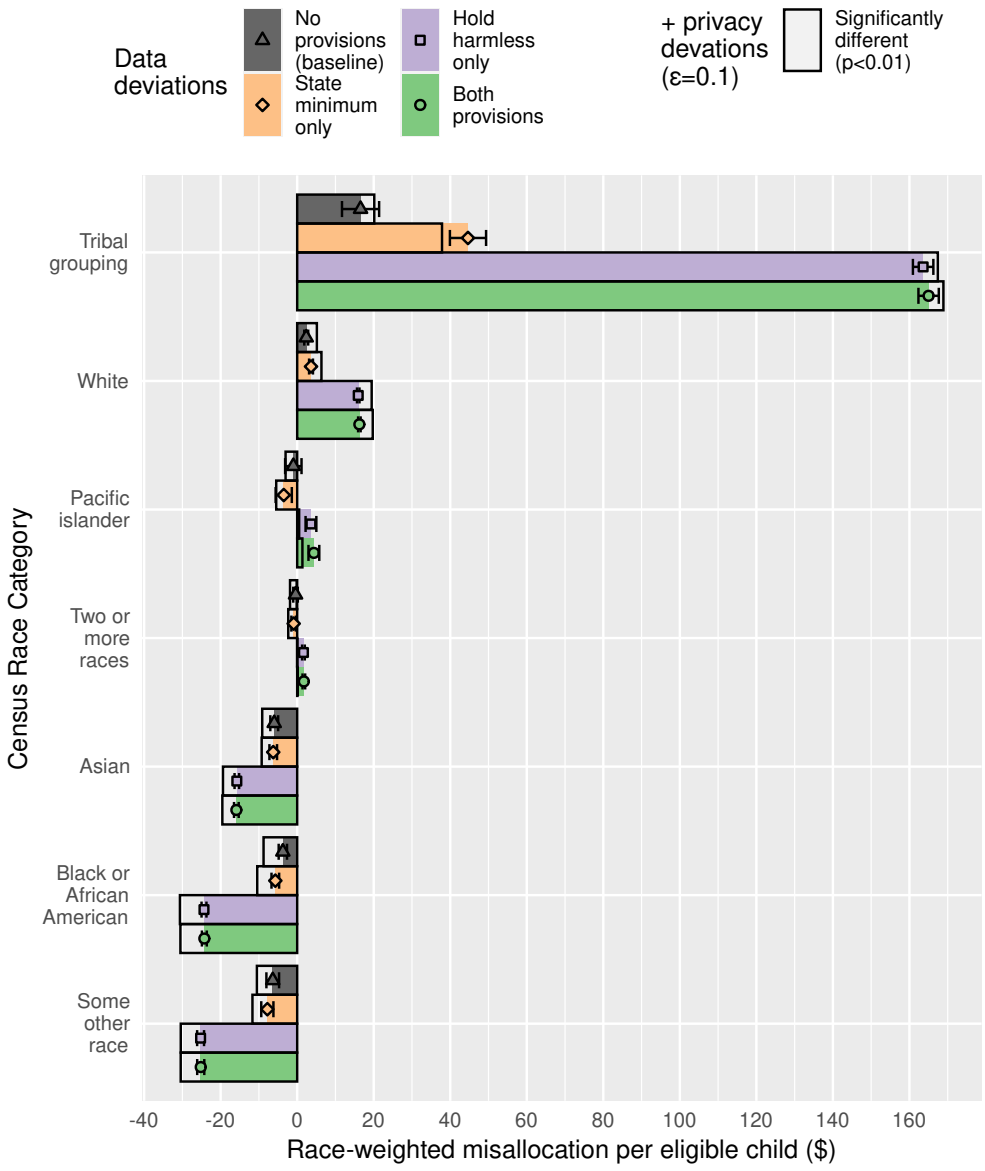


Figure 1.2: Expected misallocation by racial group. Expected misallocation borne by the average formula-eligible child in a given census group nationwide is shown (assuming each child in a district is affected by misallocation equally). Specifically, bars depict the nationwide sum of each district's misallocation multiplied by the proportion of respondents of a given census single race category in that district, divided by the total nationwide number of eligible children of that race (SM section 2). Averaged over 1000 trials. The colored bars indicate the race-weighted misallocation due to data deviations (data error) alone, with an error bar spanning a 90% normal confidence interval for this quantity. The additional impact of privacy deviations is significantly different ($p < 0.01$) for all groups, according to a two-sample z-test.

expected funding for less populous districts and alleviating worst-case outcomes. Averaging may even be just as effective at stabilizing funding year to year as the hold harmless provision (Zaslavsky & Schirm, 2002). We also tested requiring repeated years of ineligibility before disqualifying districts from funding, which did not change overall misallocation—likely because it permits more marginally wealthy districts to receive funding—but did reduce disparities (figs. S14 and S15).

1.5 Paying for (Private) Data

Simple policy changes can alleviate disparities in the impact of statistical uncertainty, but precisely targeted funding formulas will still have costs. Policy-makers could ensure that no school district expects to lose money because of the underlying data deviations quantified in our simulation by assigning just \$107 million (SD = \$31 million) in targeted payments to individual districts that lose funding on average across 1000 simulations. The cost of stronger privacy (using our simplified mechanism) could be much less: To compensate districts for only the expected additional lost funding due to privacy deviations, policy-makers need only distribute an extra \$41 million (SD = \$3.8 million) for stronger privacy ($\epsilon = 0.1$), or \$1.7 million (SD = \$601,000) for less added privacy ($\epsilon = 1$) (Appendix A.7). Still, a district's actual loss in any given year often greatly exceeds its expected loss, especially for less populous districts. To compensate districts for both data and privacy deviations in all but the worst 5% of our simulations, an additional \$4.7 billion would be needed in the stronger privacy setting ($\epsilon = 0.1$). The cost is greater if policy-makers wish to also compensate for the many other forms of error not quantified here, or for a stronger privacy mechanism. It may be difficult to justify or legislate funding increases to just the districts expected to lose funding. Simply increasing the total federal appropriation to Title I (benefiting all districts unequally) by \$135 million (the combined total expected loss) would only compensate for about half of expected losses. However, a \$4.7 billion increase (95% loss coverage) would compensate for nearly all total expected losses and cut total 5% quantile misallocation roughly in half. The White House's proposed 2022 allocation—a \$20 billion increase, since reduced to \$1 billion in Congress—would completely compensate for privacy and data deviations incurred under the 2019 budget, but inequalities would remain. An overall budget increase would provide "no-penalty" compensation (Pujol et al., 2020) for data and privacy deviations but would not solve issues of relative equity (though budget increases do reduce the number of held harmless districts).

1.6 Uncertainty-Aware Policy Design

The addition of noise for differential privacy exposes epistemic issues with formula design predicted by early work on census-guided federal funding even before differential privacy was first proposed (boyd & Sarathy, 2022; Zaslavsky & Schirm, 2002; National Research Council, 2000, 2003). Indeed, our results suggest that the impacts of differential privacy relative to other sources of error in census data could be minimal. But current legislation holds few allowances for the impacts of statistical uncertainty. Use of census data for the Title I formula is mandated "unless the Secretary and the Secretary of Commerce determine

that some or all of those data are unreliable or ... otherwise inappropriate" (20 U.S.C. §6333). National Research Council studies, commissioned by the Department of Education before ACS estimates were first incorporated in the SAIPE after 2005, warned against hard thresholds and hold harmless provisions (National Research Council, 2000, 2003)—but these provisions are still in effect. Recently, the Biden administration proposed a new Title I budget that includes funding to improve the poverty estimates—but there are still no measures to update the formula to handle uncertain inputs. Simply acknowledging the effects of data error could improve future policy design for both formula funding and disclosure avoidance.

Our findings come with limitations. Injected noise is just the tip of the iceberg: Many other unquantified forms of statistical uncertainty—including previous disclosure avoidance methods—affect poverty estimates in different ways (Groves & Lyberg, 2010). No confidentiality measures are directly applied to the SAIPE, but its inputs (mainly ACS and IRS data) may have hidden or unintended distortions due to swapping and other ad hoc disclosure avoidance techniques (Christ et al., 2022). By replacing other methods of disclosure avoidance, differential privacy could even reduce the amount of overall misallocation due to uncertainty. Lacking an alternative source of poverty data, we do not assess the impacts of systematic biases, including undercounts of marginalized groups. Our analysis of the Title I allocation process also leaves out several elements that could affect the applicability of our findings to the real-world distribution of funds, including small-district appeals (20 U.S.C. §6333) and district-level heterogeneity in use of funds. Temporal trends in funding, in combination with provisions like hold harmless, could compound the effects of deviations (Zaslavsky & Schirm, 2002).

Data error—from undercounts to sampling error to noise injection—will always affect evidence-based policy to some degree. In 2017, 316 federal spending programs relied on US census data to distribute over $1.5 trillion in federal funding across states, cities, and school districts (Reamer, 2020). Uncertainty in census data—including intentionally added error for privacy—will incur costs for stakeholders in those programs. But at least the quantifiable portion of those costs can be mitigated with uncertainty-aware policy design and budget increases—an avenue for compromise between targeted policy, equity, and also additional privacy.

Acknowledgments

We thank C. McKay Bowen, d, boyd, A. Cohen, W. Eddy, Z. El-Kilani, R. Ghani, P. Nanayakkara, D. Pujol, A. Roth, I. Schmutte, J. Sarathy, seminar participants at Carnegie Mellon University, Columbia University, the Simons Institute at Berkeley and the Census Bureau, and anonymous referees for their very helpful insights and feedback on this research. We also thank J. Maples, W. Sonnenberg, and T. Stephenson for their help with replicating the poverty estimation and Title I allocation processes.

Chapter 2

Estimating Research Impacts of Statistical Uncertainty and Privacy

This chapter is reproduced from:

Steed, R., Mustri, E. A. S., & Acquisti, A. (2025). *Impacts of Data Error and Differential Privacy on Findings from Social Science.*

2.1 Introduction

Government agencies and private companies are crucial sources of social science data for public research, but public releases of sensitive data carry privacy risks for individuals and groups. Recently, organizations have turned to new "privacy-preserving" methods to strengthen privacy protections in public data releases (Steed & Acquisti, 2025). One of the most popular approaches is differential privacy (DP), a framework for preserving individual privacy (Dinur & Nissim, 2003; Dwork et al., 2006) which has been adopted by Meta (King & Persily, 2020), the Wikimedia Foundation (Adeleye et al., 2023), the U.S. Census Bureau (Abowd et al., 2022), and many other organizations in government and industry (Desfontaines, 2021). These deployments aim to mitigate concerns about reconstruction and re-identification attacks aided by widely available commercial data and computing power (Abowd & Hawes, 2023; Narayanan & Shmatikov, 2010).[1]

Applications of differential privacy, however, have met with disagreements and even controversy about the appropriate balance of privacy and data usability in public statistics (boyd & Sarathy, 2022; Hotz et al., 2022; Oberski & Kreuter, 2020). Differential privacy is usually accomplished by injecting statistical noise when sharing statistics about sensitive personal data, and researchers worry that statistical noise and other distortions could impact the quality and feasibility of social science research (Hotz et al., 2022; Ruggles et al., 2019).

In this study, we investigate empirically the possible impacts of additive noise—arising from existing data error or injected for data privacy—on scientific findings from regression analyses common in the social sciences. We identified, collected, and replicated a benchmark of 177 key findings from 50 empirical studies published in economics and social science journals. We imagine a counterfactual world where the social statistics used to produce these findings were released using differentially private mechanisms.[2] Replacing the original public datasets with counterfactual versions simulated with additive data error and released with generic differentially private mechanisms, we evaluate *epistemic parity*: the principle that published findings should be replicable with a noise-infused dataset (Rosenblatt et al., 2023). Supposing that researchers did not change their methods to account for added noise, would social science studies reach the same conclusions if they instead relied on these noisy statistics?

We find that using privacy budgets typical in industry, between 61%–90% of simulated findings (over 10 simulations) still support the original claims at confidence level $\alpha = 0.1$; 10%–39% do not. These rates of epistemic disparity resulting from typical industry privacy budgets are generally larger than what economists say they are willing to accept in public data (Williams, Snoke, et al., 2024) but generally smaller than rates of disparity in large-scale replication studies in economics and psychology (Camerer et al., 2016; Open Science Collaboration, 2015; Silberzahn et al., 2018).

[1] There are many empirical examples of reconstruction attacks against public datasets, including attacks on Massachusetts Group Insurance Commission data (Sweeney, 2002), the Personal Genome project (Sweeney et al., 2013), the Netflix Prize dataset (Narayanan & Shmatikov, 2007), the Aircloak Challenge (Cohen & Nissim, 2020), and, most recently, the 2010 and 2020 Decennial Censuses (Abowd et al., 2023; Dick et al., 2023; Steed et al., 2024; Flaxman & Keyes, 2025).

[2] Specifically, we focus on the "trusted curator" paradigm in which an organization collects confidential microdata centrally, computes aggregate statistics, and adds noise to each before sharing (Clark et al., 2024).

We also analyze the antecedents of disparities due to additive noise, quantifying the impacts of privacy budget, choice of mechanism, and other characteristics of the statistics and analysis. In particular, we show that claims based on weaker initial effect sizes are more likely to be contradicted by added noise, as are more quantitatively specific claims. We explore how results from certain subfields of economics (e.g., Labor & Demographic Economics) may be most impacted, particularly fields that rely on certain types of data (e.g., common cross-sectioning variables such as age, education, and sex/gender).

Of course, differential privacy is not the only source of additive noise: social statistics already contain measurement error and other kinds of non-sampling error (Groves & Lyberg, 2010; Steed, Liu, et al., 2022). We find that even modest amounts of existing data error can alter findings. Differentially private mechanisms do exacerbate epistemic disparities—but the marginal impacts of privacy protection are smaller, especially when the magnitude of data error is large.

Recent federal policy calls for the advancement of privacy-preserving data sharing across public and private sectors (Biden, 2023; National Science and Technology Council, 2023). The impacts of these measures would be widespread. The Integrated Public Use Microdata Series (IPUMS) alone, for example, lists over 27,000 academic publications based on its catalogue of U.S. census and survey data (IPUMS, 2024); the the Surveillance, Epidemiology, and End Results (SEER) program claims more than 17,000 publications use its survey data (Penberthy, 2023); and thousands more studies rely on other public statistics published by government agencies and other organizations. Scholars have called for more research into the potential impacts of differential privacy and similar techniques on research (Oberski & Kreuter, 2020; Hotz et al., 2022; Acquisti & Steed, 2023). Our study provides an estimation of the key trade-offs involved and should serve as a useful guide for policymakers and data curators designing the next generation of privacy-preserving technologies.

Related Work

Our work builds on several studies examining the trade-offs between privacy and usability in public statistics, particularly in the context of the U.S. census (Abowd & Schmutte, 2019; Kenny et al., 2021; Pujol et al., 2020; Brummet et al., 2022; Steed, Liu, et al., 2022; Barrientos et al., 2023).

In particular, Williams, Barrientos, et al. (2024) evaluate the quality of differentially private linear regression methods with a highly controlled simulation study. Closest to our work is that of Rosenblatt et al. (2023), who explore the impacts of synthetic microdata on the results of eight social science studies. The methods evaluated in these studies—for generating DP synthetic microdata and running DP linear regressions on private data—may be critical in future microdata products but are not yet widespread. Our study examines the simpler and more common practice of infusing noise into aggregate statistics before public release, used in the 2020 Decennial Census (Abowd et al., 2022) and the Social Science Facebook URLs dataset collaboration (King & Persily, 2020).

Our method of analysis is also inspired by recent replication efforts in psychology and economics, particularly the Open Science Collaboration (Open Science Collaboration, 2015;

	Obs.	Mean	SD	Min.	Q1	Q2	Q3	Max.
Studies	**50**							
Years since publication	50	6.66	4.38	0	3.00	6.00	10.0	15.0
Num. results replicated	50	3.54	1.89	1.00	2.00	3.00	4.00	10.0
Citations (Semantic Scholar)	50	167.	514.	1.00	20.2	42.0	97.2	3520.
Results	**177**							
Model degrees of freedom	175	62.5	122.	1.00	7.00	20.0	40.0	513.
Num. vars about people	177	7.51	6.98	1.00	3.00	6.00	11.0	50.0
Proportion of personal vars. privatized	177	0.838	0.232	0.167	0.714	1.00	1.00	1.00
Answers primary or co-primary RQ	43							
Supports claim mentioned in abstract	157							
Personal data variables	**496**							
Successfully privacy protected	404							
Control var.	312							
Dependent var.	88							
Primary independent var.	33							
Instrumental var.	9							
Instrumented var.	20							
Subsetting var.	8							
Weighting var.	20							
Booleans/categorical	7							
Count, log count	55							
Mean, log mean	264							
Ratio	4							
Median	5							
More complex query	152							

Table 2.1: Descriptive statistics. Variables may be used in multiple results within the same study.

Camerer et al., 2016) and the Many Labs replication project (Klein et al., 2014).

2.2 Data

To find social science studies with working replication packages, we used 1) the Inter-University Consortium for Political and Social Research (ICPSR)'s replication package repository (Inter-University Consortium for Political and Social Research, 2025) and 2) Find Economic Articles with Data, a tool for searching English economics articles with accessible data/code supplements (Kranz, 2024). The database includes articles from journals with standardized data accessibility policies, including American Economic Association (AEA) journals, the *Review of Economics and Statistics*, and several other journals. We searched for studies using aggregate statistics by filtering for articles whose title or abstract uses one of several common, mostly geographic, aggregation levels such as "county", "school", "neighborhood", or "organization". Appendix B.2.1 lists all our search terms.

We manually reviewed every article resulting from these searches, excluding studies that did not use aggregate statistics (e.g., studies exclusively analyzing microdata). We also excluded studies that did not use *personal data*: information about individuals or households. To

construct a sample representative of contemporary research, we restricted the sample to articles published in the last 15 years that use personal data collected within the last 70 years (similar to the U.S. Census Bureau's policy of keeping data confidential for 72 years after collection).[3] We attempted to reproduce 93 studies that met these criteria[4], correcting the authors' code only as much as necessary to reproduce their results without error.[5] We were unable to replicate 43 studies, most often either because the replication package did not include the datasets necessary to reproduce the main results or because the datasets did not include enough information to implement differential privacy. Less often, our reproductions failed because of bugs or discrepancies in the replication packages.

Our current sample includes 50 successfully replicated studies, summarized in Table 2.1. For each article, we read the abstract and introduction and identified the key empirical claims made by the author. We then identified the statistical estimate or estimates (hereafter *results*) supporting those claims—in our sample, all regression coefficients—and modified the authors' code to extract those key numerical results and related statistics (standard error, t-statistic, degrees of freedom, etc.).[6] For most studies, we reproduced 2–4 results (Table 2.1). Most results we reproduced (157 of 177) substantiate a key empirical claim mentioned in the article's abstract (Table 2.1).

2.3 Methods

2.3.1 Differential Privacy

To investigate the counterfactual where the aggregate statistics in these studies were released with differential privacy, we injected noise into all statistics about personal data used to produce the key results in our sample.

We compare two approaches: 1) pure differential privacy with the Laplace mechanism (Dwork et al., 2006), a simple randomized mechanism for releasing statistics; and 2) zero-concentrated differential privacy (zCDP) (Dwork & Rothblum, 2016; Bun & Steinke, 2016) with the Gaussian mechanism, a relaxation of differential privacy with improved utility.[7] These are

[3]We also excluded 3 studies with code primarily written Matlab, which was not supported by our computing infrastructure. The final sample consisted of only Stata scripts, though a few studies included unused R and Matlab scripts.

[4]The terms "repeatability", "reproducibility", and "replicability" are often interchanged (Barba, 2018). Here, we use the term "reproduce" to mean achieving the same results with the same data and methods and "replicate" to mean achieving similar results using the same methods but different (noisy) data (Broman et al., 2017).

[5]We recorded the expected values of key results reported in the article and in every run, and we automatically checked those values against the results produced by our copy of the authors' code to check for errors introduced by our environment or modifications. All of our successful reproductions matched the authors' original results to same the level of numerical precision reported in their published article.

[6]To help standardize our analysis, we did not consider visual findings (multiple results represented by visual relation in a figure) (Rosenblatt et al., 2023). Visual findings substantiated key claims in only a small handful of the studies we found.

[7]Specifically, we used the implementation of the Laplace and Gaussian mechanisms in the `opendp` Python library (Shoemate et al., 2025).

two simple, popular, and generalizable DP mechanisms data curators might implement, though there are more optimal techniques for utility, such as data-adaptive mechanisms (Cummings et al., 2024).

In pure differential privacy, the parameter ϵ measures how much the inclusion of information about an individual in the dataset increases the risk that their personal information is revealed by published statistics (Def. 1).

Each study in our sample uses a set of ν variables (or covariates) $F(D)_{k\times\nu} = (f_1(D), f_2(D), \dots f_\nu(D))$: statistical queries $f_j : \mathcal{D}^n \to \mathbb{R}^k$ over a confidential dataset of personal information D about n individuals.[8] Each statistical query $f_j(D) = (f_j(d_1), f_j(d_2), \dots, f_j(d_k))$ is provided for k different disjoint subsets $d_i \in D$ (e.g., the total population of each region in each year).

Definition 1 (Dwork et al., 2006)**.** Let $M : \mathcal{D}^n \to \mathcal{Y}$ be a randomized algorithm. M is *ϵ-(pure) differentially private* if for all pairs of datasets $D, D' \in D^n$ which differ in only one entry and all possible outputs $T \subseteq \mathcal{Y}$:

$$\Pr[M(D) \in T] \leq e^\epsilon \Pr[M(D') \in T].$$

Pure DP with the Laplace mechanism. To simulate releasing $F(D)_{k\times\nu}$ with ϵ-(pure) differential privacy, we apply the Laplace mechanism to each query $f_j(D)$ (Def. 3).

We rely on basic composition (Dwork & Roth, 2013, Corollary 3.15) to release all the queries $F(D)_{k\times\nu}$ with ϵ-DP: we use $\frac{\epsilon}{\nu}$-DP mechanisms $M_j^{Lap}(D)$ for each query f_j such that $M^{\text{Lap}}(D) = \left(M_1^{Lap}(D), M_2^{Lap}(D), \dots, M_\nu^{Lap}(D)\right)$ is ϵ-(pure) DP.

For some but not all studies, it is possible that the same individual could contribute to multiple subsets (usually statistics from different time periods). For simplicity, we only protect privacy with respect to an individual's contribution to one subset d_i; in general, then, our mechanisms are thus ϵ differentially private with respect to the contribution of each individual within every subset. Since individuals typically contribute to only one geographic region, our "unit" of privacy is usually (individual)-(time period) (Desfontaines, 2021). With that assumption, $\Delta f_j(D) = \Delta f_j(d_i)\ \forall d_i \in D$ (equivalent to parallel composition).

Definition 2 (Dwork & Roth, 2013)**.** The *l_p-sensitivity* of a function $f : \mathcal{X}^n \to \mathbb{R}^k$ is:

$$\Delta f = \max_{D,D'} ||f(D) - f(D')||_p,$$

where D, D' are datasets which differ in exactly one entry.

Definition 3 (Dwork & Roth, 2013)**.** Let $f : \mathcal{D}^n \to \mathbb{R}^k$. The *Laplace mechanism* is defined as

$$M^{Lap}(D) = f(D) + (Y_1, \dots, Y_k),$$

where the Y_i are independent random variables drawn from the Laplace distribution $Y_i \sim Lap\left(\frac{\Delta f}{\epsilon}\right)$ (Def 6), and Δf is the l_1-sensitivity of the query function f (Def. 2).

[8] This dataset is typically not provided in the replication package; we only have access to the statistics $F(D)$ used by the original authors.

zCDP with the Gaussian mechanism. Zero-concentrated differential privacy (zCDP; Def. 4) is a relaxation of pure differential privacy that provides better accuracy (Dwork & Rothblum, 2016; Bun & Steinke, 2016). The Gaussian mechanism (Def. 5) satisfies zCDP. This mechanism was used to release statistics from the 2020 Decennial Census (Abowd et al., 2022).

zCDP uses a different privacy parameter, ρ. A mechanism that satisfies pure DP also satisfies cZDP; specifically, if M satisfies ϵ-DP, then M satisfies $\left(\epsilon\frac{\exp(\epsilon)-1}{\exp(\epsilon)+1} < \frac{\epsilon^2}{2}\right)$-zCDP (Bun & Steinke, 2016; Steinke, 2024). We use this bound for comparison between the two mechanisms hereafter. zCDP has the same basic composition property as pure DP (Bun & Steinke, 2016, Lemma 1.7), so we again apply $\frac{\rho}{\nu}$-zCDP mechanisms $M_j^{\text{Gauss}}(D)$ such that $M^{\text{Gauss}}(D) = \left(M_1^{Gauss}(D), M_2^{Gauss}(D), \ldots, M_\nu^{Gauss}(D)\right)$ is ρ-zCDP.

Definition 4 (Bun & Steinke, 2016)**.** Let $M : \mathcal{D}^n \to \mathcal{Y}$ be a randomized algorithm. M is ρ-zCDP if for all pairs of datasets $D, D' \in D^n$ which differ in only one entry and all $\alpha \in (1, \infty)$:

$$D_\alpha(M(D)||M(D')) \leq \rho\alpha,$$

where $D_\alpha(M(D)||M(D'))$ is the α-Rényi divergence between the distributions of $M(D)$ and $M(D')$ (Def. 7).

Definition 5 (Bun & Steinke, 2016)**.** Let $f : \mathcal{D}^n \to \mathbb{R}^k$. The *Gaussian mechanism* is defined as

$$M^{Gauss}(D) = f(D) + (Z_1, \ldots, Z_k),$$

where the Z_i are independent random variables drawn from the Gaussian distribution $Z_i \sim \mathcal{N}\left(\frac{(\Delta f)^2}{2\rho}\right)$, and Δf is the l_2-sensitivity of the query function f (Def. 2).

Sensitivity and post-processing. For each study in our sample, we identified all the statistics about personal data used to produce the key results. For each statistic, we defined the global sensitivity of the query used to produce each individual statistic (Def. 2). When the global sensitivity of a query was undefined (for a ratio of counts, for example), we first deconstructed the query into component queries with defined sensitivity (e.g., the numerator and denominator counts), applied the DP mechanism to each component independently, then reconstructed the query.[9] Some variables are used as regression weights and must not be negative; for those variables, we clipped the noisy values to 0. Differential privacy is preserved through these post-processing steps (Dwork & Roth, 2013, Proposition 2.1).

Because almost all of the datasets used in our sample contain only aggregate statistics, and not the confidential microdata used to construct them, we make several simplifying assumptions about the data curator that likely differ from a real-world deployment of DP.

[9]We included these pre- and post-processing computations even in control runs where no noise was injected, checking to ensure that our reconstructions matched the original data.

1. For a few statistics used in our sample, the authors' replication package did not provide enough information to define sensitivity directly or reconstruct the statistical query. For some queries on continuous data (mostly about income or wages), we were able to make a simplifying assumption about the sensitivity (e.g., that income does not exceed three standard deviations above the mean). We did not apply any privacy protection to the remaining statistics (18.5% of statistics in our sample).

2. We assume that individuals have equal influence over statistics. In reality, particularly for survey-based estimates, some subgroups may be weighted more heavily than others. A real-world deployment of differential privacy must take this into account (see, e.g., Drechsler & Bailie, 2024; Seeman et al., 2024).

3. Statistics are likely to be released as part of larger datasets, and data curators may have to split privacy loss budgets amongst many other queries. We only distribute the privacy budget ϵ or ρ over the statistics used within each study.

4. In our mechanisms, each statistic receives an equal share of the privacy budget. In real-world deployments such as the U.S. Decennial Census, data curators may tune budgets more carefully to preference certain critical statistics (Abowd et al., 2023).

5. The statistics used in these studies may be derivatives or linkages of other public datasets produced by other data curators; because these precursor datasets and pre-processing steps are usually not accessible, we apply differential privacy to the combined set of statistics used directly by the authors, as if all the data curators shared a single total privacy budget ϵ or ρ for each study.

6. We assume that the statistics used in our sample are not already subject to formal privacy protections. In reality, the data that produced these statistics may have been subject to cell suppression, random swapping, censoring, or other disclosure avoidance techniques before publication.

2.3.2 Additive Data Error

To compare the effects of DP mechanisms to the possible impacts of existing data error, we simulate possible alternative versions of the dataset by drawing counterfactual observations from a normal distribution conditioned on the original statistics.

Variance. Unfortunately, the replication datasets rarely include data-based variance estimates. Instead, for each observed statistic $x_{ij} = f_j(d_i)$, we simply assume some coefficient of variation c_{ij} and simulate sampling variances $v_{ij} = (c_{ij} x_{ij}^{1-b})^2$ in proportion to the observed statistic. The parameter b models the rate at which the standard deviation $c_{ij} x_{ij}$ diminishes with x_{ij}; by default, $b = 0$ (constant returns). We test a range of coefficients of variation $c \in \mathbb{R}^+$ and $b \in \mathbb{R}$.

Shrinkage. Let μ represent the true statistics. Simply drawing replicates from

$$\mu | x \sim \mathcal{N}(x, \text{diag}(v)) \tag{2.1}$$

is inadmissable for the true population statistics μ (Stein, 1956). Cui et al. (2023) suggest two admissable constructions using shrinkage estimation from multi-level empirical Bayesian modeling. Following Cui et al. (2023), we simulate replicates with the general form

$$\mu_{ij}|x \sim \mathcal{N}\left((1-B_{ij})x_{ij} + B_{ij}\beta_j, (1-B_{ij})v_{ij}\right), \tag{2.2}$$

where $\beta_j, B_{ij} \in [0,1]$ are functions of x_{ij} and v_{ij} (Morris & Lysy, 2012; Cui et al., 2023). This method adjusts the counterfactual estimate to account for a baseline β_j with variance reduced by $100B_{ij}\%$. The Hudson-Berger construction (Hudson, 1974; Berger, 1976) uses $\beta = 0$ and

$$B_{ij}^{HB} = \min\left(1, \frac{(k-2)/v_{ij}}{\sum_{m=1}^{k}(x_{mj}/v_{mj})^2}\right).$$

The Morris-Lysy construction (Morris & Lysy, 2012) uses $\beta_j = \bar{x}_j$ and

$$B_{ij}^{ML} = \frac{v_{ij}}{v_{ij} + \bar{v}_j^H(1-\hat{B}_j^H)/\hat{B}_j^H},$$

where $\bar{v}_j^H = k/\sum_{i=1}^{k} v_{ij}^{-1}$ is the harmonic mean of v_j, $\hat{B}_j^H = (k-3)/(k-1)\hat{\sigma}^2$, and $\hat{\sigma}^2 = (k-1)^{-1}\sum_{i=1}^{k}(x_{ij}-\bar{x}_j)^2/v_{ij}$ is the mean squared error in the observed statistic x_j.

Generally, the Hudson-Berger construction imposes more shrinkage for large v_{ij}, while Morris-Lysy imposes more shrinkage for smaller v_{ij}. We also present results for the inadmissable no-shrinkage construction (Eq. 2.1) for comparison.

2.3.3 Metrics

To evaluate the potential impact of differential privacy, we reproduced each study in our sample with datasets treated with these differentially private mechanisms. We compare these counterfactual results to the original results using several standardized metrics developed in previous replication studies (Open Science Collaboration, 2015; Rosenblatt et al., 2023; Williams, Barrientos, et al., 2024).

Epistemic parity. Following Rosenblatt et al. (2023), we test for *epistemic parity*: the principle that empirical claims should be replicable with a differentially private dataset. We define a *finding* as a comparison between a statistical estimate (a result) and one or more other statistical estimates or scalars (for example, a study may find that the coefficient on x is greater than zero when y is regressed on x). We define a *claim* as an epistemic assertion based on one or more findings (e.g. that the effect of x on y is positive) (Rosenblatt et al., 2023; Cohen et al., 2018). We use the term epistemic *dis*parity to describe cases where epistemic parity does not hold.

To evaluate epistemic parity, we express each claim as a Boolean condition. For each finding, we defined a range of possible coefficient values such that the original finding holds true (Fig. B.3.4). Epistemic parity requires that all the findings supporting a given claim remain in this range. This definition of parity is depends on our interpretation of

the authors' empirical claims—for example, most claims require only that a statistical significant coefficient have a particular sign; others rely on particular relationships between the magnitude of different coefficients; a few rely on coefficient values to fall within a particular range (Figure B.3.4).

We test whether counterfactual estimates produced from differentially private data produce epistemically similar results: estimates whose confidence intervals fall inside the region of epistemic parity at a given confidence level α. For example, if the authors' original claim depended on a coefficient having a positive value, epistemic parity would require the coefficient's confidence interval to be entirely positive at confidence level α. We except quantitative claims not originally based on statistical significance (e.g., a claim that an estimate fall between 0.1 and 0.2 despite its original confidence interval exceeding those bounds).

We call this notion of epistemic parity *strict* epistemic parity. Since strict epistemic parity is based on our subjective reading of the original authors' claims, we also test three standardized notions of epistemic parity that do not depend on the authors' claims:

- *Sign parity*: whether the counterfactual estimate $\beta^{\hat{\text{DP}}}$ has the same sign as the original estimate $\hat{\beta}$ at an equivalent level of significance for the test on $H_0 : \beta = 0$ vs. $H_A : \beta \neq 0$.
- *Sign reversal*: whether the counterfactual estimate $\beta^{\hat{\text{DP}}}$ has the *opposite* sign as the original estimate $\hat{\beta}$ at an equivalent level of significance for the test on $H_0 : \beta = 0$ vs. $H_A : \beta \neq 0$.
- *Confidence interval coverage at* α: whether the original estimate falls within the confidence interval of the counterfactual estimate at confidence level α.[10]

Bias in standardized effect size. Authors often rely on significance thresholds to make claims, but statistical significance is often an imprecise and arbitrary measure for epistemic justification. Often, only a small change is required to move an estimate from one significance level to another (Gelman & Stern, 2006). For added context, we also use a non-binary measure of epistemic difference: the absolute difference between the counterfactual effect size and the original effect size. To produce standardized effect sizes across studies, we follow Open Science Collaboration (2015, Appendix A3) in converting each result to a common effect size metric, the correlation coefficient r computed from the t-statistic and residual degrees of freedom ν for each result:

$$r = \sqrt{\frac{t^2/\nu}{t^2/\nu + 1}} \left(\text{sign}(\hat{\beta})\text{sign}(\beta^{\hat{\text{DP}}})\right). \tag{2.3}$$

r is coded as negative if the sign of the counterfactual estimate does not match the sign of the original estimate.

[10] Confidence intervals may give much better information about replicability than p-values, which can vary widely over replications of the same statistical test (Cumming, 2008).

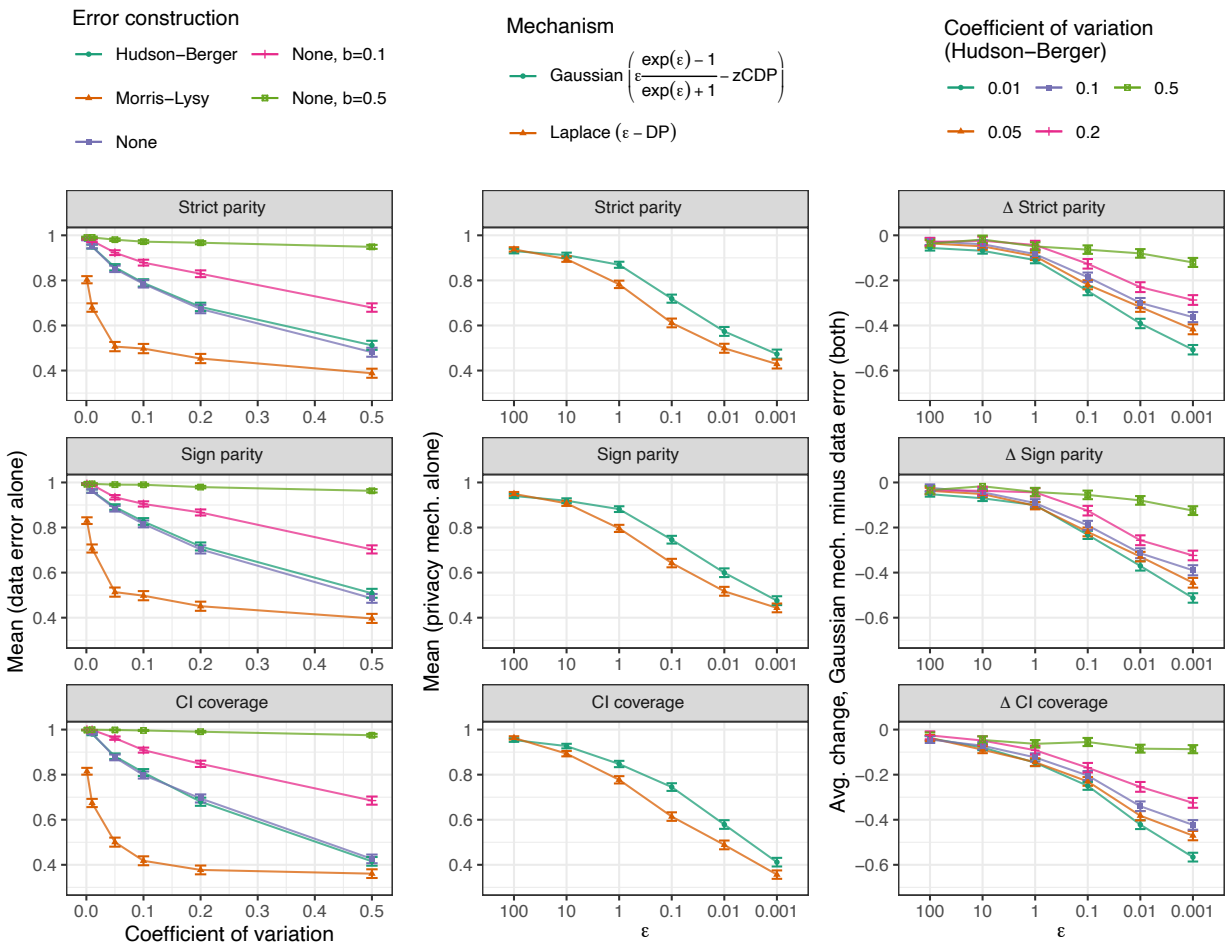


Figure 2.1: Impacts of simulated additive data error alone (left), DP mechanisms alone (middle), and the zCDP Gaussian mechanism *after* data error (right), evaluated at confidence level $\alpha = 0.1$, averaged over all results and 10 simulations. Error bars depict 90% confidence intervals. Lower values of the privacy parameter ϵ provide stronger privacy protection but require more injected noise; curators commonly use $0.1 \leq \epsilon \leq 10$.

2.4 Results

2.4.1 Impacts of noise for differential privacy

Figure 2.1 presents the impacts of noise added for differential privacy on epistemic parity, sign parity, and CI coverage evaluated at confidence level $\alpha = 0.1$. Additional metrics at additional confidence levels (effective sample size and absolute difference) are depicted in Figures B.3.1–B.3.2.

Though differential privacy advocates often prefer $\epsilon < 1$ (Dwork et al., 2019), real-world public datasets released with differential privacy tend to use values of ϵ greater than 1 and sometimes even greater than 10—Google's Community Mobility Reports use $\epsilon = 2.64$ per day, for example, and Facebook's URLs dataset was released with ($\epsilon = 1.453$, $\delta = 10^{-5}$)-DP

(Desfontaines, 2021).

At $\epsilon = 10$ with the pure DP Laplace mechanism, a relatively small amount of injected noise, we observe epistemic parity at confidence level $\alpha = 0.1$ in around 90% of simulated findings (over 10 replicates)—simulations where the results, when reproduced using differentially private data, still support the authors' original claims. The rate of parity decreases to 77% at $\epsilon = 1$ and 61% at $\epsilon = 0.1$. Almost all of the disparities we observe are originally significant results nullified by noise; the rate of originally insignificant results found to be significant after noise is relatively small (4.3% at $\epsilon = 0.1$). Using the $\epsilon\frac{\exp(\epsilon)-1}{\exp(\epsilon)+1}$-zCDP Gaussian mechanism (a relaxed privacy guarantee) noticeably increases the rate of parity, particularly for stronger privacy budgets ($\epsilon < 10$)—by 10 percentage points at $\epsilon = 1$. We observe only slightly higher rates of sign parity (matching sign and significance), a metric not based on our interpretation of authors' claims.

Are these levels of disparity acceptable to researchers? In a survey of 1,028 American Economic Association members (Williams, Snoke, et al., 2024), around 60% of economics researchers said they would be willing to accept estimates changing significance levels (from $p < 0.05$) at a rate of up to 10% before they would sacrifice access to noisy estimates based on administrative data.[11] In our sample, sign disparity (similar to significance mismatch[12]) at $\alpha = 0.05$ reaches 9% at $\epsilon = 10$ with the pure DP Laplace mechanism—a relatively weak privacy budget in typical deployments. And around 60% of economics researchers said they were willing to accept estimates switching sign (similar to sign reversal[13]) at a rate of up to 5% before they would sacrifice access to noisy estimates based on administrative data. We find that this occurs less often; even at a relatively strong budget $\epsilon = 0.1$, significant positive effects flip to significant negative effects, or vice versa, at a rate of only around 2.8% ($\alpha = 0.1$; Figure B.3.2).

On the other hand, these levels of disparity are much lower than the rates observed in large-scale replication studies (different team, different data, same methods) in economics and other fields. Replicating 110 studies from economics and political science (same data, different methods), Brodeur et al. (2024) find that only 70% of robustness checks (changing weighting, control variables, estimation methods, fixed effects, etc.) recover a significant effect in the

[11] These self-reported preferences may change with additional information and may not match revealed preferences in specific contexts. Notably, over 55.2% of respondents said they had "Never heard of the concept" of differential privacy and an additional 24.8% said they had heard of the term but were "not familiar with any of the details" (Williams, Snoke, et al., 2024).

[12] Williams, Snoke, et al. (2024) define significance mismatch as "the relative frequency with which a noisy estimate has a different statistical significance (assume 0.05 level) than the estimate without noise" (Appendix B.2.2). Respondents may have interpreted significance mismatch as sign disparity at $\alpha = 0.05$ (which would count as mismatches cases where estimates maintained significance but switched sign), but they may also have interpreted it more literally (a mismatch is any change in significance, up or down, regardless of sign). Under the latter definition, rates of significance mismatch are slightly lower (Figure B.3.1).

[13] Williams, Snoke, et al. (2024) define sign mismatch as "the relative frequency with which a noisy estimate is expected to have a different sign (positive or negative) than an estimate without noise" (Appendix B.2.2). Respondents may have interpreted sign mismatch as sign reversal (excluding initially insignificant results), but they may also have interpreted it more literally (any change in sign, regardless of significance). Under the latter definition, the rate of sign mismatch reaches 10% at $\epsilon = 1$ with the pure DP Laplace mechanism (Figure B.3.1).

same direction ($p < 0.05$). In our experiments, this rate of sign parity occurs at stronger levels of privacy between $\epsilon = 1$ and $\epsilon = 0.1$ (Laplace mechanism).[14] Similarly, a replication of 18 laboratory experiments from economics (Camerer et al., 2016) found a significant effect in the same direction ($p < 0.05$) for only 11 studies (61% sign parity)—equivalent to the effects of the Laplace mechanism with budget around $\epsilon = 0.1$. The OpenScience project in social psychology found that only 47 of 100 replicated 95% confidence intervals contained their original effect sizes (Open Science Collaboration, 2015); in our experiments, this rate of CI coverage (47%) occurs only when the privacy guarantee is quite strong, $\epsilon < 0.01$ (Laplace mechanism).

So while the rates of epistemic disparity we observe are often higher than thresholds economists say they would accept, they are often lower than the rates of disparity observed in robustness and replication experiments.

2.4.2 Antecedents of epistemic disparity

What characteristics—of studies, their results, and the data involved—are most associated with epistemic disparities? Table 2.2 summarizes a linear model of the effect of various mechanism, data, and result characteristics. We regress over the average epistemic parity at $\alpha = 0.1$ for each result ($N = 177$) over all 10 simulations, controlling for study fixed effects. We include all experimental conditions: for privacy (both mechanisms), 50 values of ϵ spaced evenly on a log scale in $[10^{-5}, 10^{3}]$.

Data characteristics. Column 1 presents the impacts of the basic privacy mechanism. Each statistic in the datasets in our sample is the product of one or more statistical queries (noisy counts, noisy sums, and noisy means). Those queries receive random noise from the Laplace and Gaussian mechanisms based on (a) the privacy budget ϵ or $\rho = \epsilon\frac{\exp(\epsilon)-1}{\exp(\epsilon)+1}$ and (b) the global sensitivity of the query (Def. 2). Table B.1 demonstrates this model in isolation; the ratio of the root mean squared deviation applied to each component query to the original range of the unnoised query is almost completely explained by (1) $\log \epsilon$, (2) the ratio of sensitivity to the original range of the statistic, and (3) the choice of mechanism. The total noise added to each statistic depends on the noise added to its component queries and on the operation used to combine them (e.g., a mean reconstructed with noisy sum over noisy count); so in Column 1, we also control for the type of statistic.

As expected, epistemic parity decreases with sensitivity and increases with ϵ. Within the values of ϵ we tested, a 10-fold increase in ϵ corresponds to a 4.1 percentage point increase in the rate of epistemic parity. This estimate is robust to the introduction of other data- and result-related controls. Using zCDP with the Gaussian mechanism corresponds to a $3.5 - 0.2 \log \epsilon$ percentage point increase over the Laplace mechanism (the Gaussian mechanism performs slightly better for $\epsilon < 10$; see Fig. 2.1).

[14] Moreover, different analysts may disagree at similar rates when replicating with the *same* dataset (same data, different team, different methods). When given the same dataset and research question, 20 of 29 analysts found a significant positive effect while 9 did not observe a significant relationship (Silberzahn et al., 2018)—equivalent to 69% sign parity, similar to the rate of successful robustness checks (Brodeur et al., 2024).

Column 2 adds characteristics of the original regression analysis used to produce the findings. For example, all else equal, each additional covariate (model degree of freedom) reduces the rate of epistemic parity by 0.4 percentage points, even after controlling for the number of statistics noised.

Especially significant is the size of the population used to produce each statistic. We proxy for this variable by labeling the time and panel indices for each regression (Figs. B.3.8–B.3.9): compared to regressions over small-population statistics (county-, tract-, block-, school district-, prefecture-level), regressions over larger-population statistics (1st divisions including states, districts, and provinces) correspond to 32.4 percentage points higher epistemic parity on average. These larger area statistics usually have the same sensitivity as their smaller counterparts (e.g., a simple population count with sensitivity 1), and thus less noise proportional to their size. However, these effects are negligible when we control for the original effect size r.

After controlling for study fixed effects, we find no significant effect based on whether the independent (treatment) or dependent (outcome) variables are noised or whether the regression used instrumental variables. However, the privacy-utility curve for findings utilizing IV regressions is noticeably steeper than for other regressions; IV estimates tend to have result in lower rates of epistemic parity (Fig. B.3.6).

Epistemic factors. Notably, some disparities appear even when very little noise is added. Under the Hudson-Berger data error construction with a coefficient of variation of just 0.1%, 2% of the findings we reproduced are no longer supported; similarly, at $\epsilon = 1000$, an extremely low privacy setting, there are still disparities in around 2% of findings. The CI coverage and difference in effect size, which do not depend on the authors' claims about their results, are less sensitive to small amounts of noise. (The rate of 90% CI coverage is 99.2% with the Laplace mechanism at $\epsilon = 1000$.)

Column 3 examines the impact of characteristics related to the strength of authors' original epistemic claims. Findings based on initially weaker evidence are more likely to be contradicted by noisy estimates; a 0.1 decrease in the original effect size corresponds to a 2.74 percentage point decrease in the rate of epistemic parity. Figures 2.2 depicts this finding graphically with results from the ϵ-DP Laplace mechanism. This finding mirrors robustness research in social science; (Brodeur et al., 2024), for example, find that half of robustness checks of point estimates significant at $0.05 < p < 0.1$ succeed, compared to 70% of robustness checks of higher-effect point estimates $p < 0.05$.

This effect is primarily attributable to attenuation, the classic observation that regression estimates deflate and standard errors inflate in the presence of measurement error (Hausman, 2001; Carroll et al., 2006). Attenuation is clearly visible in the downward shift in the distribution of effect sizes as ϵ decreases (Figs. B.3.1, B.3.3). In this ideal case, estimates with additional measurement error are systematically smaller in magnitude, below the original effect size; most of the disparities we document fit this profile (Figure 2.2a). In their robustness checks of 110 social science studies, Brodeur et al. (2024) observe significance levels 77% of the original; in their replication of 18 economics laboratory experiments, Camerer et al. (2016) similarly observe effect sizes attenuated to 66% of their original value on average.

But in noisier (e.g., lower-powered) settings where the true effect is small, measurement error may increase, not decrease, estimates even as the standard errors increase (Loken & Gelman, 2017). Selecting for significance could lead authors to make positive claims about seemingly large estimates when in fact they are drawn from the upper end of the distribution of possible effects. This phenomena could help explain why some disparities occur even when little noise is added—the claims are especially sensitive to small amounts of noise despite relatively little change to the effect size (Figure 2.2a). For $\epsilon > 1$, epistemic disparities come almost exclusively from findings with initially lower effect sizes (e.g., $r < 0.6$; Fig. 2.2a).

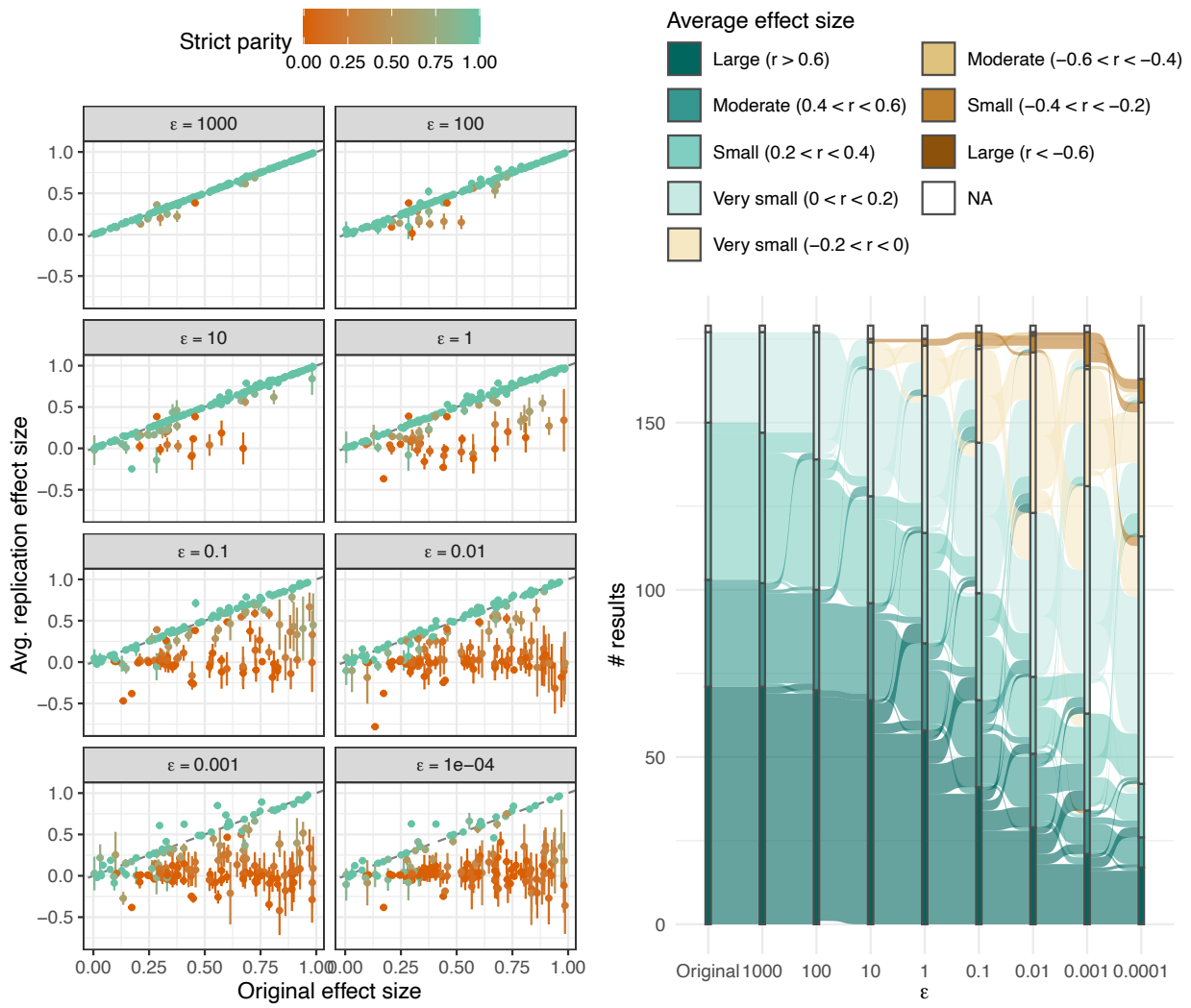
The type of epistemic claim also matters. There are almost no disparities in null findings (when the original claim requires a statistically insignificant coefficient). Most disparities occur when the original claim requires a coefficient with a certain sign or within a certain interval (Figure B.3.5).

Consequences. What do these antecedents mean for different types of research? In general, the preceding analysis suggests that studies or subfields will be particularly impacted if they use more social statistics, rely on statistics about smaller subgroups (e.g., blocks and counties), use more covariates, make more quantitatively specific claims, or rely on weaker effect sizes.

Fig. B.3.10 presents the privacy-utility curves for the 42 studies in our sample published with Journal of Economic Literature classification codes. (The analysis that follows focuses on results from the zCDP Gaussian mechanism applied to the original statistics.) Of the top-level categories with more than 10 findings, studies in Industrial Organization and Macroeconomics & Monetary Economics have noticeably higher rates of epistemic parity compared to areas such as Health, Education, & Welfare Economics and Labor & Demographic Economics, likely because those subfields tend to use more social statistics and examine smaller populations (Fig. B.3.10).

We observe similar trends in the rate of epistemic parity across types of personal data (Fig. B.3.11). Regressions using statistics about age, education, and sex/gender tend to have lower rates of epistemic parity, possibly because these variables are often crossed with other variables to create statistics about smaller subgroups. This could have consequences for studies of subgroup differences, an ongoing source of concern for researchers (Santos-Lozada et al., 2020). Election and crime statistics tend to have slightly higher rates of parity, maybe because these statistics in our datasets mostly involve large, simple counts (e.g., of votes or crimes committed).

We also cross-section the privacy utility rates by the type of data source: administrative, census, survey, or blended estimates. We find that rates of epistemic parity are already generally lower for survey products, and generally higher for blended estimates under our assumptions (Figure B.3.12). The trade-off curve for census products is nearly flat for $\epsilon > 1$; otherwise, the curves for most data types are similar to the average. However, the type of data product can have great effects on the noise required for differential privacy which are not accounted for in our mechanisms. For example, in our sensitivity analysis, we assume each person contributes equally to each statistic, but many survey products and other statistical estimations are weighted to give more influence to certain subgroups. And while applying



(a) Bars depict a 90% confidence interval. Epistemic parity computed at significance level $p < 0.1$.

(b) Horizontal bars depict the number of results whose average effect size change from one bin to another as ϵ decreases. NA indicates reproductions that failed to complete (e.g., a regression failed to converge).

Figure 2.2: Change in average standardized effect size (r) over 10 replicates as a result of applying the Laplace mechanism with increasing privacy parameter ϵ. Lower values of the privacy parameter ϵ provide stronger privacy protection but require more injected noise; curators commonly use $0.1 \leq \epsilon \leq 10$.

an algorithm to a simple random sample can yield smaller privacy loss than applying the same algorithm to the entire population, this amplification effect may not hold for other survey designs (Bun et al., 2023; Drechsler & Bailie, 2024; Seeman et al., 2024). Work is still ongoing to explore these considerations and develop optimal DP-compatible design strategies for surveys (U.S. Census Bureau, 2022; Seeman et al., 2024; Drechsler & Bailie, 2024) and small-area estimation (Seeman et al., 2020).

2.4.3 Impacts of data error

Of course, social statistics contain many different kinds of existing data error, even before data privacy protections are applied (Groves & Lyberg, 2010; Steed, Liu, et al., 2022).

We find that comparable epistemic disparities occur if statistics contain even modest amounts of additive data error. For example, U.S. Census Bureau American Community Survey guidance describes a coefficient of variation of 1.5% as "very small" and "very reliable" for aggregated count data (U.S. Census Bureau, 2018). Simulating data error with a coefficient of variation of just 1% results in sign disparity in 5% of counterfactual findings (Hudson-Berger construction). A 5% coefficient of variation increases sign disparity to 11%. Under the Morris-Lysy construction, disparities are much, much higher—even just a coefficient of variation of 0.1% results in sign disparities in 20% of findings.

These choices of the coefficient of variation may be quite conservative. Prior to 2022, Census Bureau statistical quality standard F1-6 required that the majority of key estimates have a coefficient of variation less than 30% (U.S. Census Bureau, 2013); more complicated data products, such as the school district-level Small Area Income and Poverty Estimates, are published with coefficients as high as 67% (Maples, 2008).

Again, according to self-reported preferences, researchers may find these *existing* discrepancies unacceptable. Of the 1,028 American Economic Association members surveyed, only around 60% of those surveyed said they would accept a 10% rate of changing significance (from $p < 0.05$) (Williams, Snoke, et al., 2024); the rate of sign parity reaches 13% from data error with just a 5% coefficient of variation (Hudson-Berger, $\alpha = 0.05$). Moreover, researchers may well be overconfident about the robustness of existing research. Asked to predict the robustness of 17 non-experimental American Economic Review papers, 359 economists overestimated their robustness reproducibility (46% sign parity on average at $p < 0.05$) by about 15 percentage points (Campbell et al., 2024). Similar overconfidence may extend to robustness to existing data error.

Given the potentially large effects of additive error on epistemic disparity in research based on counterfactual productions of these social statistics, it is crucial to also evaluate privacy mechanisms *on the margin* of these existing errors. The third panel of Figure 2.1 presents the marginal impact of adding the $\epsilon\frac{\exp(\epsilon)-1}{\exp(\epsilon)+1}$-zCDP Gaussian mechanism after simulating data error with the Hudson-Berger construction. Even with a small amount of data error (coeff. of variation 0.1%), a budget of $\epsilon = 10$ increases epistemic disparity by 7 percentage points in addition to the base effects of data error—less than the 9% rate of disparity estimated without accounting for data error (zCDP Gaussian mechanism). Moreover, the marginal impact decreases as the magnitude of simulated data error increases: when the coefficient of variation is 20%, still below the Census Bureau's 2022 quality standard, the privacy mechanism increases disparity by a mere 0.6 percentage points.

2.5 Discussion

At first glance, these findings paint a bleak picture—modest amounts of noise (from either differential privacy mechanisms *or* existing data error) can have noticeable impacts on the scientific findings in our sample. Should authors continue to use the same procedures, as we imagine here, the potential impacts of existing data error, much less widespread adoption of differential privacy, appear non-trivial. These findings echo replication and robustness concerns across fields of science: numerous studies have shown that results from psychology and economics tend to replicate with weaker effects and often do not support the original conclusions at rates often similar to or higher than what we observe (Klein et al., 2014; Open Science Collaboration, 2015; Camerer et al., 2016; Silberzahn et al., 2018). Other studies similarly show how findings can be sensitive to small changes in the data; Broderick et al. (2023), for example, find that the results of some economics papers can be overturned by removing less than 1% of the sample, even when t-statistics are very large.

But there is an upside: acknowledging and accounting for the impacts of existing data error clarifies the trade-offs associated with privacy protections. For example, we show that the relative impacts of differential privacy mechanisms are noticeably slimmer after accounting for the effects of existing data error. Whether these results translate to real-world deployments depends on several factors. Our privacy mechanisms are simple and generic; with more effort and expertise, their design could be optimized to the individual context of each data product. Moreover, we make several unrealistic assumptions: not all statistics in each study are likely to be made private; not all organizations may implement differential privacy for every statistic; and some statistics may remain invariant for operational or legal reasons, as in the 2020 Decennial Census (Abowd et al., 2022). Still, there are ways our setup may underestimate the impact of differential privacy: in the real world, these statistics may have to share privacy budget with other statistics, and for survey, blended, and estimated statistics, the noise required to achieve DP may be higher than we assume here. Future work may build on these empirical findings to explore optimal mechanism designs and privacy budget allocations tailored for social science research.

More importantly, this framing contradicts the assumption that differential privacy poses entirely novel problems for science. Social science data already includes measurement error, missing values, and other distortions; differential privacy may be viewed as only a modern addition to existing sources of data error (Steed, Liu, et al., 2022; Gong, 2022; Groves & Lyberg, 2010). As our results demonstrate, statistical significance is a less useful signal in noisier settings and may result in exaggerated estimates of effect size (Loken & Gelman, 2017). There are long-established methods for calibrating regression estimates to account for additive error (Carroll et al., 2006; Cook & Stefanski, 1994). Ideally, authors could "robustify" their claims with noise-aware data processing procedures. Recent work proposes method for multiple imputation and data cleaning for differentially private data specifically (Blackwell et al., 2017; Evans & King, 2023). Agarwal and Singh (2024), for example, propose a method that could recover the main results of Autor et al. (2013) (also included in our sample) with equivalent precision down to $\epsilon = 0.01$. There are already examples of this kind of adaptation in practice: authors developed new methods to account for injected noise when Meta released its differentially private Facebook URLs Dataset (Buntain et al., 2021; Evans & King, 2023).

However, these efforts are nascent and not widely known; indeed, the majority of economists are not even familiar with differential privacy (Williams, Snoke, et al., 2024).

While adopting new methods is costly, there can be benefits to scientific practice. For example, Dwork et al. (2015) and Echenique and He (2024) show that by enforcing provable stability in the outputs of statistical analyses, differentially private methods can also inhibit p-hacking and other dubious scientific techniques that rely on overfitting and adaptive data slicing. Indeed, we show that weaker initial effects and overly precise claims are less likely to replicate under differential privacy. Exploring these solutions may call into question the sharp privacy-utility trade-off commonly assumed in policy debates and point the way towards practices that are both more private and more robust.

Acknowledgments

We sincerely thank our research assistants Xingyu Chen, Annie Qian, and Donna Zhu for their contributions to our replication framework and dataset. Thanks to Zeid El-Kilani, Roy Rinberg, and Steven Wu for feedback on earlier drafts of this work.

	(1)	(2)	(3)
(Intercept)	0.937 ***	0.782 ***	0.736 ***
	(0.125)	(0.179)	(0.184)
Log epsilon	0.041 ***	0.041 ***	0.041 ***
	(0.003)	(0.003)	(0.003)
Gaussian mech. (zCDP)	0.035 ***	0.035 ***	0.035 ***
	(0.008)	(0.008)	(0.008)
Log epsilon x Gaussian mech. (zCDP)	-0.002 **	-0.002 **	-0.002 **
	(0.001)	(0.001)	(0.001)
Log sum of sensitivity	-0.019 *	-0.013 ***	-0.014 ***
	(0.009)	(0.004)	(0.003)
Log sum of sensitivity x Gaussian mech. (zCDP)	0.000	0.000	0.000
	(0.001)	(0.001)	(0.001)
Log regression sample size		0.022	0.011
		(0.014)	(0.014)
Model degrees of freedom		-0.004 ***	-0.002 ***
		(0.001)	(0.001)
Noised ind./treatment var.		-0.072	0.143
		(0.234)	(0.222)
Noised dep/outcome var.		-0.031	0.117
		(0.068)	(0.076)
Cmd: ivreg, ivreg2, xtivreg2, ivregress, ivreghdfe		0.147	0.094
		(0.126)	(0.089)
Cmd: other (arima, nbreg)		0.091	0.279
		(0.257)	(0.266)
Region: City/municipality/MSA/commuting zone		0.024	-0.215
		(0.148)	(0.148)
Region: State/district/province (1st division)		0.324 *	0.124
		(0.148)	(0.150)
Original effect size			0.274 *
			(0.122)
Claim: insignificant			0.357 ***
			(0.081)
Claim: non-zero upper/lower bound			-0.099
			(0.072)
Study FE	Yes	Yes	Yes
Implementation controls	Yes	Yes	Yes
Query type controls	Yes	Yes	Yes
N. obs.	20178	19950	19950
R squared	0.559	0.569	0.606
F statistic	417.420	385.776	431.164
p value	0.000	0.000	0.000

*** $p < 0.001$; ** $p < 0.01$; * $p < 0.05$. Robust standard errors are clustered by study.

Table 2.2: Impact of mechanism & result characteristics on average epistemic parity over 10 replicates. Implementation controls include number of variables noised and number of component queries. Dummies for "Cmd: reg, xtreg", "Claim: sig. positive/negative", and "Region: County/tract/block/district/prefecture (below 1st division)" excluded. Includes all experimental conditions for privacy (both mechanisms): 50 values of ϵ spaced evenly on a log scale in $[10^{-5}, 10^{3}]$.

Chapter 3

Algorithmic Decoupling in 'Privacy-Preserving' Analytics

This chapter is reproduced from:

Steed, R., & Acquisti, A. (2025). *Algorithmic Decoupling and the Adoption of 'Privacy-Preserving' Analytics*. https://doi.org/10.2139/ssrn.4718865

3.1 Introduction

Facing consumers' calls for privacy and policymakers' threats of regulatory action, technology leaders are endorsing "privacy-preserving" techniques for data analytics as the future of digital privacy (Egan, 2020). Organizations in industry and government—including Meta, Google, Apple, Microsoft, Wikipedia, Mozilla, the U.S. Census Bureau, and the Internal Revenue Service (Desfontaines, 2021)—are pioneering deployments of differential privacy (Dwork et al., 2006), federated learning (McMahan et al., 2017), and other algorithmic approaches to reconcile analytics and privacy.

On its face, consumers and regulators have reason to hope adoption will improve digital privacy: these mathematical, statistical, and algorithmic techniques—which we will broadly refer to as *privacy-preserving analytics* (PPA)—are used to produce useful insights from people's personal data while still preserving some technical definition of data privacy. But these techniques are complex in theory and implementation, and their actual impacts on consumer privacy and social welfare are often untested. Privacy advocates (Cyphers, 2019) and advertisers (Lomas, 2021) alike, for instance, have raised doubts about Google's privacy-preserving plans for the future of targeted advertising (Goel, 2022). PPA adoption could present new benefits to researchers and consumers—or, as some fear, it could simply preserve extractive, surveillance-based economies (Cohen, 2018; Zuboff, 2019; McGuigan et al., 2023).

Despite a flourishing technical literature in computer science and statistics, little IS, economics, or social science research has examined organizational processes driving the adoption and deployment of PPA systems. What drives organizations to adopt privacy-preserving technologies? How do these drivers inform design choices made during deployment? And how does PPA adoption change the relationship between policy and practice in the organization?

In absence of empirical studies of PPA adoption in organizations, prior research does, however, offer possible theoretical answers to these questions. Some streams of organizational research suggest that organizations innovate socially beneficial technologies in response to changing consumer expectations and to regulatory pressure, such as for environmental protection (Ashford et al., 1985). But institutional research also suggests that policies adopted to satisfy consumers and regulators may at times be merely symbolic—"decoupled" from substantive changes to daily practice (Edelman, 2016; Bromley & Powell, 2012). On this trajectory, PPA adoption could amount to little more than "privacy theater": gestures accorded deference incommensurate with their actual benefits to online privacy (Soghoian, 2011).

We investigate the motivations and decision-making processes behind PPA adoption and design across 21 large technology firms and startups, non-profits, and government agencies. We interviewed 28 executives, managers, and key contributors in technical, legal, and policy roles responsible for deciding whether to adopt and how to deploy PPA systems. The firms and agencies they work for include many of the organizations leading PPA adoption in the United States. Analyzing transcripts and documents, we used grounded theory methodology (Charmaz, 2014) to develop a model of the organizational processes that led to and shaped adoption.

Our primary finding is that, during implementation, PPA design choices were constrained and sometimes dominated by operational concerns, disconnecting algorithmic system design from both internal policies and external representations—a special form of organizational decoupling we term *algorithmic decoupling*. Both public and private sector organizations adopted PPA systems primarily to preserve existing modes of operation against new regulations or consumer expectations, though some organizations leveraged PPAs to use data in new ways. The processes involved with deciding when to use PPA techniques and interpreting privacy expectations into algorithmic designs often prioritized these managerial interests. Through infrastructure sharing and active standard-setting, these practices set the bar for future industry implementations and even for regulatory guidance. However, we also find that morally motivated privacy "champions" (Tahaei et al., 2021) constituted a countervailing force against algorithmic decoupling within the organization. They had significant leverage over the interpretation of privacy laws and internal policies on the one hand and the technical properties of PPA systems on the other, mediating between managerial, legal, and technical concerns. A subset of these practitioners—particularly in the private sector—used their influence to push for PPA adoption and defend privacy standards, often because of their own ethical and professional commitments.

Connecting our findings with research from sociology, law, and technology studies, this study makes two key contributions to research on information systems (IS), organizational behavior, and privacy. First, we conceptualize algorithmic decoupling: ways in which the aspects of organizational practice embedded in algorithmic systems may be uniquely decoupled from formal policy and external expectations and may uniquely alter those expectations in turn. Prior work shows how practitioners' interpretations of the legal environment mediate the effect of regulation (Fuller et al., 2000). Algorithmic decoupling draws on another dimension of ambiguity: practitioners' *technological* interpretations of algorithmic systems employed to fulfill formal policies. Algorithmic decoupling helps to explain why even significant investments in PPA adoption often fall short of public privacy expectations (Martin et al., 2023). While our current study focuses on algorithms for private data processing, the proliferation of algorithmic systems in other parts of the market and society suggests that the propagation of similarly motivated innovations may not be an unalloyed benefit to society, if they are overly mediated by managerial concerns. Second, our findings suggest that although algorithmic ambiguity leaves room for decoupling, it also uniquely empowers expert PPA practitioners. Where prior work focused on the role of executives' ethical commitments in preventing decoupling (Weaver et al., 1999) or the role of lawyers in framing the legal environment (Edelman, 2016), algorithmic decoupling is influenced most by technologists' commitments to privacy. Our findings offer a guide for scholars, managers, and policymakers to more critically evaluate and intervene in the use of algorithmic systems to fulfill social responsibilities.

3.2 Theoretical Background

3.2.1 Technology Adoption and Social Performance

Though organizational IS research on PPA adoption is scant, our work builds upon decades of research on the social benefits of digital innovation in IS (Yoo et al., 2010), particularly related

to sustainability (Malhotra et al., 2013; Hanelt et al., 2017). At the firm-level, much IS research is devoted to identifying the antecedents (resources, competitive environment, management style, etc.) of information technology adoption and its effects on financial and operational performance (Fichman, 2004). Less research explores how organizations adopt IS technologies to improve *social* performance, efforts to fulfill social responsibilities alongside economic gains (Davis, 1973; Carroll, 1979; Orlitzky & Benjamin, 2001)—responsibilities which, some argue, now include data privacy (Pollach, 2011). In the last decade, technology firms have developed and adopted a number of algorithmic innovations to address privacy, fairness, sustainability, and other ethical concerns with data processing and artificial intelligence (AI) practices (Bamberger & Mulligan, 2015; Hirsch et al., 2020; Metcalf et al., 2019; Morozov, 2013). Recent IS research, for example, examines the drivers and performance benefits of "green" IS practices (including smart grids and building automation) adopted to improve environmental sustainability (Seidel et al., 2013; Hanelt et al., 2017; Malhotra et al., 2013; Leidner et al., 2022; Hu et al., 2016; Loeser et al., 2017; Ketter et al., 2023).

Earlier economic theories explain these practices simply as strategic, cost-saving adaptations to changes in regulation (Oliver, 1991). Stricter environmental standards, for example, forced firms to innovate new technologies such as the catalytic converter to avoid financial penalties (Ashford et al., 1985). A wide-ranging literature from institutional theory (Scott, 2007), on the other hand, explains organizational behaviors as a product of "rational myths"—widely-accepted ideas about how organizations should act, conditioned on historical, cultural, and social context (Meyer & Rowan, 1977; DiMaggio & Powell, 1983). Adoption is mediated not only by strict economic rationality but also by the expectations of activists, competitors, investors, employees, consumers, and other stakeholders (Campbell, 2007; Boldosova, 2019; Aguilera et al., 2007; Jones, 1995). Under this theory, adopting socially beneficial technologies helps organizations maintain their social license to operate (Gunningham et al., 2004).

3.2.2 Algorithmic Decoupling as a Dimension of Organizational Decoupling

A key observation of contemporary institutional theory is that organizations facing these external pressures may partially or completely *decouple* the performance of daily practices ("performative" aspect) from their presentation in formal structures and policies ('ostensive" aspect) (Oliver, 1991; Bromley & Powell, 2012; Boxenbaum & Jonsson, 2017; Feldman & Pentland, 2003). The concept of decoupling arises from early observations that organizations can maintain contradictory institutional logics by insulating inconsistent, yet responsive, practices from one another—a phenomenon referred to in institutional theory as "loose coupling" (Weick, 1976; Meyer & Rowan, 1977; Orton & Weick, 1990; Hallett & Hawbaker, 2021). This perspective has proved useful for analyzing IS adoption (Strong & Volkoff, 2010; Chen et al., 2011). Berente and Yoo (2012), for example, use loose coupling to explain improvisational user responses to enterprise IS adoption as a resolution to the friction between abstract software and local contexts.

For organizations facing strong but ambiguous or contradictory external and internal expectations (Powell & DiMaggio, 2023; Scott, 2007), decoupling—sometimes referred to as

"organized hypocrisy" (Brunsson, 2003; Lim & Tsutsui, 2011)—serves as a buffer between internal practices and external pressures and as a key component of organizational legitimacy (Meyer & Rowan, 1977; Weber, 1978; Suchman, 1995). Several studies examine decoupling in the context of social performance efforts (Lim & Tsutsui, 2011; Schoeneborn et al., 2020; Dobbin & Kalev, 2022; Marquis & Qian, 2014; Li & Wu, 2020). For example, some firms obtained green technology certificates from the Korean government without actually implementing those technologies in daily operations (Park & Cha, 2019). Decoupling the formal adoption and espousal of these efforts from their practice allows organizations to reduce costs while avoiding legal sanctions and reputational harms (Bromley & Powell, 2012). However, nearly all studies of organizational decoupling focus on primarily non-technological practices (see, e.g., Bromley & Powell, 2012, Table 2). And while IS research explores the possibility of loose coupling as a response to institutional contradictions in enterprise system implementations (Berente & Yoo, 2012; Keller et al., 2019; Baptista et al., 2021; Chen et al., 2011), the interaction between organizational *decoupling* and IS has not been fully explored.

IS research on technology-mediated organizational change explores how organizational routines are *materially embedded* in enterprise IS (Volkoff et al., 2007). When information systems are adopted, routines, roles, and other organizational structures are constrained and modified by material aspects of the system as built (Silva & Hirschheim, 2007; Berente & Yoo, 2012). IS artifacts, then, constitute a form of embedded, often invisible, organizational regulation (de Vaujany et al., 2018; Hennigsson & Eaton, 2024)—a topic of nascent IS research agendas (Butler et al., 2023; de Vaujany et al., 2018). de Vaujany et al. (2018) call for further research on the "materialization" of rules in IT artifacts, in addition to temporal decoupling between design time and use time; this study explores the processes involved with rules materialization and the consequences for regulation & compliance. Technological embeddedness introduces the possibility of decoupling the presentation of organizational practice (its ostensive aspect) not only from its performance (performative aspect), but also from the aspect of practice embedded in information systems (material aspect).

Algorithmic decoupling helps describe this additional dimension: the gap between policy and the technical and material properties of the algorithmic system deployed to fulfill it. Research on technology and social performance often frames the adoption of technologies like the catalytic converter as uniformly implemented and categorically beneficial (Ashford et al., 1985); algorithmic decoupling accounts for the reality that technology adoption is contextual and adapted, its design mediated by the organization. In the context of PPA, privacy advocates' criticisms of several prominent, public proposals provide an early indication that organizations' claims have not been fulfilled by their technological designs (Cyphers, 2019; McGuigan et al., 2023; Martin et al., 2023); algorithmic decoupling helps to explain these shortcomings. In our study, we explore how the use of algorithmic systems complicates existing theory about the mediators of and remedies to organizational decoupling.

3.2.3 From Algorithmic Decoupling to Perverse Innovation

Decoupling makes clear that organizational responses to external pressures are not determined. Organizations and their constituents mediate the impact of the institutional environment on the adoption of new practices (Edelman, 2016; Oliver, 1991). But institutional research on

heterogeneous diffusion also explores the influence of adoption on the institutional environment in turn (Powell & DiMaggio, 2023).

In particular, organizations model their early "educated guesses" at compliance to their peers and competitors. As in the case of equal opportunity, employment law, and insider trading, these initial guesses are often legitimated by courts and policymakers and become standard practice (Edelman, 2016; Bozanic et al., 2012). After defendants began instituting grievance procedures to forestall unionization and insulate against discrimination suits, for example, courts and legal journals increasingly considered those procedures relevant to liability despite little evidence that they actually reduced complaints (Sutton & Dobbin, 1996; Edelman et al., 1999; Dobbin & Kalev, 2022). Edelman (2016) calls this phenomenon *legal endogeneity*: after organizations decide what forms of compliance are reasonable, those practices become institutionalized as rational responses to regulation. Private organizations may also engage directly in lobbying, corporate-sponsored research, and other forms of regulatory capture to promote their versions of compliance (Hillman et al., 2004; Kamieniecki, 2006). Technology firms in particular, such as Airbnb and Uber, are exemplars of "regulatory entrepreneurship": the pursuit of business models that are predicated on changing the law (Pollman & Barry, 2016).

When organizational practices are mediated by algorithms and IS, regulatory entrepreneurship may be accomplished with technological innovation. Burk (2016) uses the term "perverse innovation" to describe technological innovation directed at exploiting loopholes in formal rules. Seed producers in the E.U., for example, avoided restrictions on genetically-modified crops by replacing recombinant DNA technologies with mutagenic chemicals, an alternative approach with possibly greater health and safety risks; and the PT Cruiser was designed with the footprint of a "small truck" to allow Chrysler to avoid stricter EPA fuel efficiency requirements for "passenger cars" (Burk, 2016).

Algorithmic decoupling is perverse innovation when and if the implemented algorithmic system is not only disconnected from but contrary to expected social benefits (e.g., "privacy-preserving" technologies that increase data collection without providing substantive privacy benefits). Like other compliance practices, technological designs may set legal precedent. Algorithmic decoupling helps to explain how perverse practices may become institutional standards not only through sociolegal mechanisms but also through sociotechnical mechanisms, primarily cloud platform-dependence (Narayan, 2022; Cutolo & Kenney, 2021)and open source innovation (West & Gallagher, 2006).

3.2.4 Privacy-Preserving Analytics

This study explores organizational decoupling in the context of a burgeoning area of IS technology for social performance: privacy-preserving analytics (PPA). Privacy technology is not new—we define PPA techniques as a particular subset of privacy enhancing technologies (PETs),[1] a variety of tools used by consumers, regulators, and organizations to negotiate information privacy issues for over three decades (Goldberg, 2007). Privacy is a multifaceted,

[1]We do not include PETs for private communication or authentication (see, e.g., Domingo-Ferrer & Blanco-Justicia, 2020)—we are specifically concerned with technologies used in data analytics.

context-dependent social concept associated with a wide range of attitudes and behaviors (Dinev et al., 2015; Acquisti et al., 2015; Bélanger & James, 2020; Belanger & Crossler, 2011). Likewise, while all the systems referred to as "privacy-preserving" in our study were used in practice to govern user data processing, the methods—and the precise definition of privacy preserved—varied (McGuigan et al., 2023). This study focuses on techniques and standards used to preserve privacy in both the inputs to and the outputs of data analysis, such as secure multiparty computation (Goldreich, 2009)(which describes cryptographic protocols for distributed computing designed not to reveal private inputs), differential privacy (a formal guarantee that outputs of analysis are not sensitive to the inclusion of any one individual's information, usually accomplished by noise injection (Dwork et al., 2006)), and federated learning (which describes techniques for training machine learning models without transferring raw data off client devices (McMahan et al., 2017)). Table 3.1 lists all the PPA practices adopted by our participants.

Organizations have used differential privacy, for example, to send COVID-19 exposure notifications and auto-complete text or emojis (Apple, Google), collect telemetry (Microsoft Windows), share data with clients and researchers (Meta, LinkedIn, Microsoft, Google, U.S. Census Bureau), and more (Desfontaines, 2021). In fact, probably spurred by regulatory initiatives in the U.S. and around the world, the number of private and public sector organizations adopting and deploying PPAs has significantly increased in recent years. New startups such as Tumult Labs are offering PPA consulting services and building open-source software. And cryptographic and federated methods—such as Google's Privacy Sandbox (Goel, 2022)—may soon replace key aspects of online advertising.

With respect to privacy practices in general, there is evidence of decoupling in existing research: Waldman (2018) distinguishes CPOs' privacy myth-making efforts from their actual performance by technologists on the ground, who had little material incentive to enact new privacy agendas, and several studies critique the claims to privacy made by public PPA proposals (McGuigan et al., 2023; Tang et al., 2017; Berke & Calacci, 2022; Martin et al., 2023). But the organizational processes behind PPA adoption specifically—and the technological aspects of decoupling in general—are less understood. And the resulting impacts—on digital privacy as well as data science, social science research, policymaking, and other data-dependent processes (Abowd & Schmutte, 2019; Hotz et al., 2022)—are largely untested outside primarily theoretical research in computer science and statistics (Acquisti & Steed, 2023).

A few interview studies explore practitioners' challenges with differential privacy adoption specifically, but these studies mostly focus on usability (Dwork et al., 2019; Munilla Garrido et al., 2023; Sarathy et al., 2023; Ngong et al., 2024; Rosenblatt et al., 2024). Some critical studies evaluate PPA proposals on technical or philosophical grounds (Tang et al., 2017; Berke & Calacci, 2022; McGuigan et al., 2023; Martin et al., 2023; Smart et al., 2022) and explore challenges with communication and participation during adoption (boyd & Sarathy, 2022; Abdu et al., 2024). Other studies investigate data ethics practices (Hirsch et al., 2020) and privacy practices for artificial intelligence (AI) products (Lee et al., 2024). But little research examines organizational aspects of these technologies. By investigating this question, our study contributes to the ongoing project of documenting and describing the social impact

of PPA technologies.

3.3 Methods

This research is based on a seven-month qualitative study of PPA adoption through semi-structured interviews with practitioners—including engineers, lawyers, managers, researchers, policy experts, and executives—at technology firms, privacy-focused startups, non-profits, and government agencies. These organizations include many of the most prominent deployments of PPA to date in the United States. Research on organizational adoption of PPA is scant, but qualitative methods have a long, impactful history of helping researchers theorize about emerging phenomena in IS and management (Monteiro et al., 2022; Wiesche et al., 2017; Edmondson & Mcmanus, 2007).

3.3.1 Data Collection

Our data are comprised primarily of IRB-approved interviews with 28 individuals responsible for helping their organizations decide whether and how to implement PPA systems. Table 3.1) describes their roles and technologies used. We contacted practitioners working on PPA products or services at organizations that had considered deploying PPA, though not all have actually deployed a PPA system. (All had made it at least as far as prototyping.) We sourced interview candidates either from professional networks (18 contacted, $N = 9$ interviewed) or known to both authors through public PPA work (7 contacted, $N = 5$ interviewed). We also asked those candidates to recommend one or two others at their organization (23 contacted, $N = 14$ interviewed). SM Appendix C.3 contains additional details about our recruitment strategy.

We designed our sample to explore theoretical variation between different adoption settings, aiming for analytical generalization from case studies to theory (Eisenhardt, 1989; Lee & Baskerville, 2003). Our sample includes 21 organizations: eight technology firms ($N = 10$), six in the Fortune 500 ($N = 8$); five privacy-focused startups ($N = 6$), organizations with privacy-branded products or offering PPA as a service; four non-profits ($N = 5$); and representatives from three U.S. government agencies ($N = 3$), two responsible for federal data collection and public statistics and one regulatory agency. For large organizations with large-scale or wide-ranging PPA activities, we recruited at least two or more participants, to add alternative perspectives on the same processes.

After we analyzed this first round of data, we conducted a second round of interviews between July and August 2023 with three practitioners in legal and policy roles to validate our understanding of how private firms interact with regulators about PPA adoption. We stopped data collection when our categories reached theoretical saturation, such that further interviews would spark no new insights (Charmaz, 2014). Each participant completed a short demographic survey and participated in a 50–100 minute interview (57 minutes, median) through video conferencing, under the condition that their identities were kept confidential. Interviews centered on open-ended questions about organizational processes involved with 1) the decision to adopt PPA, including motivations and trade-offs; 2) design and deployment,

Employer	PPA practices mentioned	Roles	Participants
Startup	DP, SDL, pseudonymization, encryption, minimization, cohort analytics, k-anonymity, deletion, PII detection, other cryptography	Director/Executive, Software/Privacy Engineer	P8, P13, P15,P17, P18, P24
Other for-profit	DP, SDL, k-anonymity, l-diversity, FL, HE, SMC, synthetic data, PPML, private set intersection, encryption, access control, retention limits, cohort analytics, other cryptography	Director/Executive, Manager, Software/ Privacy Engineer, Data Scientist, Researcher	P2, P3, P4, P7, P11, P12, P14, P19, P20, P23, P25, P26, P28
Non-profit	k-anonymity, DP, minimization, deletion, retention limits, SDL, other cryptography	Director/Executive, Engineer, Researcher	P1, P9, P10, P22, P27
Government	DP, SDL, noise infusion, SMC	Director/Executive, Software Engineer	P6, P16, P21

DP: differential privacy. FL: federated learning. HE: homomorphic encryption (Gentry, 2009). SMC: secure multi-party computation. PPML: privacy-preserving machine learning (e.g., Abadi et al., 2016). SDL: statistical disclosure limitation (Matthews & Harel, 2011)(e.g., suppression, data swapping).

Table 3.1: Practitioners interviewed. Participants were given differential privacy (DP) and federated learning (FL) as examples but were allowed to name any practices they used for PPA.

especially communication, common challenges faced, and future trends. Interviews were semi-structured, co-constructed by the interviewer and the participant to allow flexibility to explore new phenomena (Charmaz, 2014). We piloted our initial interview guide using two think-aloud interviews (Willis & Artino, 2013)with colleagues and three practice interviews with volunteer junior practitioners. While we asked about PPA adoption in every interview, we adjusted the protocol to explore different areas of theoretical interest over the course of the study. (SM Appendix C.2 provides our interview guide.) We supplemented interview transcripts with internal documentation provided by participants, press releases, white papers, blogs, news articles, and other archival documents.

3.3.2 Analysis

As is common in inductive research and grounded theory (Gioia, Corley, & Hamilton, 2013; Charmaz, 2014), qualitative analysis alternated continuously between 1) first-order, primarily inductive creation of analytic codes, 2) second-order aggregation and abductive theoretical analysis, and 3) written and visual presentation of our emergent theoretical model, grounded in first-order quotations. In first-order analysis, the first author annotated transcripts with short, precise descriptions—over 2,500 unique codes—staying grounded in the participants' language and focusing on actions and processes to avoid preconceived framing (Charmaz, 2014). Early on, the second author re-coded a sample of four interviews and provided critical feedback to calibrate our coding.

In second-order analysis, we critically sorted and synthesized initial codes to draw out

hypotheses and narratives (axial & theoretical coding). We began to define tentative concepts by comparing first-order codes and excerpts and by comparing the accounts of different participants. At this stage, we adopted a theoretically agnostic stance, permitting extant concepts (such as "scaling up") only when they fit our data and first-order analysis (Charmaz, 2014). We also compared and generalized across types of organizations and industries, similar to a case study design (Eisenhardt, 1989). We gradually arranged concepts in multi-layer hierarchies and eventually three themes describing the overarching processes involved with PPA adoption (Figure 3.1) and mapped the relationships between them with written notes and iterative process diagramming (Figures C.4.1–C.4.3). As in the first-order analysis, the authors discussed and exchanged notes and diagrams describing the concepts and their relationships until a consistent process model emerged (Gioia, Corley, & Hamilton, 2013). Analysis occurred alongside data collection, and we continuously adjusted our interview guide to follow up on topics of theoretical interest (Charmaz, 2014).

3.4 Case Study: Adoption of Privacy-Preserving Analytics

In the following analysis, we trace the couplings between formal policy, its implementation, and its outcomes, and highlight where these couplings are likely to break. First, we describe the ways organizations constructed PPA adoption as an appropriate response to external privacy expectations (§3.4.1). Second, we describe the processes by which those narratives were interpreted into specific technological design choices—choices potentially decoupled from policies or outcomes (§3.4.2). Third, we describe the ways that organizations justified PPA adoption to satisfy stakeholder expectations, setting a precedent for future adoption (§3.4.3).

3.4.1 Deciding to Adopt

Organizational investment in PPA technologies specifically has increased sharply in the last decade following "loud and furious" (P4) public and regulatory pressure embodied by sweeping data privacy regulations, including the General Data Protection Regulation (GDPR) and the California Consumer Privacy Act (CCPA), coupled with publicized data privacy scandals—for example, the news that Cambridge Analytica had deceptively amassed the personal data of millions of Facebook users (Confessore, 2018).

Across the array of organizations we studied, practitioners consistently pointed to these external privacy expectations—embodied particularly by regulators, the media, and consumer advocates—as the root of their motivation to develop and deploy PPA systems. All but three practitioners named the threat of regulation—including fines, lawsuits, and, for government officials, criminal penalties—as contributing to their organizations' decision to adopt, and sixteen (especially those at for-profits and privacy startups) mentioned specific legal requirements or agreements with regulators. Seven—especially those working in policy or legal roles outside of government—also mentioned the possibility of negative media coverage and public backlash that could sour relations with external stakeholders and make it more

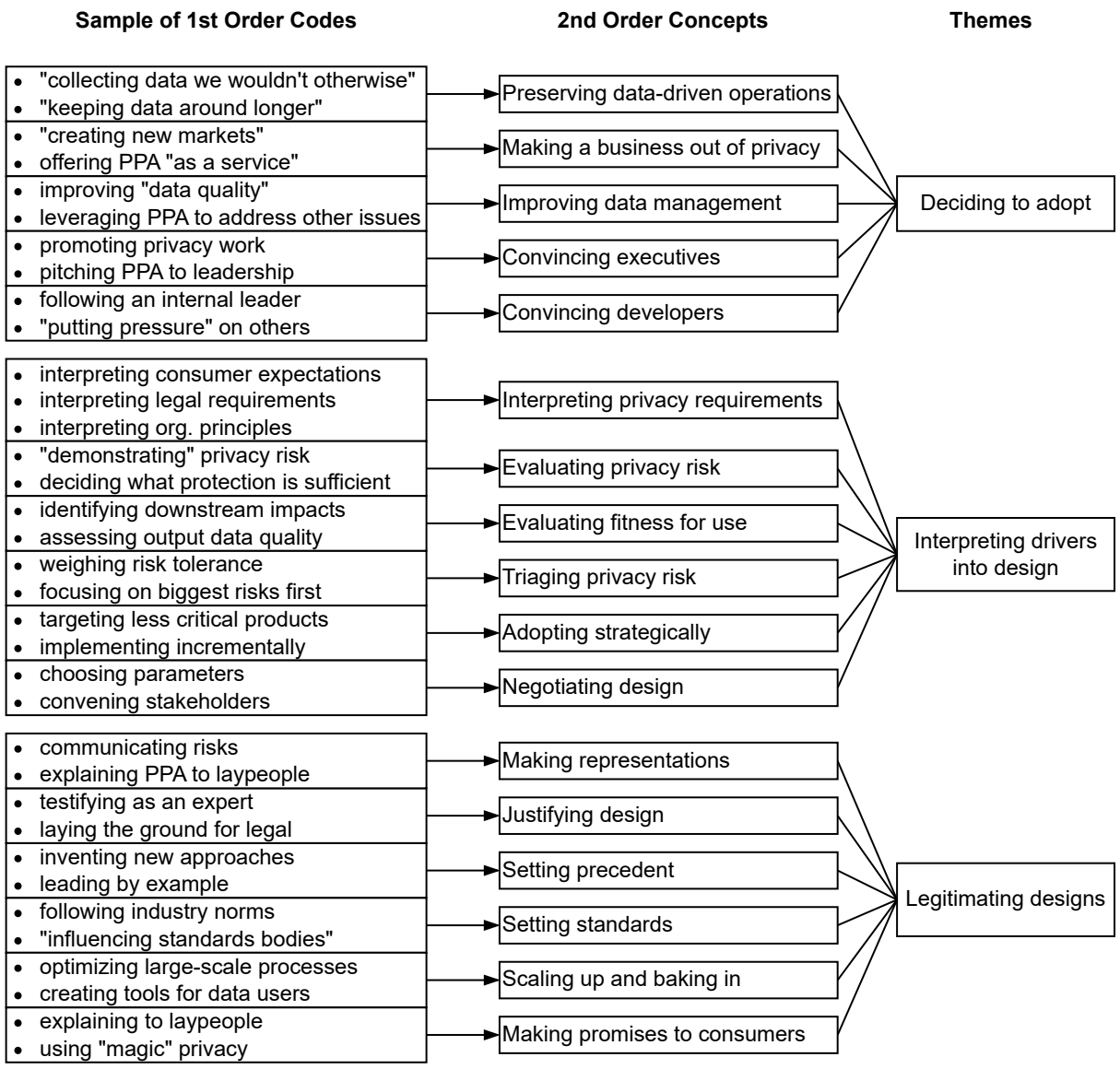


Figure 3.1: Themes and second-order concepts used in our process model, with an illustrative sample of first-order codes. Figure inspired by Gioia, Corley, and Hamilton (2013).

difficult to recruit talented employees: "protecting the brand is just as much a liability thing as protecting you from creating a compliance disaster where you have to pay millions of dollars in fines" (P10). Faced with these expectations and the increasing inadequacy of existing data governance methods against modern reconstruction, linkage, and other attacks, many organizations—such as the U.S. Census Bureau (Abowd & Hawes, 2023)—looked for new solutions.

3.4.1.1 Motivational narratives.

These expectations provided broad motivation for organizations to change privacy practices but did not prescribe PPA adoption specifically—none of our participants described a regulation or public campaign that favored new PPA technologies over long-standing, non-algorithmic alternatives such as data minimization. In the organizations we studied, PPA adoption came to be viewed as an appropriate solution through three different narratives. Practitioners and decision-makers used one or more of these narratives to convince decision-makers and justify their investment.

Preserving data-driven operations. Most commonly, organizations developed PPA to preserve a data-dependent model of operation—for example, behavioral advertising based on precise tracking and measurement. As one employee of a technology firm said, "I think a lot of folks in the industry are concerned about regulators stepping in and saying, 'Okay, here's how the web should work.' And all of a sudden, the whole business model of the web would fall apart" (P9).

Instead, organizations developed their own, overwhelmingly algorithmic, solutions. In online advertising, for example, Google developed plans to replace third party cookies, a ubiquitous technology for tracking web activity, with a suite of new techniques in their Privacy Sandbox (Goel, 2022). Meta invested in research on "novel privacy-preserving technologies" for ad measurement (Meta Research, 2019). Both organizations framed PPA adoption as a compromise between business interests and privacy concerns. In a blog post titled "A Path Forward for Privacy and Online Advertising", Meta's Chief Privacy Officer wrote: "We continue to believe personalized ads and privacy can co-exist... that's why we're investing in research and development of privacy-enhancing technologies" (Egan, 2020).

PPA adoption provided a path for organizations in our study to "collect data we wouldn't otherwise collect" (P3), "monetize data more properly" (P17), and "keep data around longer" (P26)—for example, by anonymizing data to bypass legal limitations on storage. Practitioners in both private and public organizations said they tended to "collect the data first and then figure out what it's useful for later" (P25). Some practitioners in private industry viewed this strategy for adoption as a "cloak for just collecting lots and lots of data" (P22) or "legal cover for hoovering up as much information as possible" (P15).

Making a business out of privacy. While all but one organization we studied adopted to preserve operations, thirteen—disproportionately startups—also welcomed PPA as an opportunity to access new, privacy-conscious market segments, develop more competitive marketing, or develop new services. A director recalled, "Before... we would appeal to the principles of the company. Now... you can start saying things like, 'It probably will appeal to this market' " (P10). Web browsers like DuckDuckGo and Brave have grown market shares around privacy-preserving branding. Startups like Tumult Labs or Leap Year Technologies, recently acquired by cloud data giant Snowflake, offer PPA services to businesses, non-profits, and government agencies. And even within less privacy-branded technology companies in our sample, we observed some teams relying on different narratives than others, depending on the extent to which privacy differentiated the product they worked on. Non-profit and government organizations in particular used PPA systems to share new sources of data with

researchers—for example, the U.S. Census Bureau first used differential privacy to release new data on commuting patterns (Machanavajjhala et al., 2008).

Improving data management. To combat the view that additional privacy infrastructure was "costing resources with no clear benefits" (P17), some practitioners also argued that PPA would have side benefits beyond compliance. Several practitioners argued internally that the improvements to data management required to adopt PPA would also reduce "bad science" and ultimately make for better products. PPA adoption sometimes offered practitioners a chance to bring up older, long-standing issues like data minimization: "it gives you the ability to talk about that like it's a fresh thing" (P22). However, no organization we studied adopted PPA solely to improve data management.

3.4.1.2 Convincing executives.

The primary audience for practitioners' adoption narratives was the key decision-makers within the organization—including the chief privacy officer, CEO, board members, and other "legal and risk" executives (P13). To convince executives, practitioners translated external privacy expectations into concrete business costs—one privacy engineer would "go in armed with a bucket of consumer research and case studies," including internal studies aimed at estimating "how much bad privacy can potentially cost you, based on historical data" (P3). Executive buy-in helped push forward adoption on a case-by-case basis and drove the formation of internal policies—one large technology firm in our sample, for example, integrated PPA adoption into its existing privacy review process for new features.

3.4.1.3 Convincing developers.

Participants at government agencies and more hierarchical organizations relied mostly on these top-down policies to convince developers and other employees to contribute: "your own leadership has said, 'we are doing this'—going back to [prior practices] is not an option, so let's make it work" (P6). Participants at other organizations, though, discussed the importance of a less formal "privacy culture," especially for organizations that relied on product teams to "self-forward" (P8) relevant cases for privacy or volunteer for PPA adoption.

While internal policies were often based on cost calculus, the manner of adoption was often tied up with employees' moral judgments. Six of our participants—all at for-profit technology firms and start-ups—said they considered adopting PPA because it was the "right thing" to do for users. One privacy engineer at a technology firm said:

> I think at any company from little to big you're going to find that there's some set of people who are genuinely deeply ethical people—and I have met many of those at [my organization] and they're fantastic to work with—and there are people who recognize that privacy is a business proposition... (P3)

This internal advocacy often depended on the leadership of influential privacy practitioners in centralized teams. Two participants at large technology firms observed peers "pushing hard" for particular PPA techniques—one of whom a former employee said "probably is uniquely responsible for driving adoption in the company" (P19). But when an influential

leader left, internal adoption of PPA dwindled. At least two organizations in our sample were deconstructing these central teams by the end of our study, distributing privacy professionals to product and infrastructure teams across the organization.

3.4.1.4 Adopting strategically.

Restricted to limited time and resources, most organizations we studied did not apply PPA uniformly across products and features—in fact, many adopted PPA for only one or two products or features. Instead, the rollout strategy depended on triage. For example, larger organizations had systems for prioritizing data deemed more "sensitive" or risky. Though most executives and managers agreed that it was better to start early—to promote communication and reduce disruption—they disagreed on whether it was better to target mission-critical products first (to set a precedent) or to start with "easy" use cases (to build momentum). Five privacy engineers, all in industry, said they experienced mostly the latter—adoption for only peripheral use cases. One noted that organizations "don't use the distributed machine learning approaches in things that are really mission critical... Where it really matters, they just collect a bunch of data [centrally] and make some promises around it" (P15). Another perceived a fundamental limit to adoption: "Once there's real money on the line, you get leadership involved. And someone doesn't care about protecting privacy because they're going to get promoted if you make however many billion dollars" (P19).

3.4.2 Interpreting Policies into Designs

How did these external and internal drivers of adoption inform choices made during deployment? Adopting PPA—and negotiating its design—is not yet as simple as choosing a vendor or product off the shelf. Organizations made many specific design decisions to make their new systems "privacy-preserving" and align them with internal policies and external expectations. PPA practitioners were responsible for interpreting privacy requirements and guarantees, evaluating trade-offs in proposed designs, and negotiating with product teams, lawyers, and executives to triage privacy requirements, define scope, and settle on appropriate designs.

3.4.2.1 Interpreting privacy requirements.

In private firms and public organizations alike, PPA practitioners described "interpreting" or "translating" legal requirements and internal policies into technical specifications for algorithmic systems. From a former director at a technology firm:

> The way [the organization] did GDPR was: we had this privacy legal department, they'd spent a huge amount of time with the law... We had them dump the entire law into my head and I wrote the engineering requirements! (P7)

In rare cases practitioners could reference regulatory guidance—from the European Courts, for example (Data Protection Working Party, 2014)—or "hints" from regulators about which practices would be considered unacceptable (P23). Most relied on internal policies written by executives and legal teams: "These regulations would get translated into internal policies, and

so what people inside the company care about are the internal policies... I very rarely had to care about what the actual regulations were" (P19). These internal policies were designed to satisfy both legal requirements and the social expectations of "key opinion leaders" in policy and regulation, as one policy researcher describes: "What are their expectations, and what do we need to do to meet them?" (P12) Less commonly, participants referenced users' expectations—but nearly all of our participants did not interact with data subjects.

The process of interpreting these "external mandates" (P11), as one participant called them, was not straightforward. Around half of our participants mentioned conflicting, deficient, or "unreasonable" expectations in external regulations and at least one created a "pecking order" (P23) of privacy rules to follow. Some even pointed out specific technical errors in regulatory guidance, speculating that "people who knew how this stuff really worked or could work weren't necessarily at the table" (P26) when the regulation was written. One executive at a technology firm said, "Regulations come in and kind of break what I'm doing... you [regulators] just made our system work worse and I'm very grumpy about it" (P7).

3.4.2.2 Triaging privacy risks.

When organizations did adopt PPA, simply evaluating the reduction of privacy risk provided by a particular algorithm was not straightforward. No single design met all requirements, especially for the large organizations we studied. As one executive put it, "you want to ask yourself what risk are you willing to take, how much uncertainty are you willing to live with" (P14). Another manager at a large technology firm recounted,

> When I first started working on this, I accepted the culture of 'Oh gosh, the sky is falling, it's all important, we've got to get it all!'... It turns out that we have permission to fail [on] smaller risks... [Executives] are fine with us taking misses... because they can't imagine a future in which we don't take another fine. (P23)

An engineer recounted, "The sorts of guarantees that a privacy-enhancing technology offers almost never line up with anything a lawyer would recognize... and so we wind up in a room with a whiteboard sort of scribbling frantically at each other trying to do a lingo match" (P3).

The ϵ parameter in differential privacy, for example, theoretically bounds the amount an individual's inclusion increases their risk of unwanted disclosure, but difficulty interpreting its value has contributed to heated epistemic disagreements (boyd & Sarathy, 2022; Nanayakkara & Hullman, 2022). Practitioners disagreed on what designs and parameter settings are meaningfully "privacy-preserving" for any given use case. Some practitioners in our sample believed that DP is a "gold standard" approach for managing privacy risk that provides "meaningful" privacy if implemented properly, but disagreed over whether guarantees were still meaningful after common relaxations. Others doubted further whether differential privacy is even an appropriate technical conception of privacy (see, e.g., Hotz et al., 2022; Seeman & Susser, 2023). Some fell back on "experimental" or "practical" guarantees to convince stakeholders (Dwork et al., 2019).

Without a clear conception of privacy risk, tuning the strength of privacy protections was

more art than science. Legal requirements that data be "anonymous" or "confidential", for example, have been satisficed by successively stronger technical standards in just the past decade, including the use of k-anonymity to satisfy the Health Insurance Portability and Accountability Act (HIPAA) Privacy Rule (Malin et al., 2011) and the use of differential privacy in the 2020 Decennial Census (Abowd & Hawes, 2023). As one privacy engineer at a large technology firm put it,

> There's nothing out there to guide you, so you're just kind of winging it. We've gotten as far as lining [differential privacy] up with things like GDPR's "singling out" clause, but it still doesn't give us any notion of how to do things like tuning parameters. (P3)

In practice, stakeholders with different incentives interpreted privacy risk differently. For example, Social Science One, an association of researchers partnering with Facebook to release a large dataset, argued that less stringent privacy standards satisfied the General Data Protection Regulation and Facebook's FTC consent decree, but Facebook disagreed, releasing the dataset with differential privacy (King & Persily, 2020).

3.4.2.3 Negotiating design.

These interpretations guided design negotiations between privacy practitioners, the teams responsible for implementing the PPA system, and other internal (and rarely external) stakeholders. A privacy engineer said,

> [In design meetings] there's usually a representative from the business and their job is to sit there and advocate for the continued health of the business... I think the lawyers and I are both more risk averse and the business is the counterweight to that. (P3)

For six participants, all at for-profit firms, the relationship between business operations and PPA adoption was adversarial. A former employee of a large technology firm said:

> The inherent relationship between the people on the ground doing the privacy reviews and the leadership, even within the privacy org, is very confrontational... The [product] team is trying to do the least amount of privacy that we will approve... You are fighting against different teams and organizations and you might as well act like it. (P19)

Still, privacy engineers tried to be constructive: "If people have a positive experience with privacy, they are more likely to come to us when they feel there's a problem" (P26).

As a result, design negotiations usually centered managerial and technical concerns—most participants framed the design process as maximizing "fitness for use"—including data quality, cost, efficiency, usability, and interoperability—subject to a minimum "privacy bar". Minimum standards helped reinforce boundaries during design negotiations—as one privacy engineer at a large technology firm said, "The [product] team would come in and say okay well how about this ϵ. So a lot of the time, we can really easily just point to, 'No we have a standard, it's at this ϵ, use it'" (P19). Practitioners in at least four organizations leveraged formal

processes like privacy review to resist opposing pressure from other executives: "everything goes through privacy, so we can just tell the leadership 'Hey, we're not letting this launch unless you do this'" (P11).

Because the interpretations of privacy risk and privacy guarantees could vary, standards were often flexible: "a lot of this we've just had to come up with out of thin air" (P23). A privacy engineer put it more bluntly: "The definition of what we consider to be anonymous is completely arbitrary. There's some legal things informing that but, for the most part we just made up ours" (P19). For k-anonymity, for example, another privacy engineer at a privacy-focused startup said, "We have some magic numbers in the company, like 20. $k = 20$ is like a minimum that we don't go under. But [usually we] start the discussion from $k = 1000$ or $k = 100$ at least and then go down from there" (P8). A lawyer for a technology firm described their "magic number" for k-anonymity similarly: "we kind of feel like [a k value of] 20 is usually not super necessary, but if it gets down to under 5 it's kind of a little dicey" (P28). Not all organizations had "magic numbers." Larger organizations had teams of privacy experts who helped product teams set parameters case-by-case. But some are now curtailing the authority of these teams. Recently, Meta reportedly permitted product teams, rather than privacy teams, to make the final decision about what privacy risks are acceptable (Huang, 2025).

3.4.3 Legitimating Designs

Once designs were set and PPA systems deployed, organizations still needed to convince regulators, consumer advocates, and other stakeholders that their design choices were acceptable to gain the social, economic, and legal benefits of adoption. Organizations relied again on privacy experts to decide how to represent the privacy properties of their PPA designs in legal arguments to regulators or marketing promises to consumers. And organizations scaled up their PPA practices, setting standards internally and promoting those standards to the rest of their industry, creating precedent for rational response to privacy pressures.

3.4.3.1 Making representations.

In our study, all the organizations that deployed a PPA system translated its properties into some external representation, usually a legal defense, a promise to consumers, and in some cases a policy campaign.

Justifying designs. Several practitioners—particularly those who adopted to preserve existing operations—designed PPA systems with legal justification in mind. As a policy researcher asked, "Can you back that [public statement] up with your systems in an investigation?" (P12) This judgment fell to practitioners' legal and technological expertise.

> At some point the lawyer and I just have to sit down and be like... 'This seems both ethically justifiable and probably reasonable in the eyes of the law based on some esoteric U.K. law from the 1600s...' or 'We think this is defensible and we think it's not a crappy thing to do.' (P3)

Sometimes justification had less to do with specific laws and more to do with perceptions of the regulatory climate. As one privacy engineer recalled telling internal lawyers,

> 'We don't give legal opinions, we're technical people. But, this new thing [differential privacy] is the gold standard. If a regulator says anonymization is a thing that's possible, and academia says this is the best thing you can hope for... probably you're going to be fine.' (P18)

Practitioners were prepared to rely heavily on their own expertise. A researcher at another large technology firm recounted defending their new technique internally:

> Ultimately we had a very, very senior statistician who basically just got up and told the lawyers, 'I believe that risk of leakage is very low in this model' and that's it. They didn't go into the math, they didn't look at any of the other stuff. (P20)

Making promises to consumers. Practitioners also contributed to public representations their organizations made about privacy, including public privacy policies and marketing. Most privacy engineers did not have direct contact with data subjects—they explained their work only to legal and policy teams. Those that did described the difficulty of accurately representing algorithmic protections to laypeople: "You're asking [customers] to trust an algorithm they will never understand. It's not that easy to prove to someone that these things are actually going to work" (P25).

Some organizations dealt with this opacity by choosing "easy-to-describe" systems (P15). Others, particularly non-profits and government agencies, sought to increase transparency with open source software and detailed public descriptions. Several reported making changes in response to feedback from the public, especially at non-profit organizations. The U.S. Census Bureau, for example, submitted their differentially private 2020 disclosure avoidance system to multiple rounds of public comment, commissioned reviews from organizations like JASON and MITRE, and adjusted parameters and post-processing in response (boyd & Sarathy, 2022). Even when code was public, though, our participants—and independent researchers (Dwork et al., 2019; Gong, 2022)—found it difficult to judge how exactly that code is being used from the outside, and key privacy parameters and product details were not always disclosed.

Others were less concerned about the need to comprehensively educate consumers: "[PPA is] almost like whiz-bang technology... it doesn't have to be something that everybody needs a detailed awareness of, because it just makes life easier in the background" (P12). At least three participants, in both government and industry, worried that marketing new PPA protections would "muddy the waters" (P21) by revealing flaws in previous practices. A privacy director at a technology firm pointed out a practical upside to operating PPA "in the background": firms would no longer have to ask for consent and risk "creep[ing] people out" (P7).

Several practitioners had concerns about misrepresentation. Six of our participants, mostly in private industry, feared adoption that amounted to "magic privacy" or privacy "pixie dust"—black box techniques that, when invoked in marketing or legal copy, symbolically

assure consumers and regulators of strong privacy and foreclose further inspection. An executive at a technology firm said bluntly: "most of [PPA] marketing is bullshit... they're writing checks that their tech cannot cash" (P7). The policy researcher who advocated for less detailed awareness also advocated for disclosing limitations: "Technologists always have to be really careful... because you can make this sound so much more impressive. You can make it seem like snake oil" (P12). Six other participants—who mostly worked to preserve existing products at for-profit firms—brought up the possibility that their PPA efforts were just "good theater" (P28), "privacy whitewashing" (P12), or "adoption for show" (P5). One participant mentioned Google's Privacy Sandbox as an example of this kind of proposal.

3.4.3.2 Setting precedent.

Substantive or otherwise, organizations' PPA practices became a model for others, particularly in the absence of clear industry standards. From a lawyer at a technology firm: "I've been taken aback a lot in the private sector [at] how out in the cold companies feel—like they want to do something, but they just don't know what is required" (P28). One privacy engineer at a large technology firm explained:

> An organization like [mine] really would like to avoid getting to court for every little thing. You need some kind of consistent internal standard... so that you can avoid getting into situations where you are in a public sense told, 'You have to do this.' (P19)

As a result, internal standard-setting efforts were developed in anticipation of future regulation: "Let's try to fix this two years before they make it mandatory" (P3).

Leading by example. Some organizations—particularly privacy-focused startups—aimed to actively *guide* future regulation. Executives at two privacy boutiques agreed: "We're eager to demonstrate that legislation is catching up to [us], instead of [us] catching up to legislation" (P15); "We want to be seen as thought leaders in this area as it continues to evolve" (P13). Practitioners in legal and policy teams at large private firms explicitly advocated to policymakers for regulation that would provide "safe harbor" (P2) from regulatory requirements for organizations which implemented their preferred PPA techniques. Practitioners at two large technology firms and a privacy-focused startup described steps their organizations took to influence regulation and develop relationships with policymakers—as that director put it, "glad handing" (P2)—outside the normal course of fact-sharing. Other organizations in our study worked to "influence standards bodies" (P14) such as the International Organization for Standardization (ISO), the National Institute of Standards and Technology (NIST) or the World Wide Web Consortium (W3C).

These efforts influenced other organizations. At least seven practitioners in non-profits and smaller organizations modeled their PPA systems after others' prominent deployments, particularly the use of differential privacy in the Decennial Census.

Scaling up & baking in. Second, organizations spread internal standards through developer tools and documentation. One manager said, "We bake [PPA] into infrastructure. We make it so it can't be screwed up" (P23). All but five practitioners described building or updating

software infrastructure more generally in the course of scaling up their PPA systems. And most practitioners, especially those using PPA to create new products, built or updated tools for developers—mostly software libraries and internal platforms—to increase the capacity of a limited number of PPA experts. They aimed to construct a "well-lit path" for developers. As one corporate executive explained, "You want to make it easy to do the right thing and hard to do the wrong thing" (P7).

By virtue of their market positions, some companies could lay "well-lit paths" for whole industries by sharing their infrastructure. Amazon and other cloud service companies, for example, have included PPA tooling and features in their analytics products (AWS, 2023). Twelve participants, disproportionately from smaller non-profit and government organizations, mentioned integrating external PPA software libraries or other infrastructure—built by large technology firms such as IBM or by open source communities such as OpenDP—to deploy their own PPA systems, though these tools still required expertise to use correctly.

A few of industry practitioners had concerns that these kinds of infrastructure could help firms box out competitors and control PPA development. Several mentioned Google's Privacy Sandbox and a new Apple feature that allows users to opt out of some in-app tracking used by ad brokers including Meta and Google (Morrison, 2022). Both projects drew anti-competition criticism from advertisers in France and the United Kingdom (Lomas, 2021). As one researcher at a large technology firm observed, "When [corporations] make a bid on a privacy-preserving technology, it's not just that they want to do it quietly. They want to do it spectacularly with regulations that ensconce what they've done at the expense of their competitors" (P4).

3.5 Theoretical Integration

We observed multiple points of decoupling between external privacy expectations on the one hand and the properties of deployed "privacy-preserving" systems on the other (Table 3.2). Some of these points of decoupling are consistent with existing organizational theory. When internal constituents failed to reinforce external pressure, leaders and developers often abstained from adopting PPA for core products—a commonly studied type of policy-practice decoupling (Bromley & Powell, 2012). And the trends in adoption we observed parallel the initial stages of legal endogeneity (Edelman, 2016, p. 27-41): organizations encountered ambiguous or absent privacy regulation, constructed PPA as a relevant solution, designed & implemented PPA systems to prioritize managerial concerns, and diffused those systems across industry. It is not yet clear the extent to which PPA systems will be endorsed by courts and administrative agencies, but practitioners expected it: "A lot of [adoption] is not so much 'this is what the law says' as 'differential privacy seems to make the regulators happy'" (P19). Technology companies such as Google are already advocating for regulatory exceptions for the PPA techniques they use (Google, 2022).

However, the material, technological aspect of these information systems complicates existing theory in two key ways. First, we observed less obtrusive points of decoupling arising from the algorithmic aspects of PPA systems, a new dimension of decoupling that we term *algorithmic decoupling*; second, our analysis suggests that instances of algorithmic decoupling impact law

Stage	Description of decoupling	*Representative quotation*
Deciding to adopt	PPA adoption justifies additional data collection, undercutting privacy benefits	*[There is] market demand for something that will get you legal cover for hoovering up as much information as possible... the business model isn't going to change at all. (P15)*
	Reliance on voluntary adoption results in low take-up	*We depend on people to in the company to understand the privacy context of the company and self forward to triage. (P8)*
	Leadership exempts core products from adoption	*Once there's real money on the line, you get leadership involved. And someone doesn't care about protecting privacy because they're going to get promoted if you make however many billion dollars. (P19)*
Interpreting policies into designs	Product teams negotiate for weak privacy parameters	*The [product] team is trying to do the least amount of privacy that we will approve... So it's really about how much political capital [they] have within the organization... (P19)*
	Designers misinterpret privacy guarantees	*To make [open source DP libraries] work you still have to know what you're doing... Worst case scenario you'll think you're using DP when really you're actually not. (P26)*
	PPA designs are ambiguously related to privacy requirements	*We've gotten as far as lining [differential privacy] up with things like GDPR's "singling out" clause, but it still doesn't give us any notion of how to do things like tuning parameters. (P3)*
Legitimating designs	Marketing makes overly simplified or deceptive claims	*Most of [PPA] marketing is bullshit... they're writing checks that their tech cannot cash. (P7)*

Table 3.2: Points of decoupling in PPA adoption.

and society through new mechanisms. We highlight two important mediators of algorithmic decoupling based on cross-sectional analysis.

3.5.1 Algorithmic Decoupling

Technological infrastructure—the code bases and cloud services that industry "runs on"—is inextricably linked with practice (Lampland & Star, 2009). Just as routines executed by people can become decoupled from policies and expectations, so can routines executed by algorithmic systems (Lessig, 1999). Organizational research describes how employees form an interpretation of their organization's legal environment and mobilize that "legal reading" in their everyday work (Fuller et al., 2000). Our findings add a new dimension: PPA experts also employed their own *technological readings*, interpreting technical properties alongside the social and legal environment. As Wu (2003, p. 682) writes, "The programmer is not unlike the tax lawyer, exploiting differences between stated goals of the law, and its legal or practical limits." In PPA design, legal and technological readings were jointly consequential. Particularly for recently developed techniques, modern privacy laws rarely

admit to straightforward, specific translations to technical implementation (Nissim & Wood, 2018; Balebako et al., 2014).

Our study suggests that this additional layer of technological interpretation can also contribute to decoupling. We call this dimension *algorithmic decoupling*—a gap between formal policy (e.g., an organizations' public promises about privacy) and the material aspects of organizational practice embedded in algorithmic processes (e.g., the ways personal data are processed) (Volkoff et al., 2007). Our study focuses on the professionals experts assigned to navigate this gap, translating between the ostensive and the material aspects of their organizations' privacy practices.

Compared to other organizational practices, the operation of algorithmic systems may be more easily obfuscated, subject to less scrutiny, and therefore more easily and permanently decoupled from policies and outcomes. Information technologies generally gain scale through translation and loose coupling between many layers of digital devices, networks, algorithms, and services (Faik et al., 2020; Yoo et al., 2010). The relative invisibility and complexity of these layers can make algorithmic systems inscrutable to non-experts (Burrell, 2016; Metcalf et al., 2023; Selbst et al., 2023; Jin & Salehi, 2024). For example, a major 2002 privacy transparency requirement for federal agencies was undermined by "the inaccessible idiom of technology", which impeded public participation and oversight (Bamberger & Mulligan, 2008). Moreover, algorithmic systems increase the *scale* of organizational practices involving data processing and data-driven decision-making; designers' decisions have an outsized influence on the outcomes of these processes, compared to other routines.

Algorithmic decoupling complicates prior research on organizational decoupling (Bromley & Powell, 2012; Boxenbaum & Jonsson, 2017). Prior research separates decoupling into policy-practice decoupling (symbolic adoption) and means-ends decoupling (symbolic implementation) (Bromley & Powell, 2012). Algorithmic decoupling adds an additional dimension—practitioners identified instances of algorithmic decoupling both between policies and practices (policy-practice) and between practices and outcomes (means-ends) (Figure 3.2).

Algorithmic decoupling shares certain properties with means-ends decoupling in particular: because of its inscrutability, it may be more durable than traditional policy-practice decoupling, which some argue do not withstand public scrutiny for long (Bromley & Powell, 2012). And while extensive research suggests that policy-practice decoupling may be reversed when employees reinforce societal expectations based on professional standards, moral commitments, or personal identity (Edelman, 2016; Haack et al., 2012; Turco, 2012; Gioia, Patvardhan, et al., 2013), algorithmic decoupling (and means-ends decoupling) may occur nonetheless (Bromley & Powell, 2012)—when non-expert developers misinterpret privacy guarantees, for example.

Means-ends decoupling, however, is more likely in opaque institutional fields where the effect of practices on outcomes, such as sustainability, is hard to precisely measure (Wijen, 2014). This is only partially the case for modern algorithmic systems; the material aspects of "black box" algorithmic systems are often opaque, but it sometimes possible for technical experts with access to a system to precisely measure its impacts on well-defined outcomes of concern such as discrimination and disinformation (Kroll, 2018). Crucially, algorithmic decoupling

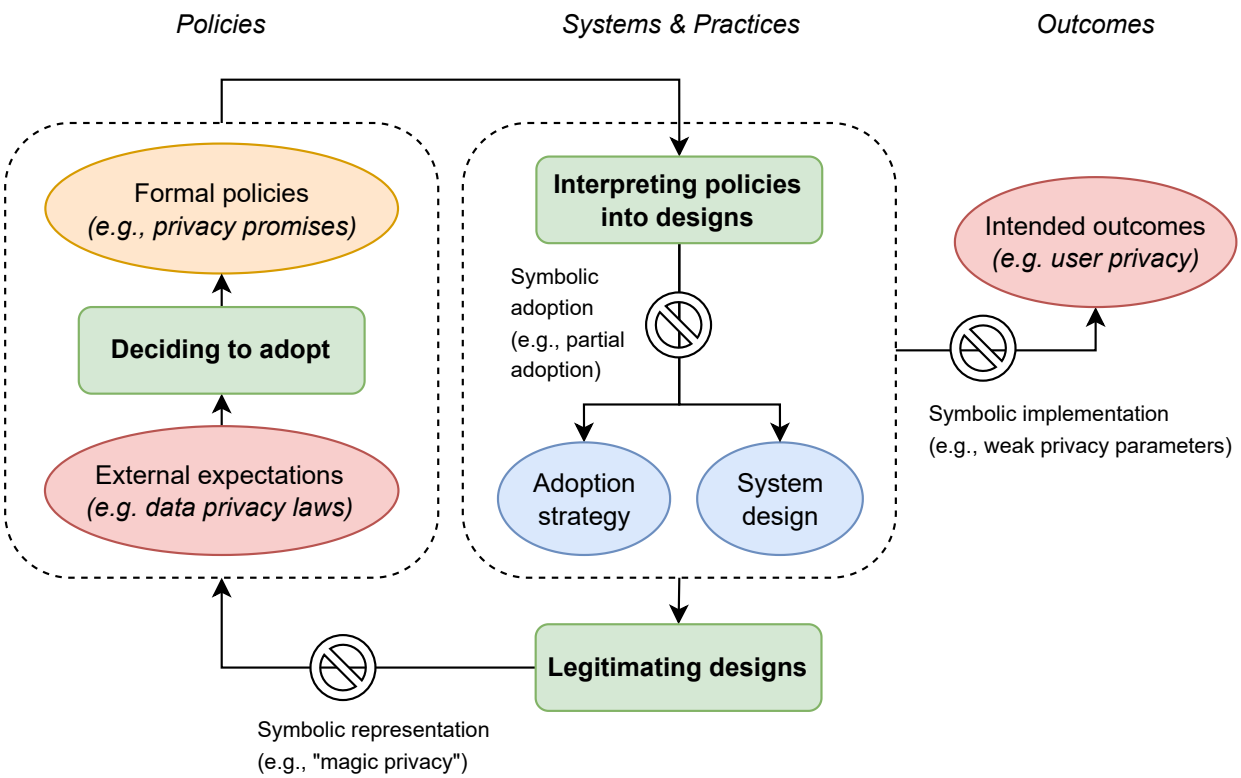


Figure 3.2: Algorithmic decoupling in our emergent process theory. Figures C.4.1–C.4.3 detail the bolded subprocesses.

involves not only disconnect between process and outcomes (as in means-ends decoupling) but also between policies and the properties of algorithmic systems used to process data. This ambiguity is especially salient when regulations focus more on procedures (e.g., standards for data processing) than outcomes, requiring practitioners to justify choices in terms system properties (e.g., the ability to "de-identify" personal data).

Studies disagree on whether decoupling is more likely in early in the adoption process, when implementation expectations are flexible for early adopters (Bromley & Powell, 2012), or later, when established symbolic cues may suffice (Tolbert & Zucker, 1983; Kennedy & Fiss, 2009). Like policy-practice decoupling, algorithmic decoupling seems more likely when technical interpretations of the underlying algorithms are uncertain or contested, such as early on in adoption. Policy-practice decoupling is not always intentional; it may arise as managers "muddle through" competing stakeholder expectations (Crilly et al., 2012). In our study, technologists muddled through technical ambiguities in addition to competing requirements. Unlike policy-practice decoupling, which is inversely related to pressure from market stakeholders, regulators, and other external constituents (Stevens et al., 2005; Okhmatovskiy & David, 2012; Bromley & Powell, 2012; Marquis & Qian, 2014) and resistance from internal constituents (Turco, 2012), algorithmic decoupling may persist in spite of scrutiny if oversight lacks technical specificity or there is no clear consensus on the acceptable algorithmic translation of social values such as privacy.

3.5.2 Algorithm-Mediated Legal Endogeneity

Algorithmic decoupling also entails new mechanisms for initial "guesses" at compliance to endogenously alter industry practices and institutional norms (Edelman, 2016). Information systems represent material standards and guidelines for the ethical treatment of personal data (Verbeek, 2006; Lampland & Star, 2009)—fundamental matters of privacy in IS, for example, are delegated to technologists and standard setting bodies (Waldman, 2018; Doty & Mulligan, 2013).

The construction of shared IS infrastructure, then, has potential for lasting influence on the institutional environment (Faik et al., 2020). In our study, for example, less-resourced developers' tendency to rely on only a few entrenched tool libraries meant that early adopters had even more influence over practices in the rest of industry. This influence grows as more developers use PPA products built into dominant cloud computing platforms such as Amazon Web Services (Narayan, 2022; Cutolo & Kenney, 2021; AWS, 2023). And algorithmic systems receive additional legal deference as courts often treat their design as a technical inevitability rather than question the design choices that lead to their creation (Selbst et al., 2023; Metcalf et al., 2023; Jin & Salehi, 2024).

3.5.3 Important Mediators of Algorithmic Decoupling

3.5.3.1 Privacy Champions.

The technical experts able to penetrate the veil of technological idiom hold special leverage over algorithmic decoupling. They had varying amounts of influence over the decision to adopt—like the success of enterprise IS adoption in general (Liang et al., 2007), decoupling depends heavily on the support or opposition of influential executives and managers and their social ties to other decoupling or non-decoupling organizations (Westphal & Zajac, 2001; Fiss & Zajac, 2004, 2006; Weaver et al., 1999). But when executives decided to adopt, the process of interpreting policies into designs was heavily dependent on PPA experts, often only one or two individuals. Nine of the organizations we studied—disproportionately smaller organizations—mentioned lack of experience as a barrier to PPA adoption and six consulted with external experts; one non-profit employee said that their PPA deployment would not have "gotten off the ground" (P19) without an external consultant from a large tech firm. Moreover, organizations in our study relied heavily on performances of expertise to justify their PPA systems to the public. The HIPAA Privacy Rule, for example, permits de-identification methods certified by a statistical expert (U.S. Department of Health and Human Services, 2012; Malin et al., 2011).

Some practitioners leveraged their central role in adoption to promote strong privacy standards and prevent decoupling. Recent work chronicles the role of privacy "champions": institutional entrepreneurs who advance privacy through informal education and daily work when official policies are missing or insufficient (Tahaei et al., 2021). Many of our participants viewed their work as aligned with external calls for privacy. The approaches these participants took to advocate for adoption mirrored stages of emergent moral leadership (Solinger et al., 2020), including a precipitating scandal, internal coalition building, negotiation with other stakeholders, and establishment of formal structures (e.g., privacy review). They

promoted PPA adoption through technological defaults—the "well-lit paths" our participants mentioned—and formed special interest groups, often centralized teams, to set and maintain privacy standards. But they also struggled to prioritize and organize around ethics in corporate environments (Lee et al., 2024), as do software engineers more generally (Widder et al., 2023; Ali et al., 2023)

3.5.3.2 Motivational Narratives.

Cross-organization analysis of our sample suggests that algorithmic decoupling is also mediated by the dominant narrative motivating adoption.

Decoupling was more difficult for organizations making a business out of privacy. Our participants reported decoupling most commonly when adopting to preserve existing data-driven operations, disproportionately at large technology firms. For example, the strategy of adopting only for less "mission critical" use cases (§3.4.1) was reported disproportionately by practitioners in private industry, as were concerns about deceptive invocations of "magic privacy" algorithmic techniques (§3.4.3). At large firms adopting mostly to comply with existing regulations, employee-led adoption efforts were more likely to falter as key privacy leaders left the company and the company broke up central privacy groups. But privacy-motivated practitioners at privacy-branded startups were more positive about PPA standards at their companies. Indeed, prior research suggests that when employees' identities are tied up with their organizations' external representations, decoupling is less likely to succeed (Turco, 2012).

Large organizations had more influence on the institutional environment. Institutionally endogenous processes such as participation in standards bodies were disproportionately discussed by practitioners at large firms. They used legal justifications and "standardization work" to convince regulators that PPA adoption was sufficient to keep existing practices compliant. Industry leaders had an advantage in influencing the institutional environment—both technology adoption and legitimacy spread through networks of influence (DiMaggio & Powell, 1983; Rogers, 2003).

3.5.4 Boundary Conditions

In our study, we analyze the adoption of a unique class of algorithmic systems for privacy-preserving analytics. Besides providing thick descriptions of important phenomena like PPA adoption (Lee & Baskerville, 2003), unconventional contexts can be useful for developing newly insightful theory (Bamberger & Pratt, 2010; Monteiro et al., 2022). We chose to study PPA adoption not only because of its potential impact on digital privacy, but also because it represents an understudied intersection of technological, managerial, and societal concerns common in a broader class of information systems technologies transforming the technology industry. Our model applies particularly to technological practices that 1) are implemented to improve social performance, 2) involve routines enacted by algorithms as well as people, and 3) have complex properties or impacts that require expertise and access to understand. Algorithmic systems in particular fit these criteria—pushes towards "responsible" artificial intelligence (AI) and ethical data science, for example, are motivated by similar institutional

pressures to PPA adoption, especially as calls for AI regulation proliferate (Hirsch et al., 2020; Lee et al., 2024). Thus, many of the mechanisms we identify in this unconventional context—for example, algorithmic decoupling—could also help explain the interaction between regulation and technology adoption in fields such as AI ethics, environmental protection, or digital social innovation (Qureshi et al., 2021).

3.6 Practical Implications

Given our findings, policymakers and privacy-minded managers face a conundrum. Within existing privacy laws, there is room for technological interpretation. But, the capacity to form and propagate these interpretations is concentrated in mostly large and mostly private organizations. These early adopters hold particular sway over the techniques that may become synonymous with privacy compliance. One practitioner, for example, described a future of "automatically invokable" mathematical techniques that assure users their privacy is preserved. It may be that the PPA systems deployed today *are* a substantive step towards privacy, and many of our participants were optimistic about their organizations' efforts. Scholars hope, for example, that the use of differential privacy could discourage dubious statistical practices such as p-hacking (Oberski & Kreuter, 2020).

But left to self-regulate, prominent early adopters may influence adoption in ways contrary to the public interest. For example, nearly all organizations in our study framed privacy risks as invasions or exploitation by state actors or other "attackers," shifting focus away from internal threats to privacy (Seeman & Susser, 2023). Privacy scholars argue that by focusing on specific technical properties such as individual anonymity or local processing, commercial PPA proposals legitimize invasive practices and foreclose more expansive privacy norms (Barocas & Nissenbaum, 2014; McGuigan et al., 2023; Martin et al., 2023; Yew et al., 2024)—a case of perverse innovation (Burk, 2016). Insights from Microsoft's differentially private Workplace Analytics tool (Bird, 2020), for example, may still be used to increase managerial control and restrict workers' autonomy (Levy, 2022). And scholars may worry that even as PPA adoption addresses real privacy concerns, it also legitimizes asymmetric economic and social relations (McGuigan et al., 2023; Veale, 2023; Viljoen, 2021). As one participant said, PPA "can't make [data collection] ethical just because it makes it private" (P15).

One remedy to managerial mediation is closer scrutiny (Edelman, 2016). Researchers and managers should investigate not only the theoretical properties of PPA techniques but also their empirical manifestations and impacts. Prior work offers guidelines for legal accountability in algorithmic systems (Selbst, 2021; Metcalf et al., 2023), particularly in areas of law where technologies are routinely deconstructed—products liability, for example (Selbst et al., 2023). And independent academic researchers are already scrutinizing the most visible PPA systems deployed by organizations such as Meta, Apple, and Google (Tang et al., 2017; Berke & Calacci, 2022; McGuigan et al., 2023; Martin et al., 2023). Some of our participants—echoed by differential privacy scholars (Cummings et al., 2024; Gong, 2022; Dwork et al., 2019)—called for additional efforts to make PPA techniques more transparent, such as a registry of PPA parameter choices (Dwork et al., 2019).

Scrutiny and transparency are not unalloyed goods—they must be accompanied by formal mechanisms for accountability (Han, 2015; Birchall, 2014). For example, Google's Privacy Sandbox included a publicly documented plan to automatically cluster users into interest groups based on their behavior. The plan drew opposition from competing browsers Firefox, Brave, and Vivaldi (Bohn, 2021) and from the Electronic Frontier Foundation, which called it "the opposite of privacy-preserving technology" (Cyphers, 2019). Google eventually proposed a new design which grouped users into fewer, purportedly less sensitive clusters (Goel, 2022). Here, scrutiny from competitors and advocates helped to make technical reforms to an unpalatable design. But the Topics API still gave advertisers the power to target users based on their behavior, leaving Google's core business model relatively untouched while hampering its competition (Lomas, 2021)—and Google has since reverted its plans to phase out third-party cookies entirely. Waldman (2018) suggests that strong regulatory interventions, such as consent orders and weighty fines, can shock organizations into more integrated privacy practices. One PPA practitioner in our study described how a "push from the outside" (P19)—for their organization, a court ruling—initiated centralized privacy review and empowered PPA practitioners to maintain higher standards.

Judges, policymakers, investigators, and managers tasked with holding PPA operators accountable should look beyond the invocation of a technique like differential privacy to examine its full technical manifestation—including parameters, definitions of sensitivity, and post-processing steps. Privacy and algorithm impact assessments optimistically could provide structure for internal advocacy and provide valuable information to outside observers (Selbst, 2021), though such assessments have had mixed effects in practice (Bamberger & Mulligan, 2008; Brandtner, 2021; Smart et al., 2022). But even as policymakers seek to advance the development of PPA technologies (e.g., National Science and Technology Council, 2023; Office of Science and Technology Policy, 2022a), they should also take care not to endorse PPA techniques outside of context and to scrutinize PPA systems implemented within "safe harbor" frameworks. Data privacy regulation has long struggled with decoupling—reviews of a "gold standard" 2000 U.S. Safe Harbor agreement for processing personal data from E.U. citizens, for example, revealed that most participating organizations failed to implement basic requirements and others made false claims about certification (Connolly, 2008). Moreover, policymakers should not assume that a solely technological solution is sufficient (Green & Viljoen, 2020)or that "magic privacy" systems even function as advertised (Raji, Kumar, et al., 2022). In addition to scrutiny, policymakers and foundations could offer funding and other support to help less-resourced organizations contribute to PPA practice.

3.7 Conclusion, Limitations, and Future Research

Our study analyzes the expert practitioners who help organizations decide to adopt, interpret PPA designs, and justify designs to external stakeholders. Through this unique perspective, we develop a theoretical model that captures interactions between institutional expectations and technology design—interactions that have great implications for digital privacy. We conceptualize new technological dimensions to theories of decoupling (algorithmic decoupling) and legal endogeneity.

Our work has limitations. First, while our sample did include organizations who considered PPA techniques but failed to deploy them, our sample does not include organizations which have never seriously considered adoption. Second, our participants' responses may be prone to social desirability bias or crafted to serve professional motives, though we do not repeat their views uncritically. Third, this study focuses on a set of formal, technical, and algorithmic approaches to privacy, and our analysis is influenced by our own backgrounds in computer science, economics, and public policy (see SM Appendix C.1 for further reflection). Research from disciplines outside of computer science and statistics could further elucidate the social trade-offs facing practitioners and the downstream consequences of adoption for privacy standards. Moreover, our sample—like the population of PPA practitioners—is predominantly American, white, and cisgender male. Our study thus examines dominant perspectives in PPA work, but those perspectives do not encompass all possible paths for PPA development—for example, organizations' PPA plans rarely accounted for documented inequities in access to privacy (Skinner-Thompson, 2020; Allen, 2022; Madden et al., 2017). Researchers and policymakers could imagine alternative trajectories for PPA development that elevate public interests—for example, as a means to facilitate algorithm auditing (Xu & Zhang, 2021).

As large technology firms promote PPA techniques as the "future of personalized advertising" (Egan, 2020) and the digital economy, we hope our study will inspire further rigorous empirical investigation and theorization about the ways regulators and consumer advocates may shape the development of socially-motivated innovation in the public interest.

Acknowledgements

We thank our participants for their time and insight. Thanks also to those who provided feedback on earlier versions of this research: Taya Cohen, Carrie Leana, Anna Mayo, Roy Rinberg, seminar participants at Carnegie Mellon University, and participants at the the 2023 USENIX Conference on Privacy Engineering Practice and Respect and the 2023 Privacy Law Scholars Conference—including Nikita Aggarwal, Jason Cronk, Elizabeth Edenberg, Thomas Haley, Cameron Kerry, Anne Klinefelter, Siona Listokin, Scott Skinner-Thompson, Jeremy Seeman, Alexis Shore, Harry Surden, and Alexandra Wood. This study was approved by the Institutional Review Board (IRB) at Carnegie Mellon University in 2021, #00000327.

Part II

Evaluating other machine learning systems

Chapter 4

Measuring Social Biases in Unsupervised Image Generation

This chapter is reproduced from:

Steed, R., & Caliskan, A. (2021). Image Representations Learned With Unsupervised Pre-Training Contain Human-like Biases. *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 701–713. https://doi.org/10.1145/3442188.3445932

4.1 Introduction

Can machines learn social biases from the way people are portrayed in image datasets? Companies and researchers regularly use machine learning models trained on massive datasets of images scraped from the web for tasks from face recognition (Hill, 2020) to image classification (Sun et al., 2017). To reduce costs, many practitioners use state-of-the-art models "pre-trained" on large datasets to help solve other machine learning tasks, a powerful approach called *transfer learning* (Tan et al., 2018). For example, HireVue used similar state-of-the-art computer vision and natural language models to evaluate job candidates' video interviews, potentially discriminating against candidates based on race, gender, or other social factors (Harwell, 2019). In this paper, we show how models trained on unlabeled images scraped from the web embed human-like biases, including racism and sexism.

Where most bias studies focus on supervised machine learning models, we seek to quantify learned patterns of implicit social bias in unsupervised image representations. Studies in supervised computer vision have highlighted social biases related to race, gender, ethnicity, sexuality, and other identities in tasks including face recognition, object detection, image search, and visual question answering (Buolamwini & Gebru, 2018a; Kay et al., 2015; Raji, Gebru, et al., 2020; Wilson et al., 2019; Manjunatha et al., 2019; Nex & Remondino, 2014). These algorithms are used in important real-world settings, from applicant video screening (Harwell, 2019; Raghavan et al., 2020) to autonomous vehicles (Geiger et al., 2012; Nex & Remondino, 2014), but their harmful downstream effects have been documented in applications such as online ad delivery (Sweeney, 1997) and image captioning (Hendricks et al., 2018).

Our work examines the growing set of computer vision methods in which no labels are used during model training. Recently, pre-training approaches adapted from language models have dramatically increased the quality of unsupervised image representations (Donahue & Simonyan, 2019; Bachman et al., 2019; He et al., 2020; Chen, Kornblith, Norouzi, & Hinton, 2020; Chen, 2020; Chen, Radford, et al., 2020; Misra & Van Der Maaten, 2020; Carion et al., 2020). With *fine-tuning*, practitioners can pair these general-purpose representations with labels from their domain to accomplish a variety of supervised tasks like face recognition or image captioning. We hypothesize that 1) like their counterparts in language, these unsupervised image representations also contain human-like social biases, and 2) these biases correspond to stereotypical portrayals of social group members in training images.

Results from natural language support this hypothesis. Several studies show that word embeddings, or representations, learned automatically from the way words co-occur in large text corpora exhibit human-like biases (Bolukbasi et al., 2016; Caliskan et al., 2017; Garg et al., 2018). Word embeddings acquire these biases via statistical regularities in language that are based on the co-occurrence of stereotypical words with social group signals. Recently, new deep learning methods for learning context-specific representations sharply advanced the state-of-the-art in natural language processing (NLP) (Devlin et al., 2018; Peters et al., 2018; Radford et al., 2019). Embeddings from these pre-trained models can be fine-tuned to boost performance in downstream tasks such as translation (Erhan et al., 2009, 2010). As with static embeddings, researchers have shown that embeddings extracted from contextualized

language models also exhibit downstream racial and gender biases (Zhao et al., 2017; Basta et al., 2019; Tan & Celis, 2019; Guo & Caliskan, 2020).

Recent advances in NLP architectures have inspired similar unsupervised computer vision models. We focus on two state-of-the-art, pre-trained models for image representation, iGPT (Chen, Radford, et al., 2020) and SimCLRv2 (Chen, 2020). We chose these models because they hold the highest fine-tuned classification scores, were pre-trained on the same large dataset of Internet images, and are publicly available. iGPT, or Image GPT, borrows its architecture from GPT-2 (Radford et al., 2019), a state-of-the-art unsupervised language model. iGPT learns representations for pixels (rather than for words) by pre-training on many unlabeled images (Chen, Radford, et al., 2020). SimCLRv2 uses deep learning to construct image representations from ImageNet by comparing augmented versions of the training images (Chen, Kornblith, Norouzi, & Hinton, 2020; Chen, 2020).

Do these unsupervised computer vision models embed human biases like their counterparts in natural language? If so, what are the origins of this bias? In NLP, embedding biases have been traced to word co-occurrences and other statistical patterns in text corpora used for training (Caliskan et al., 2017; Brunet et al., 2019; Blodgett et al., 2020). Both our models are pre-trained on ImageNet 2012, the most widely-used dataset of curated images scraped from the web (Russakovsky et al., 2015). In image datasets and image search results, researchers have documented clear correlations between the presence of individuals of a certain gender and the presence of stereotypical objects; for instance, the category "male" co-occurs with career and office related content such as ties and suits whereas "female" more often co-occurs with flowers in casual settings (Kay et al., 2015; Wang, Narayanan, & Russakovsky, 2020). As in NLP, we expect that these patterns of bias in the pre-training dataset will result in implicitly embedded bias in unsupervised models, even without access to labels during training.

This paper presents the Image Embedding Association Test (iEAT), the first systematic method for detecting and quantifying social bias learned automatically from unlabeled images.

- We find statistically significant racial, gender, and intersectional biases embedded in two state-of-the-art unsupervised image models pre-trained on ImageNet (Russakovsky et al., 2015), iGPT (Chen, Radford, et al., 2020) and SimCLRv2 (Chen, 2020).
- We test for 15 previously documented human and machine biases that have been studied for decades and validated in social psychology and conduct the first machine replication of Implicit Association Tests (IATs) with picture stimuli (Greenwald et al., 1998).
- In 8 tests, our machine results match documented human biases, including 4 of 5 biases also found in large language models. The 7 tests which did not show significant human-like biases are from IATs with only small samples of picture stimuli.
- With 16 novel tests, we show how embeddings from our model confirm several hypotheses about intersectional bias from social psychology (Ghavami & Peplau, 2013).
- We compare our results to statistical analyses of race and gender in image datasets. Unsupervised models seem to learn bias from the ways people are commonly portrayed

in images on the web.

- We present a qualitative case study of how image generation, a downstream task utilizing unsupervised representations, exhibits a bias towards the sexualization of women.

4.2 Related Work

Various tests have been constructed to quantify bias in unsupervised natural language models (Caliskan et al., 2017; Zhao et al., 2017; Basta et al., 2019; May et al., 2019), but to our knowledge, there are no principled tests for measuring bias embedded in *unsupervised* computer vision models. Wang, Narayanan, and Russakovsky (2020) develop a method to automatically recognize bias in visual datasets but still rely on human annotations. Our method uses no annotations whatsoever. In NLP, there are several systematic approaches to measuring unsupervised bias in word embeddings (Caliskan et al., 2017; May et al., 2019; Tan & Celis, 2019; Guo & Caliskan, 2020; Bommasani et al., 2020; Kurita et al., 2019). Most of these tests take inspiration from the well-known IAT (Greenwald et al., 1998; Greenwald et al., 2003). Participants in the IAT are asked to rapidly associate stimuli, or exemplars, representing two target concepts (e.g. "flowers" and "insects") with stimuli representing evaluative attributes (e.g. "pleasant" and "unpleasant") attribute (Greenwald et al., 1998). Assuming that the cognitive association task is easier when the strength of implicit association between the target concept and attributes is high, the IAT quantifies bias as the latency of response (Greenwald et al., 1998) or the rate of classification error (Nosek & Banaji, 2001). Stimuli may take the form of words, pictures, or even sounds (Nosek, Greenwald, & Banaji, 2007), and there are several IATs with picture-only stimuli (Nosek, Greenwald, & Banaji, 2007).

Notably, Caliskan et al. (2017) adapt the heavily-validated IAT (Greenwald et al., 1998) from social psychology to machines by testing for the mathematical association of word embeddings rather than response latency. They present a systematic method for measuring language biases associated with social groups, the Word Embedding Association Test (WEAT). Like the IAT, the WEAT measures the effect size of bias in static word embeddings by quantifying the relative associations of two sets of target stimuli (e.g., {"woman," "female"} and {"man," "male"}) that represent social groups with two sets of evaluative attributes (e.g., {"science," mathematics"} and {"arts," "literature"}). For validation, two WEATs quantify associations towards flowers vs. insects and towards musical instruments vs. weapons, both accepted baselines Greenwald et al. (1998). Greenwald et al. (1998) refer to these baseline biases as "universally" accepted stereotypes since they are widely shared across human subjects and are not potentially harmful to society. Other WEATs measure social group biases such as sexist and racist associations or negative attitudes towards the elderly or people with disabilities. In any modality, implicit biases can potentially be prejudiced and harmful to society. If downstream applications use these representations to make consequential decisions about human beings, such as automated video job interview evaluations, machine learning may perpetuate existing biases and exacerbate historical injustices Raghavan et al., 2020; De-Arteaga et al., 2019.

The original WEAT (Caliskan et al., 2017) uses *static* word embedding models such as

word2vec (Mikolov et al., 2013) and GloVe Pennington et al., 2014, each trained on Internet-scale corpora composed of billions of tokens. Recent work extends the WEAT to *contextualized* embeddings: dynamic representations based on the context in which a token appears. May et al. (2019) insert targets and attributes into sentences like "This is a[n] ¡word¿" and applying WEAT to the vector representation for the whole sentence, with the assumption that the sentence template used is "semantically bleached" (such that the only meaningful content in the sentence is the inserted word). Tan and Celis (2019) extract the contextual word representation for the token of interest before pooling to avoid confounding effects at the sentence level; in contrast, Bommasani et al. (2020) find that pooling tends to improve representational quality for bias evaluation. Guo and Caliskan (2020) dispense with sentence templates entirely, pooling across n word-level contextual embeddings for the same token extracted from random sentences. Our approach is closest to these latter two methods, though we pool over images rather than words.

4.3 Approach

In this paper, we adapt bias tests designed for contextualized word embeddings to the image domain. While language transformers produce contextualized *word* representations to solve the next *token* prediction task, an image transformer model like iGPT generates *image* representations to solve the next *pixel* prediction task (Chen, Radford, et al., 2020). Unlike words and tokens, pixels do not explicitly correspond to semantic concepts (objects or categories) as words do. In language, a single token (e.g. "love") corresponds to the target concept or attribute (e.g. "pleasant"). But in images, no single pixel corresponds to a semantically meaningful concept. To address the abstraction of semantic representation in the image domain, we propose the Image Embedding Association Test (iEAT), which modifies contextualized word embedding tests to compare pooled image-level embeddings. The goal of the iEAT is to measure the biases embedded during unsupervised pre-training by comparing the relative association of image embeddings in a systematic process. Chen, Radford, et al. (2020) and Chen, Kornblith, Norouzi, and Hinton (2020) show through image classification that unsupervised image features are good representations of object appearance and categories; we expect they will also embed information gleaned from the common co-occurrence of certain objects and people and therefore contain related social biases.

Our approach is summarized in Figure 4.1. The iEAT uses the same formulas for the test statistic, effect size d, and p-value as the WEAT (Caliskan et al., 2017), described in Section 4.3.3. Section 4.3.1 summarizes our approach to replicating several different IATs; Section 4.3.2 describes several novel intersectional iEATs. Section 4.3.3 describes our test statistic, drawn from embedding association tests like the WEAT.

4.3.1 Replication of Bias Tests

In this paper, we validate the iEAT by replicating as closely as possible several common IATs. These tests fall into two broad categories: valence tests, in which two target concepts are tested for association with "pleasant" and "unpleasant" images; and stereotype tests, in

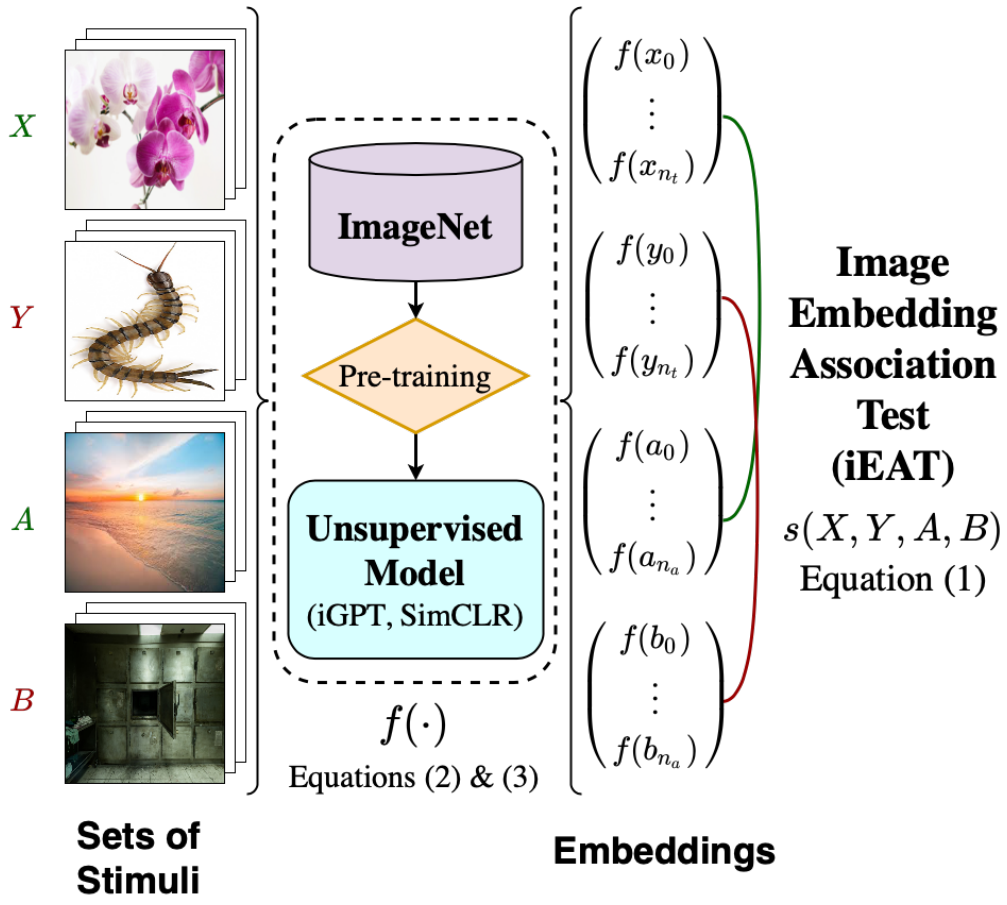


Figure 4.1: Example iEAT replication of the Insect-Flower IAT (Greenwald et al., 1998), which measures the differential association between flowers vs. insects and pleasantness vs. unpleasantness.

which two target concepts are tested for association with a pair of stereotypical attributes (e.g. "male" vs. "female" "career" vs. "family"). To closely match the ground-truth human IAT data and validate our method, our replications use the same concepts as the original IATs (listed in Table 4.1). Because some IATs rely on verbal stimuli, we adapt them to images, using image stimuli from the IATs when available. When no previous studies use image stimuli, we map the non-verbal stimuli to images using the data collection method described in Section 4.5.

Many of these bias tests have been replicated for machines in the language domain; for the first time, we also replicate tests with image-only stimuli, including the Asian and Native American IATs. Most of these tests were originally administered in controlled laboratory settings (Greenwald et al., 1998; Greenwald et al., 2003), and all except for the Insect-Flower IAT have also been tested on the Project Implicit website at http://projectimplicit.org (Nosek et al., 2002; Greenwald et al., 2003, 2009). Project Implicit has been available worldwide for over 20 years; in 2007, the site had collected more than 2.5 million IATs. The average effect sizes (which are based on samples so large the power is nearly 100%) for these tests are reproduced in Table 4.1. To establish a principled methodology, all the IAT verbal and original image stimuli for our bias tests were replicated exactly from this online IAT platform

(Nosek, Smyth, et al., 2007). We will treat these results, along with the laboratory results from the original experiments (Greenwald et al., 1998), as ground-truth for human biases that serve as validation benchmarks for our methods (Section 4.6).

4.3.2 Intersectional iEATs

We also introduce several new tests for intersectional valence bias and bias at the intersection of gender stereotypes and race. Intersectional stereotypes are often even more severe than their constituent stereotypes (Crenshaw, 1990). Following Tan and Celis (2019), we anchored comparison on White males, the group with the most representation, and compared against White females, Black males, and Black females, respectively (Table 4.2). Drawing on social psychology (Ghavami & Peplau, 2013), we pose three hypotheses about intersectional bias:

- *Intersectionality hypothesis*: tests at the intersection of gender and race will reveal emergent biases not explained by the sum of biases towards race and gender alone.
- *Race hypothesis*: biases between racial groups will be more similar to differential biases between the men than between the women.
- *Gender hypothesis*: biases between men and women will be most similar to biases between White men and White women.

4.3.3 Embedding Association Tests

Though our stimuli are images rather than words, we can use the same statistical method for measuring biased associations between image representations (Caliskan et al., 2017) to quantify a standardized effect size of bias. We follow Caliskan et al. (2017) in describing the WEAT here.

Let X and Y be two sets of target concepts embeddings of size N_t, and let A and B be two sets of attribute embeddings of size N_a. For example, the Gender-Career IAT tests for the differential association between the concepts "male" (A) and "female" (B) and the attributes "career" (X) and "family" (Y). Generally, experts in social psychology and cognitive science select stimuli that are typically representative of various concepts. In this case, A contains embeddings for verbal stimuli such as "boy," "father," and "man," while X contains embeddings for verbal stimuli like "office" and "business." These linguistic, visual, and sometimes auditory stimuli are proxies for the aggregate representation of a concept in cognition. Embedding association tests use these unambiguous stimuli as semantic representations to study biased associations between the concepts being represented. Since the stimuli are chosen by experts to most accurately represent concepts, they are not polysemous or ambiguous tokens. We use these expert-selected stimuli as the basis for our tests in the image domain.

The test statistic measures the differential association of the target concepts X and Y with

the attributes A and B

$$s(X, Y, A, B) = \sum_{x \in X} s(x, A, B) - \sum_{y \in Y} s(y, A, B) \tag{1}$$

where $s(w, A, B)$ is the differential association of w with the attributes, quantified by the cosine similarity of vectors

$$s(w, A, B) = \text{mean}_{a \in A} \cos(w, a) - \text{mean}_{b \in B} \cos(w, b)$$

We test the significance of this association with a permutation test[1] over all possible equal-size partitions $\{(X_i, Y_i)\}_i$ of $X \cup Y$ to generate a null hypothesis as if no biased associations existed. The one-sided p-value measures the unlikelihood of the null hypothesis

$$p = Pr[s(X_i, Y_i, A, B) > s(X, Y, A, B)]$$

and the effect size, a standardized measure of the separation between the relative association of X and Y with A and B, is

$$d = \frac{\text{mean}_{x \in X} s(x, A, B) - \text{mean}_{y \in Y} s(y, A, B)}{\text{std}_{w \in X \cup Y} s(w, A, B)}$$

A larger effect size indicates a larger differential association; for instance, the large effect size d in Table 4.1 for the gender-career bias example above indicates that in human respondents, "male" is strongly associated with "career" attributes compared to "female," which is strongly associated with "family" attributes. Note that these effect sizes cannot be directly compared to effect sizes in human IATs, but the significance levels *are* uniformly high. Human IATs measure individual people's associations; embedding association tests measure the aggregate association in the representation space learned from the training set. In general, significance increases with the number of stimuli; an insignificant result does not necessarily indicate a lack of bias.

One important assumption of the iEAT is that categories can be meaningfully represented by groups of images, such that the association bias measured refers to the categories of interest and not some other, similar-looking categories. Thus, a positive test result indicates only that there is an association bias between the corresponding samples' sets of target images and attribute images. To generalize to associations between abstract social concepts requires that the samples adequately represent the categories of interest. Section 4.5 details our procedure for selecting multiple, representative stimuli, following validated approaches from prior work Greenwald et al., 1998.

We use an adapted version of May et al. (2019)'s Python WEAT implementation. All code, pre-trained models, and data used to produce the figures and results in this paper can be accessed at github.com/ryansteed/ieat.

[1] We use an exact, non-parametric permutation test over all possible partitions. There are no normality assumptions about the distribution of the null hypothesis.

4.4 Computer Vision Models

To explore what kinds of biases may be embedded in image representations generated in unsupervised settings, where class labels are not available for images, we focus on two computer vision models published in summer 2020, iGPT and SimCLRv2. We extract representations of image stimuli with these two pre-trained, unsupervised image representation models. We choose these particular models because they achieve state-of-the-art performance in *linear evaluation* (a measure of the accuracy of a linear image classifier trained on embeddings from each model). iGPT is the first model to learn from pixel co-occurrences to generate image samples and perform image completion tasks.

4.4.0.1 Pre-training Data

Both models are pre-trained on ImageNet 2012, a large benchmark dataset for computer vision tasks (Russakovsky et al., 2015).[2] ImageNet 2012 contains 1.2 million annotated images of 200 object classes, including a person class; even if the annotated object is not a person, a person may appear in the image. For this reason, we expect the models to be capable of generalizing to stimuli containing people (Russakovsky et al., 2013, 2015). While there are no publicly available pre-trained models with larger training sets, and the "people" category of ImageNet is no longer available, this dataset is a widely used benchmark containing a comprehensive sample of images scraped from the web, primarily Flickr (Russakovsky et al., 2015). We assume that the portrayals of people in ImageNet are reflective of the portrayal of people across the web at large, but a more contemporary study is left to future work. CIFAR-100, a smaller classification database, was also used for linear evaluation and stimuli collection (Krizhevsky, 2009).

4.4.0.2 Image Representations

Both models are *unsupervised*: neither use any labels during training. Unsupervised models learn to produce embeddings based on the implicit patterns in the entire training set of image features. Both models incorporate neural networks with multiple hidden layers (each learning a different level of abstraction) and a projection layer for some downstream task. For linear classification tasks, features can be drawn directly from layers in the base neural network. As a result, there are various ways to extract image representations, each encoding a different set of information. We follow Chen, Radford, et al. (2020) and Chen, Kornblith, Norouzi, and Hinton (2020) in choosing the features for which linear evaluation scores are highest such that the features extracted contain high-quality, general-purpose information about the objects in the image. Below, we describe the architecture and feature extraction method for each model.

[2]Both models were tested on the Tensorflow version of ILSVRC 2012, available at https://www.tensorflow.org/datasets/catalog/imagenet2012.

4.4.1 iGPT

The Image Generative Pre-trained Transformer (iGPT) model is a novel, NLP-inspired approach to unsupervised image representation. We chose iGPT for its high linear evaluation scores, minimalist architecture, and strong similarity to GPT-2 (Radford et al., 2019), a transformer-based architecture that has found great success in the language domain. Transformers learn patterns in the way individual tokens in an input sequence appear with other tokens in the sequence (Vaswani et al., 2017). Chen, Radford, et al. (2020) apply a structurally simple, highly parameterized version of the GPT-2 generative language pre-training architecture (Radford et al., 2019) to the image domain for the first time. GPT-2 uses the "contextualized embeddings" learned by a transformer to predict the next token in a sequence and generate realistic text (Radford et al., 2019). Rather than autoregressively predict the next entry in a sequence of tokens as GPT-2 does, iGPT predicts the next entry in a flattened sequence of pixels. iGPT is trained to autoregressively complete cropped images, and feature embeddings extracted from the model can be used to train a state-of-the-art linear classifier (Chen, Radford, et al., 2020).

We use the largest open-source version of this model, iGPT-L 32x32, with $L = 48$ layers and embedding size 1536. All inputs are restricted to 32x32 pixels; the largest model, which takes 64x64 input, is not available to the public. Original code and checkpoints for this model were obtained from its authors at github.com/openai/image-gpt. iGPT is composed of L blocks

$$\begin{aligned} n^l &= \text{layer\_norm}(h^l) \\ a^l &= h^l + \text{multihead\_attention}(n^l) \\ h^{l+1} &= a^l + \text{mlp}(\text{layer\_norm}(a^l)) \end{aligned}$$

where h^l is the input tensor to the l^{th} block. In the final layer, called the *projection head*, Chen, Radford, et al. (2020) learn a projection from $n^L = \text{layer\_norm}(h^L)$ to a set of logits parameterizing the conditional distributions across the sequence dimension. Because this final layer is designed for autoregressive pixel prediction, the final layer may not contain the optimal representations for object recognition tasks. Chen, Radford, et al. (2020) obtain the best linear classification results using embeddings extracted from a middle layer - specifically, somewhere near the 20th layer (Chen, Radford, et al., 2020). A linear classifier trained on these features is much more accurate than one trained on the next-pixel embeddings (Chen, Radford, et al., 2020). Such "high-quality" features from the middle of the network f^l are obtained by average-pooling the layer norm across the sequence dimension:

$$f^l = \langle n_i^l \rangle_i \tag{2}$$

Chen, Radford, et al. (2020) then learn a set of *class* logits from f^l for their fine-tuned, supervised linear classifier, but we will just use the embeddings f^{20}. In general, we prefer these embeddings over embeddings from other layers for two reasons: 1) they can be more closely compared to the SimCLRv2 embeddings, which are also optimal for fine-tuning a linear classifier; 2) we hypothesize that embeddings with higher linear evaluation scores will also be more likely to embed biases, since stereotypical portrayals typically incorporate certain objects and scenes (e.g. placing men with sports equipment). In Appendix D.3, we try another embedding extraction strategy and show that this hypothesis is correct.

4.4.2 SimCLR

The Simple Framework for Contrastive Learning of Visual Representations (SimCLR) (Chen, Kornblith, Norouzi, & Hinton, 2020; Chen, Kornblith, Swersky, et al., 2020) is another state-of-the-art unsupervised image classifier. We chose SimCLRv2 because it has a state-of-the-art open source release and for variety in architecture: unlike iGPT, SimCLRv2 utilizes a traditional neural network for image encoding, ResNet (He et al., 2016). SimCLRv2 extracts representations in three stages: 1) data augmentation (random cropping, random color distortions, and Gaussian blur); 2) an encoder network, ResNet (He et al., 2016); 3) mapping to a latent space for contrastive learning, which maximizes agreement between the different augmented views (Chen, Kornblith, Norouzi, & Hinton, 2020). These representations can be used to train state-of-the-art linear image classifiers (Chen, Kornblith, Norouzi, & Hinton, 2020; Chen, Kornblith, Swersky, et al., 2020). We use the largest pre-trained open-source version (the model with the highest linear evaluation scores) of SimCLRv2 (Chen, Kornblith, Swersky, et al., 2020), obtained from its authors at github.com/google-research/simclr. This pre-trained model uses a 50-layer ResNet with width 3× and selective kernels (which have been shown to increase linear evaluation accuracy), and it was also pre-trained on ImageNet (Russakovsky et al., 2015).

As with iGPT, we extract the embeddings identified by Chen, Kornblith, Norouzi, and Hinton (2020) as "high-quality" features for linear evaluation. Following (Chen, Kornblith, Norouzi, & Hinton, 2020), let $\tilde{x}_i$ and $\tilde{x}_j$ be two data augmentations (random cropping, random color distortion, and random Gaussian blur) of the same image. The base encoder network $f(\cdot)$ is a network of L layers

$$h_i = f(\tilde{x}_i) = \text{ResNet}(\tilde{x}_i) \tag{3}$$

where $h_i \in \mathbb{R}^d$ is the output after the average pooling layer. During pre-training, SimCLRv2 utilizes an additional layer: a projection head $g(\cdot)$ that maps h_i to a latent space for contrastive loss. The contrastive loss function can be found in (Chen, Kornblith, Norouzi, & Hinton, 2020).

After pre-training, Chen, Kornblith, Norouzi, and Hinton (2020) discard the projection head $g(\cdot)$, using the average pool output $f(\cdot)$ for linear evaluation. Note that the projection head $g(h)$ is still necessary for pre-training high-quality representations (it improves linear evaluation accuracy by over 10%); but Chen, Kornblith, Norouzi, and Hinton (2020) find that training on h rather than $z = g(h)$ also improves linear evaluation accuracy by more than 10%. We follow suit, using h_i (the average pool output of ResNet) to represent our image stimuli, which has dimensionality $2,048$. High dimensionality is not a great obstacle; association tests have been used with embeddings as large as $4,096$ dimensions (May et al., 2019).

4.5 Stimuli

To replicate the IATs, we systematically compiled a representative set of image stimuli for each of the concepts, or categories, listed in Table 4.1. Rather than attempting to specify

and justify new constructs, we adhere as closely as possible to stimuli defined and employed by well-validated psychological studies. For each category (e.g. "male" or "science") in each IAT (e.g. Gender-Science), we drew representative images from either 1) the original IAT stimuli, if the IAT used picture stimuli (Nosek, Smyth, et al., 2007), 2) the CIFAR-100 dataset (Krizhevsky, 2009), or 3) a Google Image Search.

This section describes how we obtained a set of images that meaningfully represent some target concept (e.g. "male") or attribute (e.g. "science") as it is normally, or predominantly, portrayed in society and on the web. We follow the stimuli selection criteria outlined in foundational prior work to collect the most typical and accurate exemplars (Greenwald et al., 1998; Greenwald et al., 2003). For picture-IATs with readily available image stimuli, we accept those stimuli as representative and exactly replicate the IAT conditions, with two exceptions: 1) the weapon-tool IAT picture stimuli include outdated objects (e.g. cutlass, Walkman), so we chose to collect an additional, modernized set of images; 2) the disability IAT utilizes abstract symbols, so we collected a replacement set of images of real people for consistency with the training set. For IATs with verbal stimuli, we use Google Image Search as a proxy for the predominant portrayal of words (expressed as search terms) on the web (described in Section 4.5.1). Human IATs employ the same philosophy: for example, the Gender-Science IAT uses common European American names to represent male and female, because the majority of names in the U.S. are European American (Nosek et al., 2002). We follow the same approach in replicating the human IATs for machines in the vision domain.

One consequence of the stimuli collection approach outlined in Section 4.5.1 is that our test set will be biased towards certain demographic groups, just as the Human IATs are biased towards European American names. For example, Kay et al. (2015) showed that in 2015, search results for powerful occupations like CEO systematically under-represented women. In a case like this, we would expect to underestimate bias towards minority groups. For example, since we expect Gender-Science biases to be higher for non-White women, a test set containing more White women than non-White would exhibit lower overall bias than a test set containing an equal number of stimuli from white and non-White women. Consequently, tests on Google Image Search stimuli would be expected to result in under-estimated stereotype-congruent bias scores. While under-representation in the test set does not pose a major issue for measuring normative concepts, we cannot use the same datasets to test for intersectional bias. For those iEATs, we collected separate, equal-sized sets of images with search terms based on the categories White male, White female, Black male, and Black female, since none of the IATs specifically target these intersectional groups.

4.5.1 Verbal to Image Stimuli

One key challenge of our approach is representing social constructs and abstract concepts such as "male" or "pleasantness" in images. A Google Image Search for "pleasantness" returns mostly cartoons and pictures of the word itself. We address this difficulty by adhering as closely as possible to the verbal IAT stimuli, to ensure the validity of our replication. In verbal IATs, this is accomplished with "buckets" of verbal exemplars that include a variety of common-place and easy-to-process realizations of the concept in question. For example, in

the Gender-Science IAT, the concept "male" is defined by the verbal stimuli "man," "son," "father," "boy," "uncle," "grandpa," "husband," and "male" (Xu et al., 2014). To closely match the representations tested by these IATs, we use these sets of words to search for substitute image stimuli that portray one of these words or phrases. For the vast majority of exemplars, we were able to find direct visualizations of the stimuli as an isolated person, object, or scene. For example, Figure 4.1 depicts sample image stimuli corresponding to the verbal stimuli "orchid" (for category "flower"), "centipede" ("insect"), "sunset" ("pleasant"), and "morgue" ("unpleasant").[3]

We collected images for each verbal stimulus from either CIFAR-100[4] or Google Image Search according to a systematic procedure detailed in Appendix D.2. This procedure controls for image characteristics that might confound the category we are attempting to define (e.g. lighting, background, dominant colors, placement) in several ways: 1) we collected more than one for each verbal stimulus, in case of idiosyncrasies in the images collected; 2) for stimuli referring to an object or person, we chose images that isolated the object or person of interest against a plain background, unless the object filled the whole image; 3) when an attribute stimulus refers to a group of people, we chose only images where the target concepts were evenly represented in the attribute images;[5] 4) for the picture-IATs, we accepted the original image stimuli to exactly reconstruct the original test conditions. We also did not alter the original verbal stimuli, relying instead on the construct validity of the original IAT experiments.[6] For each verbal stimulus, Appendix D.2 lists corresponding search terms and the precise number of images collected. All the images used to represent the concepts being tested are available at github.com/ryansteed/ieat.

4.5.2 Choosing Valence Stimuli

Valence, the intrinsic pleasantness or goodness of things, is one of the principal dimensions of affect and cognitive heuristics that shape attitudes and biases (Greenwald et al., 1998). Many IATs quantify implicit bias by comparing two social groups to the valence attributes "pleasant" vs. "unpleasant." Here, positive valence will denote "pleasantness" and negative valence will denote "unpleasantness." The verbal exemplars for valence vary slightly from test to test. Rather than create a new set of image stimuli for each valence IAT, we collected one, large consolidated set from an experimentally validated database (Bellezza et al., 1986) of low and high valence words (e.g. "rainbow," "morgue") commonly used in the valence IATs. To quantify norms, (Bellezza et al., 1986) asked human participants to rate these non-social

[3] In the original IATs, the category set sizes N_t and N_a range from 5-15 exemplars. We collected $n \approx 5$ images for each exemplar such that N_t and N_a are 30-50. Significance could be increased by including more stimuli, at the risk of diluting the test set with less-representative images from farther down in the search results.

[4] We first check for test images in CIFAR-100 because iGPT performs well in out-of-sample linear evaluation on this dataset (Chen, Kornblith, Norouzi, & Hinton, 2020).

[5] For example, for the "family" attribute in the Gender-Career test, we chose only images of families with equal numbers of men and women.

[6] One exception: the Gender-Career IAT used specific male- and female-sounding names, rather than general exemplars like "man" or "father" as in the Gender-Science IAT. We use the general exemplars for both tests.

words for "pleasantness" and "imagery" in a controlled laboratory setting. Because some of the words for valence do not correspond to physical objects, we collected images for verbal stimuli with high valence and imagery scores. We used the same procedure as for all the other verbal stimuli (described above in Section 4.5.1). The full list of verbal valence stimuli can be found in Appendix D.1.

4.6 Evaluation

We evaluate the validity of iEAT by comparing the results to human and natural language biases measured in prior work. We obtain stereotype-congruent results for baseline, or "universal," biases. We also introduce a simple experiment to test how often the iEAT incorrectly finds bias in a random set of stimuli.

Predictive Validity. We posit that iEAT results have predictive validity if they correspond to ground-truth IAT results for humans or WEAT results in word embeddings. In this paper, we validate the iEAT by replicating several human IATs as closely as possible (as described in Section 4.5) and comparing the results. We find that embeddings extracted from at least one of the two models we test display significant bias for 8 of the 15 ground-truth human IATs we replicate (Section 4.7). The insignificant biases are likely due to small sample sizes. We also find evidence supporting each of the intersectional hypotheses listed in Section 4.3.2, which have also been empirically validated in a study with human participants (Ghavami & Peplau, 2013).

Baselines. As a baseline, we replicate a "universal" bias test presented in the first paper introducing the IAT (Greenwald et al., 1998): the association between flower vs. insects and pleasant vs. unpleasant. If human-like biases are encoded in unsupervised image models, we would expect a strong and statistically significant flower-insect valence bias, for two reasons: 1) as Greenwald et al. (1998) conjecture, this test measures a close-to-universal baseline human bias; 2) our models (described in Section 4.4) achieve state-of-the-art performance when classifying simple objects including flowers and bees.[7] The presence of universal bias and absence of random bias suggests our conclusions are valid for other social biases.

Specificity. Prior work on embedding association tests does not evaluate the false positive rate. To validate the specificity of our significance estimation, we created 1,000 random partitions of $X \cup Y \cup A \cup B$ from the flower-insect test to evaluate true positive detection. Our false positive rate is roughly bounded by the p-value: 10.3% of these random tests resulted in a false positive at $p < 10^{-1}$; 1.2% were statistically significant false positives at $p < 10^{-2}$.

4.7 Experiments and Results

In correspondence with the human IAT, we find several significant racial biases and gender stereotypes, including intersectional biases, shared by both iGPT and SimCLRv2 when

[7] A linear image classifier trained on iGPT embeddings reaches 88.5% accuracy on CIFAR-100; SimCLRv2 embeddings reach 89% accuracy (Chen, Radford, et al., 2020).

pre-trained on ImageNet.

4.7.1 iEATs

Effect sizes and p-values from the permutation test for each bias type measurement are reported in Table 4.1 and interpreted below.

4.7.1.1 Widely Accepted Biases

First, we apply the iEAT to the widely accepted baseline Insect-Flower IAT, which measures the association of insects and flowers with pleasantness and unpleasantness, respectively. As hypothesized, we find that embeddings from both models contain significant positive biases in the same direction as the human participants, associating flowers with pleasantness and insects with unpleasantness, with $p < 10{-}1$ (Table 4.1). Notably, the magnitude of bias is greater for SimCLRv2 (effect size 1.69, $p < 10^{-3}$) than for iGPT (effect size 0.34, $p < 10^{-1}$). In general, SimCLRv2 embeddings contain stronger biases than iGPT embeddings but do not contain as many kinds of bias. We conjecture that because SimCLRv2 transforms images before training (including color distortion and blurring) and is more architecturally complex than iGPT (Chen, Kornblith, Norouzi, & Hinton, 2020), its embeddings become more suitable for concrete object classification as opposed to implicit social patterns.

4.7.1.2 Racial Biases

Both models display statistically significant racial biases, including both valence and stereotype biases. The racial attitude test, which measures the differential association of images of European Americans vs. African Americans with pleasantness and unpleasantness, shows no significant biases. But embeddings extracted from both models exhibit significant bias for the Arab-Muslim valence test, which measures the association of images of Arab-Americans vs. others with pleasant vs. unpleasant images. Also, embeddings extracted with iGPT exhibit strong bias large effect size (effect size 1.26, $p < 10{-}2$) for the Skin Tone test, which compares valence associations with faces of lighter and darker skin tones. These findings relate to anecdotal examples of software that claim to make faces more attractive by lightening their skin color. Both iGPT and SimCLRv2 embeddings also associate White people with tools and Black people with weapons in both classical and modernized versions of the Weapon IAT.

4.7.1.3 Gender Biases

There are statistically significant gender biases in both models, though not for both stereotypes we tested. In the Gender-Career test, which measures the relative association of the category "male" with career attributes like "business" and "office" and the category "female" with family-related attributes like "children" and "home," embeddings extracted from both models exhibit significant bias (iGPT effect size 0.62, $p < 10^{-2}$, SimCLRv2 effect size 0.74, $p < 10^{-3}$). This finding parallels Kay et al. (2015)'s observation that image search results for powerful occupations like CEO systematically under-represented women. In the Gender-Science test, which measures the association of "male" with "science" attributes like math and engineering

and "female" with "liberal arts" attributes like art and writing, only iGPT displays significant bias (effect size 0.44, $p < 10^{-1}$).

4.7.1.4 Other Biases

For the first time, we attempt to replicate several other tests measuring weight stereotypes and attitudes towards the elderly or people with disabilities. iGPT displays an additional bias (effect size 1.67, $p = 10^{-4}$) towards the association of thin people with pleasantness and overweight people with unpleasantness. We found no significant bias for the Native American or Asian American stereotype tests, the Disability valence test, or the Age valence test. For reference, significant age biases have been detected in static word embeddings; the others have not been tested because they use solely image stimuli (Caliskan et al., 2017). Likely, the target sample sizes for these tests are too low; all three of these tests use picture stimuli from the original IAT, which are all limited to fewer than 10 images. Replication with an augmented test set is left to future work. Note that lack of significance in a test, even if the sample size is sufficiently large, does not indicate the embeddings from either model are definitively bias-free. While these tests did not *confirm* known human biases regarding foreigners, people with disabilities, and the elderly, they also did not *contradict* any known human-like biases.

4.7.2 Intersectional Biases

4.7.2.1 Intersectional Valence

Intersectional valence tests with the iGPT embeddings are the most consistent with social psychology, exhibiting results predicted by the intersectionality, race, and gender hypotheses listed in Section 4.3 (Ghavami & Peplau, 2013). Overall, iGPT embeddings contain a positive valence bias towards White people and a negative valence bias towards Black people (effect size 1.16, $p < 10^{-3}$), as in the human Race IAT (Nosek, Smyth, et al., 2007). As predicted by the race hypothesis, the same bias is significant but less severe for both White males vs. Black males (iGPT effect size 0.88, $p < 10^{-2}$) and White males vs. Black females (iGPT effect size 0.83, $p < 10^{-2}$), and the White female vs. Black female bias is insignificant; in general, race biases are more similar to the race biases between men. We hypothesize that as in text corpora, computer vision datasets are dominated by the majority social groups (men and White).

As predicted by the gender hypothesis, our results also conform with the theory that females are associated with positive valence when compared to males (Eagly et al., 1991), but only when those groups are White (iGPT effect size 0.79, $p < 10^{-2}$); there is no significant valence bias for Black females vs. Black males. This insignificant result might be due to the under-representation of Black people in the visual embedding space. The largest differential valence bias of all our tests emerges between White females and Black males; White females are associated with pleasant valence and Black males with negative valence (iGPT effect size 1.46, $p < 10^{-3}$).

4.7.2.2 Intersectional Stereotypes

We find significant but contradictory intersectional differences in gender stereotypes (Table 4.2). For Gender-Career stereotypes, the iGPT-encoded bias for White males vs. Black females is insignificant though there is a bias (effect size 0.81, $p < 10^{-3}$) for male vs. female in general. There is significant Gender-Career stereotype bias between embeddings of White males vs. White females (iGPT effect size 0.97, $p < 10^{-3}$), even higher than the general case; this result conforms to the race hypothesis, which predicts gender stereotypes are more similar to the stereotypes between Whites than between Blacks. The career-family bias between White males and Black males is reversed; embeddings for images of Black males are more associated with career and images of White men with family (iGPT effect size 0.89, $p < 10^{-2}$). One explanation for this result is under-representation; there are likely fewer photos depicting Black men with non-stereotypical male attributes.

Unexpectedly, the intersectional test of male vs. female (with equal representation for White and Black people) reports no significant Gender-Science bias, though the normative test (with unequal representation) does (Table 4.1). Nevertheless, race-science stereotypes do emerge when White males are compared to Black males (iGPT effect size 0.49, $p < 10^{-1}$) and, to an even greater extent, when White males are compared to Black females (iGPT effect size 0.80, $p < 10^{-2}$), confirming the intersectional hypothesis (Ghavami & Peplau, 2013). But visual Gender-Science biases do not conform to the race hypothesis; the gender stereotype between White males and White females is insignificant, though the overall male vs. female bias is not.

4.7.3 Origins of Bias

4.7.3.1 Bias in Web Images

Do these results correspond with our hypothesis that biases are learned from the co-occurrence of social group members with certain stereotypical or high-valence contexts? Both our models were pre-trained on ImageNet, which is composed of images collected from Flickr and other Internet sites (Russakovsky et al., 2015). Yang et al. (2020) show that the ImageNet categories unequally represent race and gender; for instance, the "groom" category may contain mostly White people. Under-representation in the training set could explain why, for instance, White people are more associated with pleasantness and Black people with unpleasantness. There is a similar theory in social psychology: most bias takes the form of in-group favoritism, rather than out-group derogation (Hewstone et al., 2002). In image datasets, favoritism could take the form of unequal representation and have similar effects. For example, one of the exemplars for "pleasantness" is "wedding," a positive-valence, high imagery word (Bellezza et al., 1986); if White people appear with wedding paraphernalia more often than Black people, they could be automatically associated with a concept like "pleasantness," even though no explicit labels for "groom" and "White" are available during training.

Likewise, the portrayal of different social groups in context may be automatically learned by unsupervised image models. Wang, Narayanan, and Russakovsky (2020) find that in OpenImages (also scraped from Flickr) (Kuznetsova et al., 2018), a similar benchmark classification dataset, a higher proportion of "female" images are set in the scene "home or

(a) Cropped image of an artificial ("Generated Photos," 2021) White male face.

(b) 8 random autoregressive completions of the cropped image. 6 depict career-related attire.

Figure 4.2: Example of career associations in image completion of a male face with iGPT, pre-trained on ImageNet.

hotel" than "male" images. "male" is more often depicted in "industrial and construction" scenes. This difference in portrayal could account for the Gender-Career biases embedded in unsupervised image embeddings. In general, if the portrayal of people in Internet images reflects human social biases that are documented in cognition and language, we conclude that unsupervised image models could automatically learn human-like biases from large collections of online images.

4.7.3.2 Bias in Autoregression

Though the next-pixel prediction features contained very little significant bias, they may still propagate stereotypes in practice. For example, the incautious and unethical application of a generative model like iGPT could produce biased depictions of people. As a qualitative case study, we selected 5 male- and 5 female-appearing artificial faces from a database ("Generated Photos," 2021) generated with StyleGAN (Karras et al., 2019). We decided to use images of non-existent people to avoid perpetuating any harm to real individuals. We cropped the portraits below the neck and used iGPT to generate 8 different completions (with the temperature hyperparameter set to 1.0, following Chen, Radford, et al. (2020)). We found that completions of woman *and* men are often sexualized: for female faces, 52.5% of completions featured a bikini or low-cut top; for male faces, 7.5% of completions were shirtless or wore low-cut tops, while 42.5% wore suits or other career-specific attire. One held a gun. This behavior might result from the sexualized portrayal of people, especially women, in internet images (Graff et al., 2013) and serves as a reminder of computer vision's controversial history with Playboy centerfolds and objectifying images (Iozzio, 2016). To avoid promoting negative biases, Figure 4.2 shows only an example of male-career associations in completions of a GAN-generated face.

4.8 Discussion

By testing for bias in unsupervised models pre-trained on a widely used large computer vision dataset, we show how biases may be learned automatically from images and embedded in general-purpose representations. Not only do we observe human-like biases in the majority of

our tests, but we also detect 4 of the 5 human biases replicated in natural language (Caliskan et al., 2017). Caliskan et al. (2017) show that artifacts of the societal status quo, such as occupational gender statistics, are imprinted in online text and mimicked by machines. We suggest that a similar phenomenon is occurring for online images. One possible culprit is confirmation bias (Schweiger et al., 2014), the tendency of individuals to consume and produce content conforming to group norms. Self-supervised models exhibit the same tendency (Arazo et al., 2020).

In addition to confirming human and natural language machine biases in the image domain, the iEAT measures visual biases that may implicitly affect humans and machines but cannot be captured in text corpora. Foroni and Bel-Bahar (2010) conjecture that in humans, picture-IATs and word-IATs measure different mental processes. More research is needed to explore biases embedded in images and investigate their origins, as Brunet et al. (2019) suggest for language models. Tenney et al. (2019) show that contextual representations learn syntactic and semantic features from the context. Voita et al. (2019) explain the change of vector representations among layers based on the compression/prediction trade-off perspective. Advances in this direction would contribute to our understanding of the causal factors behind visual perception and biases related to cognition and language acquisition.

Our methods come with some limitations. The biases we measure are in large part due to patterns learned from the pre-training data, but ImageNet 2012 does not necessarily represent the entire population of images currently produced and circulated on the Internet. Additionally, ImageNet 2012 is intended for object detection, not distinguishing people's social attributes, and both our models were validated for non-person object classification.[8] The largest version of iGPT (not publicly available) was pre-trained on 100 million additional web images (Chen, Radford, et al., 2020). Given the financial and carbon costs of the computation required to train highly parameterized models like iGPT, we did not train our own models on larger-scale corpora. Complementary iEAT bias testing with unsupervised models pre-trained on an updated version of ImageNet could help quantify the effectiveness of dataset de-biasing strategies.

A model like iGPT, pre-trained on a more comprehensive private dataset from a platform like Instagram or Facebook, could encode much more information about contemporary social biases. Clearview AI reportedly scraped over 3 billion images from Facebook, YouTube, and millions of other sites for their face recognition model (Hill, 2020). Dosovitskiy et al. (2021) recently trained a very similar transformer model on Google's JFT-300M, a 300 million image dataset scraped from the web (Sun et al., 2017). Further research is needed to determine how architecture choices affect embedded biases and how dataset filtering and balancing techniques might help (Wang, Qinami, et al., 2020; Wang, Zhao, et al., 2019). Previous metric-based and adversarial approaches generally require labeled datasets (Wang, Zhao, et al., 2019; Wang, Narayanan, & Russakovsky, 2020; Wang, Qinami, et al., 2020). Our method avoids the limitations of laborious manual labeling.

Though models like these may be useful for quantifying contemporary social biases as they

[8] Recently, Yang et al. (2020) proposed updates to improve fairness and representation in the ImageNet "person" category that could change our results.

are portrayed in vast quantities of images on the Internet, our results suggest the use of unsupervised pre-training on images at scale is likely to propagate harmful biases. Given the high computational and carbon cost of model training at scale, transfer learning with pre-trained models is an attractive option for practitioners. But our results indicate that patterns of stereotypical portrayal of social groups do affect unsupervised models, so careful research and analysis are needed before these models make consequential decisions about individuals and society. Our method can be used to assess task-agnostic biases contained in a dataset to enhance transparency (Gebru et al., 2018; Mitchell et al., 2019a), but bias mitigation for unsupervised transfer learning is a challenging open problem.

4.9 Conclusions

We develop a principled method for measuring bias in unsupervised image models, adapting embedding association tests used in the language domain. With image embeddings extracted by state-of-the-art unsupervised image models pre-trained on ImageNet, we successfully replicate validated bias tests in the image domain and document several social biases, including severe intersectional bias. Our results suggest that unsupervised image models learn human biases from the way people are portrayed in images on the web. These findings serve as a caution for computer vision practitioners using transfer learning: pre-trained models may embed all types of harmful human biases from the way people are portrayed in training data, and model design choices determine whether and how those biases are propagated into harms downstream.

Table 4.1: iEAT tests for the association between target concepts X vs. Y (represented by n_t images each) and attributes A vs. B (represented by n_a images each) in embeddings generated by an unsupervised model. Effect sizes d represent the magnitude of bias, colored by conventional small (0.2), medium (0.5), and large (0.8). Permutation p-values indicate significance. Reproduced from Nosek, Smyth, et al. (2007), the original human IAT effect sizes are all statistically significant with $p < 10^{-8}$; they can be compared to our effect sizes in sign but not in magnitude.

	X	Y	A	B	n_t	n_a	Model	iEAT d	iEAT p	IAT d
Age†	Young	Old	Pleasant	Unpleasant	6	55	iGPT	0.42	0.24	1.23
							SimCLR	0.59	0.16	1.23
Arab-Muslim	Other	Arab-Muslim	Pleasant	Unpleasant	10	55	iGPT	0.86	0.02	0.33
							SimCLR	1.06	$< 10^{-2}$	0.33
Asian§	European American	Asian American	American	Foreign	6	6	iGPT	0.25	0.34	0.62
							SimCLR	0.47	0.21	0.62
Disability†	Disabled	Abled	Pleasant	Unpleasant	4	55	iGPT	-0.02	0.53	1.05
							SimCLR	0.38	0.34	1.05
Gender-Career	Male	Female	Career	Family	40	21	iGPT	0.62	$< 10^{-2}$	1.1
							SimCLR	0.74	$< 10^{-3}$	1.1
Gender-Science	Male	Female	Science	Liberal Arts	40	21	iGPT	0.44	0.02	0.93
							SimCLR	-0.10	0.67	0.93
Insect-Flower	Flower	Insect	Pleasant	Unpleasant	35	55	iGPT	0.34	0.07	1.35
							SimCLR	1.69	$< 10^{-3}$	1.35
Native§	European American	Native American	U.S.	World	8	5	iGPT	-0.33	0.73	0.46
							SimCLR	-0.19	0.65	0.46
Race†	European American	African American	Pleasant	Unpleasant	6	55	iGPT	-0.62	0.85	0.86
							SimCLR	-0.57	0.83	0.86
Religion	Christianity	Judaism	Pleasant	Unpleasant	7	55	iGPT	0.37	0.25	-0.34
							SimCLR	0.36	0.26	-0.34
Sexuality	Gay	Straight	Pleasant	Unpleasant	9	55	iGPT	-0.03	0.52	0.74
							SimCLR	0.04	0.47	0.74
Skin-Tone†	Light	Dark	Pleasant	Unpleasant	7	55	iGPT	1.26	$< 10^{-2}$	0.73
							SimCLR	-0.19	0.71	0.73
Weapon§	White	Black	Tool	Weapon	6	7	iGPT	0.86	0.07	1.0
							SimCLR	1.38	$< 10^{-2}$	1.0
Weapon (Modern)	White	Black	Tool	Weapon	6	9	iGPT	0.88	0.06	N/A
							SimCLR	1.28	0.01	N/A
Weight†	Thin	Fat	Pleasant	Unpleasant	10	55	iGPT	1.67	$< 10^{-3}$	1.83
							SimCLR	-0.30	0.74	1.83

§ Originally a picture-IAT (image-only stimuli). † Originally a mixed-mode IAT (image and verbal stimuli).

Table 4.2: iEAT tests for the association between intersectional group X vs. Y (represented by n_t images each) and attributes A vs. B (represented by n_a images each) in embeddings produced by an unsupervised model. Effect sizes d represent the magnitude of bias, colored by conventional small (0.2), medium (0.5), and large (0.8). Permutation p-values indicate significance.

	X	Y	A	B	n_t	n_a	d	p
Gender-Career (MF)	Male	Female	Career	Family	40	21	0.81	$< 10^{-3}$
Gender-Career (WMBF)	White Male	Black Female			20	21	0.20	0.27
Gender-Career (WMBM)	Black Male	White Male	Career	Family	20	21	0.89	$< 10^{-2}$
Gender-Career (WMWF)	White Male	White Female			20	21	0.97	$< 10^{-3}$
Gender-Science (MF)	Male	Female	Science	Liberal Arts	40	21	0.00	0.50
Gender-Science (WMBF)	White Male	Black Female			20	21	0.80	$< 10^{-2}$
Gender-Science (WMBM)	White Male	Black Male	Science	Liberal Arts	20	21	0.49	0.06
Gender-Science (WMWF)	White Male	White Female			20	21	-0.37	0.88
Valence (BFBM)	Black Female	Black Male	Pleasant	Unpleasant	20	55	0.17	0.29
Valence (BW)	White	Black			40	55	1.16	$< 10^{-3}$
Valence (FM)	Female	Male	Pleasant	Unpleasant	40	55	0.39	0.04
Valence (WFBF)	White Female	Black Female			20	55	1.51	$< 10^{-3}$
Valence (WFBM)	White Female	Black Male	Pleasant	Unpleasant	20	55	1.46	$< 10^{-3}$
Valence (WMBF)	White Male	Black Female			20	55	0.83	$< 10^{-2}$
Valence (WMBM)	White Male	Black Male	Pleasant	Unpleasant	20	55	0.88	$< 10^{-2}$
Valence (WMWF)	White Female	White Male			20	55	0.79	$< 10^{-2}$

Chapter 5

Gaps and Opportunities in AI Audit Tooling

This chapter is reproduced from:

Ojewale, V., Steed, R., Vecchione, B., Birhane, A., & Raji, I. D. (2025). Towards AI Accountability Infrastructure: Gaps and Opportunities in AI Audit Tooling. *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, 1–29. https://doi.org/10.1145/3706598.3713301

5.1 Introduction

Despite increasing policy enthusiasm,[1] the execution of effective *AI audits* remains practically difficult. Often defined as independent evaluations of the performance, fairness, legality, or safety of deployed AI systems, the maturity of the audit ecosystem in the technology sector lags far behind other industries such as finance and healthcare (Raji, Smart, et al., 2020; Raji, Xu, et al., 2022).[2] AI audits are often inconsistent and unreliable (Ryan-Mosley, 2023), and the lack of access and visibility to many AI systems leaves auditors without the information needed to make adequate and truly independent assessments (Holstein et al., 2019; Terzis et al., 2024).

In the face of these challenges, practitioners often rely on tools—software, frameworks, and other resources—to support their AI audit work. Past research in human-computer interaction (HCI) and social computing has developed and studied a host of fairness, explainability, and other toolkits that inform such evaluations (Bellamy et al., 2018; Smith-Renner et al., 2020; Kaur et al., 2020; Bertrand et al., 2023; Deng et al., 2023; Woodruff et al., 2018; Brown et al., 2019; Madaio et al., 2020; DeVos et al., 2022; Lee & Singh, 2021; Holstein et al., 2019; Deng et al., 2022; Amershi et al., 2015; Wong et al., 2023; Ehsan & Riedl, 2020). Governments across the globe are developing their own tools and making use of existing resources as part of their enforcement regimes for AI governance (Kaye & Dixon, 2023).[3] In the U.S. alone, several recently proposed "AI innovation" bills are explicitly geared towards investing in tooling and resource development for AI auditing.[4] Similarly, in the E.U., enforcement reports for the Digital Services Act emphasize the importance of AI audit tooling for effective oversight enforcement (Klinger & Ohme, 2023).

Despite these developments, research has yet to properly map, taxonomize, and understand the full scope of tooling needed to meaningfully support AI audit practitioners. HCI research has identified many practical challenges facing practitioners (Holstein et al., 2019; Costanza-Chock et al., 2022; Brown et al., 2019; Madaio et al., 2020; DeVos et al., 2022) and valuable recent work critically examines some of the toolkits involved (Lee & Singh, 2021; Deng et al., 2022; Wong et al., 2023; Berman et al., 2024). However, the auditing process involves more

[1] AI audits have been featured in several recent U.S. congressional bills (Algorithmic Accountability Act of 2022, 2019; Lenhart, 2023) and state efforts (Perrigo, 2023; Stop Discrimination by Algorithms Act of 2021, 2021), and the practice is regularly mentioned in AI governance proposals internationally Galindo et al., 2021, from the E.U. Digital Services Act to a municipal hiring bill passed in New York City (A Local Law to Amend the Administrative Code of the City of New York, in Relation to Automated Employment Decision Tools, 2021).

[2] This dearth of generalized AI audit guidance is remedied only partially by recent efforts from government advisory bodies like the U.K.'s Information Commissioner's Office (ICO) Information Commissioner's Office, 2023, the U.S. National Institute of Standards and Technology Tabassi, 2023 and others (Office of Science and Technology Policy, 2022b).

[3] Examples include the U.S. AI Safety Institute's Inspect, the U.S. National Institute of Standards and Technology's ARIA & Dioptra, Singapore's AI Verify, and the U.S. National Science Foundation's Artificial Intelligence Research Resource (NAIRR) Pilot.

[4] Examples include the "CREATE AI Act of 2023", "VET Artificial Intelligence Act", "Promoting United States Leadership in Standards Act of 2024", "TEST AI Act of 2023", "Artificial Intelligence Research, Innovation, and Accountability Act of 2023", "Artificial Intelligence Public Awareness and Education Campaign Act", and "Future of Artificial Intelligence Innovation Act of 2024".

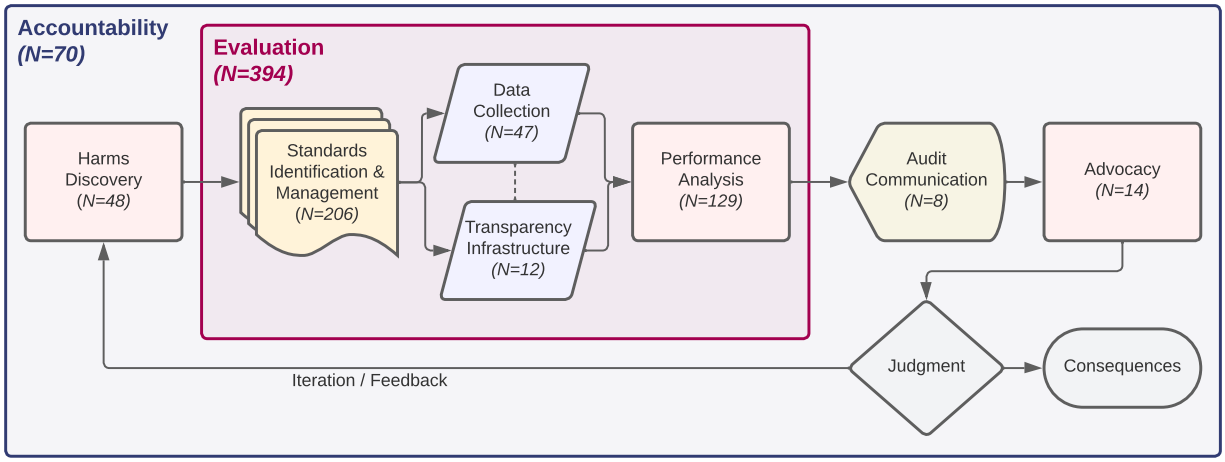


Figure 5.1: Stages of the tool-supported audit process surfaced in our survey of AI audit tooling. We taxonomize tools by the stage of the AI audit process in which they are used. Tools may be used in multiple stages.

than just a performance analysis of the AI product or model. A thorough evaluation alone is not sufficient to hold key stakeholders responsible for system-wide behavior (Goodman & Tréhu, 2022; Raji, Xu, et al., 2022)—auditors also need tools that support key components of the accountability process, including target identification, standardization of practice, communication, and advocacy (Costanza-Chock et al., 2022).

In this study, we compare audit practitioners' tooling needs to the current landscape of AI audit tooling to understand challenges to accountability and potential opportunities for expanding the scope of HCI research and tool development.

1. **RQ1: What tools do practitioners need to conduct AI audits?**
2. **RQ2: What tools currently exist to support AI audit work?**
3. **RQ3: Do existing tools support practitioners' needs?**

To investigate RQ1, we interviewed 35 AI audit practitioners—employed at 24 organizations, including tech companies, startups, government agencies, non-profits, consulting firms, and academic institutions—about the tools they use and how they support the audit process. To investigate RQ2, we conducted a landscape analysis of 435 existing audit tools (Fig. 5.1).[5] In the interviews, we asked practitioners about gaps they encountered and compared their responses to the current tooling landscape (RQ3).

We find that while there currently exist many tools to support audit work, particularly for evaluating AI systems and managing standards, these tools often fell short of helping auditors achieve accountability in practice. Practitioners found some tools—such as open source tools for data collection—empowering, but tools for tasks beyond evaluation, such as discovering harms, communicating audit results, and advocating for subsequent changes, were much less common. The practitioners we interviewed often adapted existing tools or

[5]Note that our sample is not an exhaustive list of all AI audit tools.

built their own from scratch to relieve the tedious or difficult tasks in their particular audit workflows. Auditors envisioned tools to help them access high-quality, uncompromised data, apply consistent and holistic standards and methods, and ensure audit integrity.

Our results identify challenges for every stage of the AI audit process as well as opportunities for HCI researchers, policymakers, and practitioners to pursue an alternative vision for tool development that supports rigor, inclusion, independence, and accountability. We conclude with a summary of research and design opportunities for the HCI community and other stakeholders that could help push the landscape of AI audit tools beyond evaluation and towards infrastructure for meaningful accountability.

5.2 Related work

AI auditing. The practice of AI auditing, or algorithm auditing—a term first formally proposed in 2014 by Sandvig et al. (2014) to describe methods for detecting discrimination in online platforms—has expanded over the last decade (Vecchione et al., 2021). Researchers have since used the term to encompass not just field studies for detecting discrimination with causal estimation (Metaxa et al., 2021) but also any kind of independent assessment of an automated or data-defined system (Digital Regulation Cooperation Forum, 2022; Raji, Smart, et al., 2020). Early audit studies of facial recognition systems and criminal risk assessments, (Angwin et al., 2016; Buolamwini & Gebru, 2018b), for example, resulted in widespread advocacy and, at times, even changes to or recalls of the audited systems (Raji & Buolamwini, 2022; Sherwin & Bhandari, 2019; Spinks, 2020; Sheard, 2021). However, not all audits succeed in holding system builders and operators accountable (Goodman & Tréhu, 2022; Watkins et al., 2021; Birhane et al., 2024).

Following Birhane et al. (2024), we use the term *AI audit* (Def. 9) to refer to "any independent assessment of an identified audit target via an evaluation of articulated expectations with the implicit or explicit objective of accountability". Accountability is used here in the legal-political sense to mean consequential judgment of a systems' behaviors and downstream impacts Bovens, 2007. Consequential judgment distinguishes an audit from a simple evaluation or assessment (Fig. 5.1).

More recently, corporations and policymakers have focused on what we refer to as *internal* AI audits (Def. 10) conducted by teams of employees or contractors separate from the product and engineering teams (European Parliament & Council of the European Union, 2016; Raji, Xu, et al., 2022). These internal audits, afforded access in cooperation with audit targets, may enable accountability if designed properly (Raji, Smart, et al., 2020; Desai & Kroll, 2017), but they may also result in false assurances ("audit washing") (Raji, Xu, et al., 2022; Goodman & Tréhu, 2022) or foreclose on key remedies such as abandonment or disgorgement (Def. 12) (Sloane, Moss, & Chowdhury, 2022; Li, 2022; Johnson et al., 2024). As a result, there is an important role for *external* AI audits (Def. 10): investigations conducted by civil society, journalists, lawyers, regulators, and other third-party actors. These external audits are typically voluntary research studies and investigations into deployed AI systems. However, as our findings reveal, external auditors faced significant hurdles to accountability, including lack of access to the systems they aimed to evaluate. The experiences and challenges of

those doing external audit work has not been the primary focus of HCI research. Our work provides references for understanding their tooling and resource needs Bandy, 2021; Birhane et al., 2024; Costanza-Chock et al., 2022.

While these prior studies define the practice of AI auditing and some of the components of accountability, no study yet examines the full range of *tools* and technical infrastructure used to support AI auditing and accountability.

Infrastructure & accountability. A rich field of literature in science, technology and society (STS) describes infrastructure as not merely physical or technological, but also as holding social and relational power Winner, 1980; Star, 1999. Infrastructure is "something that other things 'run on"' (Lampland & Star, 2009): the invisible and axiomatic basis of tools, standards and frameworks that uphold and shape more complex existing processes, services, and engineered artifacts. Embedded in social systems, technological infrastructures encode standards, norms, and guidelines for social organization (Verbeek, 2006; Lampland & Star, 2009).

Accountability also requires infrastructure. Across various industries, the common accountability practice of auditing has long relied on rituals, organizational processes and tools in order to make consistent, inter-operable and reliable judgments Power, 1999, as well as provide broader access to a larger range of stakeholder participants in the audit process Birhane et al., 2022. Audit tooling, then, represents not only a mechanism for maintenance and consistency of audit integrity but also a key capacity-building intervention to lower the barriers for broader participation.

Past work on AI audit tools. Recent research in human-computer interaction (HCI), social computing, and cooperative design documents the experiences of practitioners evaluating AI and the challenges they face (Costanza-Chock et al., 2022; Holstein et al., 2019; Brown et al., 2019; Madaio et al., 2020; DeVos et al., 2022; Lee & Singh, 2021). For example, Holstein et al. (2019) documents the *practical* and *technical* difficulties faced by internal auditors of ML systems trying to identify and improve fairness, while other studies have examined the *organizational* challenges and barriers these practitioners face (Rakova et al., 2021; Madaio et al., 2020; Widder et al., 2023; Selbst et al., 2019; Costanza-Chock et al., 2022). Several HCI studies have since specifically examined the ways AI audit practitioners make use of various tools to address these issues (Deng et al., 2022; Lee & Singh, 2021; Amershi et al., 2015).

While some studies of AI audit tools focus on performance analysis (Amershi et al., 2015; Harvey et al., 2024) or user-driven grassroots auditing (DeVos et al., 2022; Deng et al., 2023), most HCI studies focus specifically on tools for assessing fairness and interpretability (Bellamy et al., 2018; Smith-Renner et al., 2020; Kaur et al., 2020; Bertrand et al., 2023; Woodruff et al., 2018; Madaio et al., 2020; Lee & Singh, 2021; Deng et al., 2022). Lee and Singh (2021), for example, compare six prominent open source fairness toolkits along various criteria related to practitioners' needs. Deng et al. (2022) documented the ways practitioners learn about and use two prominent fairness toolkits, AI Fairness 360 and Fairlearn (Weerts et al., 2024). And in a survey of 152 audit practitioners, Costanza-Chock et al. (2022) found that 62% of practitioners used existing tools like AI Fairness 360, Scikit Fairness, or Parity, though only

7% of respondents used a standardized framework for their overall audit protocol.

However, less work examines tools for assessing harms beyond fairness, including refusal of medical services (Waldman, 2024), privacy violations, and other types of harm (Shelby et al., 2023). And fairness toolkits do not typically support other necessary steps of an audit such as data collection. Our study aims to expand HCI research by looking beyond fairness evaluation toolkits to examine other kinds of tools involved in AI audits, such as tools for basic performance analysis or data collection.

We also build on more general critiques of "responsible AI" toolkits. Wong et al. (2023), for example, examine documentation from 27 toolkits for "AI Ethics", finding that these resources employ a narrow technical framing that fails to involve more diverse stakeholders or reckon with the non-technical dimensions of AI ethics work; Kaye and Dixon (2023) survey and critiques the tools currently used for AI governance by governments across the globe; and Berman et al. (2024) call for more evaluations of the effectiveness of responsible AI tools.

As yet, however, a comprehensive survey of auditors' practical needs relative to the landscape of available tools is lacking. We extend these analyses and lay out a more complete view on what we refer to as *AI audit tools*: software, interfaces, code, benchmarks, frameworks, and other artifacts used by auditors in the AI audit process (Def. 11).[6]

5.3 Methodology

To better understand the kinds of tools auditors use and where those tools fall short, we conducted 27 semi-structured interviews with 35 auditors across 24 organizations employing internal and external AI auditors. We use the term AI broadly to include tools applied to any AI-advertised product or model, including automated decision systems (ADS), algorithmic recommendation systems, large machine learning base models, generative AI products and more (see Fig. E.5.4). In parallel, to better understand existing tools, we curated a dataset of 435 tools designed or used for AI auditing and developed a taxonomy of the audit tool landscape based on our findings.

5.3.1 Interview methodology

We conducted 27 interviews with a total of 35 audit tool builders and practitioners, representing diverse backgrounds such as engineering, law, journalism, advocacy, policy, and academic research across North America ($N = 22$) and Europe ($N = 5$) (Table 5.1). These practitioners have all participated in internal or external audit work; many have also built tools for AI auditing. We used purposive sampling and snowball sampling methods to recruit participants. We began by contacting practitioners in our professional networks who had conducted notable AI audit work and were active in AI audit communities. Occasionally,

[6]We include in this definition tools that may also be used for other "responsible AI" efforts, such as internal benchmarking, that do not meet our criteria for an audit (Def 9). Auditing is an institutional arrangement—selecting the right tools does not guarantee operational independence, for example.

Table 5.1: Participants' organizations and titles at the time of interview. Some titles are summarized for anonymity. Participants in the same interview are grouped in parentheses.

Employer	Roles of Interviewees	Participants
Large tech for-profit	Director of Policy Research, VP of Research, Data Science Mgr., Research Eng., Researcher	P5, P8, P11, P14, P19, P27
Tech startup	Co-Founder, CEO, Chief Scientist	P4, P12
Government agency	Tech Policy Principal/Mgr./Assoc./Advisor, Research Fellow	(P21, P28-35), P25
University	Assoc./Asst. Professor, Postdoc. Fellow, Data Scientist	P3, P10, P13, P17, P20
Research non-profit	Co-Founder, Director, Research Scientist	P9, P16, P18
Civil society non-profit	Director, Head of Analytics, Statistician, Researcher, Policy Fellow	P1, P2, P6, P22, P24
Non-profit news org.	Opinion Writer, Data Journalist	P7, P15
Law/consulting for-profit	Policy Director, Mgr., Consultant	(P23, P28), P27

participants were referred to us by a colleague or professional contact at another organization. Our sample encompassed both *internal* and *external* auditors employed by for-profit tech companies, AI evaluation startups, research and civil society non-profits, universities, and government agencies.

Interviews followed a semi-structured format and lasted 30–60 minutes. Our questions centered on 1) the specific tools and methods practitioners built or employed and 2) common obstacles and unmet needs. (The full interview protocol is included in Appendix E.4.) Participants had the option to remain anonymous and skip questions at their discretion, though none skipped a question. Our protocol was approved by three university IRBs. To analyze the interview data, we transcribed each interview and annotated the transcripts with manual codes. Our coding approach followed an inductive methodology, allowing patterns and themes to emerge from the data (Wolcott, 1994; Glaser & Strauss, 2017). We employed a combination of descriptive coding, which captured the content of the interviews, and values coding, which captured the attitudes and beliefs expressed by participants. Through collaborative sessions and memo writing, we organized these codes and related quotes into key insights, presented in §5.4.

5.3.2 Tool Taxonomy

Initial search. To taxonomize the landscape of tools available to support AI audit work, we first developed an initial list of tools and tool-building organizations by searching for tools mentioned in a dataset of published audits from academic audit studies, news articles, government reports and frameworks, white papers from civil society organizations, law firm

reports, and case files (Birhane et al., 2024) as well as existing lists of tools such as (Hickock, 2023) (see Appendix E.3.1 for details). Our initial search was conducted in August 2022. We also included specific tools and sources that were mentioned in our interviews with practitioners. After we had collected an initial list of 143 tools, we developed an initial taxonomy by clustering tools into 21 initial categories based on their intended or actual uses in AI audit work. This approach served as a starting point for category development rather than as an exhaustive inventory.

Theoretical sampling. Next, we expanded our initial set of tools with two kinds of additional theoretical sampling. With targeted keyword searches on English Google and GitHub, we searched explicitly for additional tools in areas where we had fewer examples until theoretically fresh examples of tools ceased to arise.[7] Our search queries were descriptors from our initial categories or descriptors used by tools already collected—alone and combined with terms like "audit tool" or "responsible AI" (see Appendix E.3.2). We expanded our list of sources based on our initial taxonomy. (For example, to find tools in our initial "Participatory" category, we searched in the proceedings of participatory AI workshops). We also followed links and references in our initial sample of tools to identify additional, similar tools (snowball sampling). With these methods, we added 181 tools between August and October 2022, and we continued to update the dataset with 102 more tools through September 2024. In September 2024, we ran our search queries again for the top-level categories, adding 9 more tools.

These searches surfaced new examples which we used to expand and re-define our categories. We iteratively revised our taxonomy twice more to accommodate new examples, integrate findings from the interview study, and clarify or expand our initial categories. We did not explicitly ask interview participants about these categories, but we did incorporate the tools they used and their descriptions of audit tool use while developing the taxonomy. Our final taxonomy groups tools into 30 main categories with 27 subcategories (Table 5.2) grounded in the properties of the tools we found and shaped by our interviews with and experiences as AI audit practitioners. We sorted these categories into 7 "stages" of the tool-assisted audit process (Fig. 5.1).

This stage of our search was designed to iteratively refine conceptual categories with the goal of generalizing from empirical descriptions to theoretical insights (Lee & Baskerville, 2003). So while our search for tools was systematic, it was not exhaustive. The sample represented in our dataset does not include every tool that could support AI audits. Likewise, though our sample is designed to represent commonly used and commonly built tools in a theoretically representative set of categories, numeric descriptions of our sample may not reflect the statistical distribution of all AI audit tools used in practice. In particular, our theoretical sampling strategy deliberately over-represents some less common kinds of tooling (such as tools for advocacy). And because our search relied on public materials, we likely over-represent open tooling over proprietary tooling (though our dataset includes both). As a result, while our dataset reflects a structured and iterative search, it should not be viewed as a comprehensive or perfectly representative sample.

[7] In this stage, we aimed for theoretical saturation, in the style of grounded theory (Charmaz, 2014).

Landscape analysis. To analyze the qualities of tools across our taxonomy, we also manually labeled each tool with several tags describing the tool's documentation and function, including license (open source or not), organization type (for-profit, non-profit, government, or academic), and other characteristics. One author created the labels and at least one other author reviewed each label for agreement. We also supplemented our dataset with funding & employment data from Crunchbase (accessed in September 2023), activity data from Github (September 2024), and citation data from Google Scholar API (September 2023). Detailed methods and additional analysis can be found in Appendix E.5, and a full interactive version of our dataset can be viewed at tools.auditing-ai.com.[8]

5.4 Results

While there exist many tools for aiding AI auditing, practitioners found existing tools inadequate in multiple ways. Practitioners struggled to independently access high-quality data about system behavior, apply consistent and holistic standards and methods, ensure audit integrity, involve affected stakeholders, and collaborate across disciplines.

In particular, though we found many tools for evaluating the performance of AI systems, current tooling did not always help practitioners reach their accountability goals. First, while our survey of AI audit tools ($N = 435$), revealed a wide-ranging landscape of resources built by a variety of academic, for-profit, non-profit, and government organizations (Table 5.2), we primarily surfaced tools for *evaluation*, particularly tools for Standards Identification & Management ($N = 206$) and Performance Analysis ($N = 129$). Tools for other stages of the audit process crucial to accountability—Harms Discovery ($N = 48$), Audit Communication ($N = 8$), Advocacy ($N = 14$), and model/data transparency ($N = 12$)—were much less common in our sample.

Second, while we found many freely available and open source tools (77.9% of our dataset), auditors highlighted the messy, context-specific nature of their actual audit tool use: "*Many approaches were not necessarily principled. They were quite ad hoc*" (P20). Even when open source tools existed, auditors often preferred to build their own tooling solutions: "*if we tried to use the existing stuff, it would just complicate that process*" (P25). For some, existing tools were inadequate for the complexity and scale of the systems being evaluated: "*Most often we try to use open source tools, but that's very different than a data pipeline that... curates data on millions of [users] every day*" (P3). In each stage of the audit process, auditors encountered practical challenges and development gaps between their needs and the landscape of available tools.

In the remainder of this section, we detail the challenges practitioners faced in each stage, compare their experiences to the existing landscape of AI audit tools, and discuss the implications of our findings for the practice and study of AI auditing.

[8]All the code for our analysis and resulting plots—as well as instructions for obtaining supplemental data—can be accessed at github.com/ryansteed/oat-analysis.

Table 5.2: High-level description of the tool taxonomy categories. (Visit tools.auditing-ai.com for an interactive visualization).

Stage	Categories *(Subcategories)*	*N*	Purpose	Examples
Harms Discovery	Education / Awareness *(community education, visioning)*, Incident Reporting *(incident databases, intake forms, bug bounties, hotlines)*, Target Identification *(algorithm visibility)*	48	Help auditors identify and prioritize audit targets and harms to investigate.	ACLU Wa.'s Algorithm Equity Toolkit, AI Incident Database, Algorithm Tips
Standards Identification & Mgmt.	Goal Articulation *(principle statements, standards formulation)*, Self-Assessment *(checklists, grading)*, Documentation *(single stage, continuous, licenses)*, Regulatory Awareness *(discovery, monitoring)*, Methods Design, Participatory Standards-Setting	206	Help auditors identify and formulate principles and norms to guide their investigations.	AI-RFX Procurement Framework, Microsoft's AI Fairness Checklist, Model Cards (Mitchell et al., 2019b), Queensland's Community Engagement Toolkit (Queensland Government, 2017), Community Jury
Transparency Infrastructure	Structured/API Access, Secure & Private Sharing *(federated learning)*, Model/Data Exchange	12	Help auditors interact with and analyze proprietary information about the data or model with centralized infrastructure.	NIST's Face Recognition Vendor Test (Ngan et al., 2020), Google AI Test Kitchen (Warkentin & Woodward, 2022), Airbnb's Project Lighthouse (Airbnb, 2020)
Data Collection	Field Data Collection *(scraping, donation, interviews/surveys, compelled disclosure)*, Bot Deployment, Simulation	47	Help auditors collect information about a model's interactions with its subjects.	Mozilla's YouTube Regrets (Mozilla Foundation, 2021), Tracking Exposed, Selenium, Meta's Web-Enabled Simulation (Ahlgren et al., 2020)
Performance Analysis	Accuracy Evaluation *(A/B testing, benchmarks, adversarial testing, monitoring)*, Explainability *(models, training data)*, Fairness, Qualitative Analysis	129	Help auditors evaluate and explain model behavior through the calculation of performance metrics.	Weights & Biases, Meta's DynaBench, Foolbox, Fairlearn, IBM's AI Fairness 360, Hugging Face's ROOTS (Piktus et al., 2023), Google PAIR's Language Interpretability Tool
Audit Communication	Dataset Visualization, Audit Reporting	8	Help auditors communicate the results of an audit to a broader audience.	Google PAIR's FACETS
Advocacy	Organizing/Resistance, Community Spaces, Legal Search	14	Help organize community action and other accountability measures in response to discovered harms.	Gigbox, Para, Adnauseam, Benefits Tech Advocacy Hub

5.4.1 Harms Discovery

Auditing an AI system first requires identifying the system that should be subject to scrutiny and identifying its potential harms. This task can be especially difficult for *external* auditors who may not know where AI systems are in use or what their impacts might be. Tools for Harms Discovery ($N = 48$) help identify and select targets for audits and support the identification, characterization, and prioritization of potential harms to investigate. This category includes tools for Education & Awareness (to engage affected stakeholders in articulating harms), Incident Reporting (to gather reports of algorithmic harms from users and the public, e.g., through bug bounties or incident databases (Charlie Pownall, 2021)), and Target Identification (e.g., the Algorithm Tips database contains a list of deployed systems in the U.S.). Compared to other stages of our taxonomy (Fig. 5.2), nonprofits

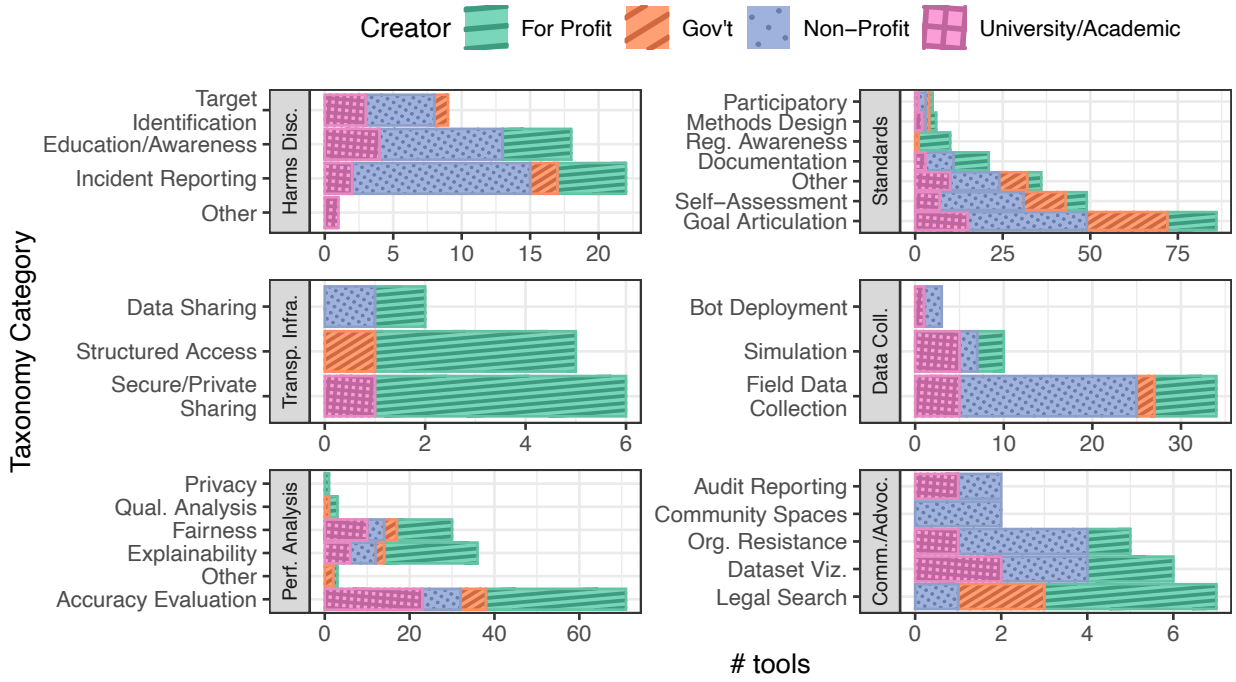


Figure 5.2: Number of tools in each category within each stage of our taxonomy, grouped by type of organization. Tools may be used in multiple stages. Note that the scales differ—the Standards and Performance Analysis stages contain many more tools than the others. Nonprofit and university/academic developers account for relatively more Harms Discovery and Data Collection tools. For-profit developers contribute relatively more Performance Analysis and Transparency Infrastructure tools.

contributed significantly to creating and maintaining these types of tools in our dataset (79.2% not-for-profit; see Fig. E.5.6).

Facilitating more participatory audits. Auditors recognized that to comprehensively identify AI-related harms, they must engage with those directly impacted. Participation in harms discovery had two main benefits for auditors. First, participation helped auditors anticipate a broader range of possible harms: "*Different types of biases are going to manifest, and accordingly it requires ... diverse groups from society, to understand their experiences and expectations in these settings and how they can be impacted*" (P20). Second, participation helped make audits more "context-dependent" and inclusive by providing thorough understanding of how an AI system interacts with impacted groups. Participation also helps instill confidence in the audit process and foster engagement with subsequent accountability efforts. Our tool survey surfaced some tools for participatory incident reporting. This includes the bug bounty platform HackerOne, which Twitter used to crowd audit its image cropping algorithm (Twitter, 2021), and the AI Incident Database, a collection of reports of AI harms. However, we found fewer tools designed specifically for identifying and collaborating with affected users, such as the American Civil Liberties Union (ACLU) of Washington's Algorithmic Equity Toolkit (Barghouti et al., 2020).

Avoiding participation washing. Participants also recognized the challenge of designing methodologies and tools without exploitation, tokenization, or other forms of "participation washing" (Sloane, Moss, Awomolo, & Forlano, 2022), highlighting the need for fair compensation and participatory algorithmic development, in addition to participatory auditing. One participant emphasized the importance of "*an iterative process of doing longer-term, ongoing auditing or observation of algorithmic behavior, and then using that to feed into tweaks, or changes, or even big shifts in where and how [audit] outcomes are used*" (P17).

Implications. Limited access to information about AI systems poses a fundamental barrier to conducting comprehensive audits. Without a clear understanding of which AI systems require auditing, there is a risk of overlooking critical systems that have significant societal implications. While we did find multiple popular databases for recording and collating incidents of harm (such as the AI Incident Database), these databases record harms after they have already occurred, often rely on second-hand reports, and may not delve deeply into causes of harms or impacts (Turri & Dzombak, 2023).[9]

Several participants proposed mandating that corporations disclose AI system use, including information about model versions, anticipated use cases, expected number of users, and past audit results. Researchers and policymakers may also explore mechanisms for centralized, proactive documentation and mandatory, standardized incident reporting for both private firms and government agencies (Turri & Dzombak, 2023). This ensures that current federal AI transparency requirements are actually implemented (Lawrence et al., 2023). Additionally, leveraging mechanisms such as Freedom of Information Act (FOIA) requests could facilitate access to information held by public institutions or government agencies regarding the use of AI systems. Future work could also develop and study systems for fair, inclusive community participation in auditing, the path most often suggested by practitioners for identifying systems and their harms.

5.4.2 Standards Identification & Management

Auditors also used tools to formulate principles and norms to guide their investigations. While HCI research has not traditionally included frameworks and guidelines as tools for AI, Standards Identification & Management was a key focus for participants and comprised the largest collection of tools in our dataset. Standards Identification & Management ($N = 206$), includes tools for Goal Articulation (e.g., broad principles statements), Self-Assessment (more specific procedural assessment tools, such as Microsoft's AI Fairness Checklist (Madaio et al., 2020)), Documentation (e.g., Model Cards (Mitchell et al., 2019b)), Regulatory Awareness (tools, often paid services, for discovering and monitoring relevant regulations), Methods Design (standards for audit methodology), and Participatory Standard-Setting (methods for developing standards in collaboration with affected groups, such as Microsoft's Community Jury (Cass et al., 2022)).

This category includes both internal organization standards and principles and formal national or international standards. Tools such as NIST's Risk Management Framework (RMF) and

[9]Note: The creators of the AI and Algorithmic Incident and Controversies Repository dispute the characterization of their tool in Turri and Dzombak (2023).

relevant ISO standards are referenced extensively by auditing organizations (AI, 2024), and regulators in Singapore and other nations have invested in similar tools as part of compliance and oversight frameworks (Infocomm Media Development Authority & AI Verify Foundation, 2023; Kaye & Dixon, 2023). While the weight and enforceability of these standards differ, they are united by a shared goal of defining audit methodologies and expectations for system performance.

Need for more context-specific standard. Despite the large number of standards and evaluation frameworks surfaced in our tool survey, auditors still felt that evaluation frameworks needed refinement. Auditors emphasized the importance of standardized evaluation frameworks that provide clarity and consistency: "*I think standardization is a big [concern]...*" (P23) Many wished to streamline the auditing process by offering predefined structures and templates for assessment, which are essential for conducting audits effectively and facilitating communication. Most commonly found were goal-articulating "principle statements" ($N = 86$), self-assessment checklists ($N = 49$), and similar documents, while methods for participatory standard-setting ($N = 5$) were comparatively rare in our dataset. The Standards Identification tools we found were particularly general in their applications, compared to other categories of tools—the principle statements, checklists, and similar resources we found were usually developed without a specific kind of target system in mind (Fig. E.5.4). Some participants found these tools too broad to easily apply:

> *I think what I've seen is that companies and institutions... really, really struggle to understand, 'What should we even do when it comes to auditing or evaluating the use of machine learning in our organizations?' And while a template or a checklist is not the right answer, a lot of them don't even know where to start... And so, (it helps) when you have a tool that...has some built in frameworks.* (P8)

Need for more standards beyond fairness. Some participants also thought that assessments templates and checklists were focused too narrowly on fairness assessment. One civil society auditor said, "*I would love some guidance on audits generally and tools that describe the non-fairness components of an audit...*" (P6) Another wished for resources that covered criteria such as explainability, privacy, and transparency: "*Currently we have to.. find open source tools and put them together ourselves, and you need to have expertise to know what to look for*" (P21).

Need for clear and consistent regulatory guidance. Despite some participants' desire for official frameworks like NIST's Risk Management Framework (RMF) (Tabassi, 2023), multiple participants commented on the difficulty of harmonizing current or expected regulatory guidance with practical implementation. Regulatory guidance itself may function as a tool, providing a framework for compliance with emerging policies. One civil society auditor asked:

> *What are we evaluating for? And the question of, even when we have some kind of legal or other benchmark in mind, what are the metrics, and what are the benchmarks and other ways in which to evaluate, technical and otherwise, which also remain quite unclear? We're seeing the ready adoption of audit language into policies, so that just kind of makes us nervous... What are we auditing for?* (P2)

Several emphasized the necessity of regulatory entities being more forthcoming in defining best practices. Some sought the guidance of the regulatory bodies in the domains they operate to establish their own frameworks but often struggle to translate industry expectations into meaningful standards for evaluation. Multiple participants felt there was a "*culture or communications gap between the legal and compliance people on one side and the engineers on the other*" (P12). As a result, some hoped for more collaborative approaches instead of command-and-control prescriptions: "*Half the time we reach out [to regulators], and there's just no one to contact... it's just a black hole...* " (P13).

Implications. Standards for evaluating AI systems must be simultaneously holistic, context-specific, inclusive, and compatible with practice. Some of the tools we found made advances in one or more of these dimensions, but few accomplished all three. Microsoft's AI Fairness checklist, for example, was co-designed with practitioners (Madaio et al., 2020) in an effort to be more compatible with practical challenges, but many of the other standards frameworks we found did not obviously consult practitioners and fewer involved affected stakeholders. Likewise, while NIST's AI RMF includes safety, security, reliability, transparency, explainability, and privacy, in addition to fairness, its guidance for specific evaluation techniques remains fairly broad (Tabassi, 2023).

Research could continue to explore how regulatory standards could be translated into concrete metrics and other effective guidance for industry (Wachter et al., 2021; Guha et al., 2023). One participant at a large tech company, for example, preferred a "*must, could, should*" (P14) structure for regulatory guidance: a non-technical legal minimum ("must") accompanied by more precise technical paths to compliance ("should") and a set of ideal best practices for high performers and innovators ("could").

5.4.3 Data Collection & Transparency Infrastructure

Gathering empirical evidence is a key step in AI auditing, but often poses the most significant challenge in practice. When model operators were unwilling or unable to release relevant documentation and other evidence, auditors turned to two main classes of tools to help.

Tools for Transparency Infrastructure ($N = 12$) are interfaces and databases hosted by model operators that allow controlled access to relevant data. This category includes tools for Structured or Application Programming Interface (API) Access (tools that allow auditors to interact with models and live systems, such as Google's AI Test Kitchen (Warkentin & Woodward, 2022)), tools for Data Sharing (platforms or trusts for hosting models and related data, such as the Gig Economy Data Hub ("Gig Economy Data Hub," 2021)), and tools for Secure & Private sharing (tools that help mitigate concerns with sharing data, such as Airbnb's Project Lighthouse (Airbnb, 2020)).

More commonly, though, tools for Data Collection ($N = 47$), helped *external* auditors in particular gather information *outside* auditee-controlled interfaces. These tools help auditors gather data about model behavior, including relevant information not routinely collected by model operators. This category includes tools for Field Data Collection, which collect data from real systems and real users—including tools for Data Donation (such as Mozilla's YouTube Regrets project (Mozilla Foundation, 2021)), Data Scraping (e.g., Tracking

Exposed (Agosti, 2023)), Interviews/Surveys, and Compelled Transparency (e.g. tools such as MuckRock, which facilitates public records requests). We also found tools for Simulation (e.g. Meta's Web Enabled Simulation platform for simulating interactions on Facebook (Ahlgren et al., 2020)) and Bot Deployment (tools used for sock puppet auditing (Bandy, 2021), such as Selenium or Appium), both used to test systems with artificial or semi-artificial interactions.

Need for uncompromised data access. Data collection tools aim to address one of the challenges most frequently mentioned by our participants: the difficulty of accessing data and other vital information required to conduct meaningfully independent audits. Transparency Infrastructure tools provide external auditors with controlled access to models and data—especially for online platforms, in our dataset (Fig. E.5.4)—but they require investment from model operators. Despite court orders and regulations like Article 40 of the Digital Service Act (Leerssen, 2023) that require the construction of transparency tools, one participant noted that key APIs used for auditing are becoming more costly and undependable:

> *There's a direct impact on shutting off the access to information that affects people doing audits... we saw that with Reddit charging for their API and shutting down... Twitter charging astronomical, now, amounts for their API. It's because everybody is scraping public Reddit and public Twitter to train large [AI] models...* (P24)

They wished for "*access to platform data... a context under which people can do controlled experiments, using data that is provided by platforms directly*" (P24).

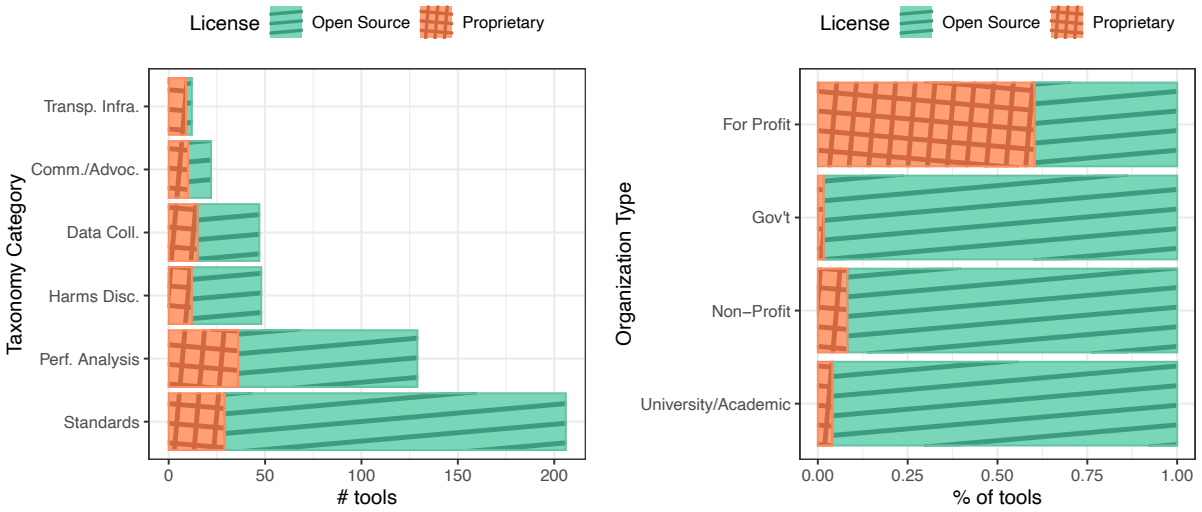
Auditors also said that corporate control over APIs undermined their independence in conducting audits which aligns with our definition of an audit (Def. 9). A civil society auditor said,

> *I wish I had something to force people to give me their data... Part of the problem of auditing in general is that the only people who get let in are usually the people who are willing to say nice things about whatever the technology is being audited.* (P6)

In practice, participants reported that key details, such as data sampling methods, data provenance, model versioning, metrics, and design justifications, were often omitted or only partially disclosed. And currently, the vetting process for API use often requires the auditor to disclose their intent for the evaluation in advance, which may compromise the integrity of the study.

> *I'm very much concerned that what's going to happen is.. [platforms have] given all this access, and there will actually be... more of a cover up than there is now... They have so much power in this conversation to just share whatever information they want.* (P7)

Instead, one participant's ideal was an "*inspectability API*" (P7), a required, standardized interface to allow researchers to interact with online platforms, including the ability to test different profiles, geographies, and other variables needed to evaluate disparate treatment, misinformation, and other algorithmic harms. Similarly, an auditor at a startup wished for centralized data archive available to the public:



(a) Number of open source tools in each taxonomy stage. Tools may be used in multiple stages.

(b) Percentage of open source tools by organization type.

Figure 5.3: Tool licensing by taxonomy stage (top) and by organization type (bottom).

> *If I really had to paint my perfect vision, it would be an independent database archive, or whatever, of all the relevant data. And then... different parts of society can tap into it... I think what you want is a lot of different innovative organizations and people and builders taking this data and building useful things with it, rather than a single one.* (P12)

Challenges with independent data collection. Rather than rely on the model operator to provide access, auditors—especially *external* auditors—often turned to tools for Data Collection to obtain evidence themselves, sometimes developing and sharing their own tools and processes. Unlike the tools we found for Transparency Infrastructure, which were mostly not open source (25.0% open source in our dataset; see Fig. 5.3), tools for Data Collection were much more likely to be available under an open source license (68.1% open source in our dataset). Tools for Data Collection—most not built specifically for AI auditing—also comprised the most popular Github repositories in our dataset (Table E.2). For example, auditors used Selenium ("Selenium," 2023), a popular collection of open source tools for browser automation, to simulate user profiles while scraping data (known as a "sock puppet" audit (Bandy, 2021)).

While this approach gave auditors more freedom, it could also take more effort. Some of the tools we found—such as the Markup's Citizen Browser (The Markup, 2022), a data donation platform—were built from scratch to collect specific kinds of data for auditing. Existing tools for data scraping were helpful but often required extensive adaptation:

> *We almost always have to build custom scrapers to collect data... There's some templates right out there for these scrapers, and then you usually have to customize them. And then there's a huge amount of work to keep them alive... They break all the time with all these edge cases. And so they're really a pain. I don't really*

> *know that there's a way to solve that.* (P7)

Despite these difficulties, external auditors—particularly the journalists in our sample—still saw advantages in independent data collection. The same journalist noted:

> *There's a lot of requests for inside access, as if that's the only way to do this type of thing, but in reality inside access can actually be a trap... your view inside the room is actually really limited. And so I have come to believe... that actually doing analysis from the outside can often be way more revealing... I usually never have insight into the algorithm itself, but I can do analysis on the outputs. And to me, that's the right place for a journalist to operate, because the outputs are the real-life impact.* (P7)

Challenges with understanding and processing data. Even when they had access to the data they wanted, participants noted that basic challenges involved with managing and analyzing data required more labor than any other task. A government auditor said, "*90% of the work is figuring out what different tables are, and what different columns are, and working out whether it makes sense to join certain things. And then figuring out some meaningful metrics that we can draw from that data...*" (P25). Participants also spent lots of time reviewing and requesting additional documents from audit targets.

Multiple participants wished for tools to help with tasks such as data collection and cleaning: "*most of the value of data infrastructure is literally cleaning data*" (P3). One auditor hoped for innovation in data quality management: "*There are custom scrapers, a lot of human data quality work, and one thing that I have really wanted to do and never been able to do is try to figure out ways to get that data quality work done in more interesting ways*" (P7).

Data quality concerns intersected with concerns about audit integrity. "*I think we often have to worry about... [whether] what we see is what we think it is*" (P15). With recent datasets scaling up to staggering sizes, this concern has become more acute and auditors commented on how manual analysis was no longer feasible: "*Given how much data we are able to process, we need new methods to analyze the data curation process and what kind of problems data comes with. And then we need tools to detect what is synthetic, what is real*" (P20).

Risk of retaliation. Some external auditors also worried that external data collection tools—particularly tools for data scraping or data donation—may violate terms of service set by platforms and result in legal liability or retaliation. Auditors expressed concerns about legal risks from existing laws and regulations, particularly the Computer Fraud and Abuse Act (CFAA) and other privacy laws. Under the CFAA, for example, an auditor who violates a platforms' terms of service may be held criminally liable.

It is difficult for auditors to determine what methods—such as public data scraping or even data donation—may be deemed "unauthorized" and criminal under the CFAA unless the audited organization grants authorization. In *Sandvig v. Barr* (2019), for example, the ACLU sued to allow researchers to set up false accounts (sock puppets) to audit computer algorithms (*Sandvig v. Bar*, 2020; American Civil Liberties Union, 2019). In 2021, Facebook used exactly this term of art—"unauthorized"—several times in its justification for disabling the accounts of a group of researchers auditing its advertising algorithms with a data donation

tool called Ad Observer (Clark, 2021; Edelson & McCoy, 2021). Facebook initially insinuated that disabling the researchers' accounts was required by an Federal Trade Commission (FTC) consent decree, backtracking only after the FTC called the statement inaccurate (Levine, 2021). One civil society auditor said,

> *Even if the legal risks have not been acted upon as much, there's cases everyone points to in terms of attacks against researchers. It's a matter of time before it ramps up. As soon as our work becomes threatening enough, that's when it all really starts.* (P24)

As a result, auditors engaged in external data collection had to take great steps to guard against liability and retaliation. One journalist said, "*[The CFAA] is an incredible legal hangover for the type of work I do... how much lawyering I need to even get one tool off the ground is insane*" (P7). Auditors also expressed hesitation about reforms that give platforms more control over what data is released. The same journalist continued,

> *Exemption from [the CFAA] would honestly be more helpful than these platform access roles that the E.U. is claiming that they're going to offer [e.g., in the Digital Services Act], which I am very skeptical about... I just feel like the history of these things is that when platforms have been required to provide API access, they have somehow always made it impossible to do real accountability.* (P7)

Even auditors hired internally may assume some degree of personal risk. One civil society auditor noted, for example:

> *Another really frequent sort of question that I get [from organizations]... is what is my liability around doing this kind of [audit work]? And frankly, to your earlier question about building in-house versus contracting, that's another main [reason to contract]... it's like, okay, we're still going to do this thing, but just sort of outsource it. So I think that just remains like a really open question that people doing this kind of work are carrying a lot of legal risk in doing so.* (P24)

Implications. Future research could explore tools and processes for not only facilitating access to data—especially independently, through scraping or simulation—but also for ensuring data quality and integrity. Participants specifically wished for more tooling for data donation and user-driven auditing, a nascent area of research in human-computer interaction (Lam et al., 2023; DeVos et al., 2022; Deng et al., 2023). Auditors also faced challenges common to data work in general, and research on practices surrounding data quality and data integrity may be applied specifically to discrimination testing and AI auditing methods.

Other challenges were more particular to auditing work. Future work could explore how auditors request information and interact with model operators. Transparency Infrastructure in particular is a nascent area of tooling that may become more common in auditing practice as AI regulation develops and as barriers to external data collection mount. Independent research may help guide these tools into more trustworthy mechanisms for disclosure, even as policymakers can ensure platform-controlled tools are not the only avenue for scrutiny.

5.4.4 Performance Analysis

Tools for Performance Analysis ($N = 129$) are designed to help auditors evaluate and explain model behavior, usually through the calculation of quantiative metrics related to accuracy/safety ($N = 71$), explainability ($N = 36$), or fairness ($N = 30$). This category includes tools for Fairness Evaluation, Accuracy Evaluation (including tools for A/B testing, benchmarking, and model monitoring, such as Meta's Dynabench or the Linux Foundation's Adversarial Robustness Toolbox), Explainability (tools for explaining the behavior of a model, such as IBM's AI Explainability 360, or for exploring training data, such as Hugging Face's ROOTS search tool (Piktus et al., 2023)), and Qualitative Analysis.

Concerns about methodological integrity. Despite the many tools developed for Performance Analysis—including the most popular AI-specific Github repositories in our dataset (e.g., OpenAI Evals (OpenAI, 2023); see Table E.2)—practitioners expressed a need for more robust, well-vetted tools and methodologies. Internal auditors in particular had concerns about the validity, reproducibility, transparency, and trustworthiness of the methods used in popular Performance Analysis tools. One auditor at a tech startup said, "*I'm still not convinced of the validity, even, of some of those methods*" (P4) used in tools for monitoring and validation.

For example, the most popular Performance Analysis tool we found on Github is SHAP (SHapley Additive exPlanations), a game-theoretic method for measuring feature importance in a model (Lundberg & Lee, 2017) (see Table E.2). But as Kumar et al. (2020) argue, Shapley values are prone to misuse and may be unsuitable for normative evaluation. Interpretability and explainability methods promoted by popular tools vary widely in their goals (Lipton, 2018) and, like many "snake oil" AI products (Kaltheuner, 2021; Narayanan & Kapoor, 2024; Stark & Hutson, 2021), may encourage false confidence in their users (Ghassemi et al., 2021). Yet explainability tools such as SHAP are often suggested in official regulatory guidance (Kaye & Dixon, 2023).

Some participants put methodological deficiencies down to differences in the rigor employed by the various disciplines involved with auditing. An auditor in civil society said, "*The bar of the kind of threshold of... validity of findings and novelty of findings is much higher in academia than it is for civil society*" (P24). Participants had concerns about maintenance, effectiveness of automated monitoring processes, and the efficacy of synthetic data for representing real users instead of functioning as "*an academic exercise*" (P4): "*You can perturb all the different inputs you want. But they might not be realistic combinations of features for people who are actually using the system*" (P25).

Need for inspectable, reproducible methods. To allay methodological concerns, several participants emphasized the importance of open-sourcing tools for others in the community to inspect. Some of our industry participants had reproducibility in mind when designing evaluation procedures: "*You want to iterate... but you know that also makes the results less reproducible. And are you being deceptive then, if you [refer in published evaluations] to a model that's different from the one that people analyze?*" (P5) However, the Performance Analysis tools we found (48.1% of which were built by for-profit organizations) were disproportionately *not* open source compared to other tools in our dataset, especially tools for Explainability

(Fig. E.5.3).

Need for more analysis tools beyond fairness & explainability. Similar to Standards Identification & Management, our participants wished for tools to help evaluate a broader spectrum of criteria for AI systems. The Performance Analysis tools we found—the most popular of which were built and maintained by disproportionately large, for-profit firms (Figures E.5.6)—often focused on a narrow set of technical fairness definitions and explainability methods popularized in academic literature. The tools we surveyed in this category often overlooked entire other areas of concern, including the basic functionality of the model (Raji, Kumar, et al., 2022) as well as other methods of evaluation. For example, tools specifically devoted to qualitative—as opposed to quantitative—analysis were much harder to find ($N = 3$) and rarely mentioned by our participants.

Implications. While there are multiple studies on the use of tools for fairness evaluation (Holstein et al., 2019; Lee & Singh, 2021; Deng et al., 2022) and explainability (Bertrand et al., 2023; Wang, Yang, et al., 2019; Liao et al., 2020; Kaur et al., 2020; Smith-Renner et al., 2020; Kim et al., 2023), fewer studies examine how practitioners evaluate other criteria such as basic functionality (Raji, Kumar, et al., 2022), safety, privacy, or recourse, just as fewer tools exist for this purpose. Future work could develop and investigate tools for a broader range of evaluation criteria. Future work could also explore practitioners' standards for audit tooling (Kaye & Dixon, 2023), and policymakers may develop standards that require academic peer review or vetting by regulatory bodies for audit tooling.

Moreover, research must examine further how tools may contribute to "audit washing," the use of auditing procedures to legitimize unethical practices (Goodman & Tréhu, 2022). A tool for accuracy evaluation, for example, may be used to analyze the accuracy of dubious technology for predicting "criminality" or "trustworthiness" without questioning underlying ethical issues with these applications (Stark & Hutson, 2021; Wang et al., 2023). In general, tools may claim to provide auditing capabilities—using terms such as fairness, safety, or explainability—while failing to conduct evaluations that meaningfully contend with power dynamics and institutional barriers to accountability (Wong et al., 2023).

5.4.5 Audit Communication & Advocacy

Some of the most crucial accountability work of an AI audit comes after empirical evaluation is complete. We found two emerging sets of tools that begin to address this important stage of AI auditing: tools for Audit Communication, to effectively translate audit results to a broader audience, and tools for Advocacy, for reporting and campaigning for consequential outcomes in response to audit results. Tools to facilitate Audit Communication were the rarest in our dataset ($N = 8$), and consist mostly of tools for Dataset Visualization (e.g. Google's FACETS ("Facets - Know Your Data," 2023)) and Audit Reporting (e.g., the ACLU's repository of NYC Local Law 144 hiring bias audit reports (Gerchick & Madubuonwu, 2023, August 9/2024)). We found more tools for Advocacy ($N = 14$)—including Community Spaces (e.g., the Benefits Tech Advocacy Hub ("Benefits Tech Advocacy Hub," 2023), which facilitates collaboration between advocates who oppose algorithm-based cuts to public benefits), tools for Organizing/Resistance (e.g., the Algorithmic Ecology framework (Stop

LAPD Spying Coalition & Free Radicals, 2020), a tool for mapping the non-technical dimensions of algorithmic impact), and tools for Legal Search (e.g., the generic case database Westlaw often used to identify relevant precedent for legal redress)—but still fewer than we found in other stages of our taxonomy.

These types of tools were also mentioned less often in our interviews, compared to preceding stages of our taxonomy. Still, auditors wanted their evaluation work to inform consequential judgments, in line with our definition of an audit (Def. 9). As one put it, "*in the business of designing audits, it should be as important to design the consequences and penalties that accompany these audits*" (P2). For the auditors we spoke to, tooling in this stage of auditing was mostly aspirational.

Tools and resources for community building. Auditors especially wished for resources that would bring together the diverse, interdisciplinary groups involved in auditing, similar to the few tools for Community Spaces we found in our tool survey. As one auditor put it, "*We have to have people in the accountability business*" (P7). Auditors hoped greater communication could help unite the profession around policy developments, shared language, standards, and goals that could improve the impact of their work. One auditor at a tech startup said, "*[NIST] has AI guidelines that come out, and we work with them... we send in comments, we give talks, all that kind of stuff. I think that's an important part of the auditing community*" (P4). Another described a workshop attended by civil society, academics, and consulting firms to help prepare for legislation in the European Union (P23). One auditor hoped that communication could lead to shared tooling: "*How do we bring these interdisciplinary communities together so that we can use tools together?*" (P20)

Implications. Communicating audit findings, lessons, and insights learned can help build trust and validate audit findings, recommendations, and subsequent interventions. Audit report repositories could expand the forum holding model operators accountable, allowing policymakers, journalists, and other public stakeholders to engage with evaluations more easily. Embracing public evaluation results also helps audit practitioners to learn from each other's experiences. Despite these benefits, tooling to support these stages is rare. While we found some domain-specific spaces where auditors can interact—the Benefits Tech Advocacy Hub ("Benefits Tech Advocacy Hub," 2023), for example—we found few audit reporting tools or tools for facilitating communication between auditors, journalists, and activists. Future design work and research could explore these emergent categories. Academic research could also explore in more detail the specific mechanisms of audit communication that are most likely to result in meaningful change—such as including concrete demands for action (Raji & Buolamwini, 2022)—and imagine new tools to support those mechanisms.

5.5 Discussion

The HCI community has historically contributed key research to the design and development of *AI audit tools* and helped define the concept of AI auditing Sandvig et al., 2014; Lee and Singh, 2021; Holstein et al., 2019; Wong et al., 2023. In this section, we specifically discuss the important takeaways for that community in particular, as well as broader lessons for other stakeholders, including policymakers, audit practitioners, and funders.

5.5.1 Moving beyond evaluation, towards accountability

Costanza-Chock et al. (2022) found that over 65% of surveyed AI audit practitioners felt that "accountability" (defined as a "commitment from auditee to address problems covered by audit within set time") was a top unmet need in their AI auditing work. This echoes a theme repeated several times throughout our interviews—AI auditors care deeply about accountability but struggled to achieve it.

Despite searching deliberately for non-evaluation tools, we found more than *five times* as many tools in the evaluation stages of the AI audit process as we did tools for harms discovery, audit communication, or advocacy. Perhaps unsurprisingly, these are also the stages of the audit process that participants described as most requiring contextual awareness and typically under-studied participatory and community engagement methods. Research and development related to these and other practical challenges could bolster practitioners' accountability efforts. Promising new directions for HCI research and policy include:

- **Studying and developing tools for harms discovery, audit communication, and advocacy**. Our tool survey identifies several neglected categories of tools—particularly tools for Incident Reporting, Education/Awareness, Target Identification, and Audit Communication—that are worthy subjects for future HCI research. For instance, promising recent research explores the existing limitations (Turri & Dzombak, 2023) and educational applications (Feffer et al., 2023) of incident reporting databases, but little work explores complementary technical infrastructure, such as, for example, the AI inventories often used by journalists (such as Algorithm Tips) and previously required for federal agencies (Biden, 2023). Likewise, auditors envisioned audit report databases as accountability tools to facilitate the amalgamation and communication of audit findings to key stakeholders, mirroring interventions such as the U.S. Security and Exchange Commission (SEC) EDGAR database for financial accounting audits (Raji, Xu, et al., 2022). These gaps in development also present meaningful opportunities for further investment and institutionalization by policy-makers and funders.

- **Validating existing tools in practice**. Audits have limited impact if their results are not reliable or meaningfully connected to real world requirements (Raji & Buolamwini, 2022). Unreliable performance or accuracy analysis tools that fail to meaningfully assess the audit target operate as misleading "rubber stamps" for vendors and lead to "audit washing" Goodman and Tréhu, 2022, posing a serious challenge to the legitimacy of audit results. There is a growing opportunity for HCI researchers to explore ways that AI audit practices interact with existing accountability processes such as litigation or regulatory compliance. For instance, several tools surfaced in our survey (e.g., from Holistic.ai, Credo.ai) were explicitly marketed for use for NYC Local Law 144 compliance, but studies of audit practice suggest the produced measures may not be reliable or legally compatible (Xiang & Raji, 2019; Groves et al., 2024; Wright et al., 2024). Researchers might also investigate the validity and effectiveness of government-sponsored tooling (Kaye & Dixon, 2023) and court-mandated transparency infrastructure (e.g., Facebook Ad Library, built after settlements with civil rights groups (Sandberg, 2019)).

- **Developing participatory methods for audit work**.

  Given the broader calls for a participatory turn in AI development Delgado et al., 2023; Kulynych et al., 2020; Birhane et al., 2022, it is no surprise that participation is an increasing focus for AI audit practitioners and audit tool developers as well. Recent HCI work on user auditing (Deng et al., 2023), such as the "WeAudit" tool DeVos et al., 2022, exemplifies this shift and demonstrates the possibility of designing for a more participatory AI audit process Shen et al., 2021, 2022; Deng et al., 2023. Policymakers can also further emphasize participation as a requirement in audit guidance and invest in tools that support participatory methods.

- **Open & reproducible practices for AI audit tools**. Some participants were concerned about the efficacy of many AI audit tools—particularly tools whose methods were not made available for public scrutiny. (Tools we found for Performance Analysis were less likely to be open source; Fig. 5.3). Making AI audit tools publicly available enables both external collaboration and third-party validation. Open tooling practices may also contribute to knowledge-sharing, standards-setting, transparency, accessibility, and trust, but can have complex interactions with power and oversight that are worthy of further study (Widder et al., 2024). Researchers, policymakers, foundations, and other stakeholders developing audit tools should prioritize open practices. Policymakers could also consider requiring that published audit reports include clear explanations of auditors' methods and tools.

- **Independence in audit tool use and protection from retaliation**. Power dynamics between auditors and the audited have a critical impact on accountability Raji, Xu, et al., 2022; Birhane et al., 2024. Participants had audit results blocked from publication (P1) or unduly restricted in scope (P13) due to interventions from audit targets. Audit target retaliation and censorship was raised as a risk to both *internal* auditors (P5), who face the threat of firings, social dismissal or professional demotion Widder et al., 2023; Boag et al., 2022, and *external* auditors (P7), who face the threat of legal action under existing privacy and anti-hacking laws Urman et al., 2024; Raji, Xu, et al., 2022. HCI research has explored software engineers' attempts to act on ethics concerns in the face of similar risks (Widder et al., 2023; Tahaei et al., 2021); further work could explore auditors' experiences specifically. Policymakers and stakeholders can take steps to provide protections for auditors through legal reforms (Longpre et al., 2024) or legal funds such as the Coalition for Independent Technology Research.

5.5.2 Moving beyond *ad hoc* toolkits, towards shared infrastructure

Participants agreed on the need for shared infrastructure that supports the auditing process. As one participant said, tools are "*a superpower for journalists, and something that really is the future of accountability... There really needs to be some sort of public infrastructure [for auditing]*" (P7). But developing high quality audit tools—even when adapting existing open source tools—took resources, and participants noted a lack of long-term investment: "*We need more funding for this space... especially when it comes to infrastructure*" (P20).

Currently, much of the funding for even open source audit tools comes from private, for-profit organizations. In our landscape analysis, we found that even the free tools currently dominating the audit tooling landscape were often built by large, for-profit tech companies (Fig. 5.2)—for example, 9 of the 12 Transparency Infrastructure tools we found were built by for-profit organizations. Audit tools like these can hold great power over the audit process:

> *Our determination of the performance of the algorithm carries a lot of weight within the organization. It's not like somebody else could just throw it and be like "Oh, we'll go ask somebody else"... because we built the infrastructure, so we have that lever... If you can control the data sources or the ways to integrate algorithms into the data sources, that gives you power.* (P3)

The external auditors we interviewed were especially skeptical of the data provided from these tools and aware of their unreliability, citing the shutdowns of transparency infrastructures like Reddit and Twitter's APIs (P24) as well as Facebook's CrowdTangle tool (P12). Open source tools (e.g., for data scraping) can sometimes fill the gap, but do not always cover the scope and complexity of practical audit work.

Our participants' aspirations for tooling envision another path—a path towards lasting public infrastructure that gives auditors additional levers to hold model operators accountable (Marda et al., 2024). Directions for HCI research and development include:

- **Tool catalogs & other shared infrastructure**. Tool selection was a major source of uncertainty for practitioners; as one expert suggested, "*it's not just about creating a multitude of auditing tools but also about fostering decision support frameworks that empower practitioners to make informed choices based on the context they are dealing with*" (P25). In addition to HCI work evaluating the efficacy of audit tools (Lee & Singh, 2021; Deng et al., 2022; Wong et al., 2023; Kaye & Dixon, 2023; Berman et al., 2024), future research could explore frameworks that assist audit practitioners in identifying and choosing between tools at each stage in an audit. Policymakers can invest in these frameworks and publish catalogs of vetted tools for auditors to reference, similar to the OECD's Tools for Trustworthy AI list (OECD, 2021). In general, shared AI inventories, AI incident databases, AI audit report registries, tool catalogs, regulatory guidance, and other centralized repositories or common transparency infrastructure could increase awareness, accessibility, and knowledge sharing, particularly if the audit community and affected stakeholders are empowered to not only utilize this infrastructure but also contribute to its development.

- **Institutionalized tool maintenance & funding**. Currently, the burden of building and maintaining tools to address technical debt falls on the auditors, raising concerns about the sustainability of auditing efforts. One participant suggested putting expiration dates on tools to avoid future inaccuracies (P7). They also hoped to find "*funders who are on board for longer tool maintenance projects*" (P7), but another noted that attempts to raise support by connecting "*performance issues and ML to downstream business KPIs*" (P4) was difficult when talking about "*nebulous things, like fairness and bias*" (P4). Policymakers and foundations should set aside funding and resources

for long-term tooling projects to ensure that audit tools are high quality, long-lasting, and address the wide range of needs identified in this study. Some initiatives, such as the Mozilla Technology Fund ("Mozilla Technology Fund (MTF)," 2024), the U.K. AI Safety Institute's Systemic AI Safety Grants ("Fast Grants," 2024), and the French AI Action Summit Public Interest fund (AI Action Summit, 2024) are already positioned to make these investments. Researchers and practitioners should include long-term maintenance plans with newly developed tools, possibly in collaboration with civil society organizations such as the Linux Foundation, which hosts several of the open source tools we found (IBM's AI Fairness 360 (Bellamy et al., 2018), for example).

5.5.3 Ongoing Impact

This work was completed as part of the Open Source Audit Tooling (OAT) project at the Mozilla Foundation and has already had demonstrable impact in policy engagement and funding. After submitting public comments, these findings were cited several times in the U.S. National Telecommunication and Information Administration (NTIA) "Artificial Intelligence Accountability Policy Report" (Goodman, 2024), the "Summary report on the call for evidence on the Delegated Regulation on data access" for the E.U. Digital Services Act (Leerssen, 2023), and the U.K. AI Safety Institute's "International AI Safety Report" (Bengio et al., 2025). OAT team members presented these findings to regulators at the U.K.'s OfCom and the U.S. Federal Trade Commission. OAT team members also participated in advising the selection of two rounds of Mozilla Technology Fund (MTF) awardees, including 5 teams in an inaugural cohort (2022), and 8 teams in a follow-up cohort (2023) focused on AI audit tooling.

5.6 Conclusion

Ideally, AI audit studies will translate into tangible outcomes of accountability, but this outcome is far from certain. In order for the audit process to truly be feasible and effective, we—researchers, policymakers, and audit practitioners—need to invest in the infrastructure required for accountability. This will require a full effort on multiple fronts, including everything from the design and development of new tools; to new community infrastructure, communication standard-setting; to considering advocacy for certain policy positions. We cannot accept the minimum from AI auditing—we must push the boundaries of this practice until it becomes the meaningful mechanism of accountability it has the potential to be.

Acknowledgments

Thanks to our participants for volunteering their time and insight. Thanks also to seminar participants at Carnegie Mellon University and the Northeast HCI Meeting for their feedback on an earlier version of this work.

Bibliography

Abadi, M., Chu, A., Goodfellow, I., McMahan, H. B., Mironov, I., Talwar, K., & Zhang, L. (2016). Deep Learning with Differential Privacy. *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*, 308–318. https://doi.org/10.1145/2976749.2978318

Abdu, A. A., Chambers, L. M., Mulligan, D. K., & Jacobs, A. Z. (2024). Algorithmic Transparency and Participation through the Handoff Lens: Lessons Learned from the U.S. Census Bureau's Adoption of Differential Privacy. *The 2024 ACM Conference on Fairness, Accountability, and Transparency*, 1150–1162. https://doi.org/10.1145/3630106.3658962

Abowd, J. M., Adams, T., Ashmead, R., Darais, D., Dey, S., Garfinkel, S. L., Goldschlag, N., Kifer, D., Leclerc, P., Lew, E., Moore, S., Rodríguez, R. A., Tadros, R. N., & Vilhuber, L. (2023). *The 2010 Census Confidentiality Protections Failed, Here's How and Why*. arXiv: 2312.11283 [cs, econ, stat]. https://doi.org/10.48550/arXiv.2312.11283

Abowd, J. M., Ashmead, R., Cumings-Menon, R., Garfinkel, S., Heineck, M., Heiss, C., Johns, R., Kifer, D., Leclerc, P., Machanavajjhala, A., Moran, B., Sexton, W., Spence, M., & Zhuravlev, P. (2022). The 2020 Census Disclosure Avoidance System TopDown Algorithm. *Harvard Data Science Review*. https://doi.org/10.1162/99608f92.529e3cb9

Abowd, J. M., & Hawes, M. B. (2023). Confidentiality Protection in the 2020 US Census of Population and Housing. *Annual Review of Statistics and Its Application*, *10*(1), 119–144. https://doi.org/10.1146/annurev-statistics-010422-034226

Abowd, J. M., Kifer, D., Moran, B., Ashmead, R., Leclerc, P., Sexton, W., Garfinkel, S., & Machanavajjhala, A. (2019). *Census TopDown: Differentially Private Data, Incremental Schemas, and Consistency with Public Knowledge*. U.S. Census Bureau. https://systems.cs.columbia.edu/private-systems-class/papers/Abowd2019Census.pdf

Abowd, J. M., & Schmutte, I. M. (2019). An Economic Analysis of Privacy Protection and Statistical Accuracy as Social Choices. *American Economic Review*, *109*(1), 171–202.

Acquisti, A., Brandimarte, L., & Loewenstein, G. (2015). Privacy and human behavior in the age of information. *Science*, *347*(6221), 509–514. https://doi.org/10.1126/science.aaa1465

Acquisti, A., & Steed, R. (2023). Learning to Live with Privacy-Preserving Analytics. *Communications of the ACM*, *66*(7), 24–27. https://doi.org/10.1145/3597173

Adeleye, T., Berghel, S., Desfontaines, D., Hay, M., Johnson, I., Lemoisson, C., Machanavajjhala, A., Magerlein, T., Modena, G., Pujol, D., Simmons-Marengo, D., & Triedman,

H. (2023). *Publishing Wikipedia usage data with strong privacy guarantees*. arXiv: 2308.16298 [cs]. https://doi.org/10.48550/arXiv.2308.16298

Agarwal, A., & Singh, R. (2024). *Causal Inference with Corrupted Data: Measurement Error, Missing Values, Discretization, and Differential Privacy*. arXiv: 2107.02780 [cs, econ, math, stat]. https://doi.org/10.48550/arXiv.2107.02780

Agosti, C. (2023). *Tracking Exposed Manifesto*. Tracking Exposed. https://tracking.exposed/manifesto

Aguilera, R. V., Rupp, D. E., Williams, C. A., & Ganapathi, J. (2007). Putting the S Back in Corporate Social Responsibility: A Multilevel Theory of Social Change in Organizations. *The Academy of Management Review*, *32*(3), 836–863. https://www.jstor.org/stable/20159338

Ahlgren, J., Berezin, M. E., Bojarczuk, K., Dulskyte, E., Dvortsova, I., George, J., Gucevska, N., Harman, M., Lämmel, R., Meijer, E., Sapora, S., & Spahr-Summers, J. (2020). WES: Agent-based User Interaction Simulation on Real Infrastructure. *Proceedings of the IEEE/ACM 42nd International Conference on Software Engineering Workshops*, 276–284. https://doi.org/10.1145/3387940.3392089

AI, B. A. B. L. (2024). AI and algorithm auditor certification. https://courses.babl.ai/p/ai-and-algorithm-auditor-certification?affcode=616760_ujptkhyg

AI Action Summit. (2024). *Public interest AI*. elysee.fr. https://www.elysee.fr/en/sommet-pour-l-action-sur-l-ia/public-interest-ai

Airbnb. (2020). *A new way we're fighting discrimination on Airbnb - Resource Centre*. Airbnb Resource Centre. https://www.airbnb.ca/resources/hosting-homes/a/a-new-way-were-fighting-discrimination-on-airbnb-201

Algorithmic Accountability Act of 2022 (2019). https://www.congress.gov/117/bills/hr6580/BILLS-117hr6580ih.pdf

Ali, S. J., Christin, A., Smart, A., & Katila, R. (2023). Walking the Walk of AI Ethics: Organizational Challenges and the Individualization of Risk among Ethics Entrepreneurs. *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, 217–226. https://doi.org/10.1145/3593013.3593990

Allen, A. (2022). Dismantling the "Black Opticon": Privacy, Race Equity, and Online Data-Protection Reform. *Yale Law Journal Forum*. https://scholarship.law.upenn.edu/faculty_scholarship/2803

American Civil Liberties Union. (2019). *Sandvig v. Barr — Challenge to CFAA Prohibition on Uncovering Racial Discrimination Online*. American Civil Liberties Union. https://www.aclu.org/cases/sandvig-v-barr-challenge-cfaa-prohibition-uncovering-racial-discrimination-online

American Community Survey – Education Tabulation (ACS-ED). (2015–2019). Education and Demographic Estimates, National Center for Education Statistics. https://nces.ed.gov/programs/edge/Demographic/ACS

Amershi, S., Chickering, M., Drucker, S. M., Lee, B., Simard, P., & Suh, J. (2015). ModelTracker: Redesigning Performance Analysis Tools for Machine Learning. *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, 337–346. https://doi.org/10.1145/2702123.2702509

Angwin, J., Larson, J., Mattu, S., & Kirchner, L. (2016). Machine Bias [magazine]. *ProPublica*. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

Arazo, E., Ortego, D., Albert, P., O'Connor, N. E., & McGuinness, K. (2020). Pseudo-labeling and confirmation bias in deep semi-supervised learning. *Proceedings of the International Joint Conference on Neural Networks*, 1–8. https://doi.org/10.1109/IJCNN48605.2020.9207304

Ashford, N. A., Ayers, C., & Stone, R. F. (1985). Using Regulation to Change the Market for Innovation. *Harvard Environmental Law Review*, *9*(2), 419–466. https://dspace.mit.edu/handle/1721.1/1555
Accepted: 2002-08-05T20:06:51Z.

Autor, D. H., Dorn, D., & Hanson, G. H. (2013). The China Syndrome: Local Labor Market Effects of Import Competition in the United States. *American Economic Review*, *103*(6), 2121–2168. https://doi.org/10.1257/aer.103.6.2121

AWS. (2023). *Differential Privacy - AWS Clean Rooms*. Amazon Web Services, Inc. https://aws.amazon.com/clean-rooms/differential-privacy/

Bachman, P., Hjelm, R. D., & Buchwalter, W. (2019). Learning representations by maximizing mutual information across views. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d Alché-Buc, E. Fox, & R. Garnett (Eds.), *Advances in neural information processing systems* (pp. 15535–15545, Vol. 32). Curran Associates, Inc. https://proceedings.neurips.cc/paper/2019/file/ddf354219aac374f1d40b7e760ee5bb7-Paper.pdf

Balebako, R., Marsh, A., Lin, J., Hong, J., & Cranor, L. F. (2014). The Privacy and Security Behaviors of Smartphone App Developers. *NDSS Symposium*.

Bamberger, K. A., & Mulligan, D. K. (2008). Privacy Decisionmaking in Administrative Agencies. *The University of Chicago Law Review*, *75*(1), 75–107. https://www.jstor.org/stable/20141901

Bamberger, K. A., & Mulligan, D. K. (2015). *Privacy on the ground: Driving corporate behavior in the United States and Europe*. The MIT Press.

Bamberger, P. A., & Pratt, M. G. (2010). Moving forward by looking back: Reclaiming unconventional research contexts and samples in organizational scholarship. *Academy of Management Journal*, *53*(4), 665–671. https://doi.org/10.5465/AMJ.2010.52814357

Bandy, J. (2021). Problematic Machine Behavior: A Systematic Literature Review of Algorithm Audits. *Proceedings of the ACM on Human-Computer Interaction*, *5*, 74:1–74:34. https://doi.org/10.1145/3449148

Baptista, J., Wilson, A. D., & Galliers, R. D. (2021). Instantiation: Reconceptualising the role of technology as a carrier of organisational strategising. *Journal of Information Technology*, *36*(2), 109–127. https://doi.org/10.1177/0268396220988550

Barba, L. A. (2018). *Terminologies for Reproducible Research*. arXiv: 1802.03311. https://doi.org/10.48550/arXiv.1802.03311

Barghouti, B., Bintz, C., Dailey, D., Epstein, M., Guetler, V., Herman, B., Jobe, P. O., Katell, M., Krafft, P., Lee, J., Narayan, S., Putz, F., Raz, D., Robick, B., Tam, A., Woldu, A., & Young, M. (2020). Algorithmic Equity Toolkit. https://www.aclu-wa.org/AEKit

Barocas, S., & Nissenbaum, H. (2014). Big Data's End Run around Anonymity and Consent. In H. Nissenbaum, J. Lane, S. Bender, & V. Stodden (Eds.), *Privacy, Big Data, and*

the Public Good: Frameworks for Engagement (pp. 44–75). Cambridge University Press. https://doi.org/10.1017/CBO9781107590205.004

Barrientos, A. F., Williams, A. R., Snoke, J., & Bowen, C. M. (2023). A Feasibility Study of Differentially Private Summary Statistics and Regression Analyses with Evaluations on Administrative and Survey Data. *Journal of the American Statistical Association*, *0*, 1–24. https://doi.org/10.1080/01621459.2023.2270795

Basta, C., Costa-jussà, M. R., & Casas, N. (2019). Evaluating the underlying gender bias in contextualized word embeddings. *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, 33–39. https://doi.org/10.18653/v1/W19-3805

Belanger, F., & Crossler, R. (2011). Privacy in the Digital Age: A Review of Information Privacy Research in Information Systems. *Management Information Systems Quarterly*, *35*(4), 1017–1041. https://aisel.aisnet.org/misq/vol35/iss4/12

Bélanger, F., & James, T. L. (2020). A Theory of Multilevel Information Privacy Management for the Digital Era. *Information Systems Research*, *31*(2), 510–536. https://doi.org/10.1287/isre.2019.0900

Bell, S., & Robinson, S. (2020). *Small Area Income and Poverty Estimates: 2019* (Small Area Estimates No. P30-08). U.S. Census Bureau.

Bell, W., Basel, W., Cruse, C., Dalzell, L., Maples, J., O'Hara, B., & Powers, D. (2007). *Use of ACS Data to Produce SAIPE Model-Based Estimates of Poverty for Counties* (SAIPE ACS Model Evaluation). U.S. Census Bureau.

Bell, W. R., & Schafer, J. L. (2021). *Block-Level Simulation of Non-Sampling Variability in Decennial Census Population Counts* (Memorandum for John M. Abowd). United States Department of Commerce, Economics and Statistics Administration.

Bellamy, R. K. E., Dey, K., Hind, M., Hoffman, S. C., Houde, S., Kannan, K., Lohia, P., Martino, J., Mehta, S., Mojsilovic, A., Nagar, S., Ramamurthy, K. N., Richards, J., Saha, D., Sattigeri, P., Singh, M., Varshney, K. R., & Zhang, Y. (2018). *AI Fairness 360: An Extensible Toolkit for Detecting, Understanding, and Mitigating Unwanted Algorithmic Bias*. arXiv: 1810.01943 [cs]. https://doi.org/10.48550/arXiv.1810.01943

Bellezza, F. S., Greenwald, A. G., & Banaji, M. R. (1986). Words high and low in pleasantness as rated by male and female college students. *Behavior Research Methods, Instruments, & Computers*, *18*(3), 299–303. https://doi.org/10.3758/BF03204403

Benefits Tech Advocacy Hub. (2023). Benefits Tech Advocacy Hub. https://btah.org/

Bengio, Y., Mindermann, S., Privitera, D., Besiroglu, T., Bommasani, R., Casper, S., Choi, Y., Fox, P., Garfinkel, B., Goldfarb, D., Heidari, H., Ho, A., Kapoor, S., Khalatbari, L., Longpre, S., Manning, S., Mavroudis, V., Mazeika, M., Michael, J., … Zeng, Y. (2025). *International AI Safety Report*. arXiv. AI Action Summit. http://arxiv.org/abs/2501.17805

Benjamin, R. (2020). Race After Technology: Abolitionist Tools for the New Jim Code. *Social Forces*, *98*(4), 1–3. https://doi.org/10.1093/sf/soz162

Berente, N., & Yoo, Y. (2012). Institutional Contradictions and Loose Coupling: Postimplementation of NASA's Enterprise Information System. *Information Systems Research*, *23*(2), 376–396. https://doi.org/10.1287/isre.1110.0373

Berger, J. O. (1976). Admissible Minimax Estimation of a Multivariate Normal Mean with Arbitrary Quadratic Loss. *The Annals of Statistics*, *4*(1), 223–226. https://www.jstor.org/stable/2958003

Berke, A., & Calacci, D. (2022). Privacy Limitations of Interest-based Advertising on The Web: A Post-mortem Empirical Analysis of Google's FLoC. *Proceedings of the 2022 ACM SIGSAC Conference on Computer and Communications Security*, 337–349. https://doi.org/10.1145/3548606.3560626

Berman, G., Goyal, N., & Madaio, M. (2024). A Scoping Study of Evaluation Practices for Responsible AI Tools: Steps Towards Effectiveness Evaluations. *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, 1–24. https://doi.org/10.1145/3613904.3642398

Bertrand, A., Viard, T., Belloum, R., Eagan, J. R., & Maxwell, W. (2023). On Selective, Mutable and Dialogic XAI: A Review of What Users Say about Different Types of Interactive Explanations. *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, 1–21. https://doi.org/10.1145/3544548.3581314

Biden, J. R. (2023). Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence. *The White House*. https://www.whitehouse.gov/briefing-room/presidential-actions/2023/10/30/executive-order-on-the-safe-secure-and-trustworthy-development-and-use-of-artificial-intelligence/

Birchall, C. (2014). Radical Transparency? *Cultural Studies ↔ Critical Methodologies*, *14*(1), 77–88. https://doi.org/10.1177/1532708613517442

Bird, S. (2020). *Putting differential privacy into practice to use data responsibly*. Microsoft AI Blog for Business & Tech. https://blogs.microsoft.com/ai-for-business/differential-privacy/

Birhane, A., Steed, R., Ojewale, V., Vecchione, B., & Raji, I. (2024). AI auditing: The broken bus on the road to AI accountability. *2024 IEEE Conference on Secure and Trustworthy Machine Learning (SaTML)*, 612–643. https://doi.org/10.1109/SaTML59370.2024.00037

Birhane, A., Isaac, W., Prabhakaran, V., Diaz, M., Elish, M. C., Gabriel, I., & Mohamed, S. (2022). Power to the People? Opportunities and Challenges for Participatory AI. *Proceedings of the 2nd ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization*, 1–8. https://doi.org/10.1145/3551624.3555290

Blackwell, M., Honaker, J., & King, G. (2017). A Unified Approach to Measurement Error and Missing Data: Overview and Applications. *Sociological Methods & Research*, *46*(3), 303–341. https://doi.org/10.1177/0049124115585360

Blodgett, S. L., Barocas, S., Daumé III, H., & Wallach, H. (2020). *Language (technology) is power: A critical survey of "bias" in NLP*.

Boag, W., Suresh, H., Lepe, B., & D'Ignazio, C. (2022). Tech Worker Organizing for Power and Accountability. *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, 452–463. https://doi.org/10.1145/3531146.3533111

Bohn, D. (2021). Nobody is flying to join Google's FLoC [magazine]. *The Verge*. https://www.theverge.com/2021/4/16/22387492/google-floc-ad-tech-privacy-browsers-brave-vivaldi-edge-mozilla-chrome-safari

Boldosova, V. (2019). Deliberate storytelling in big data analytics adoption. *Information Systems Journal*, *29*(6), 1126–1152. https://doi.org/10.1111/isj.12244

Bolukbasi, T., Chang, K.-W., Zou, J. Y., Saligrama, V., & Kalai, A. T. (2016). Man is to computer programmer as woman is to homemaker? Debiasing word embeddings. In D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, & R. Garnett (Eds.), *Advances in*

neural information processing systems 29 (pp. 4349–4357). Curran Associates, Inc. http://papers.nips.cc/paper/6228-man-is-to-computer-programmer-as-woman-is-to-homemaker-debiasing-word-embeddings.pdf

Bommasani, R., Davis, K., & Cardie, C. (2020). Interpreting pretrained contextualized representations via reductions to static embeddings. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 4758–4781. https://doi.org/10.18653/v1/2020.acl-main.431

Borman, G. D., & D'Agostino, J. V. (1996). Title I and Student Achievement: A Meta-Analysis of Federal Evaluation Results. *Educational Evaluation and Policy Analysis*, *18*(4), 309–326. https://doi.org/10.3102/01623737018004309

Bovens, M. (2007). Analysing and assessing accountability: A conceptual framework 1. *European law journal*, *13*(4), 447–468.

Boxenbaum, E., & Jonsson, S. (2017). Isomorphism, Diffusion and Decoupling: Concept Evolution and Theoretical Challenges. In *The SAGE Handbook of Organizational Institutionalism* (pp. 77–97). SAGE Publications Ltd. https://doi.org/10.4135/9781446280669.n4

boyd, d., & Sarathy, J. (2022). Differential Perspectives: Epistemic Disconnects Surrounding the U.S. Census Bureau's Use of Differential Privacy. *Harvard Data Science Review*. https://doi.org/10.1162/99608f92.66882f0e

Bozanic, Z., Dirsmith, M. W., & Huddart, S. (2012). The social constitution of regulation: The endogenization of insider trading laws. *Accounting, Organizations and Society*, *37*(7), 461–481. https://doi.org/10.1016/j.aos.2012.06.003

Brandtner, C. (2021). Decoupling Under Scrutiny: Consistency of Managerial Talk and Action in the Age of Nonprofit Accountability. *Nonprofit and Voluntary Sector Quarterly*, *50*(5), 1053–1078. https://doi.org/10.1177/0899764021995240

Broderick, T., Giordano, R., & Meager, R. (2023). *An Automatic Finite-Sample Robustness Metric: When Can Dropping a Little Data Make a Big Difference?* arXiv: 2011.14999 [stat]. https://doi.org/10.48550/arXiv.2011.14999

Brodeur, A., Mikola, D., Cook, N., Brailey, T., Briggs, R., de Gendre, A., Dupraz, Y., Fiala, L., Gabani, J., Gauriot, R., Haddad, J., McWay, R., Levin, J., Johannesson, M., Metson, L., Kinge, J. M., Tian, W., Wochner, T., Mishra, S., … McManus, E. (2024). *Mass Reproducibility and Replicability: A New Hope* (Working Paper No. 107). I4R Discussion Paper Series. https://www.econstor.eu/handle/10419/289437

Broman, K., Cetinkaya-Rundel, M., Nussbaum, A., Paciorek, C., Peng, R., Turek, D., & Wickham, H. (2017). *Recommendations to Funding Agencies for Supporting Reproducible Research*. American Statistical Association. https://www.amstat.org/asa/files/pdfs/POL-ReproducibleResearchRecommendations.pdf

Bromley, P., & Powell, W. W. (2012). From Smoke and Mirrors to Walking the Talk: Decoupling in the Contemporary World. *The Academy of Management Annals*, *6*(1), 483–530. https://doi.org/10.1080/19416520.2012.684462

Brown, A., Chouldechova, A., Putnam-Hornstein, E., Tobin, A., & Vaithianathan, R. (2019). Toward Algorithmic Accountability in Public Services: A Qualitative Study of Affected Community Perspectives on Algorithmic Decision-making in Child Welfare Services. *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, 1–12. https://doi.org/10.1145/3290605.3300271

Brummet, Q., Mulrow, E., & Wolter, K. (2022). The Effect of Differentially Private Noise Injection on Sampling Efficiency and Funding Allocations: Evidence From the 1940 Census. *Harvard Data Science Review*. https://doi.org/10.1162/99608f92.a93d96fa

Brunet, M.-E., Alkalay-Houlihan, C., Anderson, A., & Zemel, R. (2019). Understanding the origins of bias in word embeddings. *Proceedings of the 36th International Conference on Machine Learning*, 803–811. http://proceedings.mlr.press/v97/brunet19a.html

Brunsson, N. (2003). *The Organization of Hypocrisy: Talk, Decisions and Actions in Organizations* (2nd edition). Copenhagen Business School Pr.

Bun, M., Drechsler, J., Gaboardi, M., McMillan, A., & Sarathy, J. (2023). *Controlling Privacy Loss in Sampling Schemes: An Analysis of Stratified and Cluster Sampling*. arXiv: 2007.12674 [stat]. https://doi.org/10.48550/arXiv.2007.12674

Bun, M., & Steinke, T. (2016). *Concentrated Differential Privacy: Simplifications, Extensions, and Lower Bounds*. arXiv: 1605.02065 [cs]. https://doi.org/10.48550/arXiv.1605.02065

Buntain, C., Bonneau, R., Nagler, J., & Tucker, J. A. (2021). *Measuring the Ideology of Audiences for Web Links and Domains Using Differentially Private Engagement Data*. https://doi.org/10.2139/ssrn.3765240

Buolamwini, J., & Gebru, T. (2018a). Gender shades: Intersectional accuracy disparities in commercial gender classification. In S. A. Friedler & C. Wilson (Eds.), *Proceedings of the 1st conference on fairness, accountability and transparency* (pp. 77–91, Vol. 81). PMLR. http://proceedings.mlr.press/v81/buolamwini18a.html

Buolamwini, J., & Gebru, T. (2018b). Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. In S. A. Friedler & C. Wilson (Eds.), *Proceedings of the 1st Conference on Fairness, Accountability and Transparency* (pp. 77–91, Vol. 81). PMLR. http://proceedings.mlr.press/v81/buolamwini18a.html

Burk, D. (2016). Perverse Innovation. *William & Mary Law Review*, *58*(1), 1. https://scholarship.law.wm.edu/wmlr/vol58/iss1/2

Burrell, J. (2016). How the machine 'thinks': Understanding opacity in machine learning algorithms. *Big Data & Society*, *3*(1), 2053951715622512. https://doi.org/10.1177/2053951715622512

Butler, T., Gozman, D., & Lyytinen, K. (2023). The regulation of and through information technology: Towards a conceptual ontology for IS research. *Journal of Information Technology*, *38*(2), 86–107. https://doi.org/10.1177/02683962231181147

Caliskan, A., Bryson, J. J., & Narayanan, A. (2017). *Semantics derived automatically from language corpora contain human-like biases* (No. 6334). Science. https://doi.org/10.1126/science.aal4230

Camerer, C. F., Dreber, A., Forsell, E., Ho, T.-H., Huber, J., Johannesson, M., Kirchler, M., Almenberg, J., Altmejd, A., Chan, T., Heikensten, E., Holzmeister, F., Imai, T., Isaksson, S., Nave, G., Pfeiffer, T., Razen, M., & Wu, H. (2016). Evaluating replicability of laboratory experiments in economics. *Science*, *351*(6280), 1433–1436. https://doi.org/10.1126/science.aaf0918

Campbell, D., Brodeur, A., Dreber, A., Johannesson, M., Kopecky, J., Lusher, L., & Tsoy, N. (2024). *The Robustness Reproducibility of the American Economic Review* (Working Paper No. 124). I4R Discussion Paper Series. https://www.econstor.eu/handle/10419/295222

Campbell, J. L. (2007). Why Would Corporations Behave in Socially Responsible Ways? An Institutional Theory of Corporate Social Responsibility. *The Academy of Management Review*, *32*(3), 946–967. https://www.jstor.org/stable/20159343

Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., & Zagoruyko, S. (2020). *End-to-end object detection with transformers*.

Carroll, A. B. (1979). A Three-Dimensional Conceptual Model of Corporate Performance. *Academy of Management Review*, *4*(4), 497–505. https://doi.org/10.5465/AMR.1979.4498296

Carroll, R. J., Ruppert, D., Stefanski, L. A., & Crainiceanu, C. M. (2006). *Measurement Error in Nonlinear Models: A Modern Perspective, Second Edition* (2nd ed.). Chapman and Hall/CRC. https://doi.org/10.1201/9781420010138

Cass, D., Alberts, M., Turetz, K., Ovhal, P., Buck, A., Pratt, T., Kshirsagar, D., & Mabrey, H. (2022). *Community jury - Azure Application Architecture Guide*. Microsoft. https://learn.microsoft.com/en-us/azure/architecture/guide/responsible-innovation/community-jury/

Charlie Pownall. (2021). AI, algorithmic and automation incident and controversy repository (AIAAIC). https://www.aiaaic.org/

Charmaz, K. (2014). *Constructing grounded theory* (2nd edition). Sage.

Chen, G. H. (2020). Deep kernel survival analysis and subject-specific survival time prediction intervals. In F. Doshi-Velez, J. Fackler, K. Jung, D. Kale, R. Ranganath, B. Wallace, & J. Wiens (Eds.), *Proceedings of the 5th machine learning for healthcare conference* (pp. 537–565, Vol. 126). PMLR. http://proceedings.mlr.press/v126/chen20a.html

Chen, M., Radford, A., Child, R., Wu, J., Jun, H., Luan, D., & Sutskever, I. (2020). Generative pretraining from pixels. In H. D. III & A. Singh (Eds.), *Proceedings of the 37th international conference on machine learning* (pp. 1691–1703, Vol. 119). PMLR. http://proceedings.mlr.press/v119/chen20s.html

Chen, T., Kornblith, S., Norouzi, M., & Hinton, G. (2020). A simple framework for contrastive learning of visual representations. In H. D. III & A. Singh (Eds.), *Proceedings of the 37th international conference on machine learning* (pp. 1597–1607, Vol. 119). PMLR. http://proceedings.mlr.press/v119/chen20j.html

Chen, T., Kornblith, S., Swersky, K., Norouzi, M., & Hinton, G. (2020). *Big self-supervised models are strong semi-supervised learners*.

Chen, Y.-D., Brown, S. A., Hu, P. J.-H., King, C.-C., & Chen, H. (2011). Managing Emerging Infectious Diseases with Information Systems: Reconceptualizing Outbreak Management Through the Lens of Loose Coupling. *Information Systems Research*, *22*(3), 447–468. https://doi.org/10.1287/isre.1110.0376

Christ, M., Radway, S., & Bellovin, S. M. (2022). Differential Privacy and Swapping: Examining De-Identification's Impact on Minority Representation and Privacy Preservation in the U.S. Census. *2022 IEEE Symposium on Security and Privacy (SP)*, 1564–1564. https://doi.org/10.1109/SP46214.2022.00135

Clark, D. D., Garfinkel, S., & Claffy, K. C. (2024). *Exploring the Limits of Differential Privacy*. https://papers.ssrn.com/abstract=4911177

Clark, M. (2021). *Research Cannot Be the Justification for Compromising People's Privacy*. Meta. https://about.fb.com/news/2021/08/research-cannot-be-the-justification-for-compromising-peoples-privacy/

Cohen, A., & Nissim, K. (2020). Linear Program Reconstruction in Practice. *Journal of Privacy and Confidentiality*, *10*(1). https://doi.org/10.29012/jpc.711

Cohen, J. E. (2018). The Biopolitical Public Domain: The Legal Construction of the Surveillance Economy. *Philosophy & Technology*, *31*(2), 213–233. https://doi.org/10.1007/s13347-017-0258-2

Cohen, J. E. (2019). *Between Truth and Power: The Legal Constructions of Informational Capitalism*. Oxford University Press.

Cohen, K. B., Xia, J., Zweigenbaum, P., Callahan, T., Hargraves, O., Goss, F., Ide, N., Névéol, A., Grouin, C., & Hunter, L. E. (2018). Three Dimensions of Reproducibility in Natural Language Processing. In N. Calzolari, K. Choukri, C. Cieri, T. Declerck, S. Goggi, K. Hasida, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk, S. Piperidis, & T. Tokunaga (Eds.), *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*. European Language Resources Association (ELRA). https://aclanthology.org/L18-1025

Confessore, N. (2018). Cambridge Analytica and Facebook: The Scandal and the Fallout So Far [newspaper]. *The New York Times: U.S.* https://www.nytimes.com/2018/04/04/us/politics/cambridge-analytica-scandal-fallout.html

Connolly, C. (2008). *The US Safe Harbor - Fact or Fiction?* Galexia. Pyrmont, Australia. https://www.europarl.europa.eu/meetdocs/2009_2014/documents/libe/dv/08_galexia_safe_harbor_/08_galexia_safe_harbor_en.pdf

Cook, J. R., & Stefanski, L. A. (1994). Simulation-Extrapolation Estimation in Parametric Measurement Error Models. *Journal of the American Statistical Association*, *89*(428), 1314–1328. https://doi.org/10.2307/2290994

Cork, D. L., Citro, C. F., & Kirkendall, N. J. (2020). Panel Discussion on Key Privacy Issues: Privacy and Census Participation. *2020 Census Data Products: Data Needs and Privacy Considerations: Proceedings of a Workshop*, 129–133. https://doi.org/10.17226/25978

Cornman, S., Zhou, L., Howell, M., Phillips, J., & Young, J. (2020). *Revenues and Expenditures for Public Elementary and Secondary Education: FY 18* (Finance Tables). U.S. Department of Education.

Costanza-Chock, S., Raji, I. D., & Buolamwini, J. (2022). Who Audits the Auditors? Recommendations from a field scan of the algorithmic auditing ecosystem. *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, 1571–1583. https://doi.org/10.1145/3531146.3533213

Crenshaw, K. (1990). Mapping the margins: Intersectionality, identity politics, and violence against women of color. *Stan. L. Rev.*, *43*, 1241.

Crilly, D., Zollo, M., & Hansen, M. T. (2012). Faking It or Muddling Through? Understanding Decoupling in Response to Stakeholder Pressures. *Academy of Management Journal*, *55*(6), 1429–1448. https://doi.org/10.5465/amj.2010.0697

Crunchbase. (2023). https://www.crunchbase.com

Cui, Y., Gong, R., Hannig, J., & Hoffman, K. (2023). Technical Comment on "Policy impacts of statistical uncertainty and privacy". *Science*, *380*(6648), eadf9724. https://doi.org/10.1126/science.adf9724

Cumming, G. (2008). Replication and p Intervals: P Values Predict the Future Only Vaguely, but Confidence Intervals Do Much Better. *Perspectives on Psychological Science: A*

Journal of the Association for Psychological Science, *3*(4), 286–300. https://doi.org/10.1111/j.1745-6924.2008.00079.x

Cummings, R., Desfontaines, D., Evans, D., Geambasu, R., Huang, Y., Jagielski, M., Kairouz, P., Kamath, G., Oh, S., Ohrimenko, O., Papernot, N., Rogers, R., Shen, M., Song, S., Su, W., Terzis, A., Thakurta, A., Vassilvitskii, S., Wang, Y.-X., … Zhang, W. (2024). Advancing Differential Privacy: Where We Are Now and Future Directions for Real-World Deployment. *Harvard Data Science Review*. https://doi.org/10.1162/99608f92.d3197524

Cutolo, D., & Kenney, M. (2021). Platform-Dependent Entrepreneurs: Power Asymmetries, Risks, and Strategies in the Platform Economy. *Academy of Management Perspectives*, *35*(4), 584–605. https://doi.org/10.5465/amp.2019.0103

Cyphers, B. (2019). *Don't Play in Google's Privacy Sandbox*. Electronic Frontier Foundation. https://www.eff.org/deeplinks/2019/08/dont-play-googles-privacy-sandbox-1

Data Protection Working Party. (2014). *Opinion 05/2014 on Anonymisation Techniques* (Opinion No. WP216). https://ec.europa.eu/justice/article-29/documentation/opinion-recommendation/files/2014/wp216_en.pdf

Davis, K. (1973). The Case for and Against Business Assumption of Social Responsibilities. *Academy of Management Journal*, *16*(2), 312–322. https://doi.org/10.5465/255331

de Vaujany, F.-X., Fomin, V. V., Haefliger, S., & Lyytinen, K. (2018). Rules, Practices, and Information Technology: A Trifecta of Organizational Regulation. *Information Systems Research*, *29*(3), 755–773. https://doi.org/10.1287/isre.2017.0771

De-Arteaga, M., Romanov, A., Wallach, H., Chayes, J., Borgs, C., Chouldechova, A., Geyik, S., Kenthapadi, K., & Kalai, A. T. (2019). Bias in bios: A case study of semantic representation bias in a high-stakes setting. *Proceedings of the Conference on Fairness, Accountability, and Transparency*, 120–128.

Delgado, F., Yang, S., Madaio, M., & Yang, Q. (2023). The Participatory Turn in AI Design: Theoretical Foundations and the Current State of Practice. *Proceedings of the 3rd ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization*, 1–23. https://doi.org/10.1145/3617694.3623261

Deng, W. H., Guo, B., Devrio, A., Shen, H., Eslami, M., & Holstein, K. (2023). Understanding Practices, Challenges, and Opportunities for User-Engaged Algorithm Auditing in Industry Practice. *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, 1–18. https://doi.org/10.1145/3544548.3581026

Deng, W. H., Nagireddy, M., Lee, M. S. A., Singh, J., Wu, Z. S., Holstein, K., & Zhu, H. (2022). Exploring How Machine Learning Practitioners (Try To) Use Fairness Toolkits. *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, 473–484. https://doi.org/10.1145/3531146.3533113

Department of Education. (2022). *Fiscal Year 2022 Budget Request* (Education for the Disadvantaged). Department of Education. https://www2.ed.gov/about/overview/budget/budget22/justifications/a-ed.pdf

Desai, D. R., & Kroll, J. A. (2017). Trust but Verify: A Guide to Algorithms and the Law. *Harvard Journal of Law & Technology (Harvard JOLT)*, *31*(1), 1–64. https://heinonline.org/HOL/P?h=hein.journals/hjlt31&i=7

Desfontaines, D. (2021). *A list of real-world uses of differential privacy*. Ted is writing things. https://desfontain.es/privacy/real-world-differential-privacy.html

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). *Bert: Pre-training of deep bidirectional transformers for language understanding*.

DeVos, A., Dhabalia, A., Shen, H., Holstein, K., & Eslami, M. (2022). Toward User-Driven Algorithm Auditing: Investigating users' strategies for uncovering harmful algorithmic behavior. *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, 1–19. https://doi.org/10.1145/3491102.3517441

Dick, T., Dwork, C., Kearns, M., Liu, T., Roth, A., Vietri, G., & Wu, Z. S. (2023). Confidence-ranked reconstruction of census microdata from published statistics. *Proceedings of the National Academy of Sciences*, *120*(8), e2218605120. https://doi.org/10.1073/pnas.2218605120

Digital Regulation Cooperation Forum. (2022). *Auditing algorithms: The existing landscape, role of regulators and future outlook* (Research and analysis). Ofcom. https://www.gov.uk/government/publications/findings-from-the-drcf-algorithmic-processing-workstream-spring-2022/auditing-algorithms-the-existing-landscape-role-of-regulators-and-future-outlook

DiMaggio, P., & Powell, W. (1983). The Iron Cage Revisited: Institutional Isomorphism and Collective Rationality in Organizational Fields. *American Sociological Review*, *48*(2), 147–160. https://doi.org/10.2307/2095101

Dinev, T., McConnell, A. R., & Smith, H. J. (2015). Research Commentary—Informing Privacy Research Through Information Systems, Psychology, and Behavioral Economics: Thinking Outside the "APCO" Box. *Information Systems Research*, *26*(4), 639–655. https://doi.org/10.1287/isre.2015.0600

Dinur, I., & Nissim, K. (2003). Revealing information while preserving privacy. *Proceedings of the Twenty-Second ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems - PODS '03*, 202–210. https://doi.org/10.1145/773153.773173

Dobbin, F., & Kalev, A. (2022). *Getting to Diversity: What Works and What Doesn't*. Harvard University Press. https://doi.org/10.2307/j.ctv2t46rd6

Domingo-Ferrer, J., & Blanco-Justicia, A. (2020). Privacy-Preserving Technologies. In M. Christen, B. Gordijn, & M. Loi (Eds.), *The Ethics of Cybersecurity* (pp. 279–298, Vol. 21). Springer. https://doi.org/10.1007/978-3-030-29053-5

Donahue, J., & Simonyan, K. (2019). Large scale adversarial representation learning. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d Alché-Buc, E. Fox, & R. Garnett (Eds.), *Advances in neural information processing systems* (pp. 10542–10552, Vol. 32). Curran Associates, Inc. https://proceedings.neurips.cc/paper/2019/file/18cdf49ea54eec029238fcc95f76ce41-Paper.pdf

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., & Houlsby, N. (2021). An image is worth 16x16 words: Transformers for image recognition at scale. *International Conference on Learning Representations*. https://openreview.net/forum?id=YicbFdNTTy

Doty, N., & Mulligan, D. K. (2013). Internet Multistakeholder Processes and Techno-Policy Standards. *Journal on Telecommunications and High Technology Law*, *11*, 135–182.

Drechsler, J., & Bailie, J. (2024). *The Complexities of Differential Privacy for Survey Data*. arXiv: 2408.07006 [stat]. https://doi.org/10.48550/arXiv.2408.07006

Dwork, C., Feldman, V., Hardt, M., Pitassi, T., Reingold, O., & Roth, A. (2015). The reusable holdout: Preserving validity in adaptive data analysis. *Science*, *349*(6248), 636–638. https://doi.org/10.1126/science.aaa9375

Dwork, C., Kohli, N., & Mulligan, D. (2019). Differential Privacy in Practice: Expose your Epsilons! *Journal of Privacy and Confidentiality*, *9*(2). https://doi.org/10.29012/jpc.689

Dwork, C., McSherry, F., Nissim, K., & Smith, A. (2006). Calibrating noise to sensitivity in private data analysis. *Proceedings of the Third Conference on Theory of Cryptography*, 265–284. https://doi.org/10.1007/11681878_14

Dwork, C., & Roth, A. (2013). The Algorithmic Foundations of Differential Privacy. *Foundations and Trends® in Theoretical Computer Science*, *9*(3–4), 211–407. https://doi.org/10.1561/0400000042

Dwork, C., & Rothblum, G. N. (2016). *Concentrated Differential Privacy*. arXiv: 1603.01887 [cs]. https://doi.org/10.48550/arXiv.1603.01887

Eagly, A. H., Mladinic, A., & Otto, S. (1991). Are women evaluated more favorably than men?: An analysis of attitudes, beliefs, and emotions. *Psychology of Women Quarterly*, *15*(2), 203–216.

Echenique, F., & He, K. (2024). *Screening p-Hackers: Dissemination Noise as Bait*. arXiv: 2103.09164 [econ]. https://doi.org/10.48550/arXiv.2103.09164

Edelman, L. B. (2016). *Working Law: Courts, Corporations, and Symbolic Civil Rights*. University of Chicago Press. https://press.uchicago.edu/ucp/books/book/chicago/W/bo24550454.html

Edelman, L. B., Uggen, C., & Erlanger, H. S. (1999). The Endogeneity of Legal Regulation: Grievance Procedures as Rational Myth. *American Journal of Sociology*, *105*(2), 406–454. https://doi.org/10.1086/210316

Edelson, L., & McCoy, D. (2021). We Research Misinformation on Facebook. It Just Disabled Our Accounts. [magazine]. *The New York Times*. https://www.nytimes.com/2021/08/10/opinion/facebook-misinformation.html

Edmondson, A. C., & Mcmanus, S. E. (2007). Methodological fit in management field research. *Academy of Management Review*, *32*(4), 1246–1264. https://doi.org/10.5465/amr.2007.26586086

Egan, E. (2020). *A Path Forward for Privacy and Online Advertising*. Meta. https://about.fb.com/news/2020/10/a-path-forward-for-privacy-and-online-advertising/

Ehsan, U., & Riedl, M. O. (2020). Human-Centered Explainable AI: Towards a Reflective Sociotechnical Approach. In C. Stephanidis, M. Kurosu, H. Degen, & L. Reinerman-Jones (Eds.), *HCI International 2020 - Late Breaking Papers: Multimodality and Intelligence* (pp. 449–466). Springer International Publishing. https://doi.org/10.1007/978-3-030-60117-1_33

Eisenhardt, K. M. (1989). Building Theories from Case Study Research. *The Academy of Management Review*, *14*(4), 532–550. https://doi.org/10.2307/258557

Erhan, D., Bengio, Y., Courville, A., Manzagol, P.-A., Vincent, P., & Bengio, S. (2010). Why does unsupervised pre-training help deep learning? *Journal of Machine Learning Research*, *11*(19), 625–660. http://jmlr.org/papers/v11/erhan10a.html

Erhan, D., Manzagol, P.-A., Bengio, Y., Bengio, S., & Vincent, P. (2009). The difficulty of training deep architectures and the effect of unsupervised pre-training. In D. van Dyk

& M. Welling (Eds.), *Proceedings of the twelth international conference on artificial intelligence and statistics* (pp. 153–160, Vol. 5). PMLR. http://proceedings.mlr.press/v5/erhan09a.html

European Parliament & Council of the European Union. (2016). *Regulation (EU) 2016/679 of the European Parliament and of the Council. of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation)*. https://data.europa.eu/eli/reg/2016/679/oj

Evans, G., & King, G. (2023). Statistically Valid Inferences from Differentially Private Data Releases, with Application to the Facebook URLs Dataset. *Political Analysis*, *31*(1), 1–21. https://doi.org/10.1017/pan.2022.1

Facets - Know Your Data. (2023). FACETS. https://pair-code.github.io/facets/

Faik, I., Barrett, M., & Oborn, E. (2020). How Information Technology Matters in Societal Change: An Affordance-Based Institutional Perspective. *MIS Quarterly*, *44*(3), 1359–1390. https://doi.org/10.25300/MISQ/2020/14193

Fast Grants. (2024). The AI Safety Institute (AISI). https://www.aisi.gov.uk/grants

Feffer, M., Martelaro, N., & Heidari, H. (2023). The AI Incident Database as an Educational Tool to Raise Awareness of AI Harms: A Classroom Exploration of Efficacy, Limitations, & Future Improvements. *Proceedings of the 3rd ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization*, 1–11. https://doi.org/10.1145/3617694.3623223

Feldman, M. S., & Pentland, B. T. (2003). Reconceptualizing Organizational Routines as a Source of Flexibility and Change. *Administrative Science Quarterly*, *48*(1), 94–118. https://doi.org/10.2307/3556620

Ferraz, C., & Finan, F. (2011). Electoral Accountability and Corruption: Evidence from the Audits of Local Governments. *American Economic Review*, *101*(4), 1274–1311. https://doi.org/10.1257/aer.101.4.1274

Fichman, R. G. (2004). Going Beyond the Dominant Paradigm for Information Technology Innovation Research: Emerging Concepts and Methods. *Journal of the Association for Information Systems*, *5*(8). https://doi.org/10.17705/1jais.00054

Fiss, P. C., & Zajac, E. J. (2004). The Diffusion of Ideas over Contested Terrain: The (Non)adoption of a Shareholder Value Orientation among German Firms. *Administrative Science Quarterly*, *49*(4), 501–534. https://www.jstor.org/stable/4131489

Fiss, P. C., & Zajac, E. J. (2006). The Symbolic Management of Strategic Change: Sensegiving Via Framing and Decoupling. *Academy of Management Journal*, *49*(6), 1173–1193. https://doi.org/10.5465/amj.2006.23478255

Flaxman, A., & Keyes, O. (2025). The Risk of Linked Census Data to Transgender Youth: A Simulation Study. *Journal of Privacy and Confidentiality*, *15*(1). https://doi.org/10.29012/jpc.891

Foroni, F., & Bel-Bahar, T. (2010). Picture-IAT versus word-IAT: Level of stimulus representation influences on the IAT. *European Journal of Social Psychology*, *40*(2), 321–337. https://doi.org/10.1002/ejsp.626

FTC Staff. (2024). *A Look Behind the Screens: Examining the Data Practices of Social Media and Video Streaming Services* (FTC Staff Report). Federal Trade Commission. https://www.ftc.gov/system/files/ftc_gov/pdf/Social-Media-6b-Report-9-11-2024.pdf

Fuller, S. R., Edelman, L. B., & Matusik, S. F. (2000). Legal Readings: Employee Interpretation and Mobilization of Law. *The Academy of Management Review*, *25*(1), 200–216. https://doi.org/10.2307/259270

Galindo, L., Perset, K., & Sheeka, F. (2021). *An overview of national AI strategies and policies* (No. 14). OECD Publishing. Paris. https://www.oecd.org/en/publications/an-overview-of-national-ai-strategies-and-policies_c05140d9-en.html

Garg, N., Schiebinger, L., Jurafsky, D., & Zou, J. (2018). Word embeddings quantify 100 years of gender and ethnic stereotypes. *Proceedings of the National Academy of Sciences of the United States of America*, *115*(16), E3635–E3644. https://doi.org/10.1073/pnas.1720347115

Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J. W., Wallach, H., Daumé III, H., & Crawford, K. (2018). *Datasheets for datasets*.

Geiger, A., Lenz, P., & Urtasun, R. (2012). Are we ready for autonomous driving? the KITTI vision benchmark suite. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 3354–3361. https://doi.org/10.1109/CVPR.2012.6248074

Gelman, A., & Stern, H. (2006). The Difference Between "Significant" and "Not Significant" is not Itself Statistically Significant. *The American Statistician*, *60*(4), 328–331. https://doi.org/10.1198/000313006X152649

Generated photos. (2021). https://generated.photos

Gentry, C. (2009). Fully homomorphic encryption using ideal lattices. *Proceedings of the Forty-First Annual ACM Symposium on Theory of Computing*, 169–178. https://doi.org/10.1145/1536414.1536440

Gerchick, M., & Madubuonwu, B. W. (2024). *Tracking Automated Employment Decision Tool Bias Audits*. https://github.com/aclu-national/tracking-ll144-bias-audits

Ghassemi, M., Oakden-Rayner, L., & Beam, A. L. (2021). The false hope of current approaches to explainable artificial intelligence in health care. *The Lancet Digital Health*, *3*(11), e745–e750. https://doi.org/10.1016/S2589-7500(21)00208-9

Ghavami, N., & Peplau, L. A. (2013). An intersectional analysis of gender and ethnic stereotypes: Testing three hypotheses. *Psychology of Women Quarterly*, *37*(1), 113–127.

Gioia, D., Corley, K., & Hamilton, A. (2013). Seeking Qualitative Rigor in Inductive Research: Notes on the Gioia Methodology. *Organizational Research Methods*, *16*(1), 15–31. https://doi.org/10.1177/1094428112452151

Gioia, D. A., Patvardhan, S. D., Hamilton, A. L., & Corley, K. G. (2013). Organizational Identity Formation and Change. *Academy of Management Annals*, *7*(1), 123–193. https://doi.org/10.5465/19416520.2013.762225

Glaser, B., & Strauss, A. (2017). *Discovery of Grounded Theory: Strategies for Qualitative Research* (1st ed.). Routledge.

Goel, V. (2022). *Get to know the new Topics API for Privacy Sandbox*. Google. https://blog.google/products/chrome/get-know-new-topics-api-privacy-sandbox/

Goldberg, I. (2007). Privacy-Enhancing Technologies for the Internet III: Ten Years Later. In A. Acquisti, S. Gritzalis, C. Lambrinoudakis, & S. di Vimercati (Eds.), *Digital Privacy* (pp. 25–40). Auerbach Publications. https://doi.org/10.1201/9781420052183-7

Goldreich, O. (2009). *Foundations of Cryptography: Volume 2, Basic Applications* (1st ed.). Cambridge University Press.

Gong, R. (2022). Transparent Privacy Is Principled Privacy. *Harvard Data Science Review*. https://doi.org/10.1162/99608f92.b5d3faaa

Goodman, E. (2024). *NTIA Artificial Intelligence Accountability Policy Report*. National Telecommunication and Information Administration. https://www.ntia.gov/sites/default/files/publications/ntia_ai_report_final-3-27-24.pdf

Goodman, E. P., & Tréhu, J. (2022). *AI Audit-Washing and Accountability* (Policy Paper). German Marshall Fund of the United States. https://www.gmfus.org/sites/default/files/2022-11/Goodman%20%26%20Trehu%20-%20Algorithmic%20Auditing%20-%20paper.pdf

Google. (2022). *Request for Information on Advancing Privacy-Enhancing Technologies* (87 Fed. Reg. 35250 No. 2022–12432). https://www.nitrd.gov/rfi/2022/87-fr-35250/Google-PET-RFI-Response-2022.pdf

Graff, K. A., Murnen, S. K., & Krause, A. K. (2013). Low-cut shirts and high-heeled shoes: Increased sexualization across time in magazine depictions of girls. *Sex roles*, *69*(11–12), 571–582.

Green, B., & Viljoen, S. (2020). Algorithmic realism: Expanding the boundaries of algorithmic thought. *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 19–31. https://doi.org/10.1145/3351095.3372840

Greenwald, A. G., McGhee, D. E., & Schwartz, J. L. (1998). Measuring individual differences in implicit cognition: The implicit association test. *Journal of Personality and Social Psychology*, *74*(6), 1464–80. http://www.ncbi.nlm.nih.gov/pubmed/9654756

Greenwald, A. G., Nosek, B. A., & Banaji, M. R. (2003). Understanding and using the implicit association test: I. An improved scoring algorithm. *Journal of Personality and Social Psychology*, *85*(2), 197–216. https://doi.org/10.1037/0022-3514.85.2.197

Greenwald, A. G., Poehlman, T. A., Uhlmann, E. L., & Banaji, M. R. (2009). Understanding and using the implicit association test: III. Meta-analysis of predictive validity. *Journal of personality and social psychology*, *97*(1), 17.

Groves, L., Metcalf, J., Kennedy, A., Vecchione, B., & Strait, A. (2024). Auditing Work: Exploring the New York City algorithmic bias audit regime. *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, 1107–1120. https://doi.org/10.1145/3630106.3658959

Groves, R. M., & Lyberg, L. (2010). Total Survey Error: Past, Present, and Future. *Public Opinion Quarterly*, *74*(5), 849–879. https://doi.org/10.1093/poq/nfq065

Guha, N., Lawrence, C. M., Gailmard, L. A., Rodolfa, K. T., Surani, F., Bommasani, R., Raji, I. D., Cuéllar, M.-F., Honigsberg, C., Liang, P., & Ho, D. E. (2023). AI Regulation Has Its Own Alignment Problem: The Technical and Institutional Feasibility of Disclosure, Registration, Licensing, and Auditing. *George Washington Law Review*, *92*(6), 1473–1557.

Gunningham, N., Kagan, R. A., & Thornton, D. (2004). Social License and Environmental Protection: Why Businesses Go beyond Compliance. *Law & Social Inquiry*, *29*(2), 307–341. https://www.jstor.org/stable/4092687

Guo, W., & Caliskan, A. (2020). *Detecting emergent intersectional biases: Contextualized word embeddings contain a distribution of human-like biases.*

Haack, P., Schoeneborn, D., & Wickert, C. (2012). Talking the Talk, Moral Entrapment, Creeping Commitment? Exploring Narrative Dynamics in Corporate Responsibility Standardization. *Organization Studies*, *33*(5–6), 815–845. https://doi.org/10.1177/0170840612443630

Hallett, T., & Hawbaker, A. (2021). The case for an inhabited institutionalism in organizational research: Interaction, coupling, and change reconsidered. *Theory and Society*, *50*(1), 1–32. https://doi.org/10.1007/s11186-020-09412-2

Han, B.-C. (2015). *The Transparency Society*. Stanford University Press.

Hanelt, A., Busse, S., & Kolbe, L. M. (2017). Driving business transformation toward sustainability: Exploring the impact of supporting IS on the performance contribution of eco-innovations. *Information Systems Journal*, *27*(4), 463–502. https://doi.org/10.1111/isj.12130

Haraway, D. (1988). Situated Knowledges: The Science Question in Feminism and the Privilege of Partial Perspective. *Feminist Studies*, *14*(3), 575–599. https://doi.org/10.2307/3178066

Harvey, E., Sheng, E., Blodgett, S. L., Chouldechova, A., Garcia-Gathright, J., Olteanu, A., & Wallach, H. (2024). *Gaps Between Research and Practice When Measuring Representational Harms Caused by LLM-Based Systems*. arXiv: 2411.15662 [cs]. https://doi.org/10.48550/arXiv.2411.15662

Harwell, D. (2019). A face-scanning algorithm increasingly decides whether you deserve the job. https://www.washingtonpost.com/technology/2019/10/22/ai-hiring-face-scanning-algorithm-increasingly-decides-whether-you-deserve-job/

Hausman, J. (2001). Mismeasured Variables in Econometric Analysis: Problems from the Right and Problems from the Left. *Journal of Economic Perspectives*, *15*(4), 57–67. https://doi.org/10.1257/jep.15.4.57

He, K., Fan, H., Wu, Y., Xie, S., & Girshick, R. (2020). Momentum contrast for unsupervised visual representation learning. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9729–9738.

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 770–778. http://image-net.org/challenges/LSVRC/2015/

Heikkilä, M. (2022). The viral AI avatar app Lensa undressed me—without my consent [newspaper]. *MIT Technology Review*. https://www.technologyreview.com/2022/12/12/1064751/the-viral-ai-avatar-app-lensa-undressed-me-without-my-consent/

Hendricks, L. A., Burns, K., Saenko, K., Darrell, T., & Rohrbach, A. (2018). Women also snowboard: Overcoming bias in captioning models. *European Conference on Computer Vision*, 793–811.

Hennigsson, S., & Eaton, B. D. (2024). Governmental regulation and digital infrastructure innovation: The mediating role of modular architecture. *Journal of Information Technology*, *38*(2), 126–143. https://doi.org/10.1177/02683962221114429

Heuer, R., & Stullich, S. (2011). *Comparability of State and Local Expenditures Among Schools Within Districts: A Report From the Study of School-Level Expenditures* (Policy and Program Studies Service). U.S. Department of Education, Office of Planning, Evaluation and Policy Development. https://files.eric.ed.gov/fulltext/ED527141.pdf

Hewstone, M., Rubin, M., & Willis, H. (2002). Intergroup bias. *Annual review of psychology*, *53*(1), 575–604.

Hickock, M. (2023). *Ethical AI Frameworks, Guidelines, Toolkits*. AI Ethicist. https://www.aiethicist.org/frameworks-guidelines-toolkits

Hill, K. (2020). The secretive company that might end privacy as we know it. https://www.nytimes.com/2020/01/18/technology/clearview-privacy-facial-recognition.html

Hillman, A. J., Keim, G. D., & Schuler, D. (2004). Corporate Political Activity: A Review and Research Agenda. *Journal of Management*, *30*(6), 837–857. https://doi.org/10.1016/j.jm.2004.06.003

Hirsch, D. D., Bartley, T., Chandrasekaran, A., Norris, D., Parthasarathy, S., & Turner, P. N. (2020). *Business Data Ethics: Emerging Trends in the Governance of Advanced Analytics and AI* (Research Paper No. 628). https://papers.ssrn.com/abstract=3828239

Holstein, K., Wortman Vaughan, J., Daumé, H., Dudik, M., & Wallach, H. (2019). Improving Fairness in Machine Learning Systems: What Do Industry Practitioners Need? *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, 1–16. https://doi.org/10.1145/3290605.3300830

Hotz, V. J., Bollinger, C. R., Komarova, T., Manski, C. F., Moffitt, R. A., Nekipelov, D., Sojourner, A., & Spencer, B. D. (2022). Balancing data privacy and usability in the federal statistical system. *Proceedings of the National Academy of Sciences*, *119*(31), e2104906119. https://doi.org/10.1073/pnas.2104906119

Hu, P. J.-H., Hu, H.-f., Wei, C.-P., & Hsu, P.-F. (2016). Examining Firms' Green Information Technology Practices: A Hierarchical View of Key Drivers and Their Effects. *Journal of Management Information Systems*, *33*(4), 1149–1179. https://doi.org/10.1080/07421222.2016.1267532

Huang, K. (2025). *Meta Curbs Privacy Teams' Sway Over Product Releases*. The Information. https://www.theinformation.com/briefings/meta-curbs-privacy-teams-sway-over-product-releases

Hudson, H. M. (1974). *Empirical Bayes Estimation*. https://purl.stanford.edu/cv052xn0322

Infocomm Media Development Authority & AI Verify Foundation. (2023). *Cataloguing LLM Evaluations*. Infocomm Media Development Authority of Singapore. https://aiverifyfoundation.sg/downloads/Cataloguing_LLM_Evaluations.pdf

Information Commissioner's Office. (2023). *Annex A: Fairness in the AI Lifecycle* (Guidance on the AI auditing framework Draft guidance for consultation). Information Commissioner's Office. https://ico.org.uk/for-organisations/uk-gdpr-guidance-and-resources/artificial-intelligence/guidance-on-ai-and-data-protection/

Inter-University Consortium for Political and Social Research. (2025). *Find Data*. ICPSR. https://www.icpsr.umich.edu/web/pages/ICPSR/index.html

Iozzio, C. (2016). The playboy centerfold that revolutionized image-processing research. *The Atlantic*. https://www.theatlantic.com/technology/archive/2016/02/lena-image-processing-playboy/461970/

IPUMS. (2024). *Bibliography*. IPUMS. https://bibliography.ipums.org/citations/results?search_terms % 5Bmax_year_published % 5D = 2024 & search_terms % 5Bmin_year_published%5D=1960

Jarmin, R. (2019). *Census Bureau Continues to Boost Data Safeguards*. The United States Census Bureau. https://www.census.gov/newsroom/blogs/random-samplings/2019/07/boost-safeguards.html

Jin, A., & Salehi, N. (2024). (Beyond) Reasonable Doubt: Challenges that Public Defenders Face in Scrutinizing AI in Court. *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, 1–19. https://doi.org/10.1145/3613904.3641902

Johnson, N., Moharana, S., Harrington, C., Andalibi, N., Heidari, H., & Eslami, M. (2024). The Fall of an Algorithm: Characterizing the Dynamics Toward Abandonment. *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, 337–358. https://doi.org/10.1145/3630106.3658910

Jones, T. M. (1995). Instrumental Stakeholder Theory: A Synthesis of Ethics and Economics. *The Academy of Management Review*, *20*(2), 404–437. https://doi.org/10.2307/258852

Kaltheuner, F. (Ed.). (2021). *Fake AI*. Meatspace Press.

Kamieniecki, S. (2006). *Corporate America and Environmental Policy: How Often Does Business Get Its Way?* Stanford University Press.

Karras, T., Laine, S., & Aila, T. (2019). A style-based generator architecture for generative adversarial networks. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4401–4410.

Kaur, H., Nori, H., Jenkins, S., Caruana, R., Wallach, H., & Wortman Vaughan, J. (2020). Interpreting Interpretability: Understanding Data Scientists' Use of Interpretability Tools for Machine Learning. *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, 1–14. https://doi.org/10.1145/3313831.3376219

Kay, M., Matuszek, C., & Munson, S. A. (2015). Unequal representation and gender stereotypes in image search results for occupations. *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems - CHI '15*, 3819–3828. https://doi.org/10.1145/2702123.2702520

Kaye, K., & Dixon, P. (2023). *Risky Analysis: Assessing and Improving AI Governance Tools*. World Privacy Forum. https://www.worldprivacyforum.org/wp-content/uploads/2023/12/WPF_Risky_Analysis_December_2023_fs.pdf

Keller, R., Ollig, P., & Fridgen, G. (2019). Decoupling, Information Technology, and the Tradeoff between Organizational Reliability and Organizational Agility. https://orbilu.uni.lu/handle/10993/44524

Kennedy, M. T., & Fiss, P. C. (2009). Institutionalization, framing, and diffusion: The logic of TQM adoption and implementation decisions among U.S. Hospitals. *Academy of Management Journal*, *52*(5), 897–918. https://doi.org/10.5465/AMJ.2009.44633062

Kenny, C. T., Kuriwaki, S., McCartan, C., Rosenman, E. T. R., Simko, T., & Imai, K. (2021). The use of differential privacy for census data and its impact on redistricting: The case of the 2020 U.S. Census. *Science Advances*, *7*(41), eabk3283. https://doi.org/10.1126/sciadv.abk3283

Ketter, W., Schroer, K., & Valogianni, K. (2023). Information Systems Research for Smart Sustainable Mobility: A Framework and Call for Action. *Information Systems Research*, *34*(3), 1045–1065. https://doi.org/10.1287/isre.2022.1167

Kim, S. S. Y., Watkins, E. A., Russakovsky, O., Fong, R., & Monroy-Hernández, A. (2023). "Help Me Help the AI": Understanding How Explainability Can Support Human-AI

Interaction. *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, 1–17. https://doi.org/10.1145/3544548.3581001

King, G., & Persily, N. (2020). *Unprecedented Facebook URLs Dataset now Available for Academic Research through Social Science One*. Social Science One. https://socialscience.one/blog/unprecedented-facebook-urls-dataset-now-available-research-through-social-science-one

Klein, R. A., Ratliff, K. A., Vianello, M., Adams, R. B., Bahník, Š., Bernstein, M. J., Bocian, K., Brandt, M. J., Brooks, B., Brumbaugh, C. C., Cemalcilar, Z., Chandler, J., Cheong, W., Davis, W. E., Devos, T., Eisner, M., Frankowska, N., Furrow, D., Galliani, E. M., … Nosek, B. A. (2014). Investigating Variation in Replicability. *Social Psychology*, *45*(3), 142–152. https://doi.org/10.1027/1864-9335/a000178

Klinger, U., & Ohme, J. (2023). *Delegated Regulation on Data Access Provided for the Digital Services Act: Response to the Call for Evidence DG CNECT-CNECT F2 by the European Commission* (No. 7). Weizenbaum Institute. https://www.weizenbaum-library.de/handle/id/380

Kranz, S. (2024). *Find Economic Articles with Data*. https://ejd.econ.mathematik.uni-ulm.de/

Krizhevsky, A. (2009). *Learning multiple layers of features from tiny images*. Citeseer / University of Toronto. https://www.cs.toronto.edu/~kriz/learning-features-2009-TR.pdf

Kroll, J. A. (2018). The fallacy of inscrutability. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, *376*(2133), 20180084. https://doi.org/10.1098/rsta.2018.0084

Kulynych, Bogdan, Madras, D., Milli, S., Raji, I. D., Zhou, A., & Zemel, R. (2020). Participatory approaches to machine learning.

Kumar, I. E., Venkatasubramanian, S., Scheidegger, C., & Friedler, S. A. (2020). Problems with Shapley-value-based explanations as feature importance measures. *Proceedings of the 37th International Conference on Machine Learning*, *119*, 5491–5500.

Kurita, K., Vyas, N., Pareek, A., Black, A. W., & Tsvetkov, Y. (2019). Measuring bias in contextualized word representations. *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, 166–172. https://doi.org/10.18653/v1/w19-3823

Kuznetsova, A., Rom, H., Alldrin, N., Uijlings, J., Krasin, I., Pont-Tuset, J., Kamali, S., Popov, S., Malloci, M., Duerig, T., et al. (2018). *The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale*.

Lam, M. S., Pandit, A., Kalicki, C. H., Gupta, R., Sahoo, P., & Metaxa, D. (2023). Sociotechnical Audits: Broadening the Algorithm Auditing Lens to Investigate Targeted Advertising. *Proceedings of the ACM on Human-Computer Interaction*, *7*, 360:1–360:37. https://doi.org/10.1145/3610209

Lampland, M., & Star, S. L. (2009). *Standards and Their Stories: How Quantifying, Classifying, and Formalizing Practices Shape Everyday Life*. Cornell University Press.

Lawrence, C., Cui, I., & Ho, D. (2023). The Bureaucratic Challenge to AI Governance: An Empirical Assessment of Implementation at U.S. Federal Agencies. *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, 606–652. https://doi.org/10.1145/3600211.3604701

Lee, A. S., & Baskerville, R. L. (2003). Generalizing Generalizability in Information Systems Research. *Information Systems Research*, *14*(3), 221–243. https://doi.org/10.1287/isre.14.3.221.16560

Lee, H.-P., Gao, L., Yang, S., Forlizzi, J., & Das, S. (2024). "I Don't Know If We're Doing Good. I Don't Know If We're Doing Bad": Investigating How Practitioners Scope, Motivate, and Conduct Privacy Work When Developing AI Products. *USENIX Security Symposium*. https://sauvikdas.com/papers/47/serve

Lee, M. S. A., & Singh, J. (2021). The Landscape and Gaps in Open Source Fairness Toolkits. *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, 1–13. https://doi.org/10.1145/3411764.3445261

Leerssen, P. (2023). *Digital Services Act: Summary report on the call for evidence on the Delegated Regulation on data access*. European Commission. https://digital-strategy.ec.europa.eu/en/library/digital-services-act-summary-report-call-evidence-delegated-regulation-data-access

Leidner, D. E., Sutanto, J., & Goutas, L. (2022). Multifarious Roles and Conflicts on an Interorganizational Green Is. *MIS Quarterly*, *46*(1), 591–608. https://doi.org/10.25300/MISQ/2022/15116

Lenhart, A. (2023). *Federal AI Legislation: An Analysis of Proposals from the 117th Congress Relevant to Generative AI tools*. Institute for Data, Democracy & Politics. George Washington University. https://iddp.gwu.edu/sites/g/files/zaxdzs5791/files/2023-06/federal_ai_legislation_v3.pdf

Lessig, L. (1999). *Code and Other Laws of Cyberspace*. Basic Books, Inc.

Levine, S. (2021). *Letter from Acting Director of the Bureau of Consumer Protection Samuel Levine to Facebook*. https://www.ftc.gov/blog-posts/2021/08/letter-acting-director-bureau-consumer-protection-samuel-levine-facebook

Levy, K. (2022). *Data Driven: Truckers, Technology, and the New Workplace Surveillance*. Princeton University Press.

Li, J., & Wu, D. ((2020). Do Corporate Social Responsibility Engagements Lead to Real Environmental, Social, and Governance Impact? *Management Science*, *66*(6), 2564–2588. https://doi.org/10.1287/mnsc.2019.3324

Li, T. (2022). Algorithmic Destruction. *SMU Law Review*, *75*(3), 479. https://doi.org/10.25172/smulr.75.3.2

Liang, H., Saraf, N., Hu, Q., & Xue, Y. (2007). Assimilation of Enterprise Systems: The Effect of Institutional Pressures and the Mediating Role of Top Management. *MIS Quarterly*, *31*(1), 59–87. https://doi.org/10.2307/25148781

Liao, Q. V., Gruen, D., & Miller, S. (2020). Questioning the AI: Informing Design Practices for Explainable AI User Experiences. *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, 1–15. https://doi.org/10.1145/3313831.3376590

Lim, A., & Tsutsui, K. (2011). Globalization and commitment in corporate social responsibility: Cross-national analyses of institutional and political-economy effect. *American Sociological Review*, *77*(1), 69–98. https://doi.org/10.1177/0003122411432701

Lipton, Z. C. (2018). The Mythos of Model Interpretability: In machine learning, the concept of interpretability is both important and slippery. *Queue*, *16*(3), 31–57. https://doi.org/10.1145/3236386.3241340

A Local Law to Amend the Administrative Code of the City of New York, in Relation to Automated Employment Decision Tools (2021). https://www.nyc.gov/site/dca/about/automated-employment-decision-tools.page

Loeser, F., Recker, J., Brocke, J. vom, Molla, A., & Zarnekow, R. (2017). How IT executives create organizational benefits by translating environmental strategies into Green IS initiatives. *Information Systems Journal*, *27*(4), 503–553. https://doi.org/10.1111/isj.12136

Loken, E., & Gelman, A. (2017). Measurement error and the replication crisis. *Science*, *355*(6325), 584–585. https://doi.org/10.1126/science.aal3618

Lomas, N. (2021). *France's competition authority declines to block Apple's opt-in consent for iOS app tracking*. TechCrunch. https://techcrunch.com/2021/03/17/frances-competition-authority-declines-to-block-apples-opt-in-consent-for-ios-app-tracking/

Longpre, S., Kapoor, S., Klyman, K., Ramaswami, A., Bommasani, R., Blili-Hamelin, B., Huang, Y., Skowron, A., Yong, Z.-X., Kotha, S., Zeng, Y., Shi, W., Yang, X., Southen, R., Robey, A., Chao, P., Yang, D., Jia, R., Kang, D., ... Henderson, P. (2024). *A Safe Harbor for AI Evaluation and Red Teaming*. arXiv: 2403.04893 [cs]. https://doi.org/10.48550/arXiv.2403.04893

Luery, D. M. (2010). Small Area Income and Poverty Estimates Program. *27th Standing Committee for Regional and Urban Statistics*, 16.

Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. *Proceedings of the 31st International Conference on Neural Information Processing Systems*, 4768–4777.

Machanavajjhala, A., Kifer, D., Abowd, A., John M, Gehrke, J., & Vilhuber, L. (2008). Privacy: Theory meets Practice on the Map. *2008 IEEE 24th International Conference on Data Engineering*, 277–286. https://doi.org/10.1109/ICDE.2008.4497436

Madaio, M. A., Stark, L., Wortman Vaughan, J., & Wallach, H. (2020). Co-Designing Checklists to Understand Organizational Challenges and Opportunities around Fairness in AI. *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, 1–14. https://doi.org/10.1145/3313831.3376445

Madden, M., Gilman, M., Levy, K., & Marwick, A. (2017). Privacy, Poverty, and Big Data: A Matrix of Vulnerabilities for Poor Americans. *Washington University Law Review*, *95*(1), 53–126. https://heinonline.org/HOL/P?h=hein.journals/walq95&i=59

Malhotra, A., Melville, N. P., & Watson, R. T. (2013). Spurring Impactful Research on Information Systems and Environmental Sustainability. *MIS Quarterly*, *37*(4), 1265–1274. https://doi.org/10.25300/MISQ/2013/37:4.3

Malin, B., Benitez, K., & Masys, D. (2011). Never too old for anonymity: A statistical standard for demographic data sharing via the HIPAA Privacy Rule. *Journal of the American Medical Informatics Association*, *18*(1), 3–10. https://doi.org/10.1136/jamia.2010.004622

Manjunatha, V., Saini, N., & Davis, L. S. (2019). Explicit bias discovery in visual question answering models. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Manski, C. F. (2011). Policy Analysis with Incredible Certitude. *The Economic Journal*, *121*(554), F261–F289. https://doi.org/10.1111/j.1468-0297.2011.02457.x

Maples, J. (2008). *Calculating Coefficient of Variation for the Minimum Change School District Poverty Estimates and the Assessment of the Impact of Nongeocoded Tax Returns* (Statistics No. 2008–10). U.S. Census Bureau.

Maples, J. (2019). Small Area Estimates of the Child Population and Poverty in School Districts Using Dirichlet-Multinomial Models. *Proceedings of the American Statistical Association*, 9.

Marda, N., Sun, J., & Surman, M. (2024). *Public AI: Making AI work for everyone, by everyone*. Mozilla Foundation.

Marquis, C., & Qian, C. (2014). Corporate Social Responsibility Reporting in China: Symbol or Substance? *Organization Science*, *25*(1), 127–148. https://www.jstor.org/stable/43660871

Martin, K., Nissenbaum, H., & Shmatikov, V. (2023). *No Cookies For You!: Evaluating The Promises Of Big Tech's 'Privacy-Enhancing' Techniques*. https://papers.ssrn.com/abstract=4655228

Matthews, G., & Harel, O. (2011). Data confidentiality: A review of methods for statistical disclosure limitation and methods for assessing privacy. *Statistics Surveys*, *5*, 1–29. https://doi.org/10.1214/11-SS074

May, C., Wang, A., Bordia, S., Bowman, S. R., & Rudinger, R. (2019). On measuring social biases in sentence encoders. *Proceedings of the 2019 Conference of the North*, 622–628. https://doi.org/10.18653/v1/N19-1063

McGuigan, L., Sivan-Sevilla, I., Parham, P., & Shvartzshnaider, Y. (2023). Private attributes: The meanings and mechanisms of "privacy-preserving" adtech. *New Media & Society*. https://doi.org/10.1177/14614448231213267

McMahan, B., Moore, E., Ramage, D., Hampson, S., & Arcas, B. A. y. (2017). Communication-Efficient Learning of Deep Networks from Decentralized Data. *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, 1273–1282. https://proceedings.mlr.press/v54/mcmahan17a.html

Meta Research. (2019). *New Research Award in Privacy Preserving Tech*. https://research.fb.com/blog/2019/11/facebook-announces-new-research-awards-in-privacy-preserving-tech-at-ccs/

Metaxa, D., Park, J. S., Robertson, R. E., Karahalios, K., Wilson, C., Hancock, J., & Sandvig, C. (2021). Auditing Algorithms: Understanding Algorithmic Systems from the Outside In. *Foundations and Trends® in Human–Computer Interaction*, *14*(4), 272–344. https://doi.org/10.1561/1100000083

Metcalf, J., Moss, E., & boyd, d. (2019). Owning Ethics: Corporate Logics, Silicon Valley, and the Institutionalization of Ethics. *Social Research: An International Quarterly*, *86*(2), 449–476.

Metcalf, J., Singh, R., Moss, E., Tafesse, E., & Watkins, E. A. (2023). Taking Algorithms to Courts: A Relational Approach to Algorithmic Accountability. *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, 1450–1462. https://doi.org/10.1145/3593013.3594092

Meyer, J. W., & Rowan, B. (1977). Institutionalized Organizations: Formal Structure as Myth and Ceremony. *American Journal of Sociology*, *83*(2), 340–363. http://www.jstor.org/stable/2778293

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). *Efficient estimation of word representations in vector space*.

Misra, I., & Van Der Maaten, L. (2020). Self-supervised learning of pretext-invariant representations. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 6707–6717.

Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., Spitzer, E., Raji, I. D., & Gebru, T. (2019a). Model cards for model reporting. *Proceedings of the Conference on Fairness, Accountability, and Transparency*, 220–229. https://doi.org/10.1145/3287560.3287596

Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., Spitzer, E., Raji, I. D., & Gebru, T. (2019b). Model Cards for Model Reporting. *Proceedings of the Conference on Fairness, Accountability, and Transparency*, 220–229. https://doi.org/10.1145/3287560.3287596

Monteiro, E., Constantinides, P., Scott, S., Shaikh, M., & Burton-Jones, A. (2022). Editor's Comments: Qualitative Research Methods in Information Systems: A Call for Phenomenon-Focused Problematization. *MIS Quarterly*, *46*(4), iii–xix. https://aisel.aisnet.org/misq/vol46/iss4/4

Morozov, E. (2013). *To Save Everything, Click Here: The Folly of Technological Solutionism* (45555th edition). PublicAffairs.

Morris, C. N., & Lysy, M. (2012). Shrinkage Estimation in Multilevel Normal Models. *Statistical Science*, *27*(1), 115–134. https://doi.org/10.1214/11-STS363

Morrison, S. (2022). *The winners and losers of Apple's anti-tracking feature*. Vox. https://www.vox.com/recode/23045136/apple-app-tracking-transparency-privacy-ads

Mozilla Foundation. (2021). *YouTube Regrets*. Mozilla Foundation. https://assets.mofoprod.net/network/documents/Mozilla_YouTube_Regrets_Report.pdf

Mozilla Technology Fund (MTF). (2024). Mozilla Foundation. https://foundation.mozilla.org/en/what-we-fund/opportunities/mozilla-technology-fund-mtf/

Munilla Garrido, G., Liu, X., Matthes, F., & Song, D. (2023). Lessons Learned: Surveying the Practicality of Differential Privacy in the Industry. *Proceedings on Privacy Enhancing Technologies*, *2023*, 151–170. https://doi.org/10.56553/popets-2023-0045

Murphy, J. (2012). Title I of ESEA: The Politics of Implementing Federal Education Reform. *Harvard Educational Review*, *41*(1), 35–63. https://doi.org/10.17763/haer.41.1.gv0n223076667l75

Nader, L. (1972). *Up the Anthropologist: Perspectives Gained From Studying Up*. https://eric.ed.gov/?id=ED065375
ERIC Number: ED065375.

Nanayakkara, P., & Hullman, J. (2022). What's Driving Conflicts Around Differential Privacy for the U.S. Census. *IEEE Security & Privacy*, 2–11. https://doi.org/10.1109/MSEC.2022.3202793

Narayan, D. (2022). Platform capitalism and cloud infrastructure: Theorizing a hyper-scalable computing regime. *Environment and Planning A: Economy and Space*, *54*(5), 911–929. https://doi.org/10.1177/0308518X221094028

Narayanan, A., & Kapoor, S. (2024). *AI snake oil: What artificial intelligence can do, what it can't, and how to tell the difference*. Princeton University Press.

Narayanan, A., & Shmatikov, V. (2007). *How To Break Anonymity of the Netflix Prize Dataset*. arXiv: cs/0610105. https://doi.org/10.48550/arXiv.cs/0610105

Narayanan, A., & Shmatikov, V. (2010). Myths and fallacies of "Personally Identifiable Information". *Communications of the ACM*, *53*(6), 24–26. https://doi.org/10.1145/1743546.1743558

National Research Council. (2000). *Small-Area Income and Poverty Estimates: Priorities for 2000 and Beyond*. National Academies Press. https://doi.org/10.17226/9957

National Research Council. (2003). *Statistical Issues in Allocating Funds by Formula*. National Academies Press. https://doi.org/10.17226/10580

National Science and Technology Council. (2023). *National Strategy to Advance Privacy-Preserving Data Sharing and Analytics*. Executive Office of the President.

Nex, F., & Remondino, F. (2014). UAV for 3D mapping applications: A review. *Springer Verlag*. https://doi.org/10.1007/s12518-013-0120-x

Ngan, M., Grother, P., & Hanaoka, K. (2020). *Ongoing Face Recognition Vendor Test (FRVT) Part 6B: Face recognition accuracy with face masks using post-COVID-19 algorithms*. National Institute of Standards and Technology. https://doi.org/10.6028/NIST.IR.8331

Ngong, I. C., Stenger, B., Near, J. P., & Feng, Y. (2024). Evaluating the usability of differential privacy tools with data practitioners. *Twentieth Symposium on Usable Privacy and Security (SOUPS 2024)*, 21–40. https://www.usenix.org/conference/soups2024/presentation/ngong

Nissim, K., & Wood, A. (2018). Is privacy privacy? *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, *376*(2128), 20170358. https://doi.org/10.1098/rsta.2017.0358

Gig Economy Data Hub. (2021). Gig Economy Data Hub. https://www.gigeconomydata.org/home

Nosek, B. A., & Banaji, M. R. (2001). The GO/NO-GO association task. *Social Cognition*, *19*(6), 625–664. https://doi.org/10.1521/soco.19.6.625.20886

Nosek, B. A., Banaji, M. R., & Greenwald, A. G. (2002). Harvesting implicit group attitudes and beliefs from a demonstration web site. *Group Dynamics*, *6*(1), 101–115. https://doi.org/10.1037/1089-2699.6.1.101

Nosek, B. A., Greenwald, A. G., & Banaji, M. R. (2007). The implicit association test at age 7: A methodological and conceptual review. In J. A. Bargh (Ed.), *Automatic processes in social thinking and behavior* (pp. 265–292). Psychology Press.

Nosek, B. A., Smyth, F. L., Hansen, J. J., Devos, T., Lindner, N. M., Ranganath, K. A., Smith, C. T., Olson, K. R., Chugh, D., Greenwald, A. G., & Banaji, M. R. (2007). Pervasiveness and correlates of implicit attitudes and stereotypes. *European Review of Social Psychology*, *18*(1), 36–88. https://doi.org/10.1080/10463280701489053

Oberski, D., & Kreuter, F. (2020). Differential Privacy and Social Science: An Urgent Puzzle. *Harvard Data Science Review*, *2*(1). https://doi.org/10.1162/99608f92.63a22079

OECD. (2021). *Tools for trustworthy AI: A framework to compare implementation tools for trustworthy AI systems*. OECD. Paris. https://doi.org/10.1787/008232ec-en

Office of Science and Technology Policy. (2022a). Request for Information on Advancing Privacy-Enhancing Technologies. *Federal Register*. https://www.federalregister.gov/

documents/2022/06/09/2022-12432/request-for-information-on-advancing-privacy-enhancing-technologies

Office of Science and Technology Policy. (2022b). *Blueprint for an AI Bill of Rights*. Executive Office of the President. https://www.whitehouse.gov/ostp/ai-bill-of-rights/

Ojewale, V., Steed, R., Vecchione, B., Birhane, A., & Raji, I. D. (2025). Towards AI Accountability Infrastructure: Gaps and Opportunities in AI Audit Tooling. *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, 1–29. https://doi.org/10.1145/3706598.3713301

Okhmatovskiy, I., & David, R. J. (2012). Setting Your Own Standards: Internal Corporate Governance Codes as a Response to Institutional Pressure. *Organization Science*, *23*(1), 155–176. https://doi.org/10.1287/orsc.1100.0642

Oliver, C. (1991). Strategic Responses to Institutional Processes. *The Academy of Management Review*, *16*(1), 145–179. https://doi.org/10.2307/258610

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, *349*(6251), aac4716. https://doi.org/10.1126/science.aac4716

OpenAI. (2023). *GPT-4 Technical Report*. arXiv: 2303.08774 [cs]. https://doi.org/10.48550/arXiv.2303.08774

Orlitzky, M., & Benjamin, J. D. (2001). Corporate Social Performance and Firm Risk: A Meta-Analytic Review. *Business & Society*, *40*(4), 369–396. https://doi.org/10.1177/000765030104000402

Orton, J. D., & Weick, K. E. (1990). Loosely Coupled Systems: A Reconceptualization. *Academy of Management Review*, *15*(2), 203–223. https://doi.org/10.5465/amr.1990.4308154

Park, S., & Cha, H. (2019). Institutional decoupling and the limited implementation of certified environmental technologies. *Journal of Environmental Management*, *247*, 253–262. https://doi.org/10.1016/j.jenvman.2019.05.116

Penberthy, L. (2023). SEER Marks 50 Years of Turning Cancer Data into Discovery. *NIH Record*, *75*(3), 12. https://nihrecord.nih.gov/2023/02/03/seer-marks-50-years-turning-cancer-data-discovery

Pennington, J., Socher, R., & Manning, C. (2014). GloVe: Global vectors for word representation. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 1532–1543. https://doi.org/10.3115/v1/D14-1162

Perrigo, B. (2023). California Bill Proposes Regulating AI at State Level [magazine]. *Time*. https://time.com/6313588/california-ai-regulation-bill/

Peters, M., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., & Zettlemoyer, L. (2018). Deep contextualized word representations. *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, 2227–2237. https://doi.org/10.18653/v1/N18-1202

Piktus, A., Akiki, C., Villegas, P., Laurençon, H., Dupont, G., Luccioni, S., Jernite, Y., & Rogers, A. (2023). The ROOTS Search Tool: Data Transparency for LLMs. In D. Bollegala, R. Huang, & A. Ritter (Eds.), *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)* (pp. 304–314). Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.acl-demo.29

Pollach, I. (2011). Online privacy as a corporate social responsibility: An empirical study. *Business Ethics: A European Review*, *20*(1), 88–102. https://doi.org/10.1111/j.1467-8608.2010.01611.x

Pollman, E., & Barry, J. M. (2016). Regulatory Entrepreneurship. *Southern California Law Review*, *90*(3), 383–448. https://heinonline.org/HOL/P?h=hein.journals/scal90&i=427

Powell, W. W., & DiMaggio, P. J. (2023). The Iron Cage Redux: Looking Back and Forward. *Organization Theory*, *4*(4), 26317877231221550. https://doi.org/10.1177/26317877231221550

Power, M. (1999). *The Audit Society*. Oxford University Press. https://doi.org/10.1093/acprof:oso/9780198296034.001.0001

Powers, D., Basel, W., & O'Hara, B. (2008). *SAIPE County Poverty Models Using Data from the American Community Survey* (Research Report). U.S. Census Bureau, Small Area Estimates Branch. Washington, D.C.

Pujol, D., McKenna, R., Kuppam, S., Hay, M., Machanavajjhala, A., & Miklau, G. (2020). Fair decision making using privacy-protected data. *FAT* 2020 - Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 189–199. https://doi.org/10.1145/3351095.3372872

Queensland Government. (2017). *Community engagement toolkit for planning*. https://dilgpprd.blob.core.windows.net/general/community-engagement-toolkit.pdf

Qureshi, I., Pan, S. L., & Zheng, Y. (2021). Digital social innovation: An overview and research framework. *Information Systems Journal*, *31*(5), 647–671. https://doi.org/10.1111/isj.12362

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language models are unsupervised multitask learners. *OpenAI Blog*, *1*(8), 9.

Raghavan, M., Barocas, S., Kleinberg, J., & Levy, K. (2020). Mitigating bias in algorithmic hiring: Evaluating claims and practices. *FAT* 2020 - Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 469–481. https://doi.org/10.1145/3351095.3372828

Raji, D., Birhane, A., Vecchione, B., Steed, R., & Ojewale, V. (2024). *Comment on NIST-2023-0309* (Request for Information (RFI) Related to NIST's Assignments Under Sections 4.1, 4.5 and 11 of the Executive Order Concerning Artificial Intelligence). National Institute of Standards and Technology. https://www.regulations.gov/comment/NIST-2023-0009-0171

Raji, D., Vecchione, B., Birhane, A., Steed, R., & Ojewale, V. (2023a). *Feedback from Mozilla Open Source Audit Tooling (OAT) Project* (Delegated Regulation on data access provided for in the Digital Services Act No. F3423931). European Commission. https://ec.europa.eu/info/law/better-regulation/have-your-say/initiatives/13817-Delegated-Regulation-on-data-access-provided-for-in-the-Digital-Services-Act/F3423931_en

Raji, D., Vecchione, B., Birhane, A., Steed, R., & Ojewale, V. (2023b). *Comment on NTIA-2023-0005* (NTIA AI Accountability Request for Comment). National Telecommunications and Information Administration. https://www.regulations.gov/comment/NTIA-2023-0005-1439

Raji, I. D., & Buolamwini, J. (2022). Actionable Auditing Revisited: Investigating the Impact of Publicly Naming Biased Performance Results of Commercial AI Products. *Communications of the ACM*, *66*(1), 101–108. https://doi.org/10.1145/3571151

Raji, I. D., Gebru, T., Mitchell, M., Buolamwini, J., Lee, J., & Denton, E. (2020). Saving face: Investigating the ethical concerns of facial recognition auditing. *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 145–151.

Raji, I. D., Kumar, I. E., Horowitz, A., & Selbst, A. (2022). The Fallacy of AI Functionality. *2022 ACM Conference on Fairness, Accountability, and Transparency*, 959–972. https://doi.org/10.1145/3531146.3533158

Raji, I. D., Smart, A., White, R. N., Mitchell, M., Gebru, T., Hutchinson, B., Smith-Loud, J., Theron, D., & Barnes, P. (2020). Closing the AI accountability gap: Defining an end-to-end framework for internal algorithmic auditing. *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 33–44. https://doi.org/10.1145/3351095.3372873

Raji, I. D., Xu, P., Honigsberg, C., & Ho, D. (2022). Outsider Oversight: Designing a Third Party Audit Ecosystem for AI Governance. *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*, 557–571. https://doi.org/10.1145/3514094.3534181

Rakova, B., Yang, J., Cramer, H., & Chowdhury, R. (2021). Where Responsible AI meets Reality: Practitioner Perspectives on Enablers for Shifting Organizational Practices. *Proceedings of the ACM on Human-Computer Interaction*, *5*, 7:1–7:23. https://doi.org/10.1145/3449081

Reamer, A. (2020). *Comprehensive Accounting of Census-Guided Federal Spending (FY2017)* (Brief No. 7A). George Washington Institute of Public Policy.

Recht, B., Roelofs, R., Schmidt, L., & Shankar, V. (2019). Do ImageNet classifiers generalize to ImageNet? In K. Chaudhuri & R. Salakhutdinov (Eds.), *Proceedings of the 36th international conference on machine learning* (pp. 5389–5400, Vol. 97). PMLR. http://proceedings.mlr.press/v97/recht19a.html

Riddle, W. (2011). *Title I and High Schools: Addressing the Needs of Disadvantaged Students at All Grade Levels* (Policy Brief). Alliance for Excellent Education.

Rodriguez, R. (2021). *Disclosure Avoidance and the ACS*. https://acsdatacommunity.prb.org/p/conferences

Rogers, E. M. (2003). *Diffusion of Innovations, 5th Edition* (5th edition). Free Press.

Rooney, P. (2021). *Preliminary Fiscal Year (FY) 2021 (School Year (SY) 2021-2022) Allocations for the Title I, Part A Grants to Local Educational Agencies (LEAs) Program Authorized by the Elementary and Secondary Education Act of 1965 (ESEA).*

Rosenblatt, L., Herman, B., Holovenko, A., Lee, W., Loftus, J., McKinnie, E., Rumezhak, T., Stadnik, A., Howe, B., & Stoyanovich, J. (2023). *Epistemic Parity: Reproducibility as an Evaluation Metric for Differential Privacy*. arXiv: 2208.12700 [cs]. https://doi.org/10.48550/arXiv.2208.12700

Rosenblatt, L., Howe, B., & Stoyanovich, J. (2024). *Are Data Experts Buying into Differentially Private Synthetic Data? Gathering Community Perspectives*. arXiv: 2412.13030 [cs]. https://doi.org/10.48550/arXiv.2412.13030

Ruggles, S., Fitch, C., Magnuson, D., & Schroeder, J. (2019). Differential Privacy and Census Data: Implications for Social and Economic Research. *403 AEA Papers and Proceedings*, *109*, 403–408. https://doi.org/10.1257/pandp.20191107

Russakovsky, O., Deng, J., Huang, Z., Berg, A. C., & Fei-Fei, L. (2013). Detecting avocados to zucchinis: What have we done, and where are we going? *Proceedings of the IEEE International Conference on Computer Vision*, 2064–2071.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A. C., & Fei-Fei, L. (2015). ImageNet large scale visual recognition challenge. *International Journal of Computer Vision*, *115*(3), 211–252. https://doi.org/10.1007/s11263-015-0816-y

Ryan-Mosley, T. (2023). Why everyone is mad about New York's AI hiring law [newspaper]. *MIT Technology Review*. https://www.technologyreview.com/2023/07/10/1076013/new-york-ai-hiring-law/

Sandberg, S. (2019). *Doing More to Protect Against Discrimination in Housing, Employment and Credit Advertising*. Meta. https://about.fb.com/news/2019/03/protecting-against-discrimination-in-ads/

Sandvig, C., Hamilton, K., Karahalios, K., & Langbort, C. (2014). Auditing Algorithms: Research Methods for Detecting Discrimination on Internet Platforms. *Data and Discrimination: Converting Critical Concerns into Productive: A Preconference at the 64th Annual Meeting of the International Communication Association*, 23.

Sandvig v. Bar. https://casetext.com/case/sandvig-v-barr

Santos-Lozada, A. R., Howard, J. T., & Verdery, A. M. (2020). How differential privacy will affect our understanding of health disparities in the United States. *Proceedings of the National Academy of Sciences of the United States of America*, *117*(24), 13405–13412. https://doi.org/10.1073/pnas.2003714117

Sarathy, J., Song, S., Haque, A., Schlatter, T., & Vadhan, S. (2023). Don't Look at the Data! How Differential Privacy Reconfigures the Practices of Data Science. *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, 1–19. https://doi.org/10.1145/3544548.3580791

Schoeneborn, D., Morsing, M., & Crane, A. (2020). Formative Perspectives on the Relation Between CSR Communication and CSR Practices: Pathways for Walking, Talking, and T(w)alking. *Business & Society*, *59*(1), 5–33. https://doi.org/10.1177/0007650319845091

Schweiger, S., Oeberst, A., & Cress, U. (2014). Confirmation bias in web-based search: A randomized online study on the effects of expert information and social tags on information search and evaluation. *Journal of Medical Internet Research*, *16*(3). https://doi.org/10.2196/jmir.3044

Scott, W. R. (2007). *Institutions and Organizations: Ideas and Interests* (3rd edition). SAGE Publications, Inc.

Seeman, J., Si, Y., & Reiter, J. P. (2024). *Differentially Private Finite Population Estimation via Survey Weight Regularization*. arXiv: 2411.04236 [cs]. https://doi.org/10.48550/arXiv.2411.04236

Seeman, J., Slavkovic, A., & Reimherr, M. (2020). Private Posterior Inference Consistent with Public Information: A Case Study in Small Area Estimation from Synthetic Census Data. In J. Domingo-Ferrer & K. Muralidhar (Eds.), *Privacy in Statistical Databases*

(pp. 323–336). Springer International Publishing. https://doi.org/10.1007/978-3-030-57521-2_23

Seeman, J., & Susser, D. (2023). Between Privacy and Utility: On Differential Privacy in Theory and Practice. *ACM Journal on Responsible Computing*. https://doi.org/10.1145/3626494

Seidel, S., Recker, J., & vom Brocke, J. (2013). Sensemaking and Sustainable Practicing: Functional Affordances of Information Systems in Green Transformations. *MIS Quarterly*, *37*(4), 1275–1299. https://www.jstor.org/stable/43825792

Selbst, A. (2021). An Institutional View Of Algorithmic Impact Assessments. *Harvard Journal of Law & Technology*, *35*(117). https://papers.ssrn.com/abstract=3867634

Selbst, A., Venkatasubramanian, S., & Kumar, I. E. (2023). Deconstructing Design Decisions: Why Courts Must Interrogate Machine Learning and Other Technologies. *Ohio State Law Journal*, *85*, forthcoming. https://papers.ssrn.com/abstract=4564304

Selbst, A. D., Boyd, D., Friedler, S. A., Venkatasubramanian, S., & Vertesi, J. (2019). Fairness and Abstraction in Sociotechnical Systems. *Proceedings of the Conference on Fairness, Accountability, and Transparency*, 59–68. https://doi.org/10.1145/3287560.3287598

Selenium. (2023). Selenium. https://www.selenium.dev/

Sheard, N. (2021). *Banning Government Use of Face Recognition Technology: 2020 Year in Review*. Electronic Frontier Foundation. https://www.eff.org/deeplinks/2020/12/banning-government-use-face-recognition-technology-2020-year-review

Shelby, R., Rismani, S., Henne, K., Moon, Aj., Rostamzadeh, N., Nicholas, P., Yilla-Akbari, N., Gallegos, J., Smart, A., Garcia, E., & Virk, G. (2023). Sociotechnical Harms of Algorithmic Systems: Scoping a Taxonomy for Harm Reduction. *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, 723–741. https://doi.org/10.1145/3600211.3604673

Shen, H., Cabrera, Á. A., Perer, A., & Hong, J. (2022). *"Public(s)-in-the-Loop": Facilitating Deliberation of Algorithmic Decisions in Contentious Public Policy Domains* (1). arXiv: 2204.10814 [cs]. https://doi.org/10.48550/arXiv.2204.10814

Shen, H., DeVos, A., Eslami, M., & Holstein, K. (2021). Everyday algorithm auditing: Understanding the power of everyday users in surfacing harmful algorithmic behaviors. *Proceedings of the ACM on Human-Computer Interaction*, *5*, 1–29.

Sherwin, G., & Bhandari, E. (2019). *Facebook Settles Civil Rights Cases by Making Sweeping Changes to Its Online Ad Platform — ACLU of Northern CA*. ACLU Northern California. https://www.aclunc.org/blog/facebook-settles-civil-rights-cases-making-sweeping-changes-its-online-ad-platform

Shoemate, M., Vyrros, A., McCallum, C., Prasad, R., Durbin, P., Casacuberta Puig, S., Cowan, E., Xu, V., Ratliff, Z., Berrios, N., Whitworth, A., Eliot, M., Lebeda, C., Renard, O., & McKay Bowen, C. (2025). *OpenDP Library* (Version 0.12.1). https://github.com/opendp/opendp

Silberzahn, R., Uhlmann, E. L., Martin, D. P., Anselmi, P., Aust, F., Awtrey, E., Bahník, Š., Bai, F., Bannard, C., Bonnier, E., Carlsson, R., Cheung, F., Christensen, G., Clay, R., Craig, M. A., Dalla Rosa, A., Dam, L., Evans, M. H., Flores Cervantes, I., … Nosek, B. A. (2018). Many Analysts, One Data Set: Making Transparent How Variations in Analytic Choices Affect Results. *Advances in Methods and Practices in Psychological Science*, *1*(3), 337–356. https://doi.org/10.1177/2515245917747646

Silva, L., & Hirschheim, R. (2007). Fighting against Windmills: Strategic Information Systems and Organizational Deep Structures. *MIS Quarterly*, *31*(2), 327–354. https://doi.org/10.2307/25148794

Skinner, R. R., & Cooper, C. G. (2020). *FY2019 State Grants Under Title I-A of the Elementary and Secondary Education Act (ESEA)* (R46269). Congressional Research Service.

Skinner-Thompson, S. (2020). *Privacy at the Margins*. Cambridge University Press. https://doi.org/10.1017/9781316850350

Sloane, M., Moss, E., Awomolo, O., & Forlano, L. (2022). Participation Is not a Design Fix for Machine Learning. *Proceedings of the 2nd ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization*, 1–6. https://doi.org/10.1145/3551624.3555285

Sloane, M., Moss, E., & Chowdhury, R. (2022). A Silicon Valley love triangle: Hiring algorithms, pseudo-science, and the quest for auditability. *Patterns (New York, N.Y.)*, *3*(2), 100425. https://doi.org/10.1016/j.patter.2021.100425

Smart, M. A., Sood, D., & Vaccaro, K. (2022). Understanding Risks of Privacy Theater with Differential Privacy. *Proc. ACM Hum.-Comput. Interact.*, *6*, 342:1–342:24. https://doi.org/10.1145/3555762

Smith-Renner, A., Fan, R., Birchfield, M., Wu, T., Boyd-Graber, J., Weld, D. S., & Findlater, L. (2020). No Explainability without Accountability: An Empirical Study of Explanations and Feedback in Interactive ML. *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, 1–13. https://doi.org/10.1145/3313831.3376624

Snyder, T., Dinkes, R., Sonnenberg, W., & Cornman, S. (2019). *Study of the Title I, Part A Grant Program Mathematical Formulas Statistical Analysis Report* (Statistical Analysis Report No. 2019–016). U.S. Department of Education.

Soghoian, C. (2011). An End to Privacy Theater: Exposing and Discouraging Corporate Disclosure of User Data to the Government. *Minnesota Journal of Law, Science & Technology*, *12*(1), 191–238. https://heinonline.org/HOL/P?h=hein.journals/mipr12&i=193

Solinger, O. N., Jansen, P. G., & Cornelissen, J. P. (2020). The Emergence of Moral Leadership. *Academy of Management Review*, *45*(3), 504–527. https://doi.org/10.5465/amr.2016.0263

Sonnenberg, W. (2016). *Allocating Grants for Title I* (National Center for Education Statistics). U.S. Department of Education.

Spencer, B. D. (1982). Technical Issues in Allocation Formula Design. *Public Administration Review*, *42*(6), 524–529. https://doi.org/10.2307/976122

Spencer, B. D. (1985). Statistical Aspects of Equitable Apportionment. *Journal of the American Statistical Association*, *80*(392), 815–822. https://doi.org/10.2307/2288538

Spinks, C. N. (2020). Contemporary Housing Discrimination: Facebook, Targeted Advertising, and the Fair Housing Act. *Houston Law Review*, *57*(4), 925–952. https://houstonlawreview.org/article/12762-contemporary-housing-discrimination-facebook-targeted-advertising-and-the-fair-housing-act

Star, S. L. (1999). The Ethnography of Infrastructure. *American Behavioral Scientist*, *43*(3), 377–391. https://doi.org/10.1177/00027649921955326

Stark, L., & Hutson, J. (2021). Physiognomic artificial intelligence. *Fordham Intell. Prop. Media & Ent. LJ*, *32*, 922.

Steed, R., & Acquisti, A. (2025). *Algorithmic Decoupling and the Adoption of 'Privacy-Preserving' Analytics*. https://doi.org/10.2139/ssrn.4718865

Steed, R., & Caliskan, A. (2021). Image Representations Learned With Unsupervised Pre-Training Contain Human-like Biases. *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 701–713. https://doi.org/10.1145/3442188.3445932

Steed, R., Liu, T., Wu, Z. S., & Acquisti, A. (2022). Policy impacts of statistical uncertainty and privacy. *Science*, *377*(6609), 928–931. https://doi.org/10.1126/science.abq4481

Steed, R., Mustri, E. A. S., & Acquisti, A. (2025). *Impacts of Data Error and Differential Privacy on Findings from Social Science*.

Steed, R., Panda, S., Kobren, A., & Wick, M. (2022). Upstream Mitigation Is Not All You Need: Testing the Bias Transfer Hypothesis in Pre-Trained Language Models. *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*, 3524–3542. https://aclanthology.org/2022.acl-long.247

Steed, R., Qing, D., & Wu, Z. S. (2024). *Quantifying Privacy Risks of Public Statistics to Residents of Subsidized Housing*. arXiv: 2407.04776 [cs]. https://doi.org/10.48550/arXiv.2407.04776

Stein, C. (1956). Inadmissibility of the Usual Estimator for the Mean of a Multivariate Normal Distribution. In *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Contributions to the Theory of Statistics* (pp. 197–207, Vol. 3.1). University of California Press. https://projecteuclid.org/ebooks/berkeley-symposium-on-mathematical-statistics-and-probability/Proceedings-of-the-Third-Berkeley-Symposium-on-Mathematical-Statistics-and/chapter/Inadmissibility-of-the-Usual-Estimator-for-the-Mean-of-a/bsmsp/1200501656

Steinke, T. (2024). Tight RDP & zCDP bounds from pure DP.

Stevens, J. M., Steensma, H. K., Harrison, D. A., & Cochran, P. L. (2005). Symbolic or substantive document? The influence of ethics codes on financial executives' decisions. *Strategic Management Journal*, *26*(2), 181–195. https://ideas.repec.org//a/bla/stratm/v26y2005i2p181-195.html

Stop Discrimination by Algorithms Act of 2021 (2021). https://lims.dccouncil.gov/Legislation/B24-0558

Stop LAPD Spying Coalition & Free Radicals. (2020). *The Algorithmic Ecology: An Abolitionist Tool for Organizing Against Algorithms*. Free Radicals. https://freerads.org/2020/03/02/the-algorithmic-ecology-an-abolitionist-tool-for-organizing-against-algorithms/

Strong & Volkoff. (2010). Understanding Organization—Enterprise System Fit: A Path to Theorizing the Information Technology Artifact. *MIS Quarterly*, *34*(4), 731. https://doi.org/10.2307/25750703

Suchman, M. C. (1995). Managing Legitimacy: Strategic and Institutional Approaches. *The Academy of Management Review*, *20*(3), 571–610. https://doi.org/10.2307/258788

Sun, C., Shrivastava, A., Singh, S., & Gupta, A. (2017). Revisiting unreasonable effectiveness of data in deep learning era. *Proceedings of the IEEE International Conference on Computer Vision*, 843–852.

Sutton, J. R., & Dobbin, F. (1996). The Two Faces of Governance: Responses to Legal Uncertainty in U.S. Firms, 1955 to 1985. *American Sociological Review*, *61*(5), 794–811. https://doi.org/10.2307/2096454

Sweeney, L. (1997). Weaving technology and policy together to maintain confidentiality (vol 25, pg 2, 1997). *Journal Of Law Medicine & Ethics*, *25*(4), 327.

Sweeney, L. (2002). K-anonymity: A model for protecting privacy. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, *10*(5), 557–570. https://doi.org/10.1142/S0218488502001648

Sweeney, L., Abu, A., & Winn, J. (2013). *Identifying Participants in the Personal Genome Project by Name (A Re-identification Experiment)*. arXiv: 1304.7605 [cs]. https://doi.org/10.48550/arXiv.1304.7605

Tabassi, E. (2023). *AI Risk Management Framework: AI RMF (1.0)* (error: NIST AI 100-1). National Institute of Standards and Technology. Gaithersburg, MD. https://doi.org/10.6028/NIST.AI.100-1

Tahaei, M., Frik, A., & Vaniea, K. (2021). Privacy Champions in Software Teams: Understanding Their Motivations, Strategies, and Challenges. *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, 1–15. https://doi.org/10.1145/3411764.3445768

Tan, C., Sun, F., Kong, T., Zhang, W., Yang, C., & Liu, C. (2018). A survey on deep transfer learning. *International Conference on Artificial Neural Networks*, 270–279.

Tan, Y. C., & Celis, L. E. (2019). Assessing social and intersectional biases in contextualized word representations. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d Alché-Buc, E. Fox, & R. Garnett (Eds.), *Advances in neural information processing systems* (pp. 13230–13241, Vol. 32). Curran Associates, Inc. https://proceedings.neurips.cc/paper/2019/file/201d546992726352471cfea6b0df0a48-Paper.pdf

Tang, J., Korolova, A., Bai, X., Wang, X., & Wang, X. (2017). *Privacy Loss in Apple's Implementation of Differential Privacy on MacOS 10.12*. arXiv: 1709.02753 [cs]. https://doi.org/10.48550/arXiv.1709.02753

Tenney, I., Xia, P., Chen, B., Wang, A., Poliak, A., McCoy, R. T., Kim, N., Van Durme, B., Bowman, S. R., Das, D., et al. (2019). *What do you learn from context? probing for sentence structure in contextualized word representations*.

Terzis, P., Veale, M., & Gaumann, N. (2024). Law and the Emerging Political Economy of Algorithmic Audits. *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, 1255–1267. https://doi.org/10.1145/3630106.3658970

The Markup. (2022). *Citizen Browser*. The Markup. https://themarkup.org/series/citizen-browser

Tolbert, P. S., & Zucker, L. G. (1983). Institutional Sources of Change in the Formal Structure of Organizations: The Diffusion of Civil Service Reform, 1880-1935. *Administrative Science Quarterly*, *28*(1), 22–39. https://doi.org/10.2307/2392383

Topalova, P. (2010). Factor Immobility and Regional Impacts of Trade Liberalization: Evidence on Poverty from India. *American Economic Journal: Applied Economics*, *2*(4), 1–41. https://doi.org/10.1257/app.2.4.1

Turco, C. (2012). Difficult Decoupling: Employee Resistance to the Commercialization of Personal Settings. *American Journal of Sociology*, *118*(2), 380–419. https://doi.org/10.1086/666505

Turri, V., & Dzombak, R. (2023). Why We Need to Know More: Exploring the State of AI Incident Documentation Practices. *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, 576–583. https://doi.org/10.1145/3600211.3604700

Twitter. (2021). *Twitter Algorithmic Bias - Bug Bounty Program*. HackerOne. https://hackerone.com/twitter-algorithmic-bias

Ujifusa, A. (2019). What Each State Will Get in Federal Title I Grants for Disadvantaged Kids Next Year [newspaper]. *Education Week: Budget & Finance*. https://www.edweek.org/leadership/what-each-state-will-get-in-federal-title-i-grants-for-disadvantaged-kids-next-year/2019/04

Urman, A., Smirnov, I., & Lasser, J. (2024). The right to audit and power asymmetries in algorithm auditing. *EPJ Data Science*, *13*(1), 1–15. https://doi.org/10.1140/epjds/s13688-024-00454-5

U.S. Census Bureau. (2013). U.S. Census Bureau Statistical Quality Standards. https://www2.census.gov/about/policies/quality/quality-standards-jul2013.pdf

U.S. Census Bureau. (2018). Calculating Measures of Error for Derived Estimates. In *ACS General Handbook*. U.S. Government Printing Office, https://www.census.gov/content/dam/Census/library/publications/2018/acs/acs_general_handbook_2018.pdf

US Census Bureau. (2020). *Quantifying Relative Error in the School District Estimates*. The United States Census Bureau. https://www.census.gov/programs-surveys/saipe/guidance/district-estimates.html

U.S. Census Bureau. (2021). *2021 TIGER/Line Shapefiles (machine- readable data files)*. https://www2.census.gov/geo/pdfs/maps-data/data/tiger/tgrshp2021/TGRSHP2021_TechDoc.pdf

U.S. Census Bureau. (2022). *Disclosure Avoidance Protections for the American Community Survey*. Census Blogs. https://www.census.gov/newsroom/blogs/random-samplings/2022/12/disclosure-avoidance-protections-acs.html

U.S. Department of Health and Human Services. (2012). *Guidance on De-identification of Protected Health Information*. https://www.hhs.gov/sites/default/files/ocr/privacy/hipaa/understanding/coveredentities/De-identification/hhs_deid_guidance.pdf

U.S. Equal Employment Opportunity Commission. (2016). *Diversity in High Tech* (Special Report). https://www.eeoc.gov/special-report/diversity-high-tech

van der Klaauw, W. (2008). Breaking the link between poverty and low student achievement: An evaluation of Title I. *Journal of Econometrics*, *142*(2), 731–756. https://doi.org/10.1016/j.jeconom.2007.05.007

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł. ukasz, & Polosukhin, I. (2017). Attention is all you need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, & R. Garnett (Eds.), *Advances in neural information processing systems* (pp. 5998–6008, Vol. 30). Curran Associates, Inc. https://proceedings.neurips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf

Veale, M. (2023). *Rights for Those Who Unwillingly, Unknowingly and Unidentifiably Compute!* https://doi.org/10.31235/osf.io/4ugxd
To appear in: Hans-Wolfgang Micklitz and Giussepe Vettori (eds.), The Person and the Future of Private Law (Hart, forthcoming).

Vecchione, B., Levy, K., & Barocas, S. (2021). Algorithmic Auditing and Social Justice: Lessons from the History of Audit Studies. *Proceedings of the 1st ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization*, 1–9. https://doi.org/10.1145/3465416.3483294

Verbeek, P.-P. (2006). Materializing Morality: Design Ethics and Technological Mediation. *Science, Technology, & Human Values*, *31*(3), 361–380. https://doi.org/10.1177/0162243905285847

Viljoen, S. (2021). A Relational Theory of Data Governance. *Yale Law Journal*, *131*(2), 573–655. https://go.gale.com/ps/i.do?p=AONE&sw=w&issn=00440094&v=2.1&it=r&id=GALE%7CA690123702&sid=googleScholar&linkaccess=abs

Voita, E., Sennrich, R., & Titov, I. (2019). *The bottom-up evolution of representations in the transformer: A study with machine translation and language modeling objectives*.

Volkoff, O., Strong, D. M., & Elmes, M. B. (2007). Technological Embeddedness and Organizational Change. *Organization Science*, *18*(5), 832–848. https://doi.org/10.1287/orsc.1070.0288

Wachter, S., Mittelstadt, B., & Russell, C. (2021). Bias Preservation in Machine Learning: The Legality of Fairness Metrics Under EU Non-Discrimination Law. *West Virginia Law Review*, *123*(3), 735. https://researchrepository.wvu.edu/wvlr/vol123/iss3/4

Waldman, A. (2024). How UnitedHealth's Playbook for Limiting Mental Health Coverage Puts Countless Americans' Treatment at Risk [magazine]. *ProPublica*. https://www.propublica.org/article/unitedhealth-mental-health-care-denied-illegal-algorithm

Waldman, A. E. (2018). Designing Without Privacy. *Houston Law Review*, *55*(3). https://houstonlawreview.org/article/3880-designing-without-privacy

Wang, A., Narayanan, A., & Russakovsky, O. (2020). REVISE: A tool for measuring and mitigating bias in visual datasets. *European Conference on Computer Vision*.

Wang, A., Kapoor, S., Barocas, S., & Narayanan, A. (2023). Against Predictive Optimization: On the Legitimacy of Decision-Making Algorithms that Optimize Predictive Accuracy. *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, 626. https://doi.org/10.1145/3593013.3594030

Wang, D., Yang, Q., Abdul, A., & Lim, B. Y. (2019). Designing Theory-Driven User-Centric Explainable AI. *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, 1–15. https://doi.org/10.1145/3290605.3300831

Wang, T., Zhao, J., Yatskar, M., Chang, K.-W., & Ordonez, V. (2019). Balanced datasets are not enough: Estimating and mitigating gender bias in deep image representations. *Proceedings of the IEEE International Conference on Computer Vision*, 5310–5319.

Wang, Z., Qinami, K., Karakozis, I. C., Genova, K., Nair, P., Hata, K., & Russakovsky, O. (2020). Towards fairness in visual recognition: Effective strategies for bias mitigation. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8919–8928.

Warkentin, T., & Woodward, J. (2022). *Join us in the AI Test Kitchen*. Google. https://blog.google/technology/ai/join-us-in-the-ai-test-kitchen/

Watkins, E. A., Moss, E., Metcalf, J., Singh, R., & Elish, M. C. (2021). Governing Algorithmic Systems with Impact Assessments: Six Observations. *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, 1010–1022. https://doi.org/10.1145/3461702.3462580

Weaver, G. R., Treviño, L. K., & Cochran, P. L. (1999). Integrated and Decoupled Corporate Social Performance: Management Commitments, External Pressures, and Corporate Ethics Practices. *The Academy of Management Journal*, *42*(5), 539–552. https://doi.org/10.2307/256975

Weber, M. (1978). *Economy and Society: An Outline of Interpretive Sociology*. University of California Press.

Weerts, H., Dudík, M., Edgar, R., Jalali, A., Lutz, R., & Madaio, M. (2024). Fairlearn: Assessing and improving fairness of AI systems. *J. Mach. Learn. Res.*, *24*(1), 257:12058–257:12065.

Weick, K. E. (1976). Educational Organizations as Loosely Coupled Systems. *Administrative Science Quarterly*, *21*(1), 1–19. https://doi.org/10.2307/2391875

West, J., & Gallagher, S. (2006). Challenges of open innovation: The paradox of firm investment in open-source software. *R&D Management*, *36*(3), 319–331. https://doi.org/10.1111/j.1467-9310.2006.00436.x

Westphal, J. D., & Zajac, E. J. (2001). Decoupling Policy from Practice: The Case of Stock Repurchase Programs. *Administrative Science Quarterly*, *46*(2), 202–228. https://doi.org/10.2307/2667086

Widder, D. G., Whittaker, M., & West, S. M. (2024). Why 'open' AI systems are actually closed, and why this matters. *Nature*, *635*(8040), 827–833. https://doi.org/10.1038/s41586-024-08141-1

Widder, D. G., Zhen, D., Dabbish, L., & Herbsleb, J. (2023). It's about power: What ethical concerns do software engineers have, and what do they (feel they can) do about them? *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, 467–479. https://doi.org/10.1145/3593013.3594012

Wiesche, M., Jurisch, M., Yetton, P., & Krcmar, H. (2017). Grounded Theory Methodology in Information Systems Research. *MIS Quarterly*, *41*(3), 685–701. https://aisel.aisnet.org/misq/vol41/iss3/4

Wijen, F. (2014). Means versus Ends in Opaque Institutional Fields: Trading off Compliance and Achievement in Sustainability Standard Adoption. *Academy of Management Review*, *39*(3), 302–323. https://doi.org/10.5465/amr.2012.0218

Williams, A. R., Barrientos, A. F., Snoke, J., & Bowen, C. M. (2024). *Benchmarking DP Linear Regression Methods for Statistical Inference*.

Williams, A. R., Snoke, J., Bowen, C. M., & Barrientos, A. F. (2024). Disclosing Economists' Privacy Perspectives: A Survey of American Economic Association Members' Views on Differential Privacy and the Usability of Noise-Infused Data. *Harvard Data Science Review*. https://doi.org/10.1162/99608f92.a8fb0371

Willis, G. B., & Artino, A. R. (2013). What Do Our Respondents Think We're Asking? Using Cognitive Interviewing to Improve Medical Education Surveys. *Journal of Graduate Medical Education*, *5*(3), 353–356. https://doi.org/10.4300/JGME-D-13-00154.1

Wilson, B., Hoffman, J., & Morgenstern, J. (2019). *Predictive inequity in object detection*. http://arxiv.org/abs/1902.11097

Wilson, D. J. (2012). Fiscal Spending Jobs Multipliers: Evidence from the 2009 American Recovery and Reinvestment Act. *American Economic Journal: Economic Policy*, *4*(3), 251–282. https://doi.org/10.1257/pol.4.3.251

Winner, L. (1980). Do Artifacts Have Politics? *Daedalus*, *109*(1), 121–136. https://www.jstor.org/stable/20024652

Wolcott, H. F. (1994). *Transforming Qualitative Data: Description, Analysis, and Interpretation*. SAGE.

Wong, R. Y., Madaio, M. A., & Merrill, N. (2023). Seeing Like a Toolkit: How Toolkits Envision the Work of AI Ethics. *Proceedings of the ACM on Human-Computer Interaction*, *7*, 145:1–145:27. https://doi.org/10.1145/3579621

Wood, S. N. (2011). Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *73*(1), 3–36. https://doi.org/10.1111/j.1467-9868.2010.00749.x

Woodruff, A., Fox, S. E., Rousso-Schindler, S., & Warshaw, J. (2018). A Qualitative Exploration of Perceptions of Algorithmic Fairness. *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, 1–14. https://doi.org/10.1145/3173574.3174230

Wright, L., Muenster, R. M., Vecchione, B., Qu, T., Cai, P. (, Smith, A., Investigators, C. 2. S., Metcalf, J., & Matias, J. N. (2024). Null Compliance: NYC Local Law 144 and the challenges of algorithm accountability. *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, 1701–1713. https://doi.org/10.1145/3630106.3658998

Wu, T. (2003). When Code Isn't Law. *Va. L. Rev.*, *89*, 679. https://scholarship.law.columbia.edu/faculty_scholarship/844

Xiang, A., & Raji, I. D. (2019). *On the Legal Compatibility of Fairness Definitions*. arXiv: 1912.00761 [cs, stat]. https://doi.org/10.48550/arXiv.1912.00761

Xu, H., & Zhang, N. (2021). Implications of Data Anonymization on the Statistical Evidence of Disparity. *Management Science*. https://doi.org/10.1287/mnsc.2021.4028

Xu, K., Nosek, B., & Greenwald, A. (2014). Data from the race implicit association test on the project implicit demo website. *Journal of Open Psychology Data*, *2*(1), e3. https://doi.org/10.5334/jopd.ac

Yang, K., Qinami, K., Fei-Fei, L., Deng, J., & Russakovsky, O. (2020). Towards fairer datasets: Filtering and balancing the distribution of the people subtree in the ImageNet hierarchy. *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 547–558. https://doi.org/10.1145/3351095.3375709

Yew, R.-J., Qin, L., & Venkatasubramanian, S. (2024). You Still See Me: How Data Protection Supports the Architecture of AI Surveillance. *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, *7*, 1709–1722. https://doi.org/10.1609/aies.v7i1.31759

Yoo, Y., Henfridsson, O., & Lyytinen, K. (2010). Research Commentary: The New Organizing Logic of Digital Innovation: An Agenda for Information Systems Research. *Information Systems Research*, *21*(4), 724–735. https://www.jstor.org/stable/23015640

Zaslavsky, A. M., & Schirm, A. L. (2002). Interactions Between Survey Estimates and Federal Funding Formulas. *Journal of Official Statistics*, *18*(3), 371.

Zhao, J., Wang, T., Yatskar, M., Ordonez, V., & Chang, K. W. (2017). Men also like shopping: Reducing gender bias amplification using corpus-level constraints. *EMNLP 2017 - Conference on Empirical Methods in Natural Language Processing, Proceedings*, 2979–2989. https://doi.org/10.18653/v1/d17-1323

Zuboff, S. (2019). *The Age of Surveillance Capitalism: The Fight for a Human Future at the New Frontier of Power* (1st edition). PublicAffairs.

Appendices

Appendix A

Estimating Policy Impacts of Statistical Uncertainty and Privacy

A.1 Additional Figures

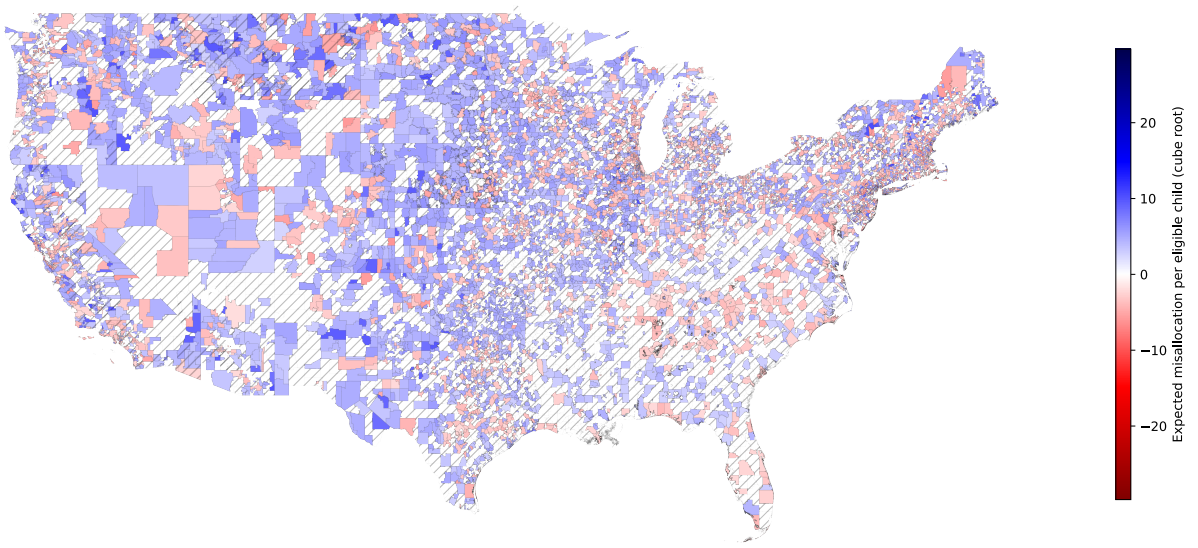
Figure A1 geographically depicts the misallocation per eligible child due to underlying data error (data deviations) and additionally injected noise (privacy deviations) at a relatively high level of privacy ($\epsilon = 0.1$). Figure A2 shows the likelihood that a district in Pennsylvania changes eligibility for any of the Title I grant types. More districts have a higher chance to *lose* eligibility than to *gain* eligibility, as most districts' poverty counts lie above the Title I thresholds. Figure A3 presents the baseline disparities in entitlements without any additional policy features. Figure A4 depicts the fitted smooth for each covariate in the relatively strong privacy setting $\epsilon = 0.1$ where demographic patterns are most visible. Note that adding post-formula provisions noticeably sharpens the effects of district median income, racial make-up, and population density on misallocation (Figure A15). Smooths and other figures for all of our experiments can be accessed at github.com/ryansteed/ieat.

A.2 Materials and Methods

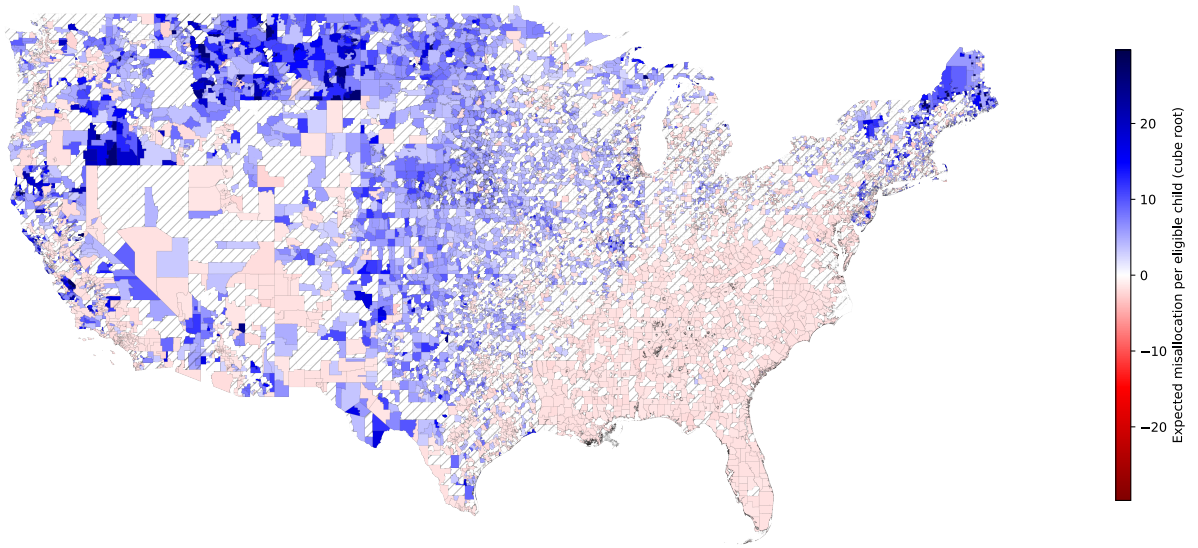
Data

We measure the impact of data and privacy deviations on Fiscal Year 2021 Title I allocations to over 13,190 local education agencies across the United States. We focus on three out of the four grant types: basic, concentration, and targeted grants. (The fourth type of grant, which accounted for 23% of Title I funds in 2015, is distributed using state-level multipliers that are not publicly available.) We calculate these allocations using the same data sources as the Department of Education, most notably the Census Bureau's Small Area Income and Poverty Estimates (SAIPE) from 2019 (Luery, 2010; Bell & Robinson, 2020), a table of counts of total population, children, and eligible children in 13,184 school districts from all fifty states.[1]

[1] Our results do not include Puerto Rico and other U.S. territories—they are excluded from the SAIPE.



(a) Expected misallocation $\mathbb{E}[y_i(x) - y_i(\mu)]$ per eligible child to school districts due to data deviations alone. Notably, Northwestern, population-sparse districts tend to benefit from uncertainty while populous Southeastern districts lose out.



(b) Expected marginal change in misallocation $\mathbb{E}[y_i(\tilde{x}) - y_i(x)]$ per eligible child after differential privacy is applied (due to both data and privacy deviations). Though gains appear to dominate the map, the districts that benefit most (and have the largest areas) tend to have small populations. Instead, losses per child in dense, populous districts increase slightly to pay for large gains per child in sparse, less-populated districts.

Figure A1: Cube root of misallocation (observed minus official) in dollars per eligible child in the continental U.S. (cube root), averaged over 1,000 trials. Blue school districts gain funding under deviations; red districts lose funding. Injected noise is drawn from Laplace mechanism with $\epsilon = 0.1$. Striped districts have mean misallocations not significantly different from zero ($p < 0.1$) using a one-sample, two-tailed z-test.

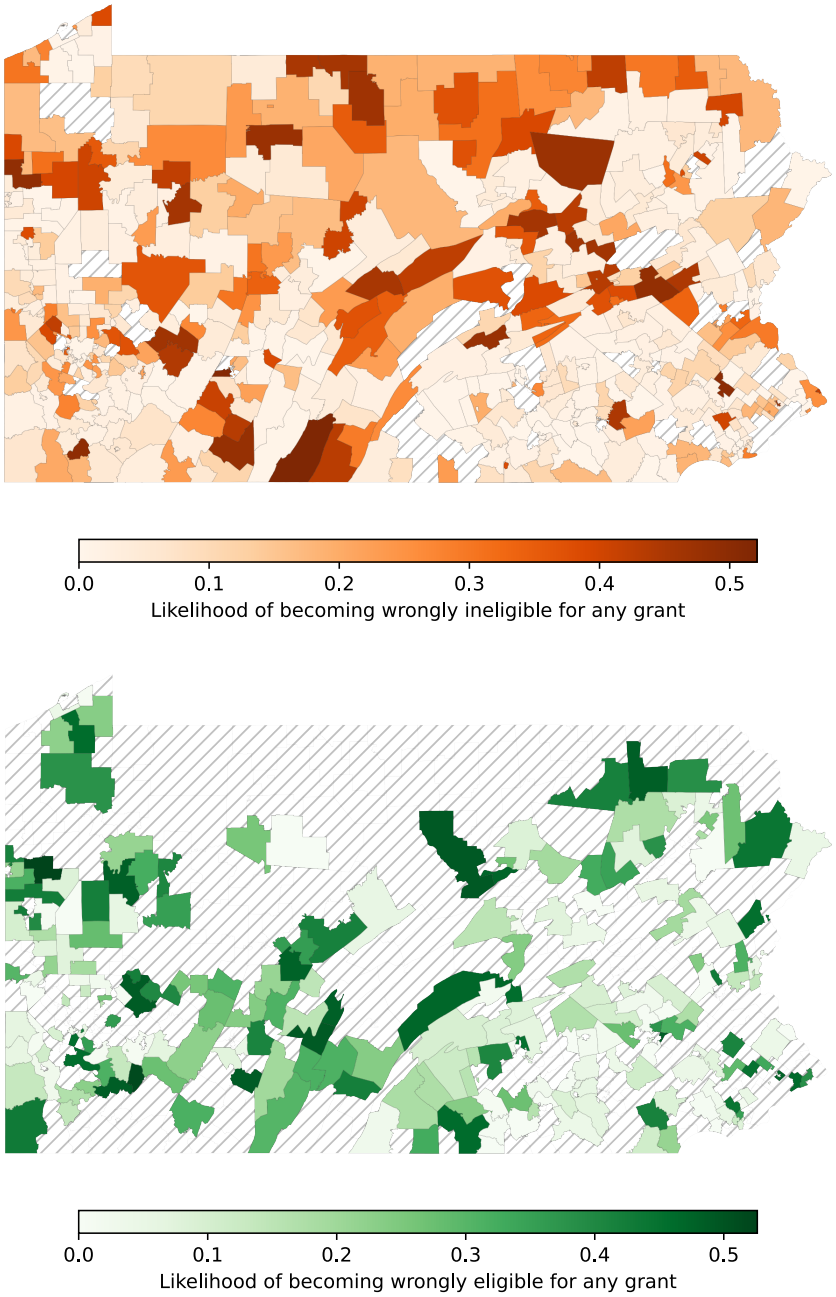


Figure A2: Likelihood of changing eligibility due to data deviations alone in Pennsylvania, computed over 1,000 trials. Striped districts have proportions not significantly different from zero ($p \geq 0.1$) using a one-sample z-test.

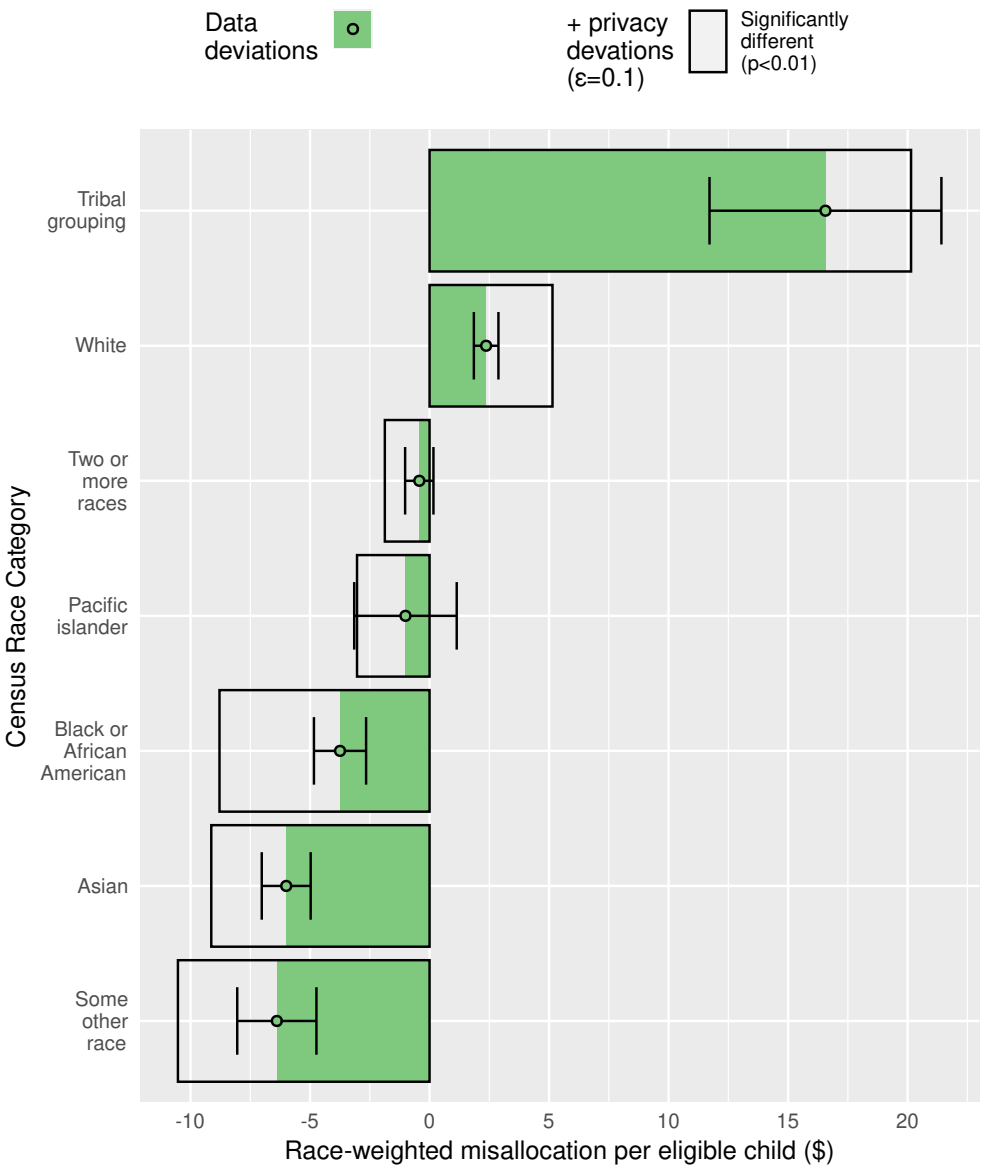


Figure A3: Bars depict the sum of misallocation multiplied by the proportion of respondents of each census single race category, divided by the sum of formula-eligible children of that race. Averaged over 1,000 trials. A black outline indicates the marginal change in misallocation due to injected noise, drawn from Laplace mechanism with $\epsilon = 0.1$. Error bars span a 90% normal confidence interval. The differences between race-weighted misallocations before and after privacy deviations are added are significantly different ($p < 0.01$) for all groups, according to a two-sample z-test.

The Census Bureau produces these estimates using a smoothed version of the most recent ACS (for SAIPE 2019, the 2018 ACS) and income tax data from the IRS (Bell et al., 2007; Powers et al., 2008). A single child can change the count of total population, children x_i, or children in poverty z_i by ± 1, so the maximum amount the table could change if one child is removed (i.e., the sensitivity for differential privacy) is $\Delta = 2$. We compare our replication of the Title I allocations to the official figures produced by the Department of Education, which also contain counts of eligible children in other categories—juvenile detention, foster homes, and those enrolled in Temporary Assistance for Needy Families (TANF) (Appendix A.8) (Rooney, 2021). For SPPE, we use the most recent education expenditure table published by the National Center for Education Statistics (NCES), from fiscal year 2018 (Cornman et al., 2020).

For testing the distribution of misallocation over demographics, we use 2015-2019 ACS demographic data aggregated at the school district level in the NCES Education Demographic and Geographic Estimates database ("American Community Survey – Education Tabulation (ACS-ED)," 2015–2019). There are 1,545 school districts with missing child-level race data; for those districts, we assume that the child demographic frequencies match the adult frequencies. We drop 14 districts (mostly unorganized territories) which also have no adult estimates. Median incomes are truncated above $250,000. For the regression analysis, we impute the means of 162 districts with missing median incomes and 148 districts with missing household sizes. Population densities are computed using land areas from the published TIGER shapefiles provided by the Census Bureau (U.S. Census Bureau, 2021). Copies of these datasets are included in our codebase at github.com/ryansteed/ieat.

Replicating Title I

Because there is no publicly available code for converting poverty counts into Title I allocations, we replicate the allocation procedure described in detail by the Department of Education (Sonnenberg, 2016; Snyder et al., 2019). Title I funds are intended to assist disadvantaged students, so the formulas for grant allocation are based on the number of eligible children in each school district. Eligible children are school-age children (5-17 years old) who either a) live in families with income at or below the poverty level, b) live in families who receive certain government assistance, c) live in institutions for neglected or delinquent children, or d) live in foster homes (Sonnenberg, 2016; Snyder et al., 2019). Funds are distributed to school districts with eligible children through four primary types of grants (Snyder et al., 2019): *basic* grants, for any LEA that qualifies; *concentration* grants, for LEAs with especially large disadvantaged populations; *targeted* grants, which are distributed according to a weighting system proportional to eligibility counts; and education finance incentive grants, which are state-level grants for state-determined distribution. We examine only the first three types of grants.

To qualify for a grant of any amount, an LEA must have a certain number of eligible students (in magnitude or in proportion of total school-age population); the qualification amounts increase from basic grants to targeted grants (Sonnenberg, 2016; Snyder et al., 2019). For basic grants, an LEA must have at least 10 eligible children *and* more than 2% of children 5-17 must be in poverty. For concentration grants, an LEA must meet the basic grant eligibility

requirements and have more than 6,500 eligible children *or* more than 15% eligible children. For targeted grants, an LEA must have at least 10 eligible children *and* more than 5% eligible children.

Grant amounts are authorized according to the following general formula:

$$\text{Auth. Amt.} = w\{\#\text{ eligible}\} * \text{adjusted SPPE}$$

where SPPE is the state per-pupil expenditure, normalized and clipped to a certain interval set by Title I legislation to truncate the tails. Adjusted SPPE refers to the SPPE provision, which sets lower and upper bounds on the SPPE coefficient for all states (Sonnenberg, 2016). For basic and concentration grants, the weights w are uniform; for targeted grants the weights are a step function of total school district population given in (Sonnenberg, 2016; Snyder et al., 2019). The "authorization" amount is the amount an LEA is eligible to receive; allocation amounts are the amount actually received, depending on the amount of federal funds available and the percentage of a state's per-pupil cost Congress agrees to fund, usually 40% (Sonnenberg, 2016). Authorization amounts are exactly proportional to allocation amounts (Sonnenberg, 2016):

$$\text{Alloc. Amt.}_i = \frac{\text{Auth. Amt.}_i}{\sum_j \text{Auth. Amt.}_j} * \text{Total Federal Appropriation}$$

The final allocation amount is much less than the authorization amount, reduced proportionally (for each grant type) to sum to the 2020 Title I appropriation (Ujifusa, 2019; Rooney, 2021). We think of these allocation amounts before post-formula provisions as *entitlements* which reflect the primary, stated goal of the legislation (to provide financial assistance to schools with poor children), which may differ from the *real* goals indicated by the final allocation amounts when the hold harmless and state minimum provisions are applied (Spencer, 1982).

There are also two special legislative provisions that modify the entitlements after they are calculated. For the majority of our results, we consider only the formula entitlements, before these provisions are applied. The hold harmless provision requires that no district lose more than some percent of its Title I funds from the preceding year, depending on the proportion of children in poverty in the district. The state minimum provision requires that no state receive less than a minimum amount for each of the four grants (Snyder et al., 2019; Skinner & Cooper, 2020). Completely satisfying both of these provisions may require several iterations. For this reason, the Title I formula cannot be expressed in closed form (Spencer, 1982). The full allocation algorithm is implemented in our codebase at github.com/ryansteed/ieat.

It should be noted that our analysis of the Title I allocation process leaves out several elements that could affect the applicability of our findings to the real-world distribution of funds, including small district appeals (20 U.S.C. §6333), district-level heterogeneity in the use and usefulness of funds (Riddle, 2011; Heuer & Stullich, 2011; Borman & D'Agostino, 1996; van der Klaauw, 2008; Murphy, 2012),[2] and temporal trends in funding, which in combination with provisions like hold harmless could compound the effects of deviations

[2]For example, about 70% of participating schools implement school-wide programs that benefit all students, rather than just Title I eligible students (Snyder et al., 2019).

(Zaslavsky & Schirm, 2002). For example, there is evidence that Title I grants generally improve educational outcomes for Title I students (Borman & D'Agostino, 1996), but for high-poverty schools especially, Title I funds may not close education gaps between poorer students and their more privileged peers (van der Klaauw, 2008)—so the educational gains and losses that come from changes in funding may also be disparately distributed.

A.3 Analysis of Variability in Outcomes

In Figure A1, we show the expected misallocation (the average over 1,000 simulation trials) to each school district across the continental U.S. While less populous districts tend to experience an average increase in allocation due to data or privacy deviations, funding to these areas is also much more volatile. Figure A5 shows the 5th percentile misallocation across all 1,000 trials (the maximum amount lost in 95% of trials). Due to data deviations alone, less populous districts in the Midwest and Northwest experience somewhat greater worst-case losses in funding than more populous districts (Figure A5a), despite gaining funding on average (Figure A1a). Similarly, injecting additional noise for privacy affects populous districts much less in the worst case than less populous districts (Figure A5b), though less populous districts gain from privacy deviations on average (Figure A1b).

A.4 Additional Categorical Analysis

We also conducted category-weighted disparity analyses for more detailed race groupings and for the ACS ethnicity question. For readability, Figure 1.2 shows an aggregation of all the single race categories in the ACS; Figure A6 shows the comparison for all of the ACS race categories. Figure A7 shows the same analysis for the ACS ethnicity question (Hispanic or non-Hispanic). Results are mostly stable (with overlapping confidence intervals) within aggregated race groups, with the exception of the Sioux and Cherokee tribal groupings (the Sioux tribal grouping gains by noticeably more than the Cherokee tribal grouping).

We also investigated the possibility that there are race- or ethnicity-based discontinuities around the current eligibility thresholds that may help explain some of the disparities we notice. (For example, if people of color are systematically grouped into districts slightly larger than the thresholds.) Figures A8 and A9 show that most districts do not lie near an eligibility threshold, except for concentration thresholds, for which each district need only lie above one of the two thresholds to receive funds. There are correlations between Whiteness and the number or rate of children in poverty. However, even accounting for this correlation, there is a noticeable dearth of majority-minority districts below the count threshold for basic and targeted grants, although there are many majority-minority districts just above the threshold and many majority-White districts just below. This pattern could also partially explain the racial disparities we document.

A.5 Regression Analysis

Figure A10 shows a simple univariate regression analysis for several covariates. These effects represent the direct distribution of misallocation over each covariate, not accounting for other covariates. Here, we include four covariates not included in Figure A4: size of renter household, citizenship, and the formula inputs, total children and children in poverty. The patterns of distribution match those reported in the multiple regression (Figure A4) but with slightly narrower confidence intervals, except for the proportion of white-only residents, which reports a negative positive effect up to 75% instead of 50%.

Because we are also interested in evaluating to what extent each covariate explains or predicts the typical misallocation to a district after accounting for other covariates (e.g., the effect of racial homogeneity after accounting for population density), we also test several multiple regression specifications. The regression takes the form

$$y_i(\tilde{x}) - y_i(\mu) = \sum_{k=1}^{d} s_k(Z_i) + \epsilon_i$$

where y is the allocation procedure conducted using either the official poverty counts μ_i or the "observed," noise-infused estimates $\tilde{x}_i$. The function s_k is a thin-plate spline and Z_i are the covariates. ϵ_i is the error term.

Table A.1 reports the OLS estimates for three specifications of our regression models: one including only the formula components (population, total child population, and population of children in poverty); another including only demographic components; and a regression including all the covariates. (We exclude renter household from the demographic regression because the effects are minimal, and we exclude the citizenship question because the data contain many outliers.) In our main results, we report variables from the demographic-only specification, because we are interested in analyzing the distribution of misallocation across demographics. When combined, the formula components tend to mediate the coefficients on the demographic variables, with some exceptions.

To estimate non-linear effects, we tried a generalized additive model, which yields a significantly lower sum of squared errors than the OLS specification (Table A.2). We used this model—summarized in Table A.3—for our main results. Where the OLS model explains about 0.02% of variance in misallocations, the GAM explains 0.1% of deviations. (The explanatory power of the model increases as we lower the number of trials; for a single trial, the model usually explains over 2% of deviations in misallocation.) Estimations were conducted using the `mgcv` package in R (Wood, 2011).

We also tested a GAM specification with marginal product smooths to capture interaction between population density, income, and Whiteness (Table A.5), which has slightly lower deviance than the baseline GAM without interactions (Table A.4), suggesting there may be some small interaction effects between racial composition and income.

Finally, we ran a regression on the misallocations due only to data deviations to see if there is a marginal difference in the estimates for just data deviations compared to data and

	Dependent variable:		
	Misallocation (deviated minus true)		
	Formula components	All variables	Demographic variables
	(1)	(2)	(3)
Log population density		−152.343	−3,560.569***
		(680.707)	(344.609)
Racial homogeneity (HHI)		23,446.180**	23,095.220***
		(10,659.330)	(6,147.777)
Proportion White		−29,914.920***	−6,367.596
		(10,849.840)	(6,255.373)
Proportion Hispanic		1,937.204	−14,012.970***
		(5,248.909)	(3,600.673)
Median income		−0.030	0.083***
		(0.036)	(0.024)
		(91.254)	(57.071)
Log total population	4,055.820***	6,689.163***	
	(460.254)	(1,027.212)	
Total # children	−2.507***	−2.518***	
	(0.132)	(0.175)	
Total # children in poverty	−7.536***	−8.230***	
	(0.670)	(0.892)	
		(27.241)	
Avg. renter's household size		8,074.700***	
		(2,181.536)	
Constant	−21,010.610***	−52,451.150***	−50,555.630***
	(4,066.151)	(18,102.130)	(6,291.561)
Observations	1,316,800	841,000	1,308,700
R^2	0.005	0.005	0.0002
Adjusted R^2	0.005	0.005	0.0002
Residual Std. Error	714,287.500 (df = 1316796)	887,137.200 (df = 840988)	718,155.900 (df = 1308693)
F Statistic	2,207.676*** (df = 3; 1316796)	391.687*** (df = 11; 840988)	53.820*** (df = 6; 1308693)
Note:			*p<0.1; **p<0.05; ***p<0.01

Table A.1: OLS estimation of the correlation between demographic and formula covariates and misallocation. Coefficients are in dollars misallocated.

	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
1	130863	7.801×10^{16}				
2	130850	7.798×10^{16}	13	3.254×10^{13}	4.30	0.0000

Table A.2: Analysis-of-deviation F-test of difference between (1) a OLS specification and (2) a GAM specification using the demographic specification. Deviations are in dollars misallocated.

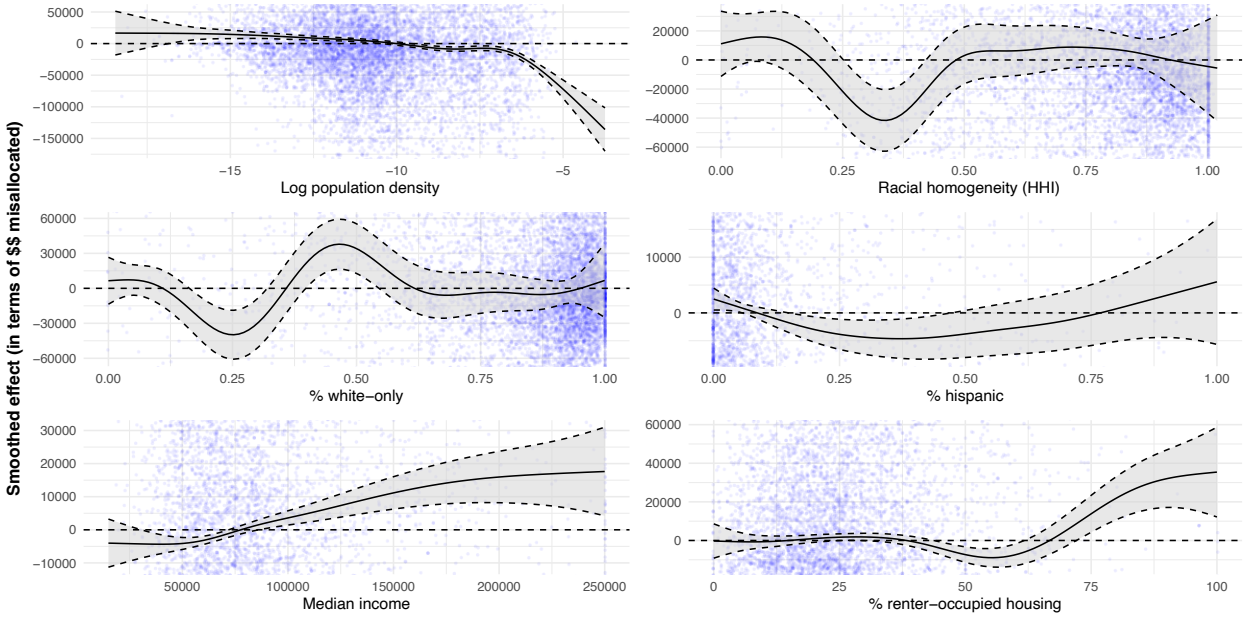
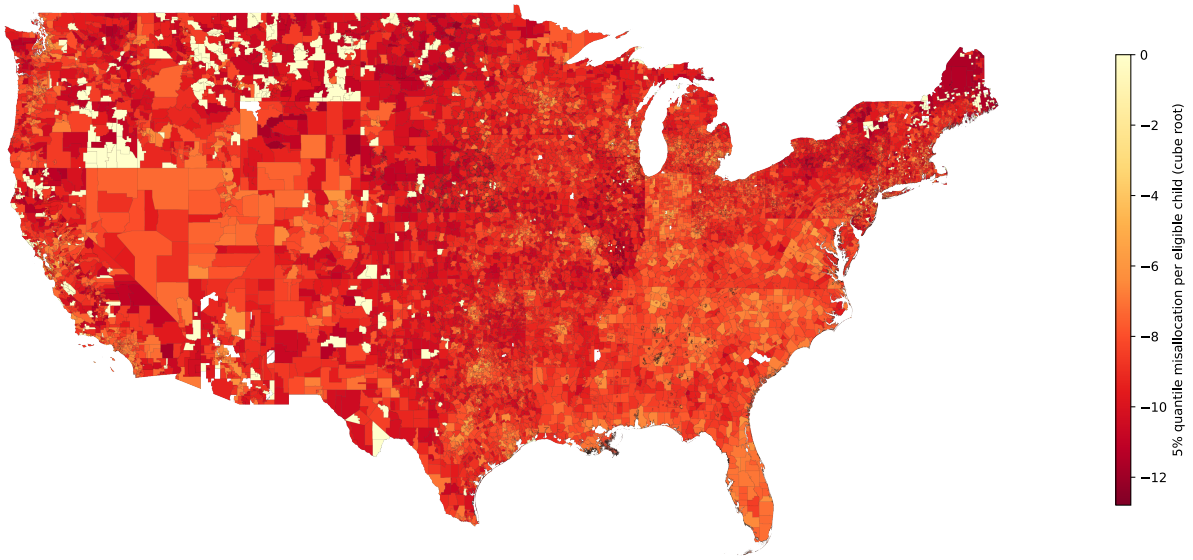


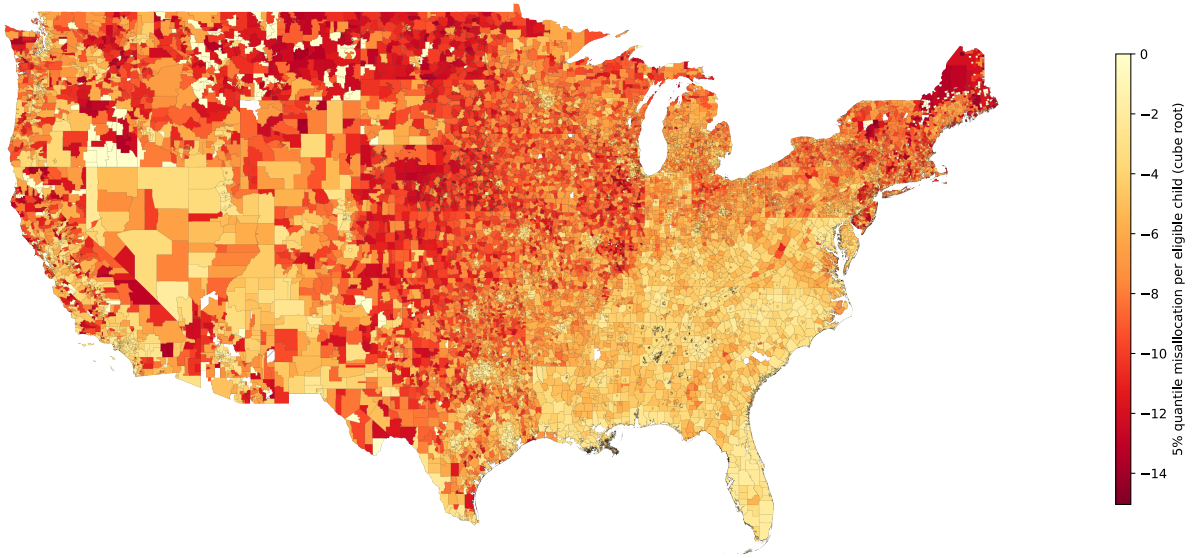
Figure A4: Model-smoothed misallocation (from both data deviations and injected noise) by covariates, with 95% confidence interval in gray, from a multivariate regression. Injected noise is drawn from a Laplace mechanism with $\epsilon = 0.1$. Positive values indicate districts that expect to benefit from combined data and privacy deviations; negative values indicate districts that expect to lose funding because of deviations.

A. parametric coefficients	Estimate	Std. Error	t-value	p-value
(Intercept)	-227.2932	2133.9246	-0.1065	0.9152
B. smooth terms	edf	Ref.df	F-value	p-value
Log population density	5.1214	6.3013	3.6347	0.0009
Racial homogeneity (HHI)	3.7221	4.7235	1.3161	0.3021
Proportion White	4.5520	5.5042	5.0941	0.0011
Proportion Hispanic	1.8194	2.2773	2.6293	0.0658
Median income	1.1889	1.3536	1.2858	0.2714
% households renting	2.2838	2.9248	1.0297	0.4089
R-sq. (adj) = 0.0008, Deviance explained = 0.09%, -REML = 1.9598e+06, Scale est. = 5.9593e+11, n = 130870				

Table A.3: GAM estimation of the correlation between demographic covariates and misallocation. edf stands for effective degrees of freedom. F-values are reported for a joint test of equality to zero across each set of spline coefficients.



(a) 5th percentile of misallocation $y_i(x) - y_i(\mu)$ per eligible child to school districts due to data deviations alone.



(b) 5th percentile of marginal change in misallocation $y_i(\tilde{x}) - y_i(x)$ per eligible child after differential privacy is applied (due to both data and privacy deviations).

Figure A5: Cube root of worst-case (5th percentile) misallocation in dollars per eligible child in the continental U.S. (cube root), averaged over 1,000 trials. Yellow districts lose minimal funding after deviations; red districts lose significant amounts of funding. Injected noise is drawn from a Laplace mechanism with $\epsilon = 0.1$.

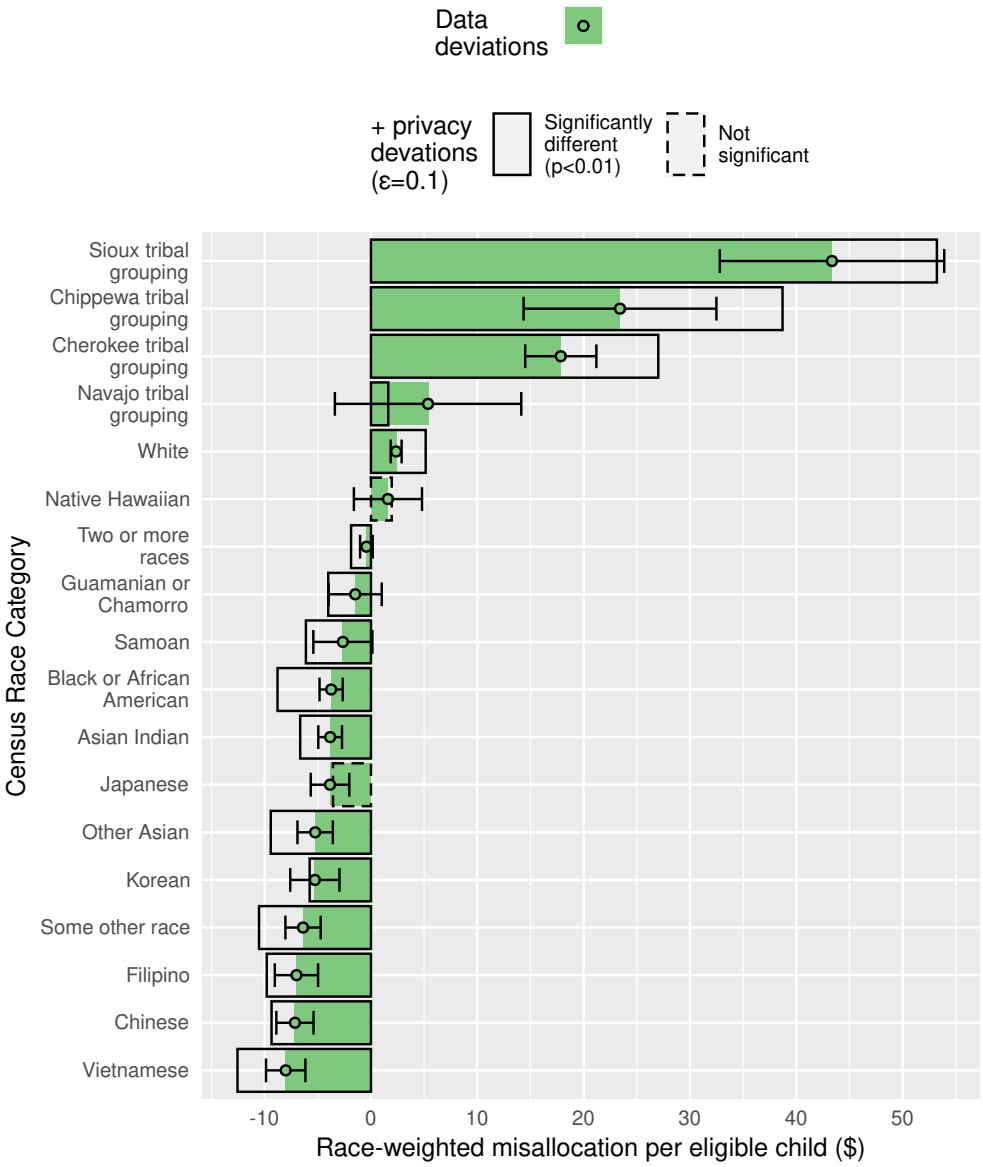


Figure A6: Sum of misallocation multiplied by the proportion of respondents of each census single race category (disaggregated), divided by the sum of formula-eligible children of that race. Averaged over 1,000 trials. A black outline indicates the marginal change in misallocation due to injected noise, drawn from Laplace mechanism with $\epsilon = 0.1$. Error bars span a 90% normal confidence interval. Dashed lines indicate a statistically insignificant difference between race-weighted data and privacy deviations using a two-sample z-test. Note that for the tribal and Pacific Islander subgroups, error in the ACS estimates could introduce an additional margin of error of up to ± $1.17.

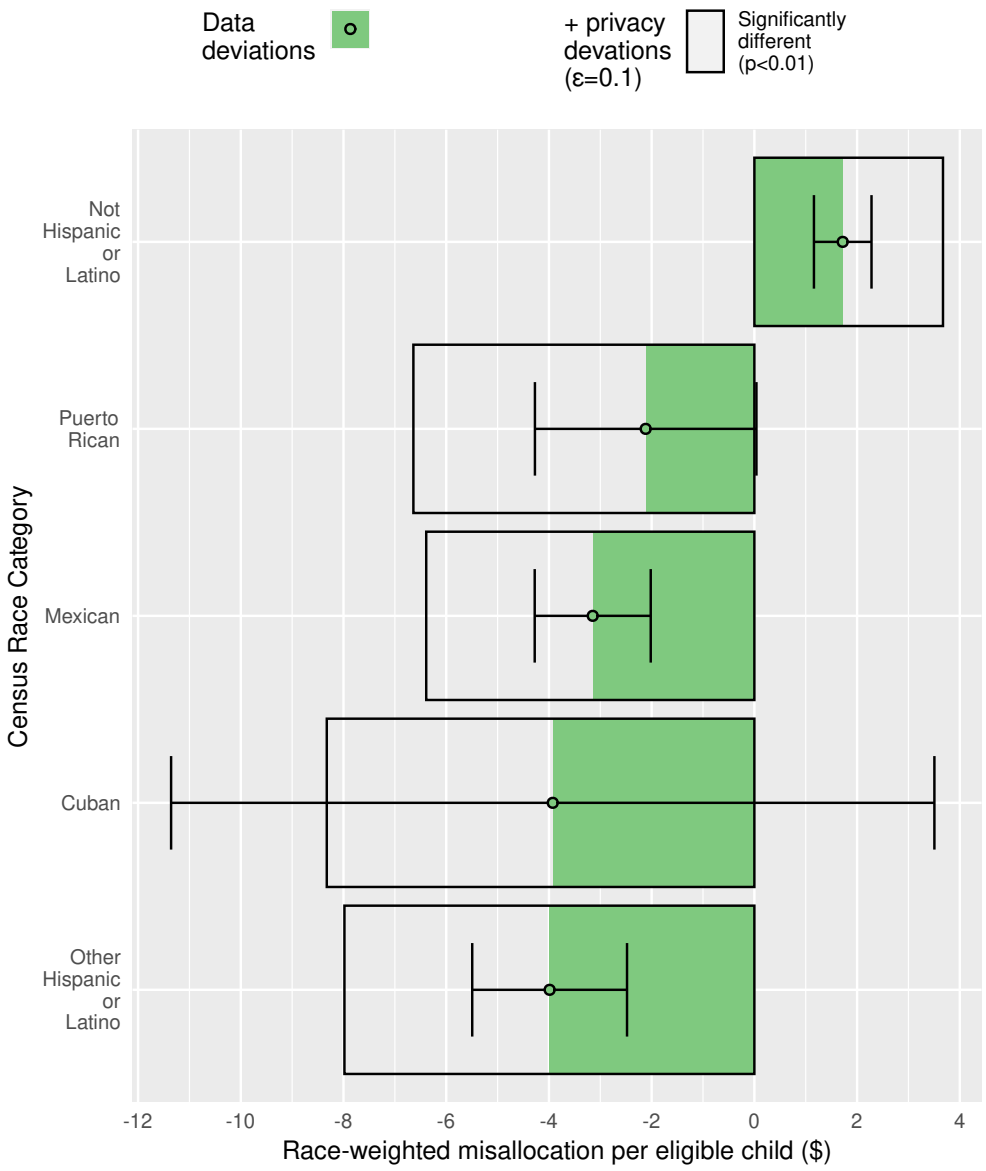
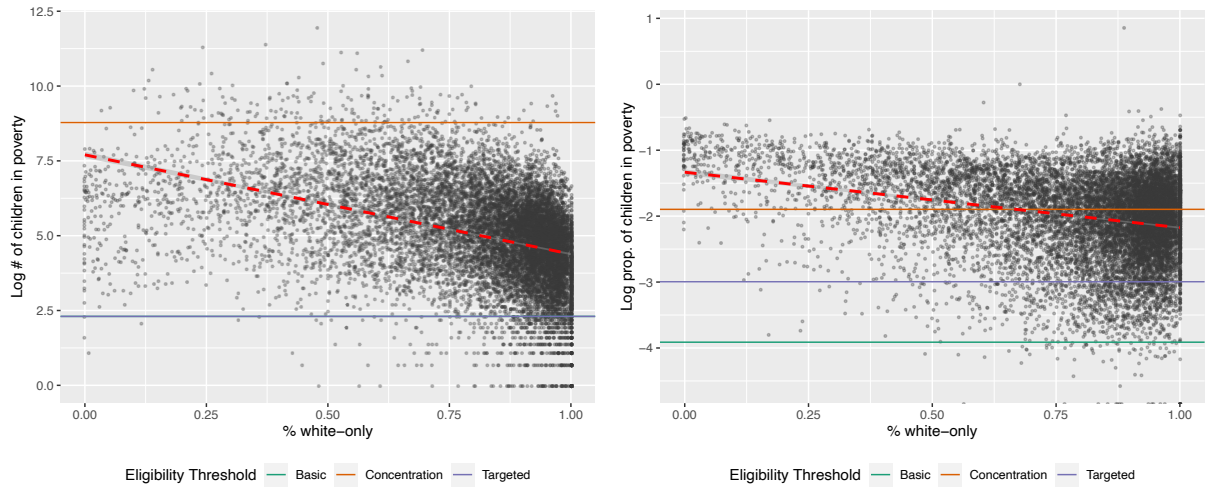


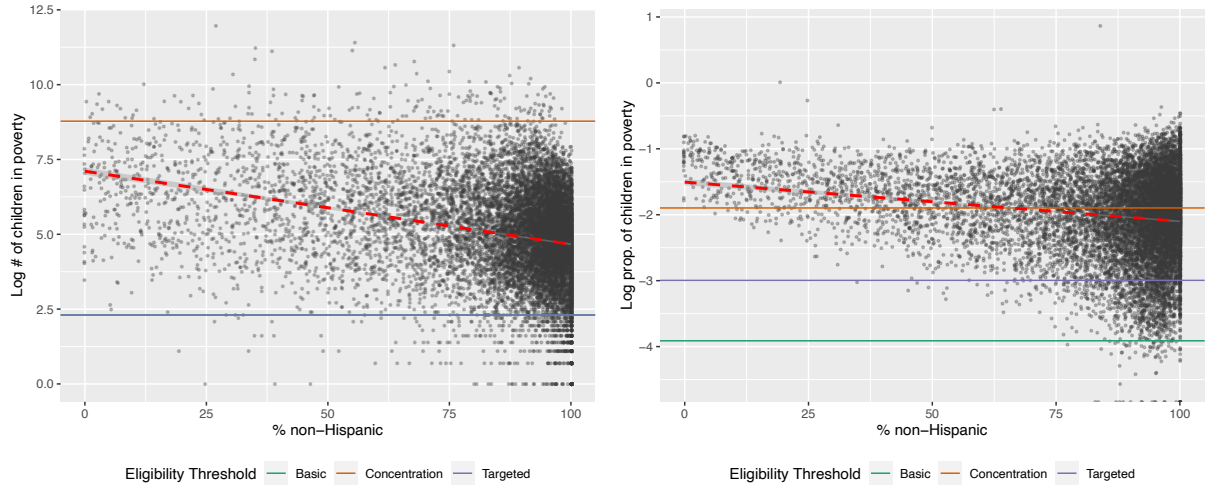
Figure A7: Sum of misallocation multiplied by the proportion of respondents of each census single ethnicity category, divided by the sum of formula-eligible children of that ethnicity. Averaged over 1,000 trials. A black outline indicates the marginal change in misallocation due to injected noise, drawn from Laplace mechanism with $\epsilon = 0.1$. Error bars span a 90% normal confidence interval. The additional impact of privacy deviations is significant ($p < 0.01$) for all groups, according to a two-sample z-test.



(a) Count thresholds.

(b) Proportional thresholds.

Figure A8: Children in poverty (as a count or a proportion of total children) by proportion of White-only children.



(a) Count thresholds.

(b) Proportional thresholds.

Figure A9: Children in poverty (as a count or a proportion of total children) by proportion non-Hispanic residents.

	Resid. Df	Resid. Dev	Df	Deviance	F	Pr(>F)
1	130,841.500	7.798×10^{16}				
2	130,833.700	7.796×10^{16}	7.811	1.668×10^{13}	3.584	0.0004

Table A.4: Analysis-of-deviation F-test of difference in sum of squared residuals between (1) a GAM specification with no tensor product smooths and a (2) a GAM specification with product smooths. Deviations are in dollars misallocated.

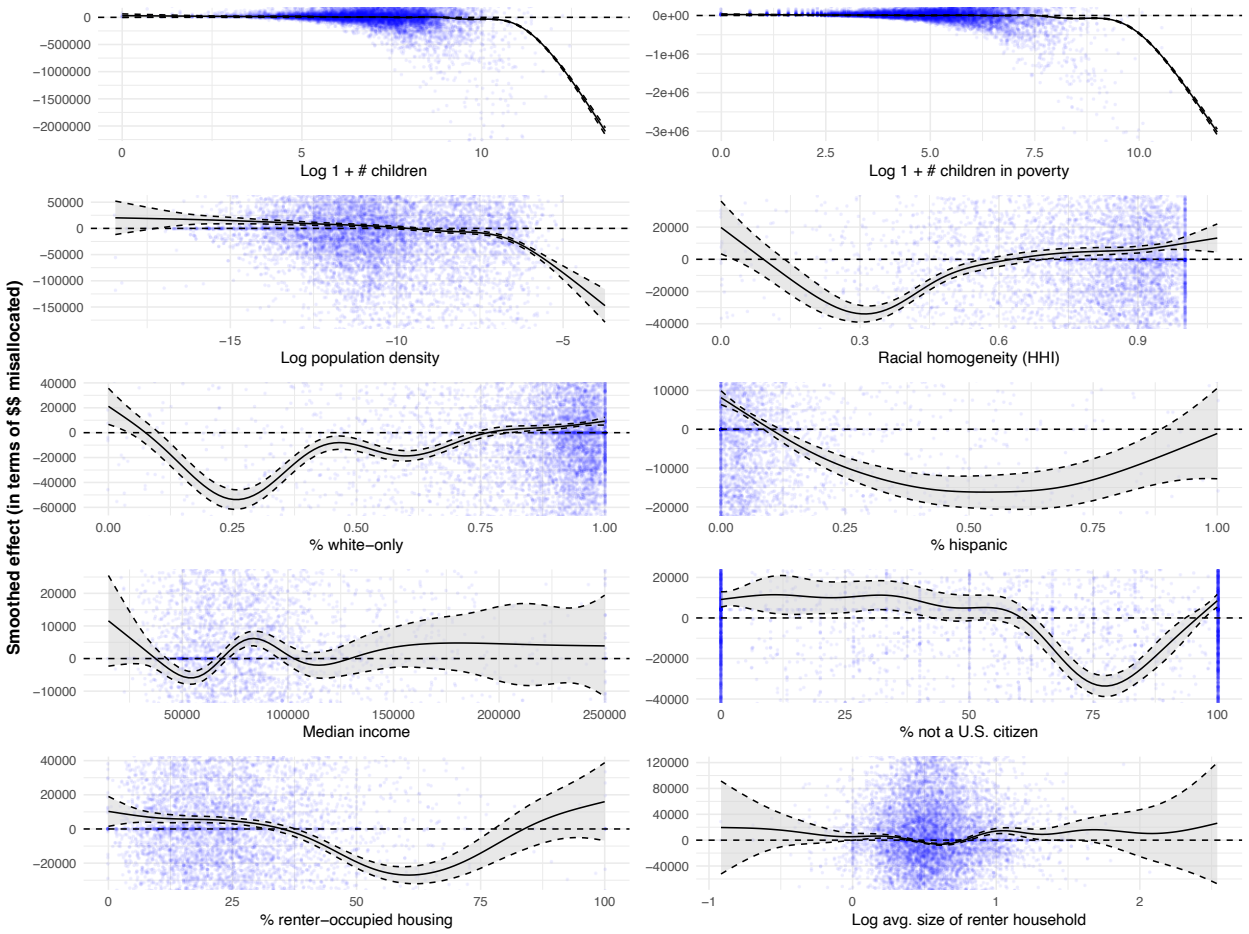


Figure A10: Model-smoothed misallocation (from both data deviations and injected noise) from independent univariate regressions, with 95% confidence interval in gray. Injected noise is drawn from a Laplace mechanism with $\epsilon = 0.1$. Positive values indicate districts that expect to benefit from combined data and privacy deviations; negative values indicate districts that expect to lose funding because of deviations.

A. parametric coefficients	Estimate	Std. Error	t-value	p-value
(Intercept)	-227.2932	2133.7902	-0.1065	0.9152
B. smooth terms	edf	Ref.df	F-value	p-value
te(prop_white,median_income_est)	16.7213	19.0180	2.7964	< 0.0001
s(log(pop_density))	5.9025	7.1219	3.9745	0.0002
s(hhi)	4.9762	6.1601	1.1466	0.3279
s(prop_hispanic)	1.6025	1.9921	2.3333	0.1119
s(renter_occupied_housing_tenure_pct)	1.0005	1.0009	0.7541	0.3854
R-sq. (adj) = 0.0009, Deviance explained = 0.11%,				
Scale est. = 5.9586e+11, n = 130870				

Table A.5: GAM estimation of the correlation between demographic covariates and misallocation. edf stands for effective degrees of freedom. F-values are reported for a joint test of equality to zero across each set of spline coefficients.

privacy deviations combined (Figure A11). The effects appear nearly identical in shape and magnitude for $\epsilon \geq 0.1$.

A.6 Policy Experiments

We tested to see how the following policy modifications might affect the distribution of misallocation. The results of these experiments are presented in Figures A12-A16.

Post-formula provisions

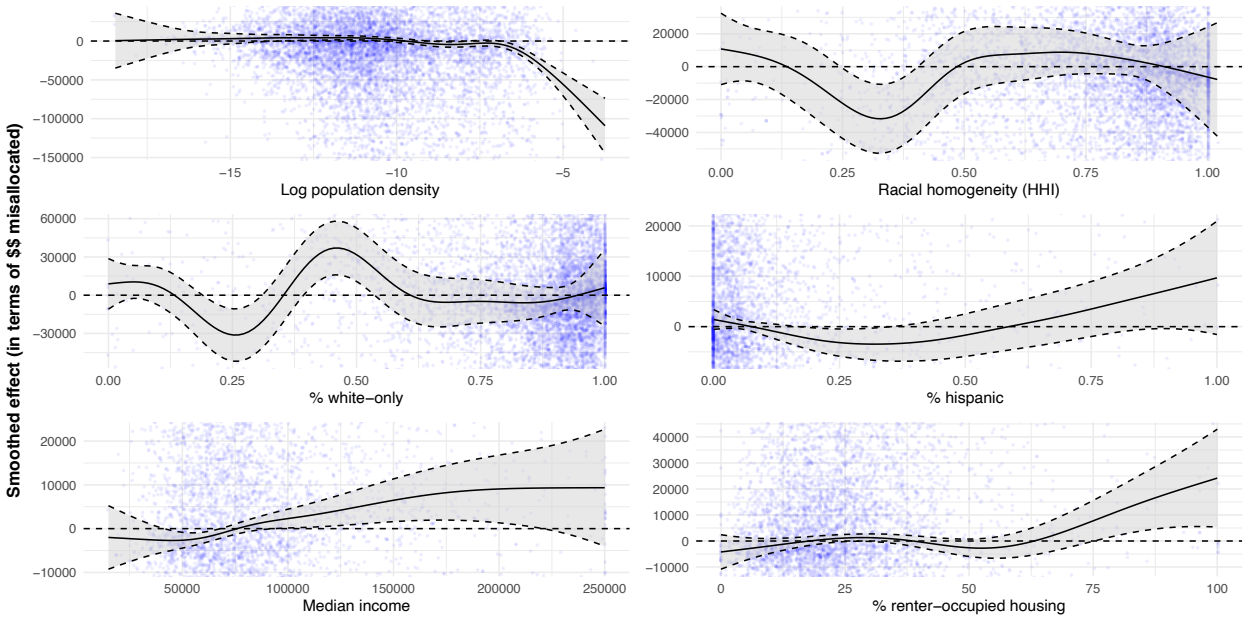
1. **No provisions (baseline).** Hold harmless provision is not applied. Regular, scaled formula entitlements used.
2. **Hold harmless.** Hold harmless provision applied. Districts cannot lose more than l times the previous year's funding. $l = 0.15$ for districts with less than 15% children in poverty; $l = 0.10$ for districts with 15-30% children in poverty; and $l = 0.05$ for other districts.
3. **State minimum.** State minimum provision applied. States cannot receive more than a minimum amount of total funding per grant type, which is the minimum of a) 25% of FY 2001 total appropriations plus 35% of the total amount allocated in excess of the total amount in 2001, and b) the average of (a) and the state's eligibility count multiplied by 150% of the national average per-pupil payment (Sonnenberg, 2016).
4. **Both provisions.** Both hold harmless and the state minimum provision are applied.

Post processing

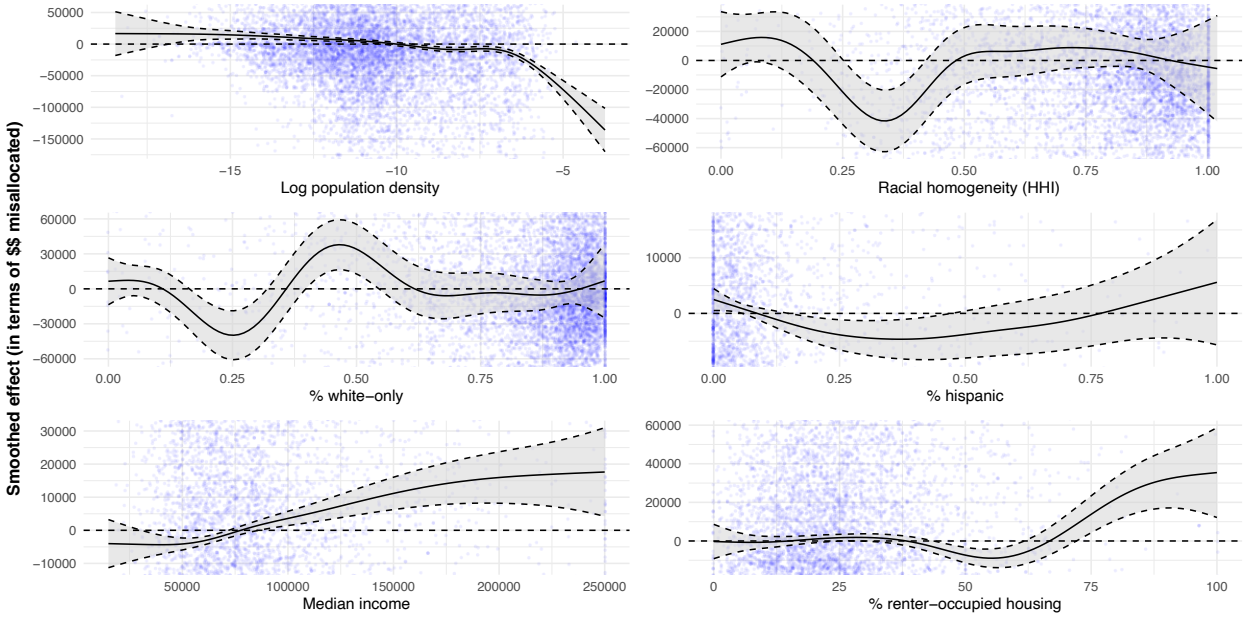
1. **No post-processing.** After applying deviations, counts are not modified.
2. **+ Clipping (baseline).** After applying deviations, negative counts are set to zero.
3. **+ Rounding.** After applying deviations, negative counts are set to zero and decimals are rounded to the nearest integer.

Moving averages

For this experiment, we assume that the ground-truth poverty estimates are the 5-year average from 2015-2019. (Using the 2019 data alone as ground truth would skew our results towards temporal disparities rather than disparities due to the effects of uncertainty.) For some districts, the SAIPE estimates already incorporate 5-year estimates alongside the single-year tax data; for others, the estimates come from a county model based on 1-year estimates (Maples, 2019). One practical implementation of this policy change could be converting all the SAIPE inputs to multi-year averages. The downside to using a moving average is that the allocations will be slower to react to trends in population. A future multi-year study could examine whether temporal population trends are large enough to exceed typical data and privacy deviations and substantially affect the usefulness of this approach.

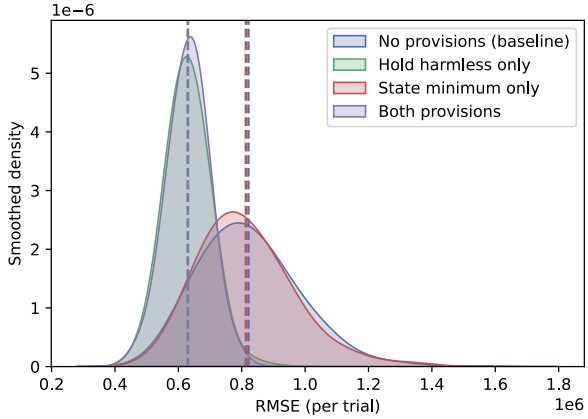


(a) Data deviations only.

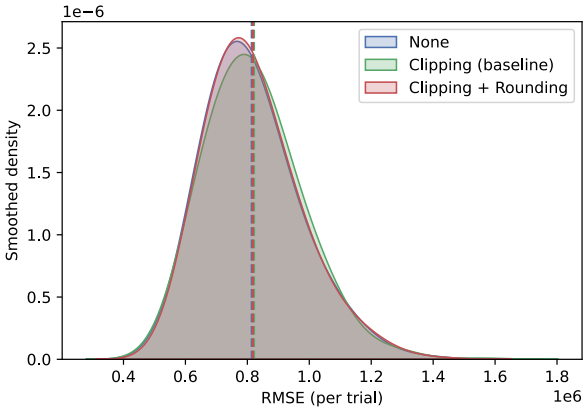


(b) Data and privacy deviations.

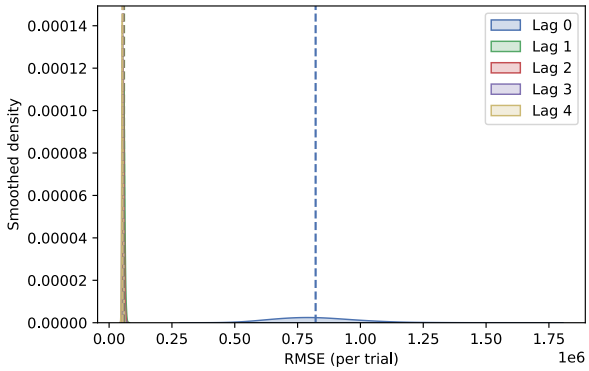
Figure A11: Effects of demographic variables on misallocation due to data *and* privacy deviations versus misallocation due to data deviation alone. Model-smoothed misallocation (from both data error and injected noise) by covariates, with 95% confidence interval in gray. Injected noise is drawn from Laplace mechanism with $\epsilon = 0.1$.



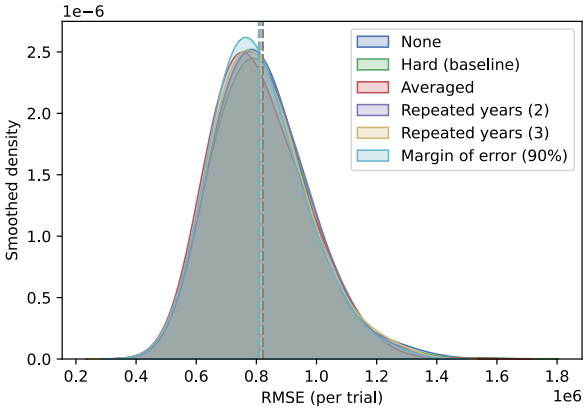
(a) With and without hold harmless provision. Pre-held harmless entitlements are ground truth.



(b) Using various post-processing mechanisms applied to the data error and injected noise: 1) no post-processing, 2) clipping negative values to zero, 3) additionally rounding to integers.

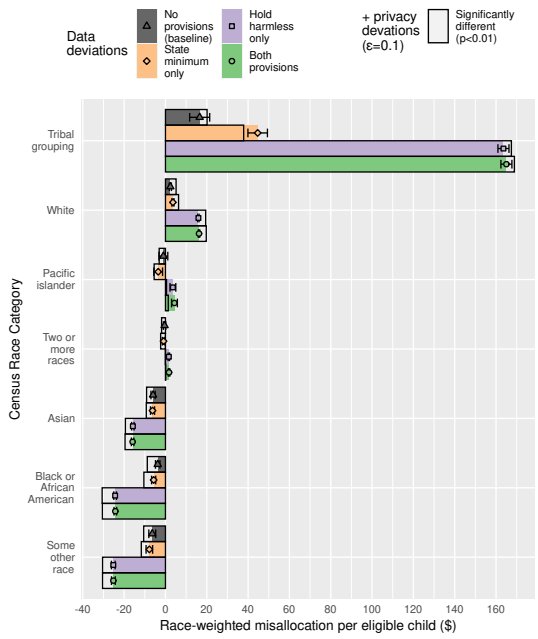


(c) Using a moving average instead of a point estimate, with the lag number indicating the number of prior years included in the average.

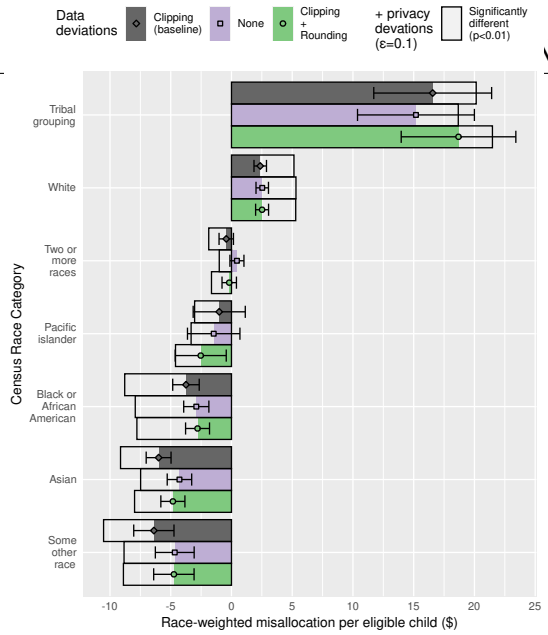


(d) Using modified eligibility thresholds for all grant types, described in Appendix A.6.

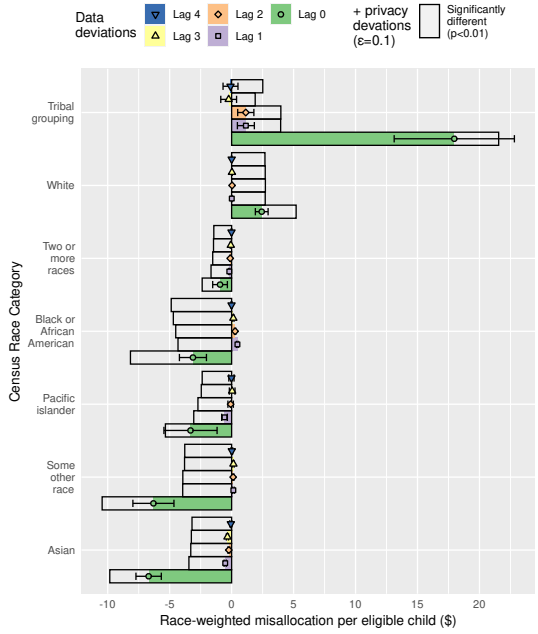
Figure A12: For each modification of Title I, root mean squared loss across 1,000 trials. Dashed line indicates average RMSE across all trials. Includes data error and injected noise drawn from Laplace mechanism with $\epsilon = 0.1$.



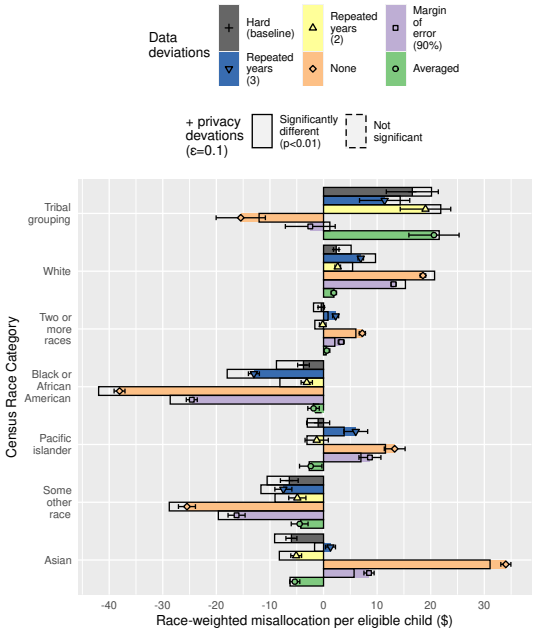
(a) With and without hold harmless provision.



(b) Using various post-processing mechanisms applied to the data error and injected noise: 1) no post-processing, 2) clipping negative values to zero, 3) additionally rounding to integers.

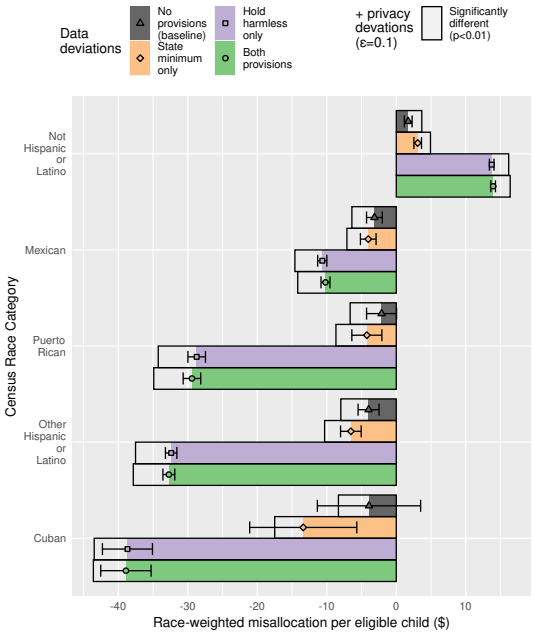


(c) Using a moving average instead of a point estimate, with the lag number indicating the number of prior years included in the average.

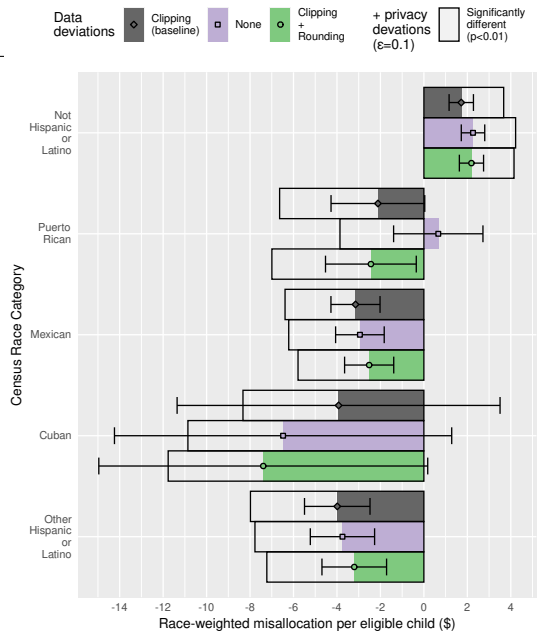


(d) Using modified eligibility thresholds for all grant types, described in Appendix A.6.

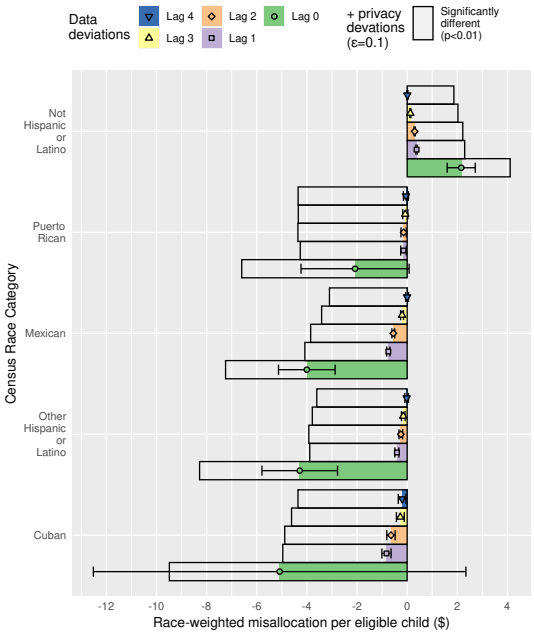
Figure A13: For each modification of the Title I procedure or privacy mechanism, race-weighted misallocation per formula-eligible child. Averaged over 1,000 trials. A black outline indicates the marginal change in misallocation due to injected noise, drawn from Laplace mechanism with $\epsilon = 0.1$. Error bars span a 90% normal confidence interval. Dashed lines indicate a statistically insignificant difference in race-weighted misallocation after privacy deviations are added, using a two-sample z-test.



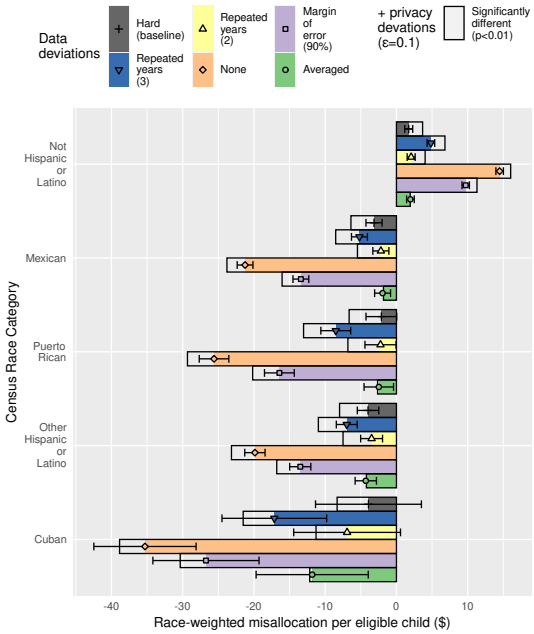
(a) With and without hold harmless provision.



(b) Using various post-processing mechanisms applied to the data error and injected noise: 1) no post-processing, 2) clipping negative values to zero, 3) additionally rounding to integers.

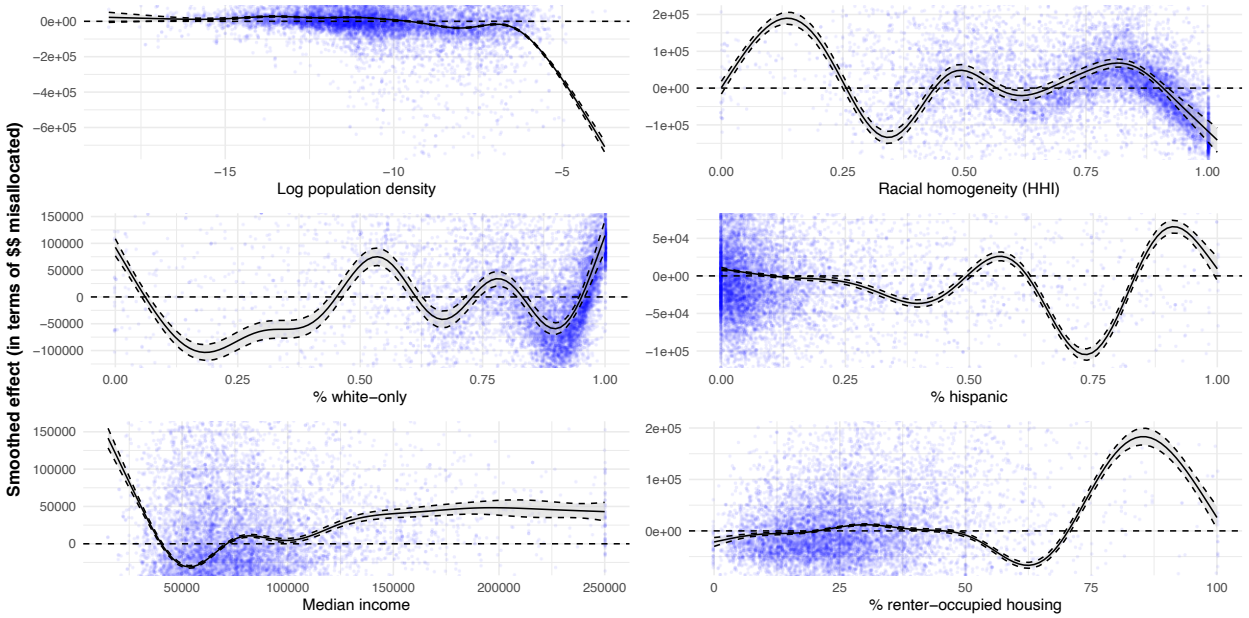


(c) Using a moving average instead of a point estimate, with the lag number indicating the number of prior years included in the average.

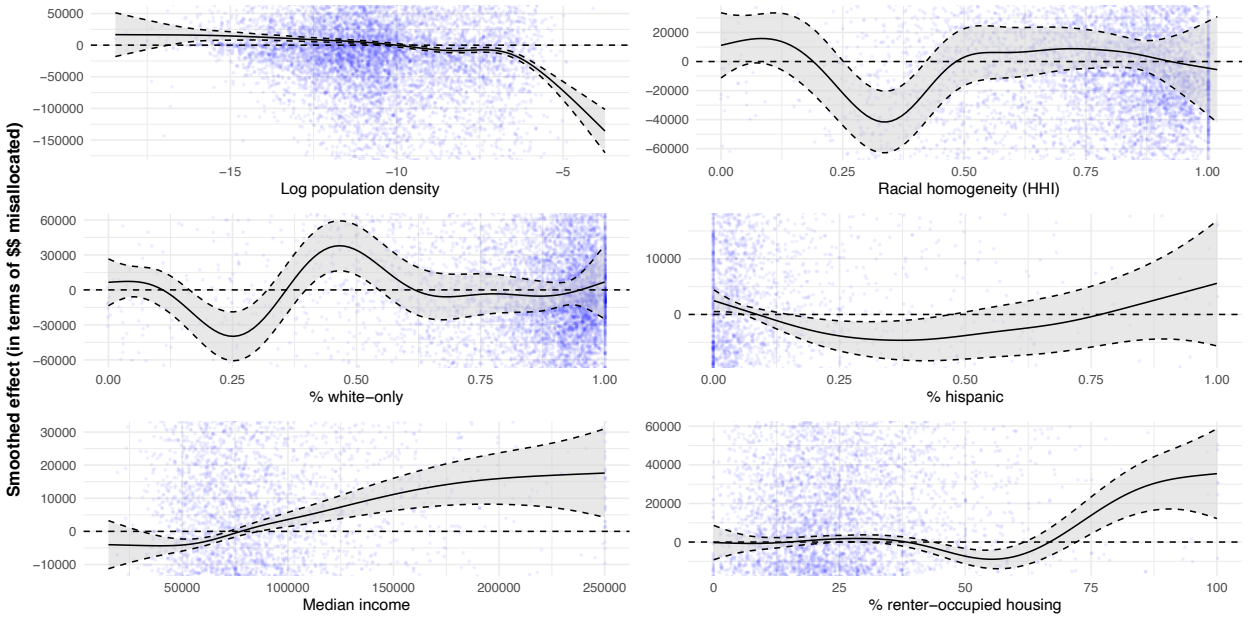


(d) Using modified eligibility thresholds for all grant types, described in Appendix A.6.

Figure A14: For each modification of the Title I procedure or privacy mechanism, ethnicity-weighted misallocation per formula-eligible child. Averaged over 1,000 trials. A black outline indicates the marginal change in misallocation due to injected noise, drawn from Laplace mechanism with $\epsilon = 0.1$. Error bars span a 90% normal confidence interval. The additional impact of privacy deviations is significant ($p < 0.01$) for all groups in all treatments, according to a two-sample z-test.



(a) Hold harmless and state minimum provisions.



(b) No provisions.

Figure A15: Effects after addition of special provisions. Model-smoothed misallocation (from both data error and injected noise) by covariates, with 95% confidence interval in gray. Injected noise is drawn from Laplace mechanism with $\epsilon = 0.1$.

1. **Single-year (baseline).** SAIPE 2019 estimates are used to determine allocations.
2. **Averaged, lag l.** The SAIPE tables from l years up to 2019 are averaged together and used to determine all allocations. The $l+1$ coefficients of variation $\{c_i^j\}_{j\in[l+1]}$ for each school district i and year j are combined into an averaged coefficient of variation

$$\bar{c}_i = \frac{\sigma_i}{\bar{x}_i} = \frac{1}{\bar{x}_i}\sqrt{\frac{\sum_{j=1}^{l+1}\sigma_i^{j2}}{(l+1)^2}},$$

where $\sigma_i^j = x_i^j c_i^j$.

Alternative thresholds

1. **Hard thresholds (baseline).** All grant eligibility thresholds are enforced as written. Districts not meeting the eligibility requirements for a given grant receive no funding for that grant.
2. **Average eligibility.** The 5-year moving average is used to determine eligibility.
3. **j-repeated ineligibility.** Districts are only counted as ineligible if they have been ineligible j years in a row; otherwise they receive the normal formula amount.
4. **α-level margin of error relaxation.** All eligibility thresholds are reduced by the district's α-level margin of error.
5. **No thresholds.** All districts are considered eligible.

Budget increases

Each proposed budget increase is distributed proportionally to each grant type.

1. **Baseline federal appropriation.** Approximately $16 billion total, $12 billion to basic, concentration, and targeted grants.
2. **+ loss.** Increase the baseline appropriation by the absolute sum of negative expected misallocation due to privacy and data deviations (under the baseline).
3. **+ α-quantile loss.** Increase the baseline appropriation by absolute sum of negative α-quantile misallocation. (This can be thought of as the reasonable worst-case loss for each district.)
4. **Biden proposal.** Increase the baseline appropriation by $20 billion, the Biden administration's proposal for 2022 (Department of Education, 2022).

A.7 Sensitivity Analysis

Privacy Mechanism

The Census Bureau has not yet made any concrete plans for disclosure avoidance in the ACS (Jarmin, 2019; Rodriguez, 2021) and the SAIPE currently does not inject noise for privacy

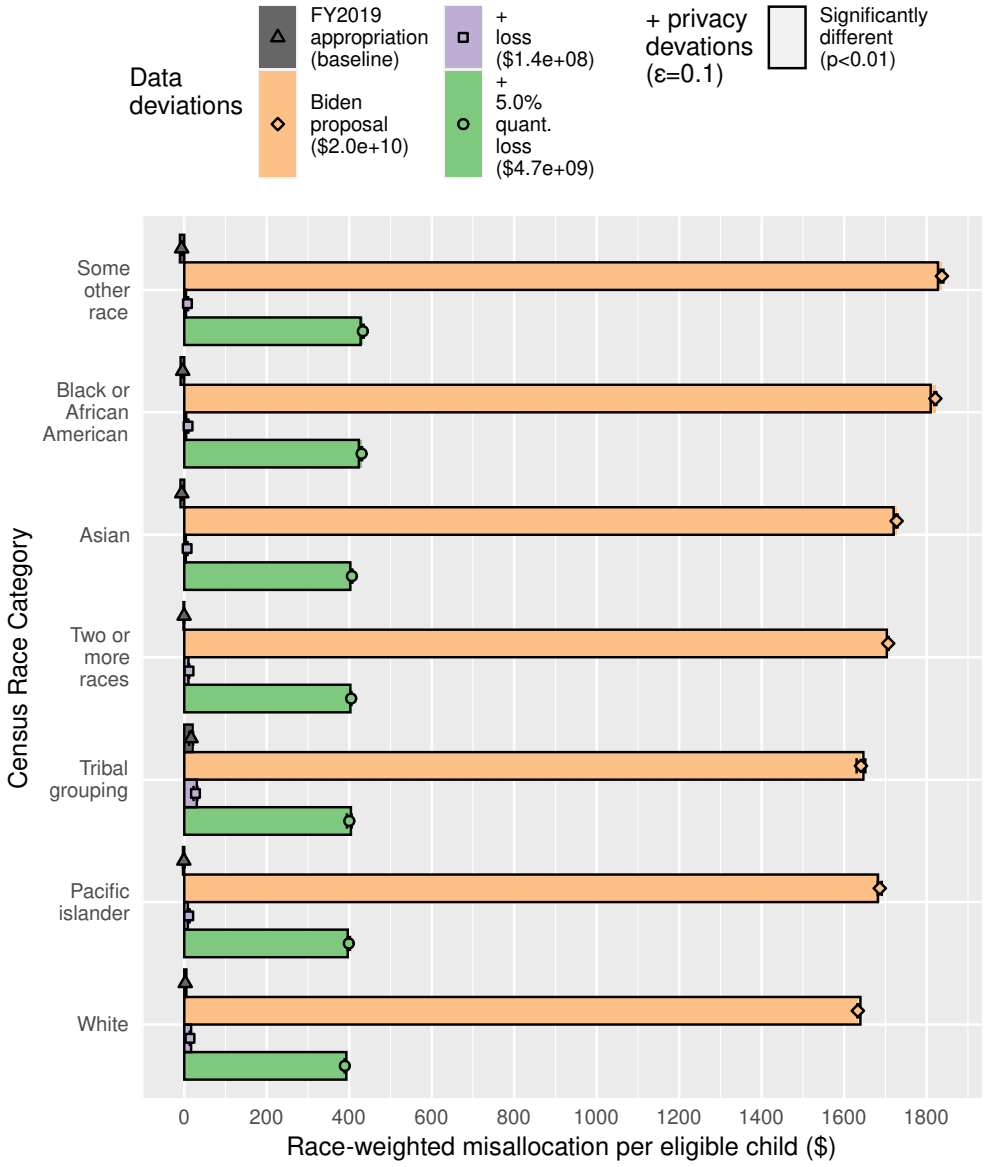


Figure A16: Change in race-weighted misallocation relative to previous budget for a series of proposed increases in federal appropriations for Title I grants, averaged over 1,000 trials. A black outline indicates the marginal change in misallocation due to injected noise, drawn from Laplace mechanism with $\epsilon = 0.1$. Error bars span a 90% normal confidence interval. The additional impact of privacy deviations is significant ($p < 0.01$) for all groups in all treatments, according to a two-sample z-test.

on top of its inputs. To fairly compare misallocations due to privacy and data deviations, we chose ϵ to provide a higher privacy guarantee than may be expected in practice. There are several good reasons to make a conservative choice for ϵ. First, the implications of privacy protection in our setting are not clear. Poverty estimates are composed of a combination of data sources, some of which include weighted samples, and the details of estimation are not all public—so it is hard to precisely define the sensitivity Δ and interpret ϵ. Instead, we try several reasonable values of ϵ and present the differences here. While prior work suggests a higher setting ($\epsilon = 2.52$) of ϵ for this problem (Abowd & Schmutte, 2019), we chose a setting low enough to provide strong privacy guarantees even if sensitivity is increased by an order of magnitude. (Privacy advocates often prefer $\epsilon < 1$ (Dwork et al., 2019).) Second, this use case may be part of a larger privacy budget in practice, where linkage between data products and queries is an issue. Though the overall privacy budget may be high, this particular use case could receive a small share of the budget. Finally, it is still not clear what privacy protections might be applied to the ACS and other data products.

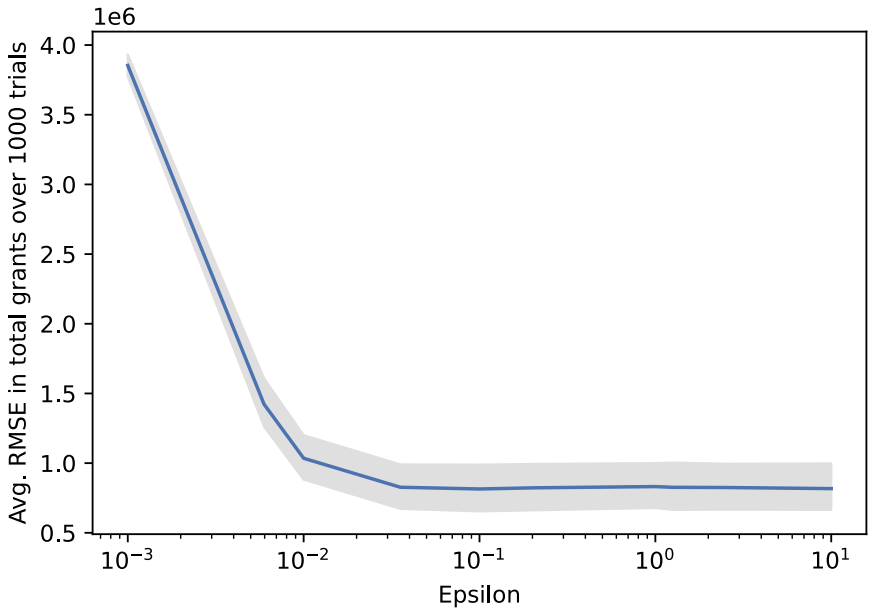
Because it is difficult to predict how the factors above will affect future privacy plans, we explore other amounts of privacy noise to see how our results might change. Figure A17 shows the privacy-utility frontier, while Figure A18 and Figure A19 depict the privacy-fairness trade-off. As ϵ increases, the marginal decrease in utility caused by privacy noise diminishes, but the marginal effect of privacy on outcome disparities increases. Both the utility and disparity trade-offs are relatively small and stable when $\epsilon \geq 0.1$. The marginal increase in average total entitlement loss (Figure 1.1) after adding privacy deviations draws closer to the total entitlement loss due to data deviations alone at $\epsilon = 0.01$ and the eventually exceeds it at $\epsilon \leq 0.01$ (Table A.6). Also notably, the effects of continuous demographic variables are much more pronounced for lower values of ϵ (Figure A20). For example, at $\epsilon = 0.01$, districts with a small Hispanic population or districts that are racially homogeneous tend to gain. The negative effect for districts with high population density is also more pronounced.

This result is not particularly surprising when we examine the magnitude of noise injected for privacy compared to the magnitude of underlying data deviations (Figure A23). The Laplace mechanism we use is invariant in population size–the variance of the privacy noise is constant. The total variance of the data deviations, on the other hand, is $(c_i\mu_i)^2$, which will be small only when μ_i is small. So, as in the 2020 Decennial Census (Bell & Schafer, 2021), the magnitude of privacy deviations is comparable to the magnitude of data deviations only in the districts with the fewest Title I eligible children (Figure A24).

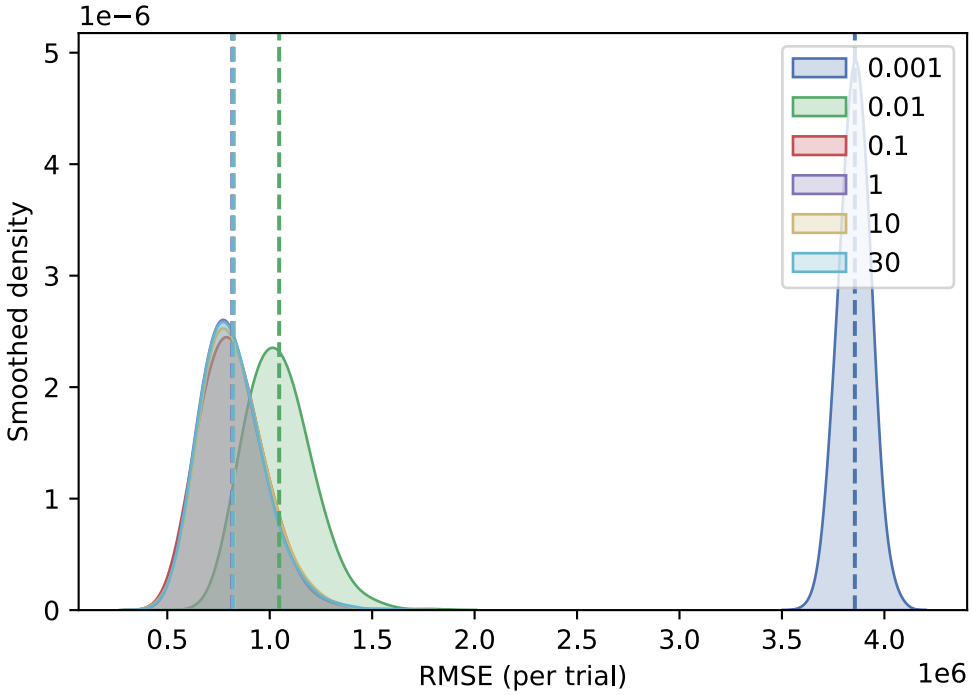
Data Error Simulation

Varying the magnitude of data deviations

Preliminary updated research (Maples, 2019) suggests that there are potentially more accurate ways of estimating poverty and that updates to the ACS and other data inputs could improve the precision of the poverty estimates. To investigate the effects of lower data error on our results, we imagine that the coefficients of variation are reduced or increased by as much as 50%, and we also try using Laplace noise instead of Gaussian noise to see if the shape of the distribution has any effect.



(a) Root mean squared error in allocations across all trials.



(b) Distribution of root mean squared error per trial. Dotted lines depict averages.

Figure A17: Combined data and privacy deviations under different ϵ settings.

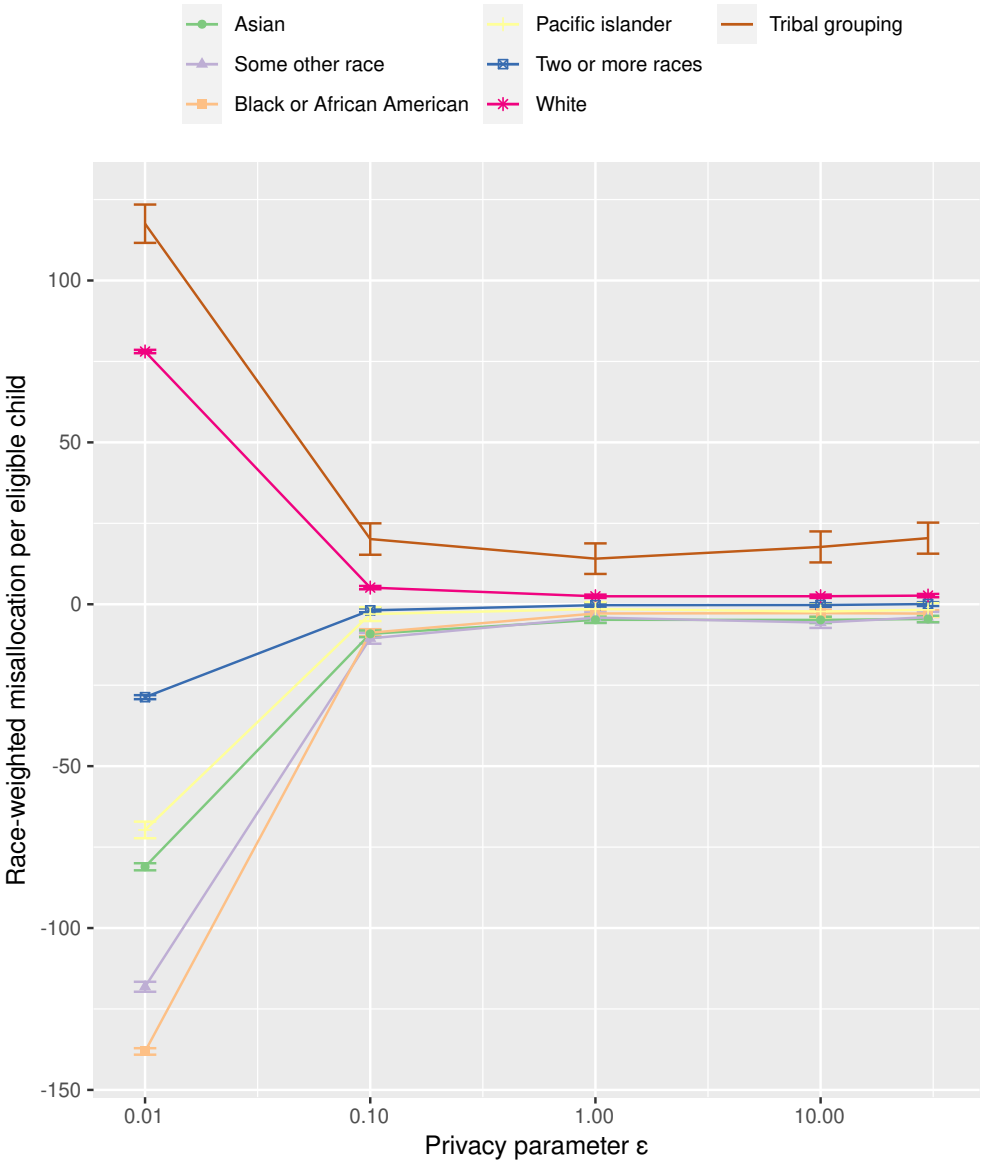


Figure A18: Race-weighted misallocation under various ϵ settings, averaged over 1,000 trials. The marginal effects of privacy deviations increase as ϵ decreases. Error bars span a 90% normal confidence interval.

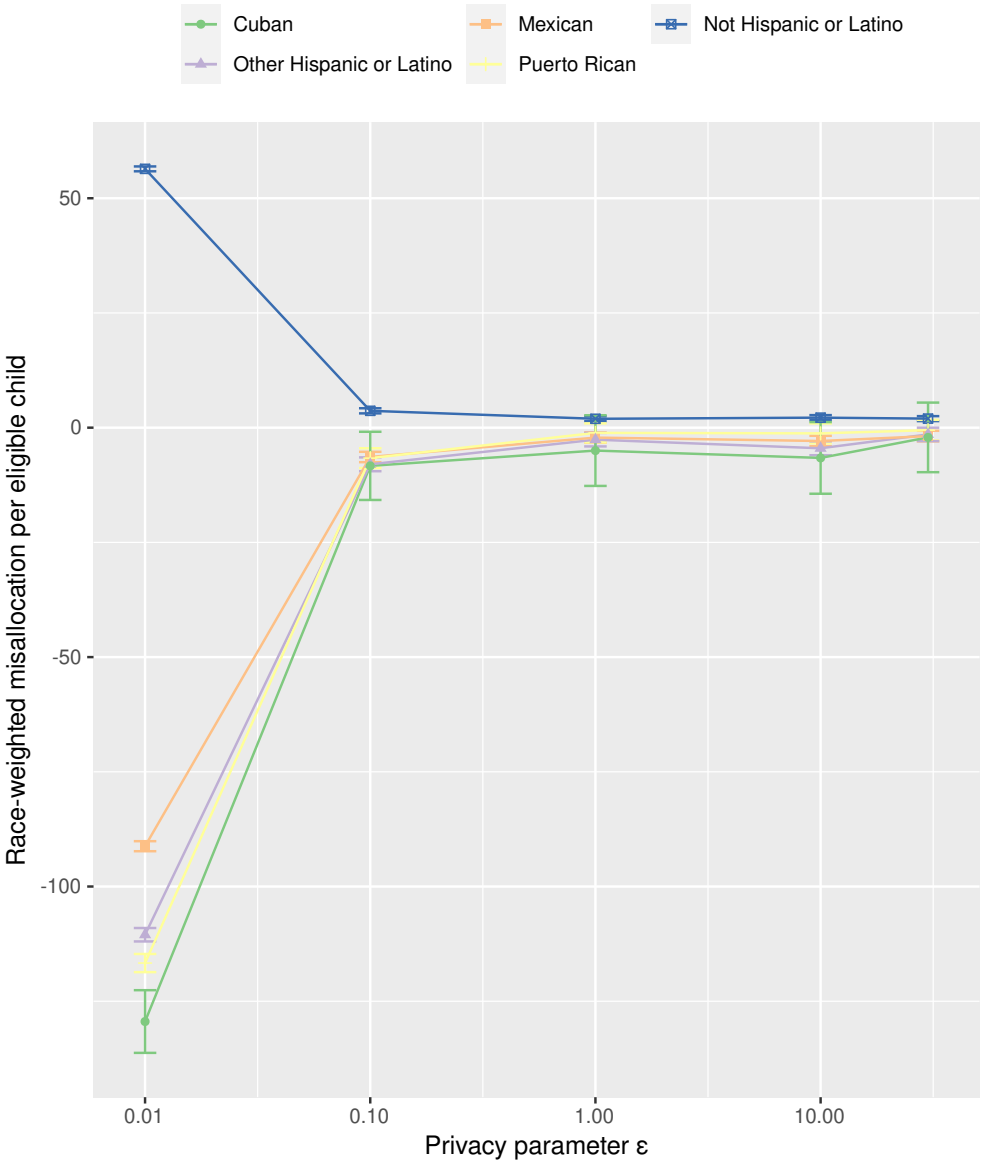
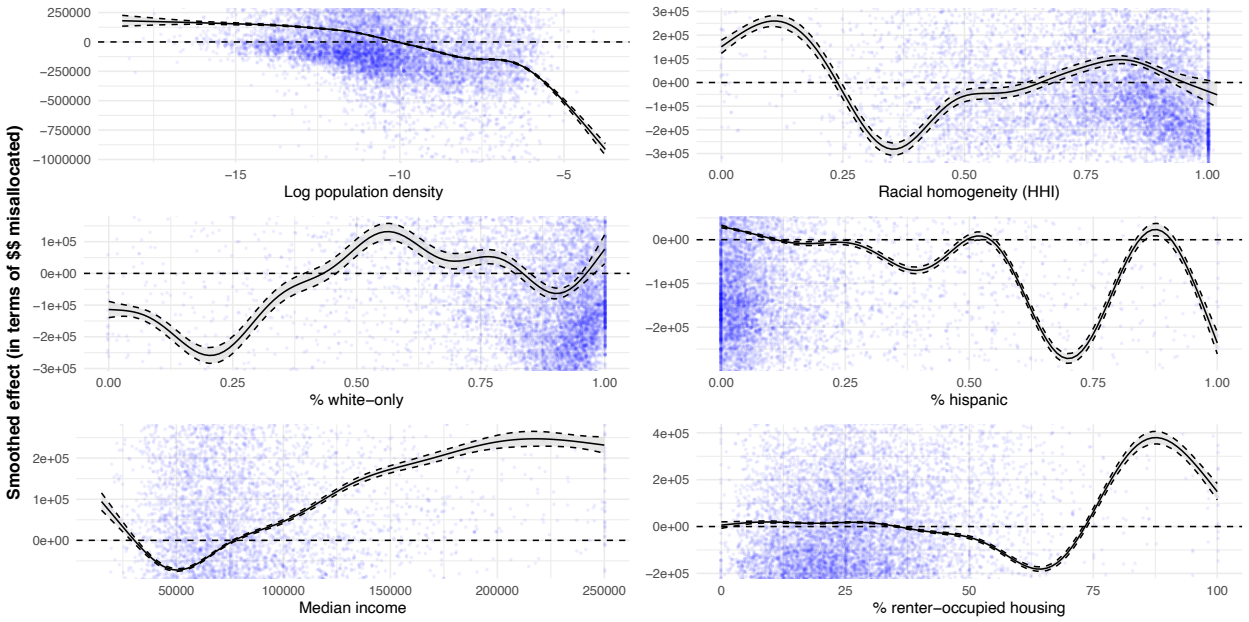
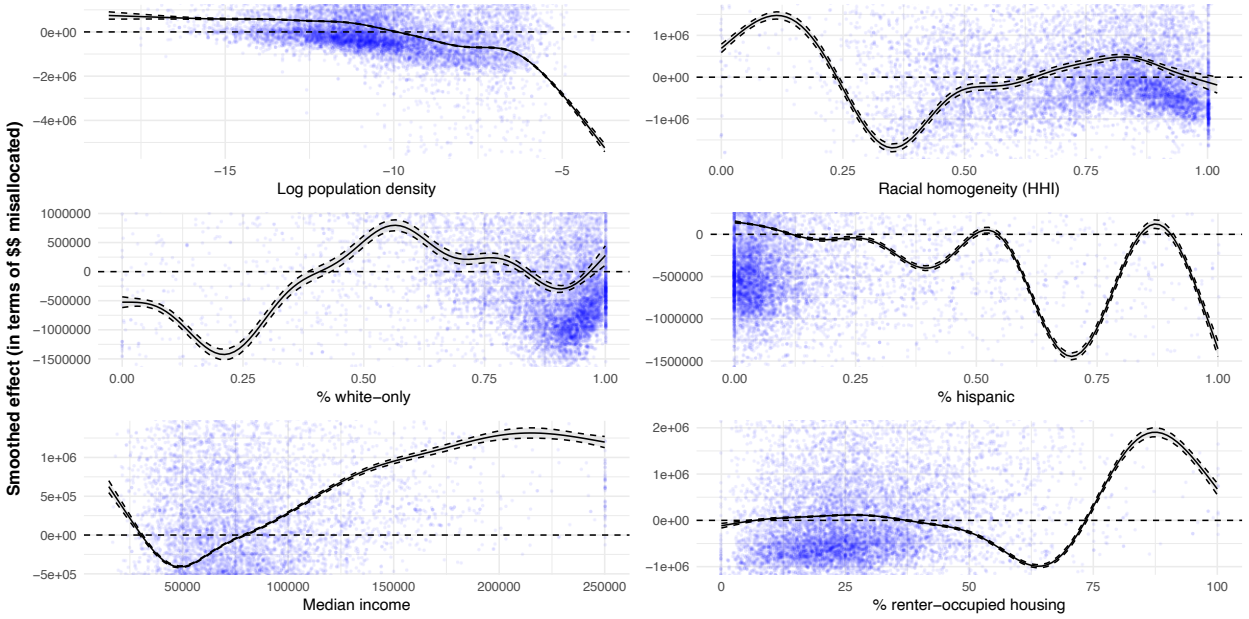


Figure A19: Ethnicity-weighted misallocation under various ϵ settings, averaged over 1,000 trials. The marginal effects of privacy deviations increase as ϵ decreases. Error bars span a 90% normal confidence interval.



(a) $\epsilon = 0.01$



(b) $\epsilon = 0.001$

Figure A20: Effects under lower ϵ settings. Model-smoothed misallocation (from both data error and injected noise) by covariates, with 95% confidence interval in gray.

ϵ for added privacy deviations	Expected total entitlement loss ($)	Sum of expected losses ($)	Sum of 5% quantile losses ($)
0.001	7.16×10^9 (4.64×10^7)	5.55×10^9 (6.85×10^7)	1.07×10^10
0.01	2.15×10^9 (3.58×10^7)	1.09×10^9 (8.70×10^7)	6.30×10^9
0.1	1.11×10^9 (3.16×10^7)	1.35×10^8 (9.30×10^7)	4.73×10^9
1	1.06×10^9 (3.04×10^7)	1.04×10^8 (9.00×10^7)	4.22×10^9
10	1.06×10^9 (3.07×10^7)	1.06×10^8 (8.66×10^7)	4.23×10^9
Data deviations alone (baseline)	1.06×10^9 (3.15×10^7)	1.07×10^8 (9.08×10^7)	4.23×10^9

Table A.6: Misallocation after combined data and privacy deviations at varying levels of ϵ.

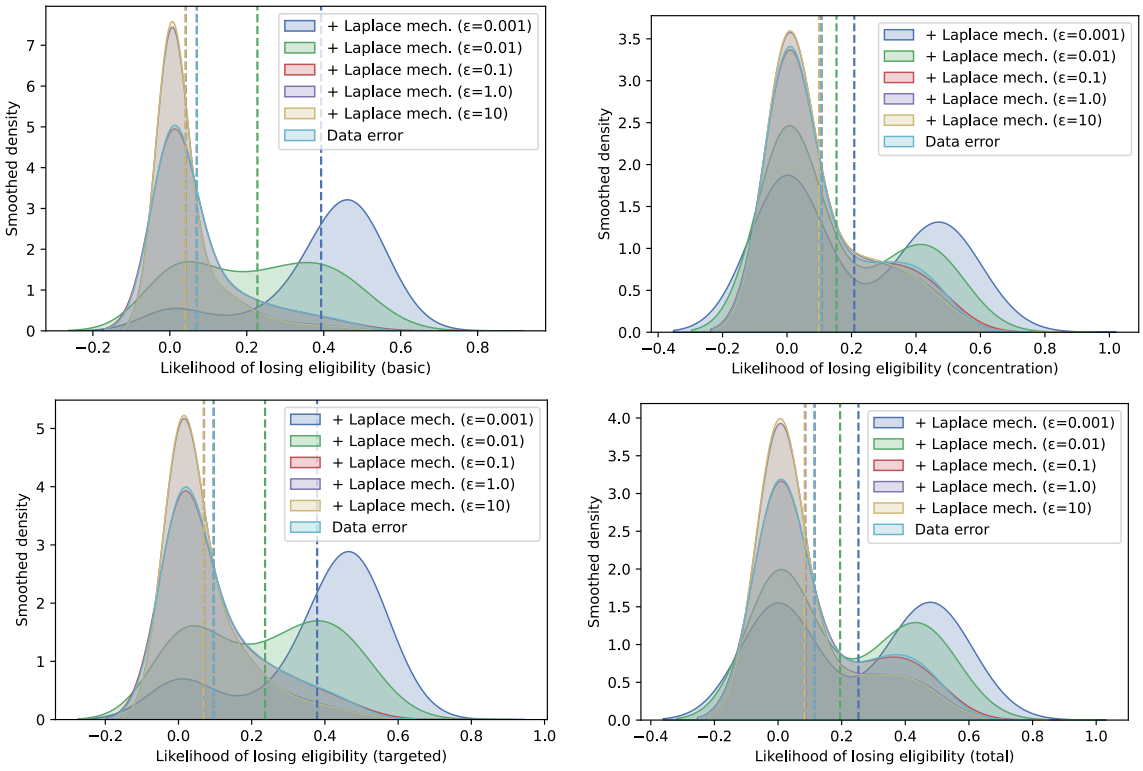


Figure A21: Likelihood of losing eligibility in each grant type, depending on the choice of privacy parameter ϵ. Higher ϵ means less privacy.

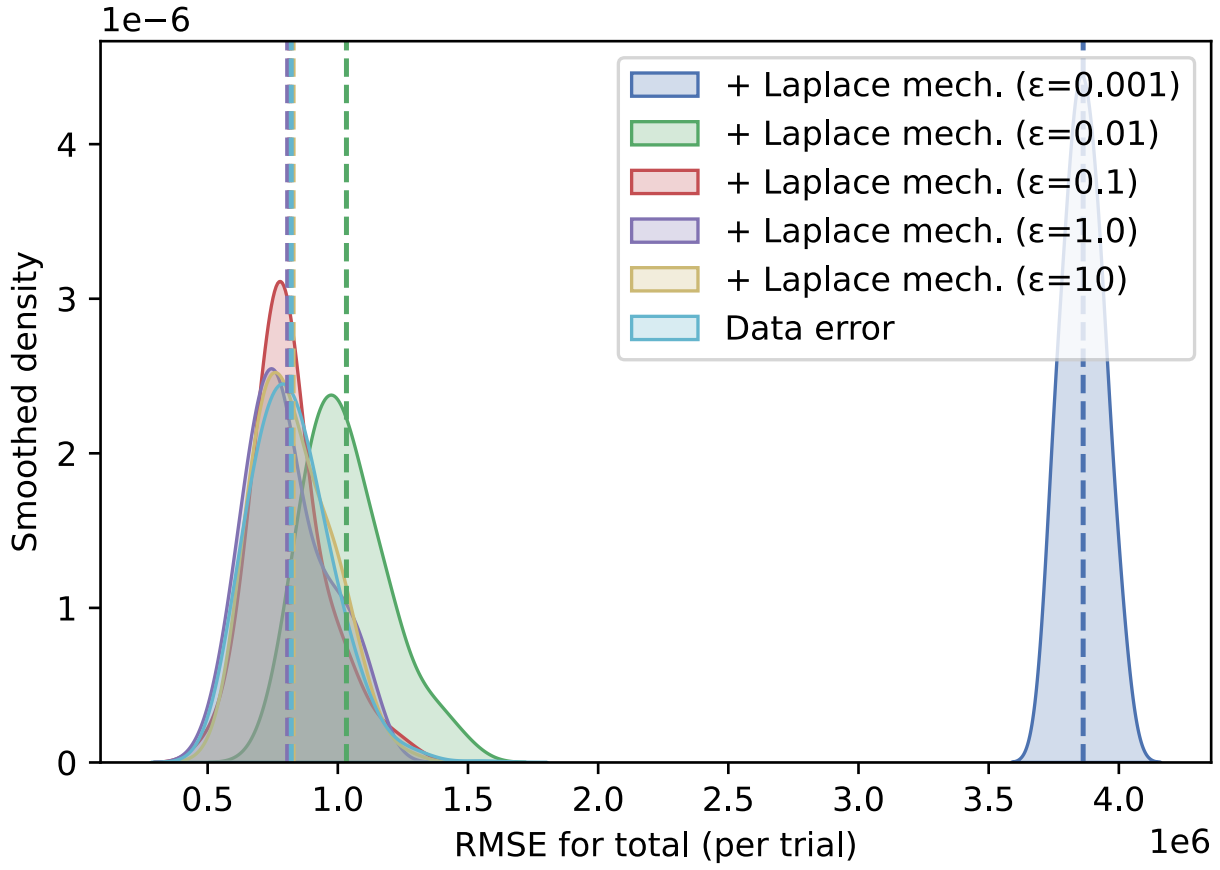
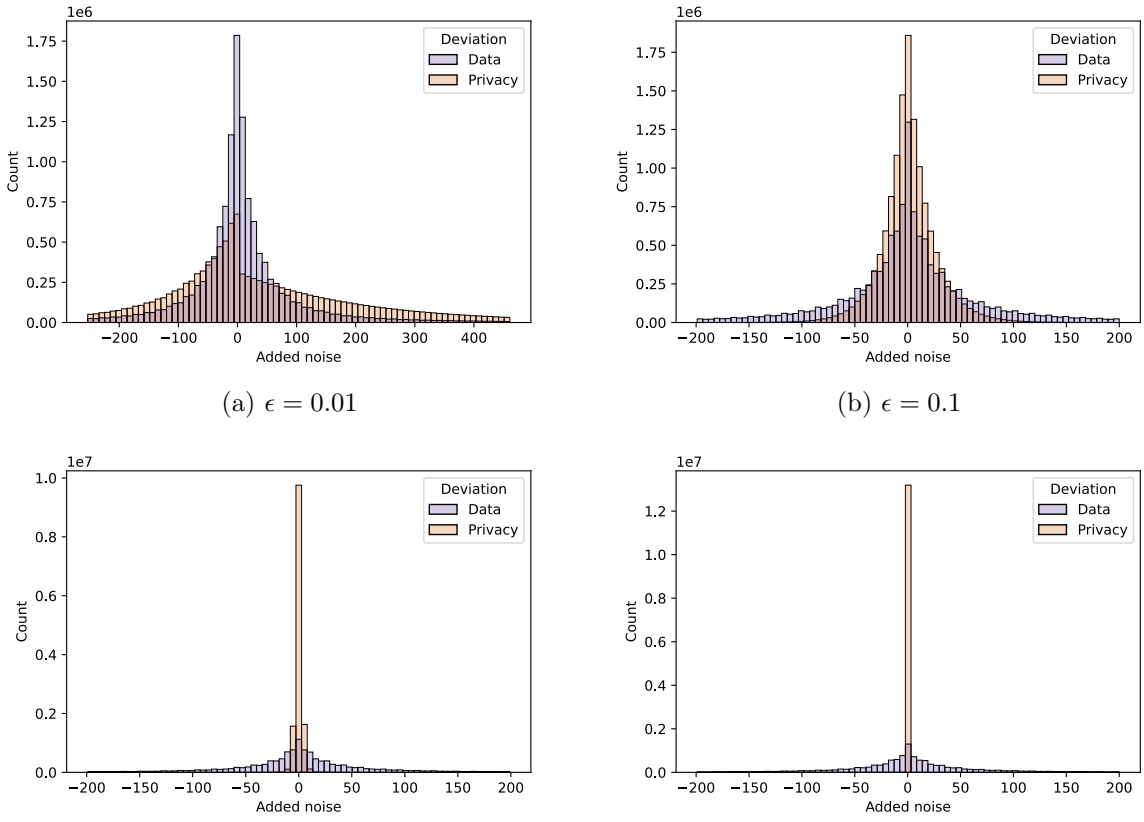


Figure A22: Root mean squared misallocation in grant funding, depending on the choice of privacy parameter ϵ. Higher ϵ means less privacy.



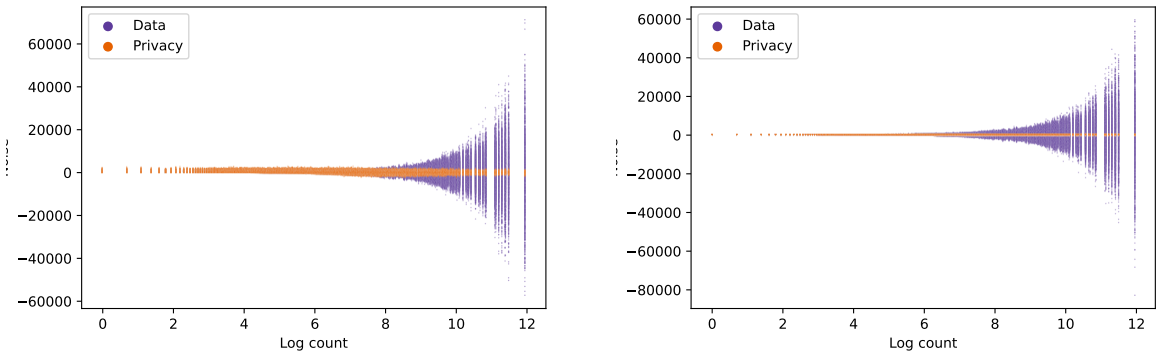
(a) $\epsilon = 0.01$

(b) $\epsilon = 0.1$

(c) $\epsilon = 1.0$

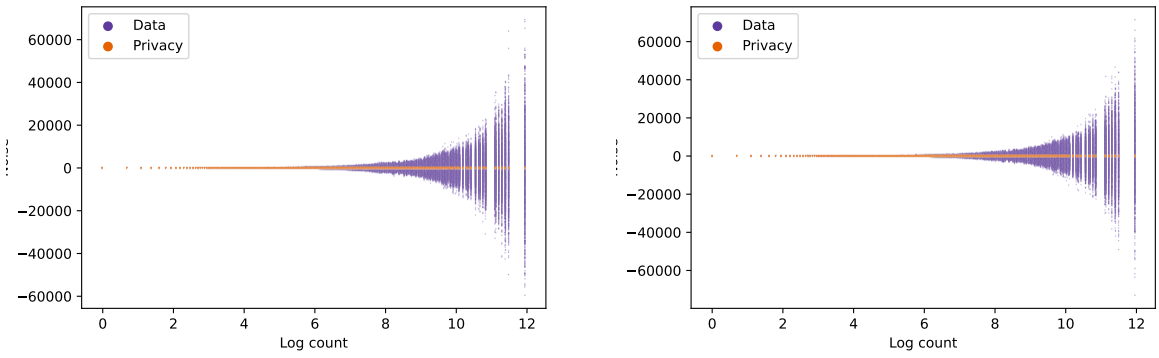
(d) $\epsilon = 10$

Figure A23: Distribution of data and privacy deviations injected in simulation, at various levels of the privacy parameter ϵ.



(a) $\epsilon = 0.01$

(b) $\epsilon = 0.1$



(c) $\epsilon = 1.0$

(d) $\epsilon = 10$

Figure A24: Distribution of data and privacy deviations injected in simulation, at various levels of the privacy parameter ϵ, by the official estimate of children in poverty.

As expected, misallocation (Figure A25) and outcome disparities (Figure A26) both increase with the amount of data error. The expected total loss due to data error alone—about $1.06 billion with the original variance estimates and a relatively high-privacy choice of $\epsilon = 0.1$—drop to $533 million when the standard deviations are halved and jump to $1.57 billion when increased by 50%. The marginal increase in losses due to privacy deviations declines with the magnitude of data deviations ($80 million, $50 million, and $33 million for 50%, 100%, and 150% of the estimated standard deviations, respectively). So, even under the most conservative estimate of data error and privacy deviations, the marginal increase in negative misallocation is still less than the losses due to underlying data error. The Laplace mechanism makes a small difference in overall losses due to data deviations, which drop by about $120 million when the error is modeled with Laplace noise instead of Gaussian with the original coefficients of variation.

Decreasing data error also decreases disparities. For conservative estimates of the data error, our finding that underlying disparities due to data error are only slightly increased by adding privacy noise does not always hold. Taking Asian students as an example, introducing differential privacy at these much lower levels of data error more than doubles race-weighted misallocation; under the default Census Bureau guidance, the increase is less noticeable.

Varying the total number of children

We also test our simplifying assumption that there are no deviations—data or privacy—in the total number of children z. The guidance published for modeling uncertainty in the SAIPE children in poverty estimate does not mention an uncertainty estimate for the total number of children per school district (US Census Bureau, 2020)—for this reason, our main analysis excludes this variable from the simulation of data and privacy deviations. To check the sensitivity of our results to this assumption, we imagine that the estimate of total children has the same coefficient of variation in each school district as the estimate of children in poverty (i.e., the standard deviation is scaled proportionally). That is, we draw $z_i \sim \mathcal{N}(\nu_i, (c_i\nu_i)^2)$, where c_i are the coefficients of variation from Table A.7 and ν_i is the official, published estimate of total children. We add privacy noise to the total number of children using the same mechanism as before, $\tilde{z}_i \sim z_i + \text{Laplace}\,(2/\epsilon)$. The privacy noise distribution is the same for both x_i and z_i, in every school district. With these assumptions, the data deviations added to the total number of children are greater in magnitude than the noise added to the number of children in poverty, so the privacy deviations are even smaller in comparison (Figure A28).

Because the estimate of total children is only used to partially determine grant eligibility, the additional impact of also noising this variable is small (Figures A29 and A30). At $\epsilon = 0.1$, also noising the estimate of total children increases the likelihood that districts near the thresholds change eligibility, but does not noticeably change RMSE (likely because RMSE weights in large districts higher, and large districts are far away from the eligibility thresholds). Noising this variable does noticeably increase disparities for certain groups that tend to concentrate in districts near the eligibility thresholds.

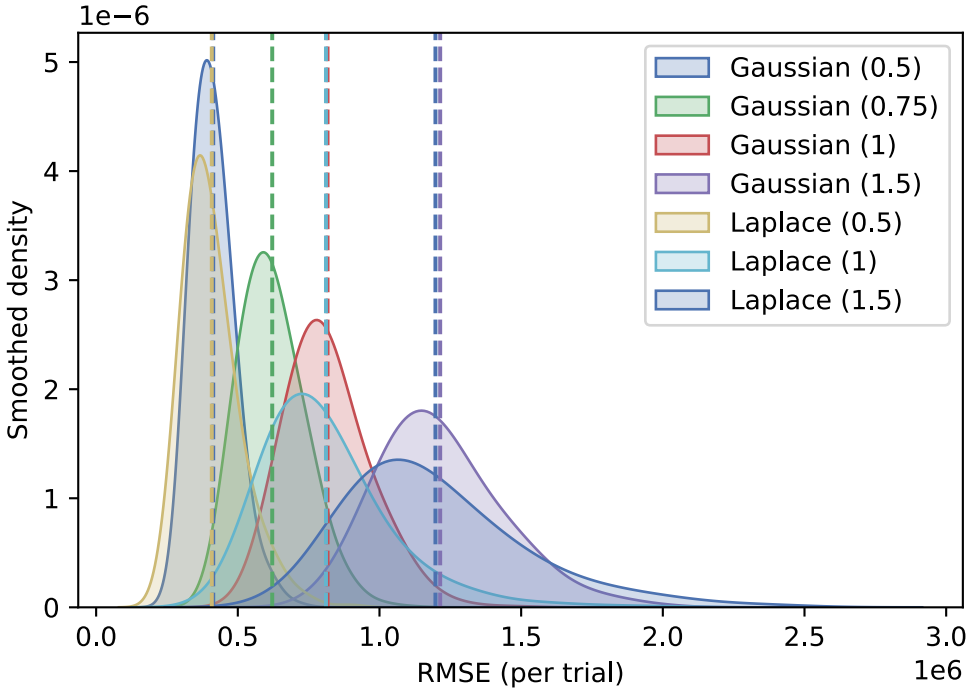


Figure A25: Distribution of root mean squared error under different data error distributions. The standard deviation of the sampling distributions was scaled by the coefficients in parentheses. Dotted lines depict averages.

Total Population of School District	Median CV
0-2,500	0.67
2,500-5,000	0.42
5,000-10,000	0.35
10,000-20,000	0.28
20,000-65,000	0.23
65,000 and up	0.15

Table A.7: Median coefficients of variation (CVs) for poverty estimates, reproduced from (US Census Bureau, 2020; Maples, 2008).

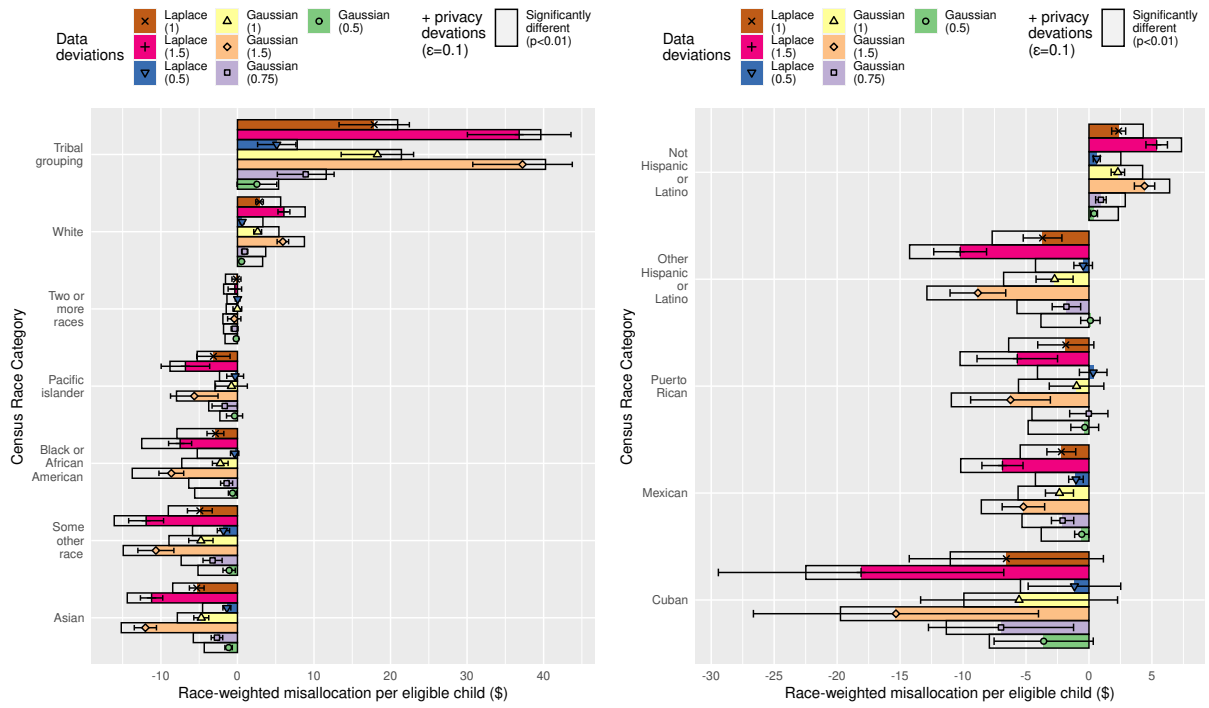
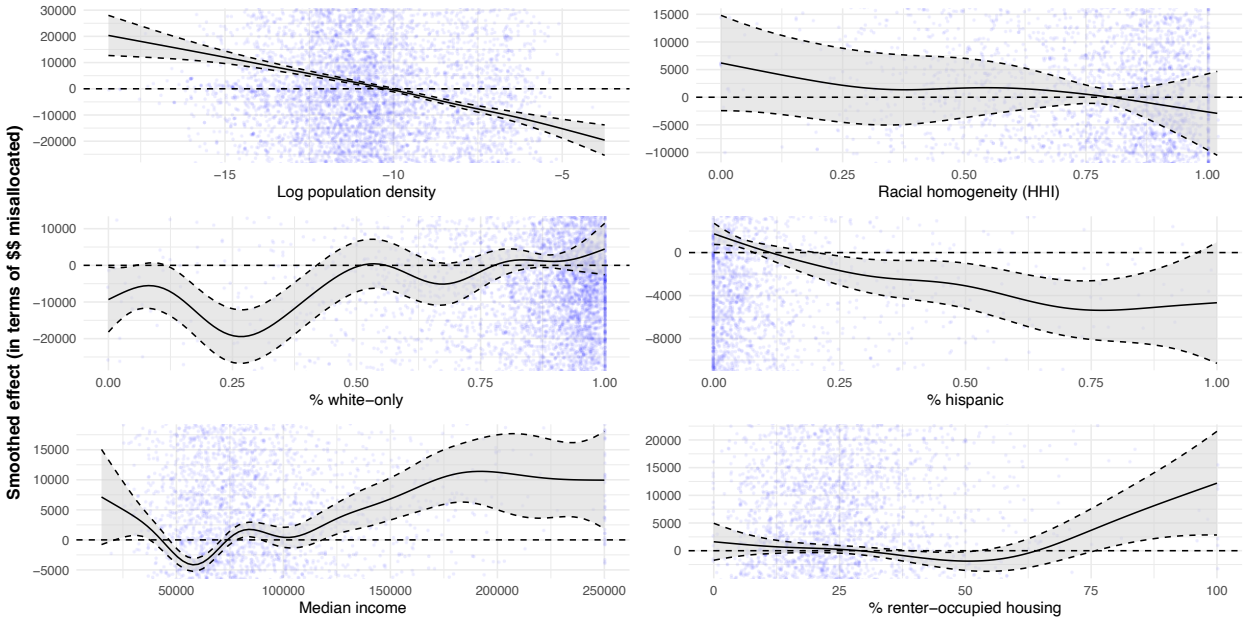
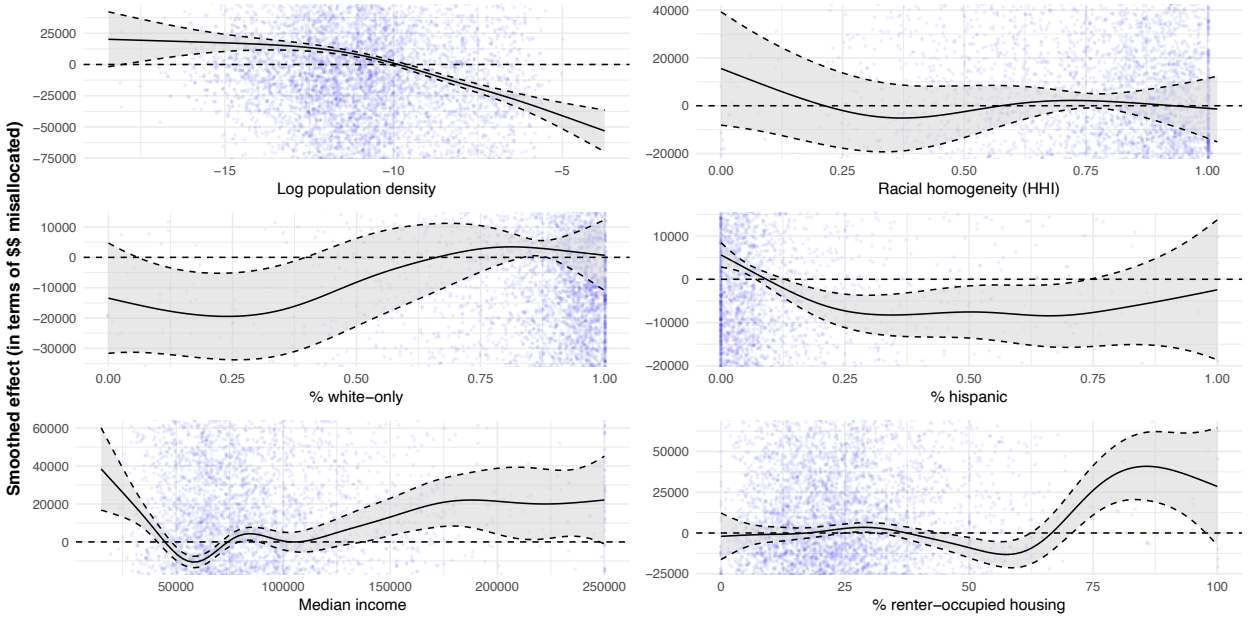


Figure A26: Race- (left) and ethnicity- (right) weighted misallocation under various amounts of either Gaussian or Laplace data error, averaged over 1,000 trials. The standard deviation of the sampling distributions was scaled by the coefficients in parentheses. A black outline indicates the marginal change in misallocation due to injected noise, drawn from Laplace mechanism with $\epsilon = 0.1$. Error bars span a 90% normal confidence interval. The additional impact of privacy deviations is significant ($p < 0.01$) for all groups, according to a two-sample z-test.

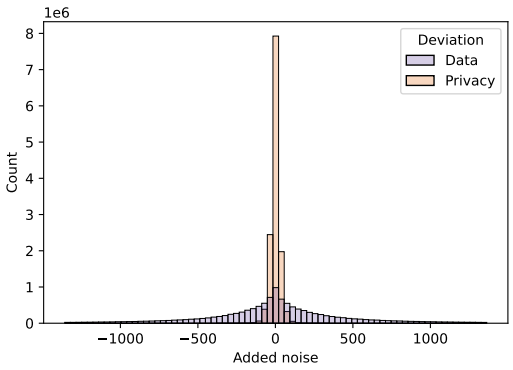


(a) Gaussian noise, standard deviation scaled by 0.5.

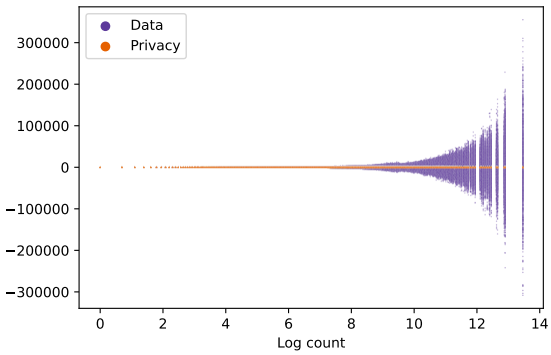


(b) Gaussian noise, standard deviation scaled by 1.5.

Figure A27: Effects under lower and higher data error. Depicts model-smoothed misallocation (from both data error and injected noise) by covariates, with 95% confidence interval in gray.



(a) Distribution of noise added to total children.



(b) Noise added to total children, plotted by total children.

Figure A28: Data and privacy deviations when the total number of children is also noised.

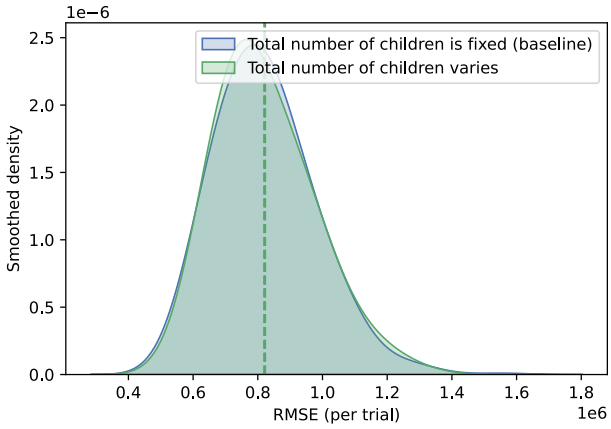


Figure A29: Distribution of root mean squared error when the total number of children is also noised. Dotted lines depict averages.

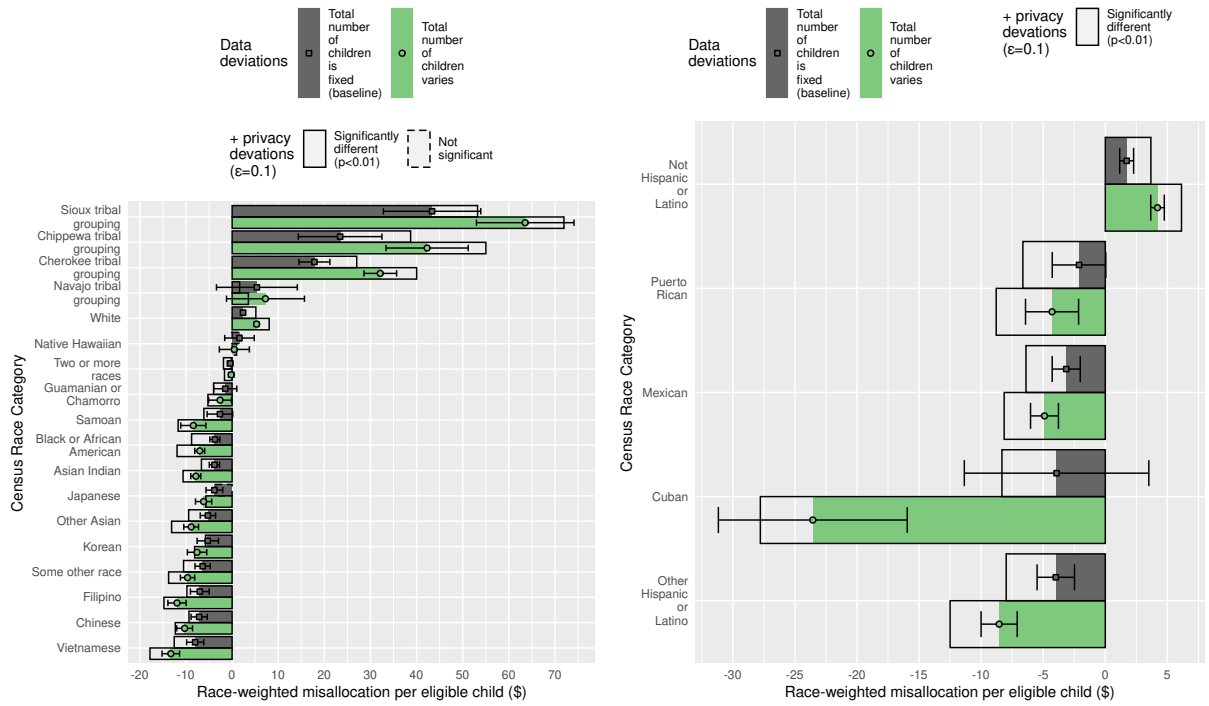
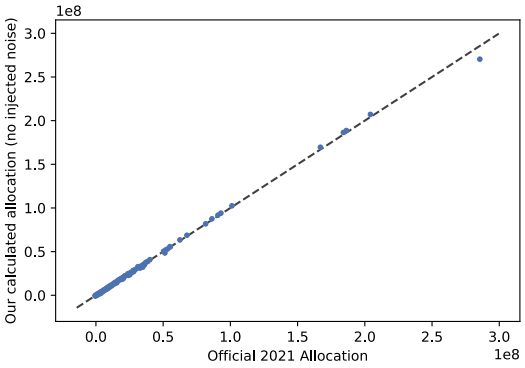


Figure A30: Race- (left) and ethnicity- (right) weighted misallocation when the total number of children is also noised, averaged over 1,000 trials. A black outline indicates the marginal change in misallocation due to injected noise, drawn from Laplace mechanism with $\epsilon = 0.1$. Error bars span a 90% normal confidence interval. Dashed lines indicate a statistically insignificant difference in race-weighted misallocation after privacy deviations are added, using a two-sample z-test.

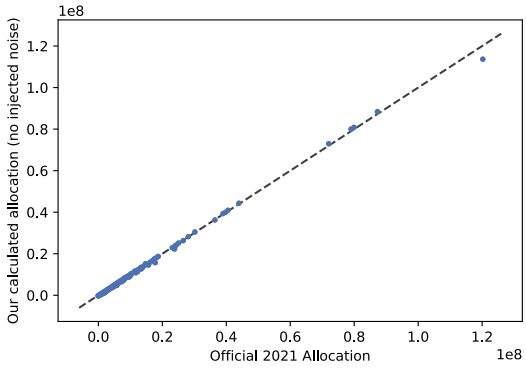
A.8 Comparison to Official Title I Allocations

To confirm that our algorithm for the Title I allocation process matches the algorithm used by the Department of Education, we compared our calculated allocation amounts after all provisions to the official figures released by the Department of Education's Office of Elementary and Secondary Education to state administrators (Rooney, 2021). The Dept. of Ed. releases three versions of the figures (preliminary, final, and final-revised); we use the version that most closely matches our calculations (the preliminary figures, because we are using the raw data sources before they have been updated by state administrators). Besides EFIG grants, we also leave out another approximately $1.3 billion in Title I LEA funding: $98,548,579 allocated for Part D Subpart 2, a provision that provides grants to districts with many children in juvenile detention; $728,460 allocated on behalf of county balances, occupied areas not assigned to a school district; and funds to U.S. territories, for which estimates were not included in our dataset.

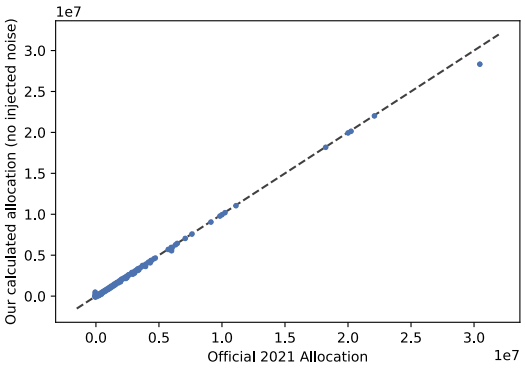
Our method approximates the official figures closely: for the average district, our estimate has an average absolute error of $20,558, with an RMSE of $162,915 (the average official allocation is about $1.2 million). There is very little systematic error in our replication (Figure A31)—with the exception of one outlier, the Los Angeles Unified School District. Most importantly, there is a strong linear correlation between our calculations and the official allocations. We therefore expect findings based on our replication of the Title I procedure to generally apply to the official implementation used by the Department of Education.



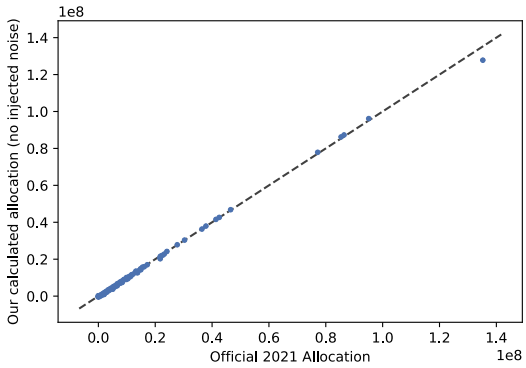
(a) Grants in total.



(b) Basic grants.



(c) Concentration grants.



(d) Targeted grants.

Figure A31: Allocation amounts calculated by our replication code (with no deviations applied), compared to the official figures. The dashed line denotes plots the frontier where our calculation and the official figures are equal.

Appendix B

Estimating Research Impacts of Statistical Uncertainty and Privacy

B.1 Additional Definitions

Definition 6 (Dwork & Roth, 2013)**.** The *Laplace distribution* (centered at 0) with scale b is the distribution with probability density function:

$$Lap(x|b) = \frac{1}{2b} \exp\left(-\frac{|x|}{b}\right).$$

Definition 7 (Bun & Steinke, 2016)**.** Let P and Q be probability distributions on Ω. For $\alpha \in (1, \infty)$, the *Rényi divergence* of order α between P and Q is:

$$D_\alpha = \frac{1}{\alpha - 1} \log \mathbb{E}_{x \sim P}\left[\left(\frac{P(x)}{Q(x)}\right)^{\alpha-1}\right]$$

B.2 Additional Methods

B.2.1 Search Terms

For each data source, we reviewed all results with abstract or title containing one of the following keywords: "state", "county", "city", "municipality", "district", "province", "block", "commuting zone", "statistical area", "prefecture", "tract", "college", "university", "school", "neighborhood".

B.2.2 Other Metrics

Significance match. Following Williams, Snoke, et al. (2024), we define significance match as the rate at which the significance level ($p < 0.01$, $p < 0.05$, $p < 0.1$, or insignificant) of the noisy estimate matches that of the original estimate, regardless of sign or magnitude.

Sign match. Following Williams, Snoke, et al. (2024), we define sign match as the rate at which the counterfactual estimate has the same sign as the original estimate, regardless of statistical significance.

Effective sample size. Following Williams, Barrientos, et al. (2024), we define the effective sample size of the counterfactual estimate as the number of observations n_{ESS} such that the counterfactual variance $\text{Var}(\hat{\beta^{\text{DP}}})$ would equal the original variance $\text{Var}(\hat{\beta})$:

$$n_{\text{ESS}} = n\frac{\text{Var}(\hat{\beta})}{\text{Var}(\hat{\beta^{\text{DP}}})}. \tag{B.1}$$

B.3 Additional Results

	(1)	(2)	(3)	(4)
(Intercept)	2.455 **	4.885 ***	3.260 ***	2.867 ***
	(0.902)	(0.156)	(0.198)	(0.046)
Log epsilon	-0.982 ***		-0.982 ***	-0.982 ***
	(0.006)		(0.006)	(0.006)
Gaussian mech. (zCDP)	-1.030 ***		-0.984 ***	-0.984 ***
	(0.148)		(0.106)	(0.106)
Log epsilon x Gaussian mech. (zCDP)	0.128 ***		0.128 ***	0.128 ***
	(0.002)		(0.002)	(0.002)
Log sensitivity		0.830 ***	0.868 ***	0.943 ***
		(0.070)	(0.041)	(0.025)
Log sensitivity x Gaussian mech. (zCDP)		0.126	0.050	0.050
		(0.094)	(0.027)	(0.027)
Study FE	No	No	No	Yes
N. obs.	39090	39090	39090	39090
R squared	0.660	0.318	0.976	0.987
F statistic	25237.581	9107.227	316006.286	59511.726
p value	0.000	0.000	0.000	0.000

*** $p < 0.001$; ** $p < 0.01$; * $p < 0.05$. Robust standard errors are clustered by study.

Table B.1: Impact of mechanism characteristics on log RMSD of noise added to each component statistical query.

	(1)	(2)	(3)
(Intercept)	0.752 ***	0.099	0.037
	(0.048)	(0.278)	(0.192)
Log coeff. of variation (a)	-0.076 ***	-0.076 ***	-0.076 ***
	(0.008)	(0.008)	(0.008)
Morris-Lysy construction	-0.229 ***	-0.229 ***	-0.229 ***
	(0.031)	(0.031)	(0.030)
Model degrees of freedom	-0.002	-0.003 *	-0.000
	(0.001)	(0.001)	(0.001)
Log regression sample size		0.051 **	0.035 **
		(0.019)	(0.012)
Noised ind./treatment var.		-0.351	-0.006
		(0.227)	(0.171)
Noised dep/outcome var.		-0.255 ***	-0.114
		(0.059)	(0.085)
Cmd: ivreg, ivreg2, xtivreg2, ivregress, ivreghdfe		0.128	0.051
		(0.127)	(0.060)
Cmd: other (arima, nbreg)		-0.399 *	-0.066
		(0.185)	(0.193)
Region: City/municipality/MSA/commuting zone		0.290	-0.072
		(0.150)	(0.115)
Region: State/district/province (1st division)		0.367 *	0.064
		(0.150)	(0.118)
Original effect size			0.424 **
			(0.137)
Claim: insignificant			0.439 ***
			(0.075)
Claim: non-zero upper/lower bound			-0.195
			(0.105)
Study FE	Yes	Yes	Yes
Query type controls	No	Yes	Yes
N. obs.	2275	2275	2275
R squared	0.458	0.483	0.548
F statistic	36.139	32.258	39.936
p value	0.000	0.000	0.000

*** $p < 0.001$; ** $p < 0.01$; * $p < 0.05$. Robust standard errors are clustered by study.

Table B.2: Impact of mechanism & result characteristics on epistemic parity. Implementation controls include the number of variables noised and the number of component queries. Dummies for Laplace mechanism (pure DP), "Cmd: reg, xtreg", "Claim: sig. positive/negative", and "Region: County/tract/block/district/prefecture (below 1st division)" excluded. Includes all experiments for additive data error (both constructions), $c \in \{10^{-3}, 10^{-2}, 0.05, 0.1, 0.2, 0.5\}$ with $b = 0$.

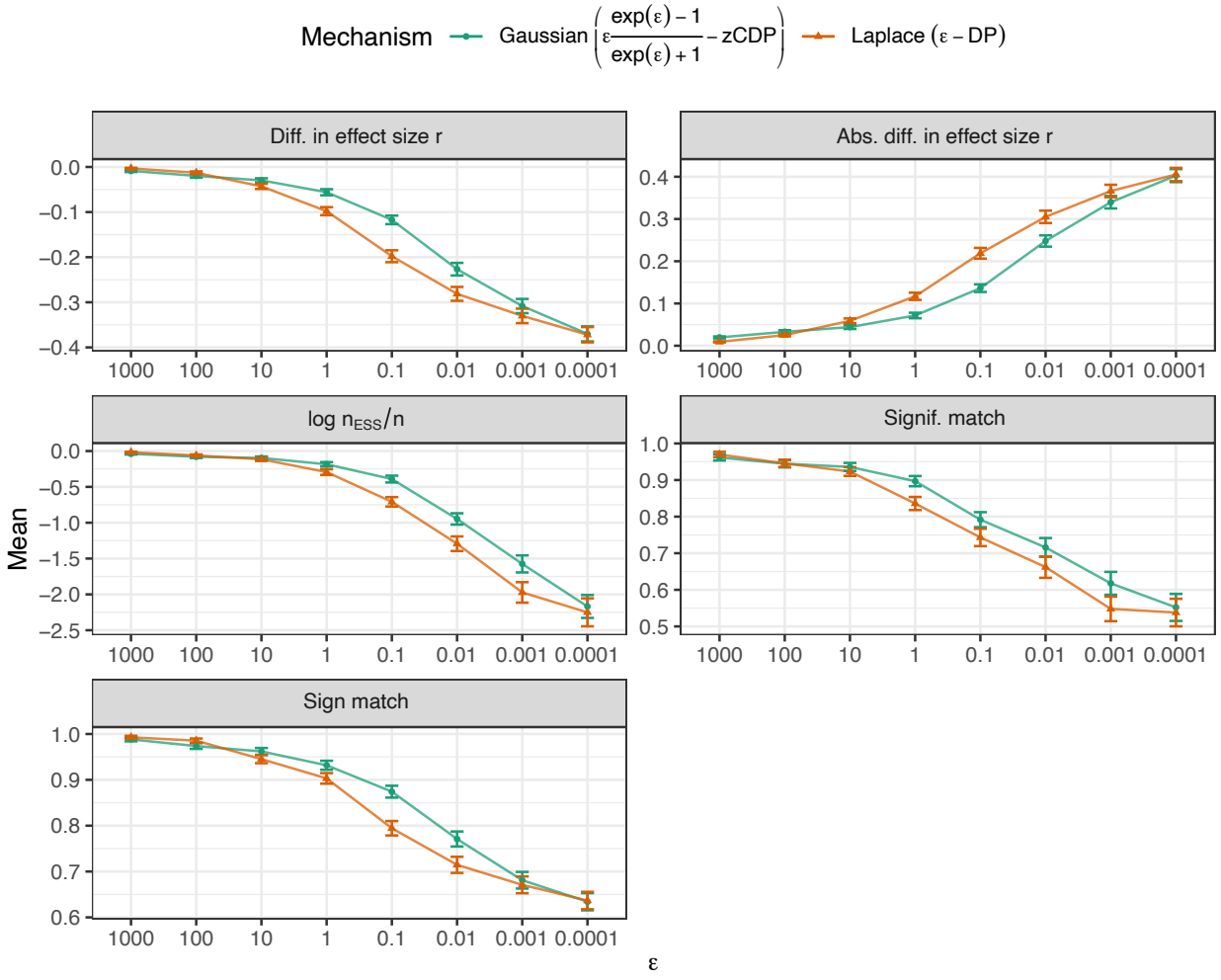


Figure B.3.1: Impacts of various settings of the privacy parameter ϵ over 10 runs of DP, averaged over all results. Lower values of the privacy parameter ϵ provide stronger privacy protection but require more injected noise; curators commonly use $0.1 \leq \epsilon \leq 10$.

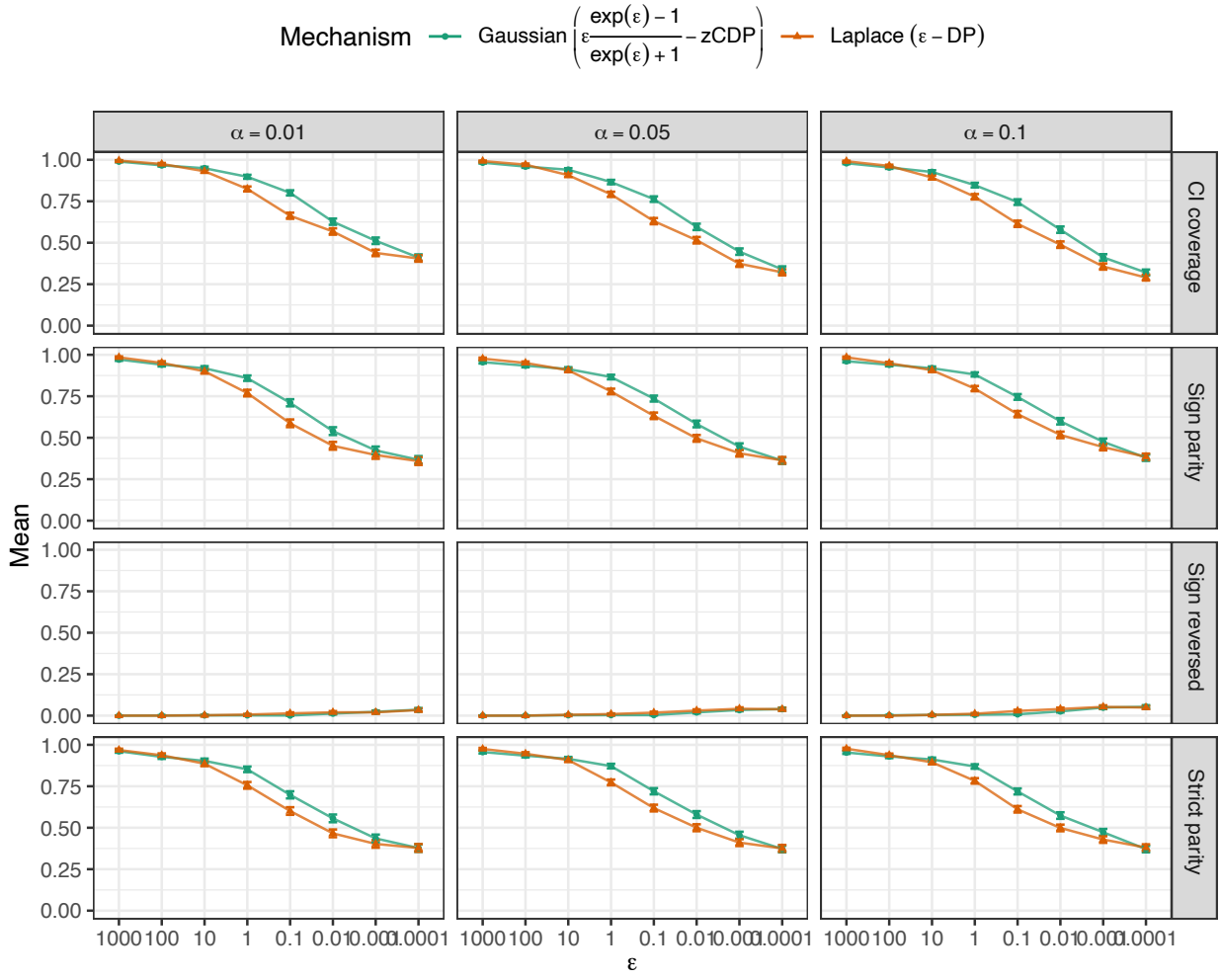


Figure B.3.2: Impacts of various settings of the privacy parameter ϵ at various significance levels α over 10 runs of DP, averaged over all results. Lower values of the privacy parameter ϵ provide stronger privacy protection but require more injected noise; curators commonly use $0.1 \leq \epsilon \leq 10$.

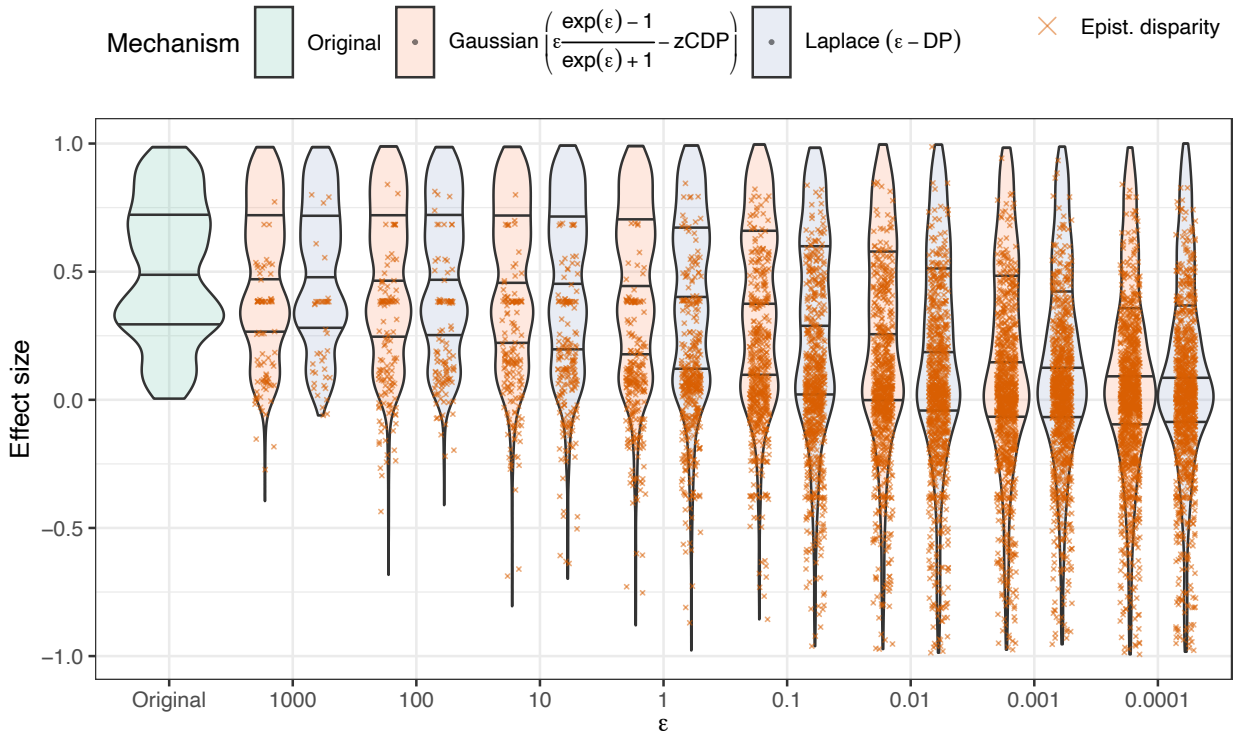


Figure B.3.3: Distribution of standardized effect sizes by study at different settings of the privacy parameter ϵ, each repeated 10 times. Lower values of the privacy parameter ϵ provide stronger privacy protection but require more injected noise; curators commonly use $0.1 \leq \epsilon \leq 10$.

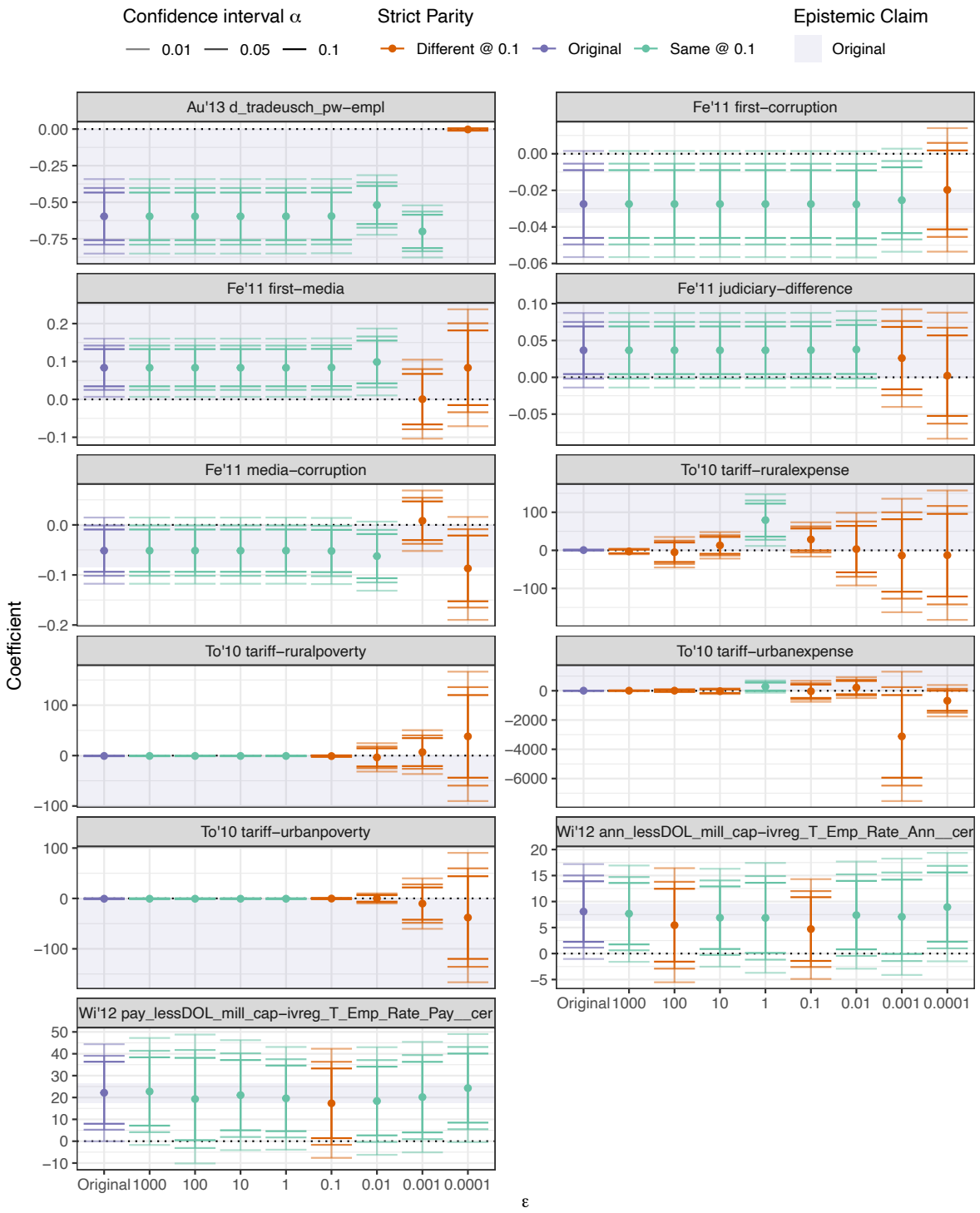


Figure B.3.4: Estimates from the first run of the ϵ-DP Laplace mechanism for a sample of studies (Autor et al., 2013; Ferraz & Finan, 2011; Topalova, 2010; Wilson, 2012). Epistemic parity colored based on confidence level $\alpha = 90$. The shaded region indicates the parity region defined by the original epistemic claim (e.g., that the coefficient is greater than zero).

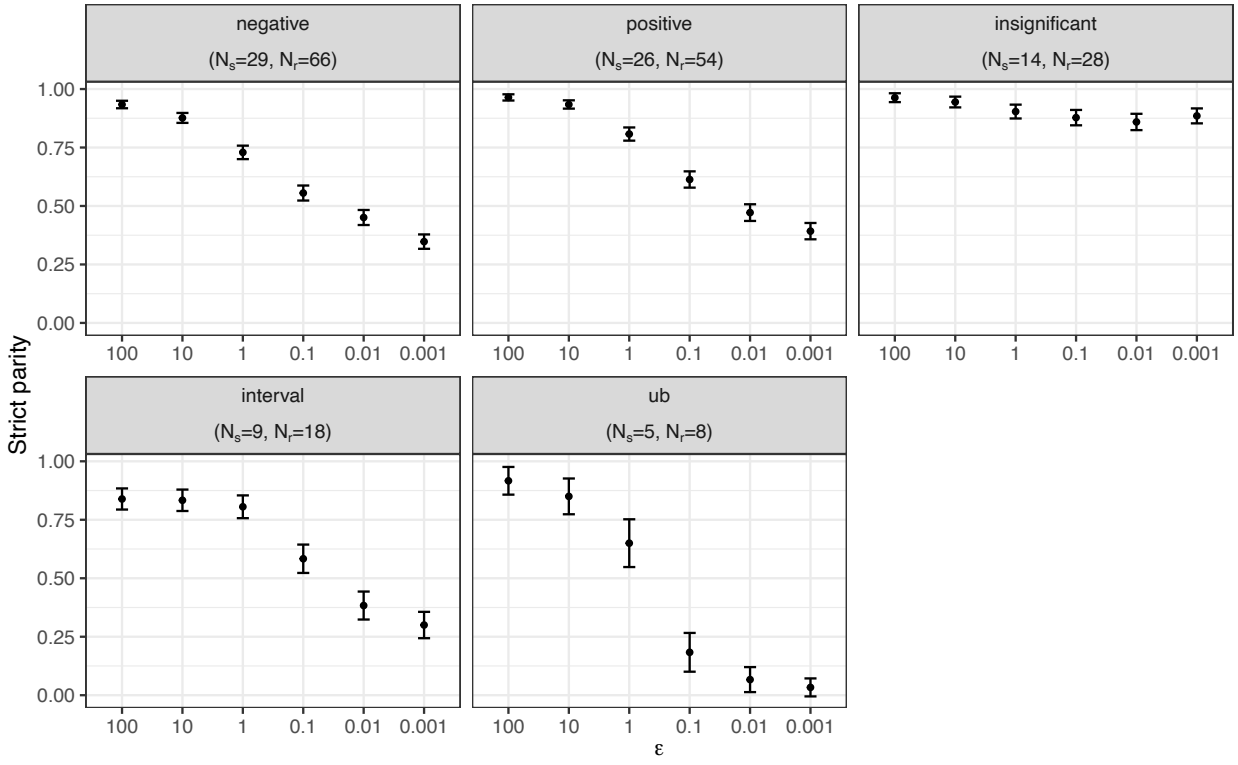


Figure B.3.5: Average epistemic parity vs. the privacy parameter ϵ across different types of epistemic claims ("ub" indicates an upper bound other than zero) over 10 runs of the ϵ-DP Laplace mechanism. Epistemic parity averaged across noised statistics. Lower values of the privacy parameter ϵ provide stronger privacy protection but require more injected noise; curators commonly use $0.1 \leq \epsilon \leq 10$. N_s is the number of studies in each category; N_r is the number of results.

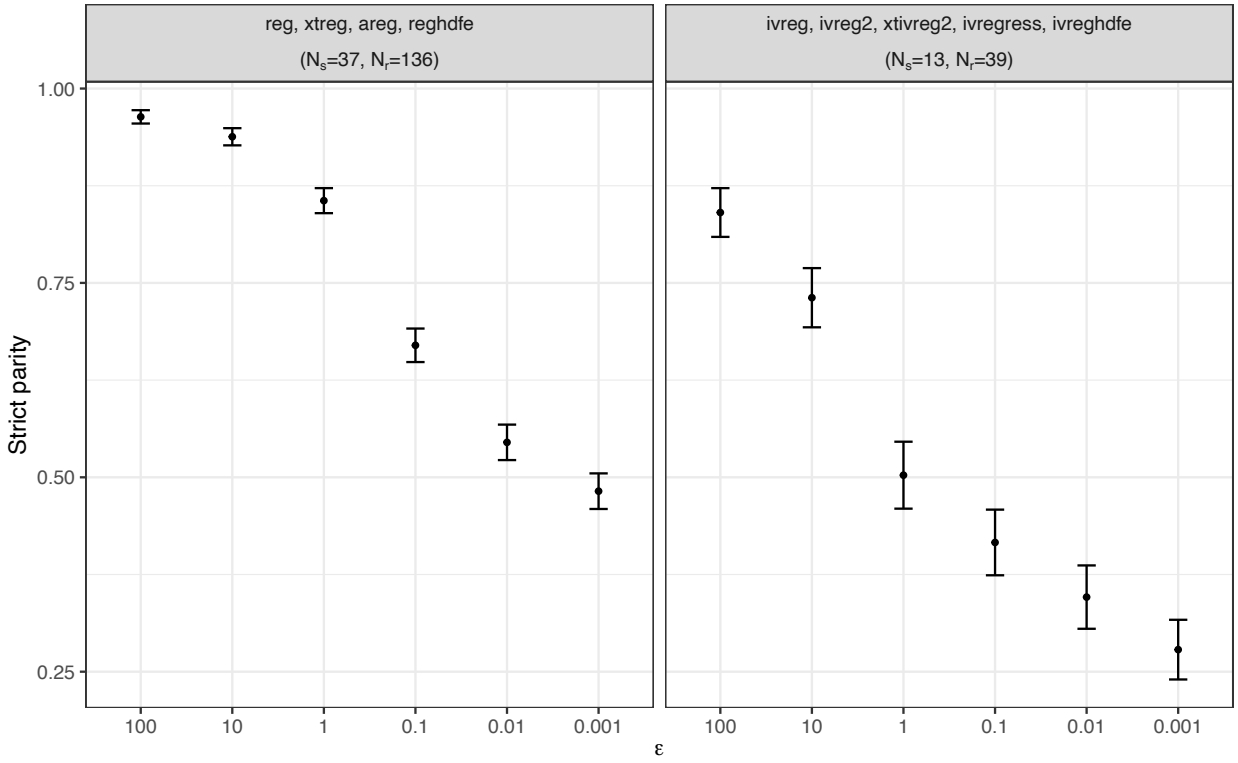


Figure B.3.6: Average epistemic parity for different Stata regression commands and different values of the privacy parameter ϵ over 10 runs of the ϵ-DP Laplace mechanism. Epistemic parity averaged across noised statistics. Lower values of the privacy parameter ϵ provide stronger privacy protection but require more injected noise; curators commonly use $0.1 \leq \epsilon \leq 10$. N_s is the number of studies in each category; N_r is the number of results.

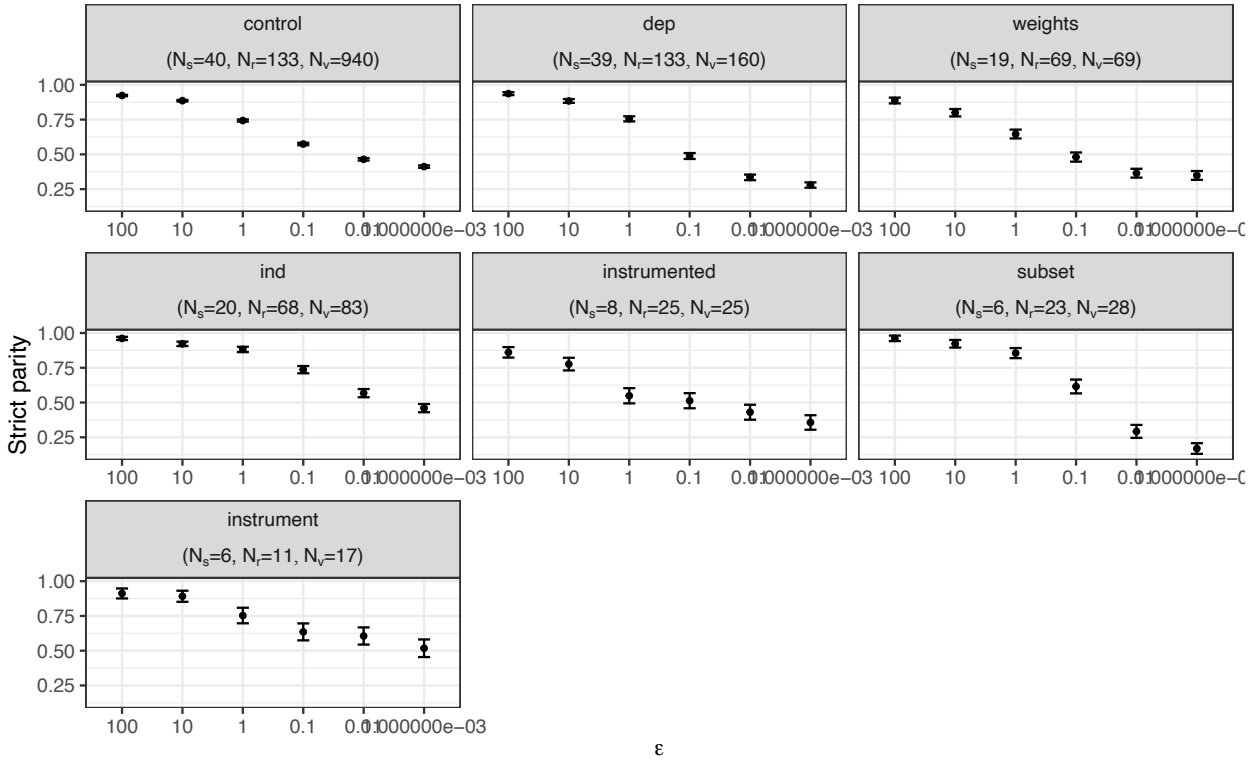


Figure B.3.7: Average epistemic parity vs. the privacy parameter ϵ over 10 runs of the ϵ-DP Laplace mechanism. Epistemic parity averaged across noised statistics and grouped by the role of each statistic in the regression analysis. Lower values of the privacy parameter ϵ provide stronger privacy protection but require more injected noise; curators commonly use $0.1 \leq \epsilon \leq 10$. N_s is the number of studies in each category; N_r is the number of results; N_v is the number of variables.

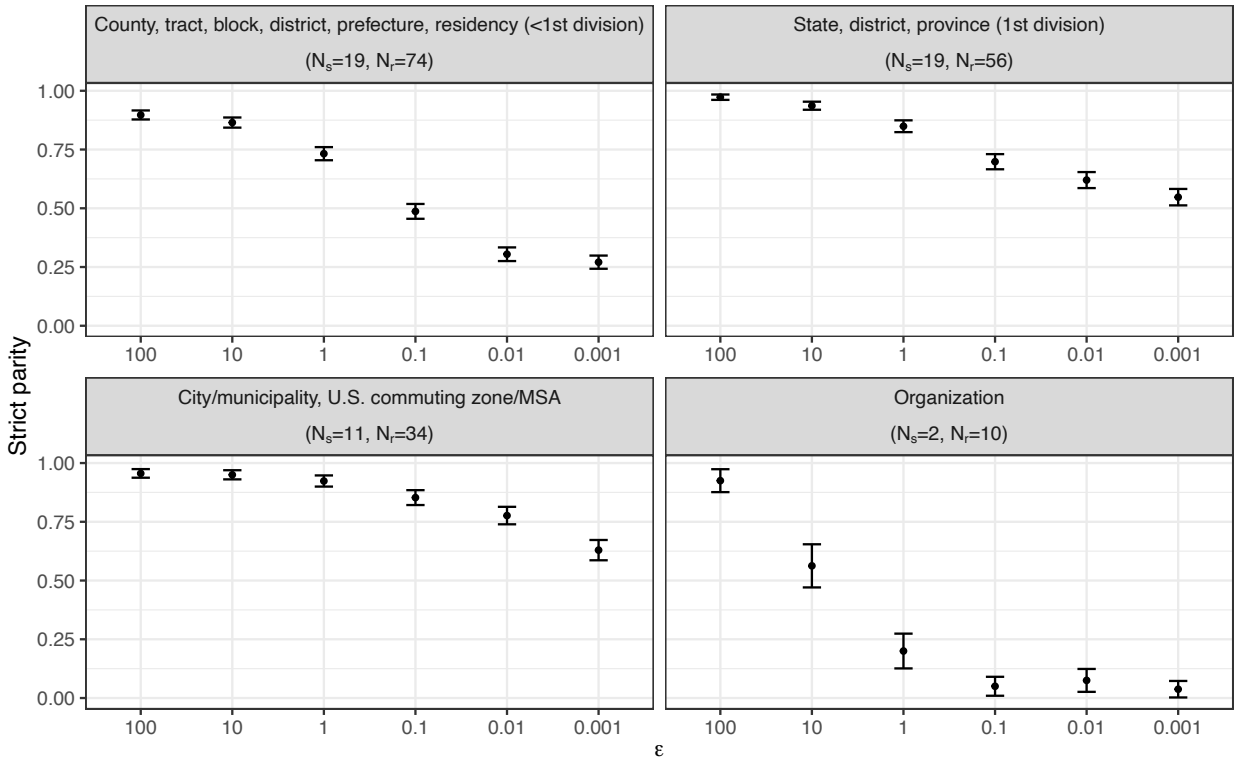


Figure B.3.8: Average epistemic parity for different aggregation levels and different values of the privacy parameter ϵ over 10 runs of the ϵ-DP Laplace mechanism. Epistemic parity averaged across noised statistics. Lower values of the privacy parameter ϵ provide stronger privacy protection but require more injected noise; curators commonly use $0.1 \leq \epsilon \leq 10$.

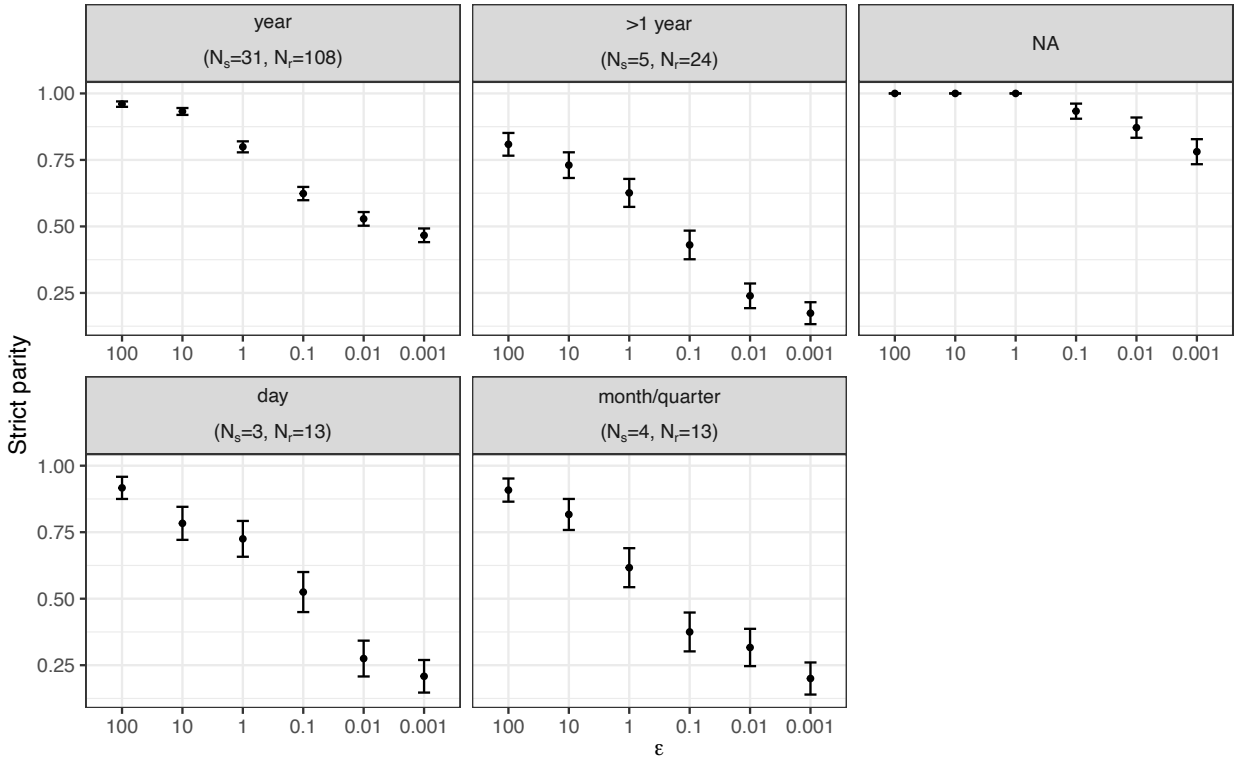


Figure B.3.9: Average epistemic parity for different aggregation levels and different values of the privacy parameter ϵ over 10 runs of the ϵ-DP Laplace mechanism. Epistemic parity averaged across noised statistics. NA indicates aggregations with no time index.Lower values of the privacy parameter ϵ provide stronger privacy protection but require more injected noise; curators commonly use $0.1 \leq \epsilon \leq 10$.

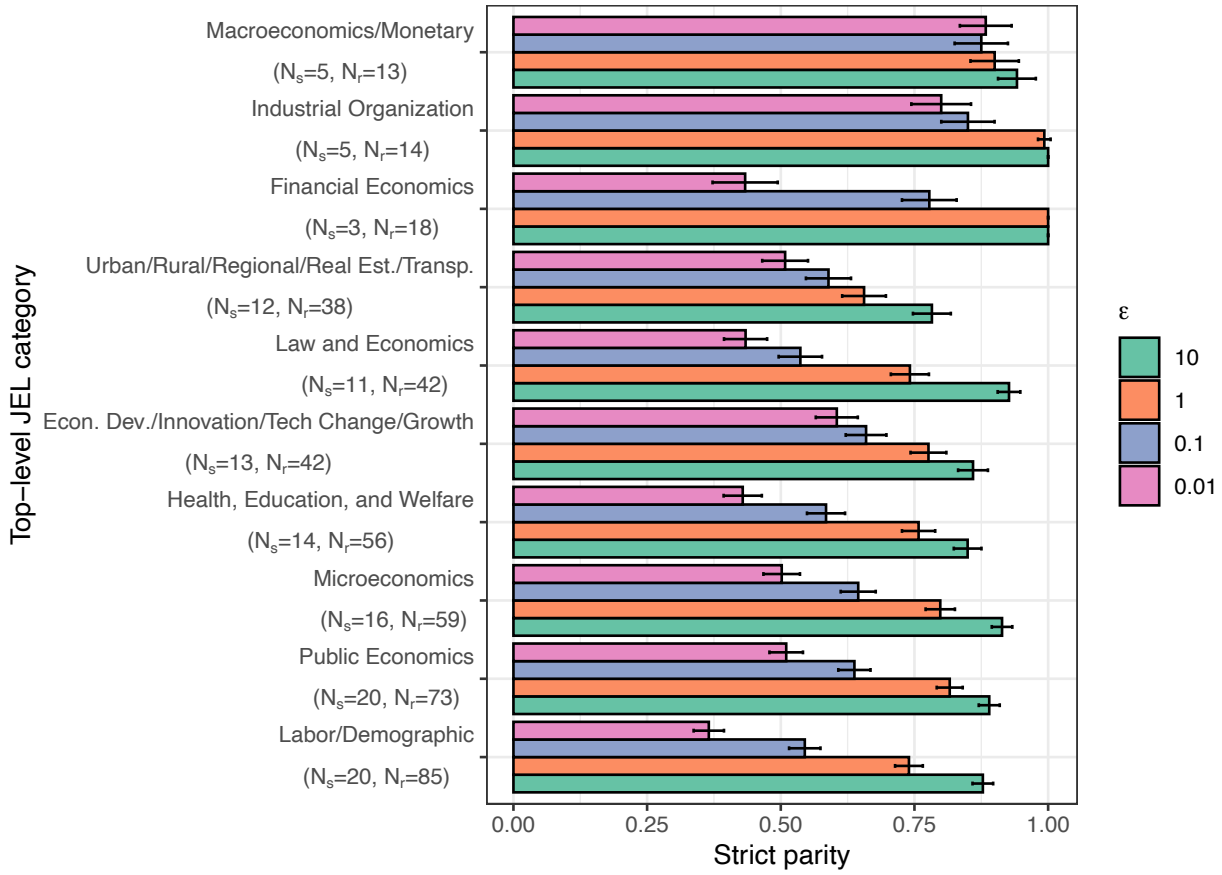


Figure B.3.10: Average epistemic parity for different JEL categories and different values of the privacy parameter ϵ over 10 runs of the ϵ-DP Laplace mechanism. Studies may appear in multiple JEL categories. Epistemic parity averaged across results. Excluding categories with less than 10 results. Lower values of the privacy parameter ϵ provide stronger privacy protection but require more injected noise; curators commonly use $0.1 \leq \epsilon \leq 10$.

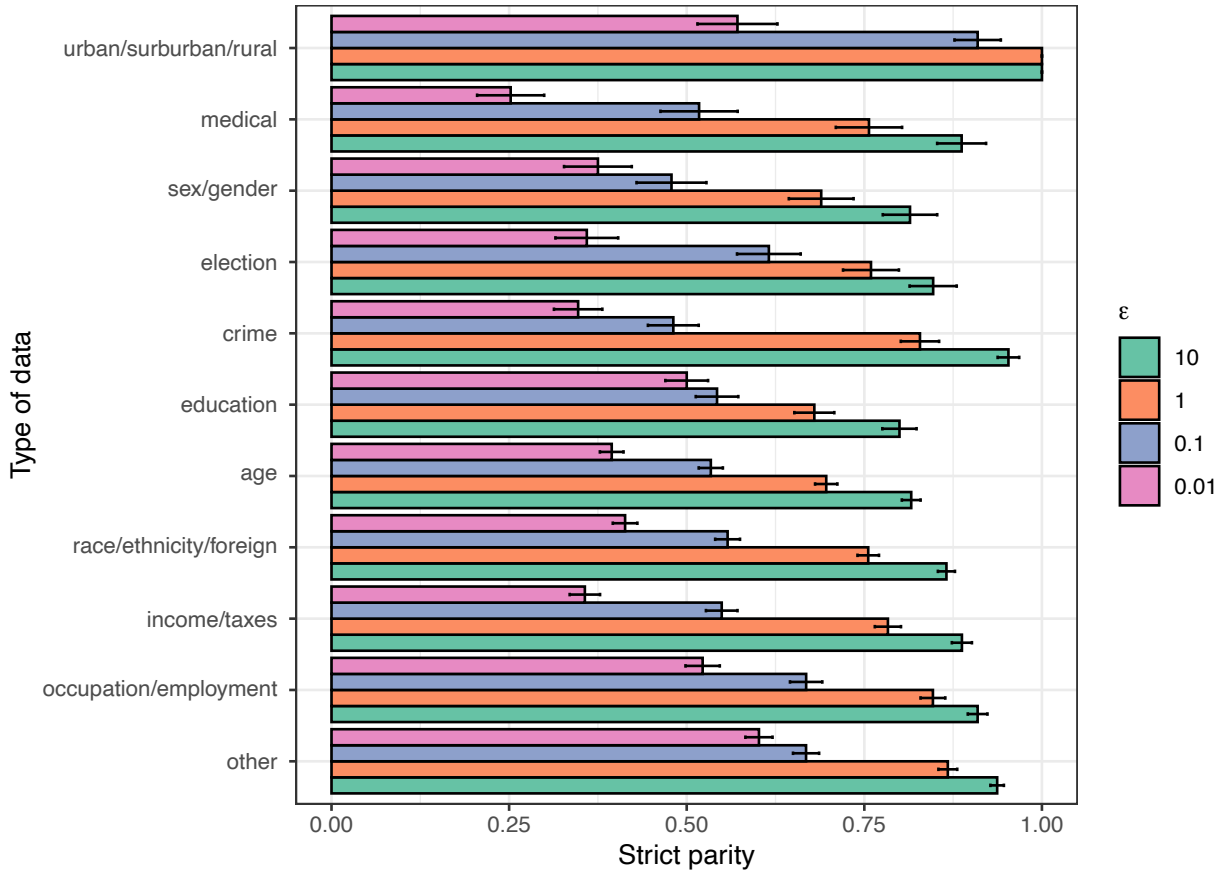


Figure B.3.11: Average epistemic parity for different types of statistics and different values of the privacy parameter ϵ over 10 runs of the ϵ-DP Laplace mechanism. Epistemic parity averaged across noised statistics. Excluding categories with less than 10 results. Lower values of the privacy parameter ϵ provide stronger privacy protection but require more injected noise; curators commonly use $0.1 \leq \epsilon \leq 10$.

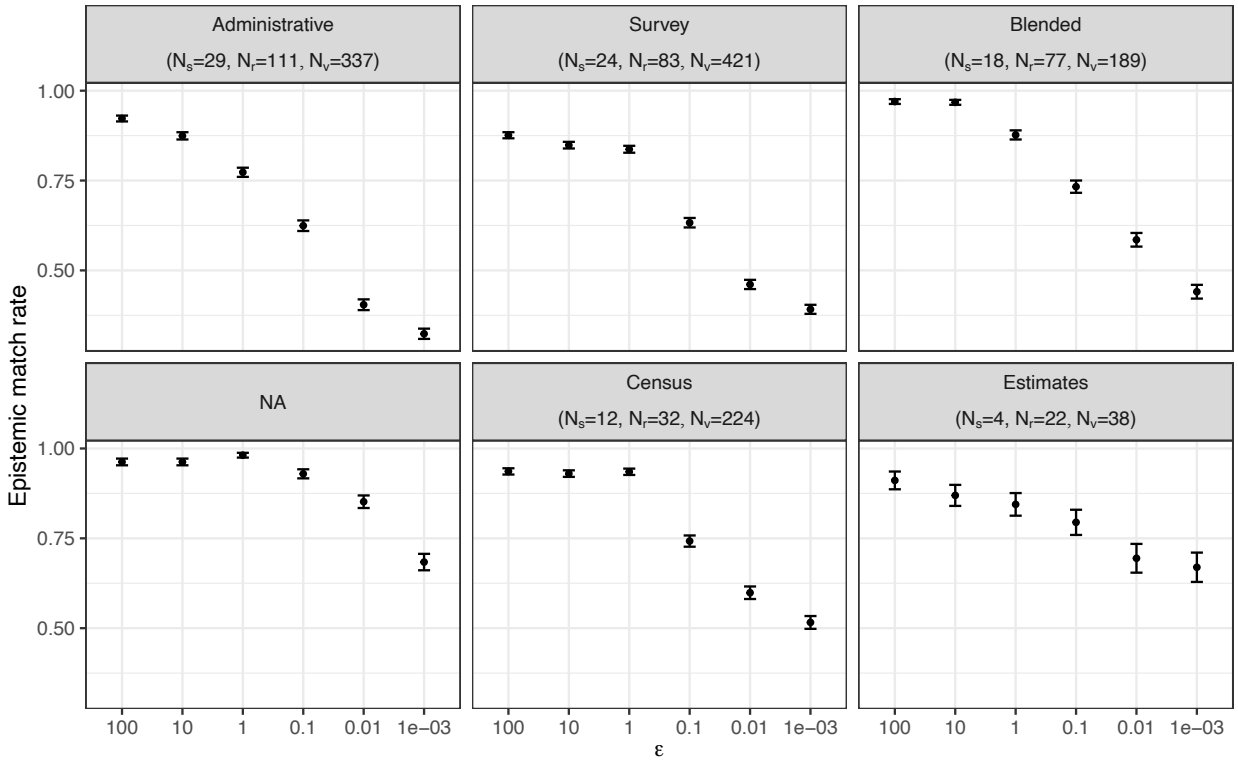


Figure B.3.12: Average epistemic parity for different types of data sources and different values of the privacy parameter ϵ over 10 runs of the ϵ-DP Laplace mechanism. Epistemic parity averaged across noised statistics. NA indicates unknown or unclear data products. Lower values of the privacy parameter ϵ provide stronger privacy protection but require more injected noise; curators commonly use $0.1 \leq \epsilon \leq 10$. N_s is the number of studies in each category; N_r is the number of results; N_v is the number of variables.

Appendix C

Algorithmic Decoupling in 'Privacy-Preserving' Analytics

C.1 Reflexivity Statement

In the tradition of ethnography and qualitative methods more generally, we tried to remain aware of our own cultural and epistemic perspectives in our work. Reflexivity is not the same as accounting and correcting for bias—research is not a "view from nowhere" (Haraway, 1988). The first author (RS) is a Ph.D. student with training in computer science and economics with experience researching the social implications of algorithms, including differentially private mechanisms. The second author (A.A.) is a tenured professor with years of research experience studying the economics of privacy and privacy-enhancing technology. Both authors are based in the United States and work at elite Western universities. Both authors have worked for tech companies in the past, and R.S. receives stipend support from Meta. Neither author has experience building and deploying PPA—rather, PPA systems have been the subject of our largely empirical research. We do not have strong political or intellectual convictions about the value or future of PPA—this study grew out of our curiosity to understand how these technologies fit into privacy practice. However, both authors have an interest and stake in the protection of online privacy. In our analysis, we considered how our backgrounds might lead us toward certain framings of our results or close us off to certain possibilities—for example, that the adoption of PPA is not inevitable. We also considered how it may also lead our interviews toward certain topics (e.g. economic trade-offs or privacy scholarship) more than others. Though we made an effort to recruit and consider perspectives other than our own, we primarily leverage our backgrounds to "study up" (Nader, 1972) and critically analyze culturally hegemonic institutions close to ourselves.

C.2 Interview Guide

The final version of our semi-structured interview guide can be viewed here. Participants were compensated with a $30 gift card or donation to a charity. Each interview was recorded (with

participant consent and IRB approval) and transcribed verbatim by the first author.

C.3 Recruitment

In the first period of data collection between July 2021 and January 2022, we calibrated each successive wave of recruitment ($N = 9$ contacted July–August, $N = 11$ September–October, $N = 3$ November) to examine adoption settings and other theoretical interests we had yet to explain with previous data (e.g., in the second wave, we included privacy-focused startups). (e.g., the process of interpreting adoption drivers into specific designs)

Because privacy is acutely important to marginalized groups (Skinner-Thompson, 2020), we deliberately aimed to include those perspectives in our sample and explicitly requested referrals to participants from underrepresented backgrounds. Still, Our sample was predominantly American (90%), white (70%), non-Hispanic (85%), heterosexual (65%) and cisgender male (55%), based on participants who chose to self-identify for each category (22 for race, ethnicity, and gender; 17 for sexuality). Our sample is similar in racial diversity to the U.S. high-technology workforce, but more diverse in gender, sexuality, and Hispanic origin (U.S. Equal Employment Opportunity Commission, 2016).

C.4 Additional Figures

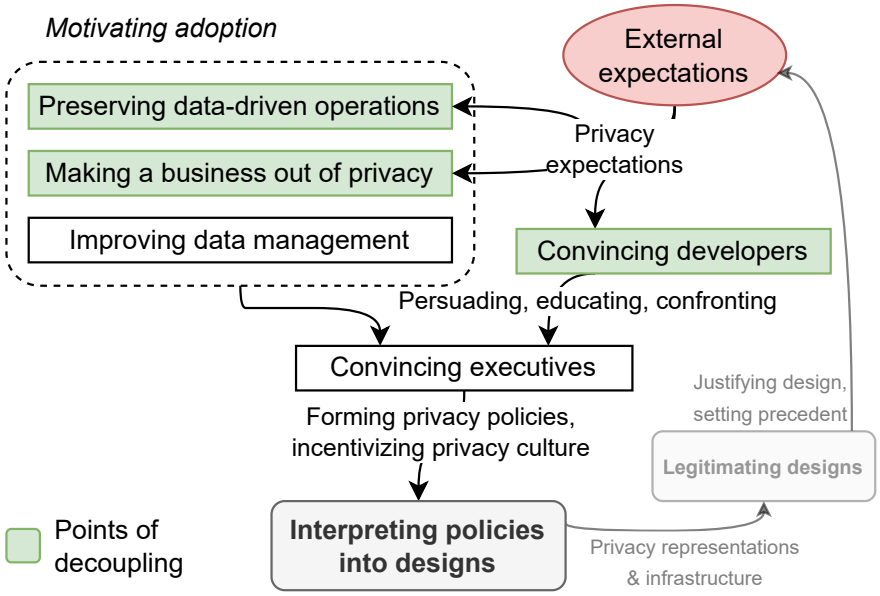


Figure C.4.1: Driving adoption. Second-order concepts are boxed. Key processes associated with managerial mediation (and possibly decoupling) or expert mediation are highlighted.

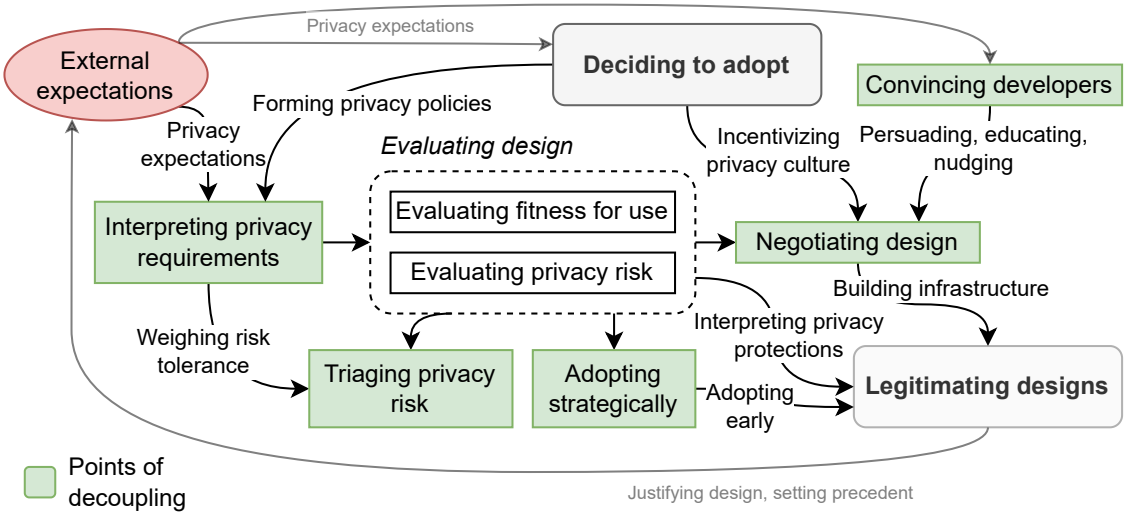


Figure C.4.2: Interpreting drivers into designs. Second-order concepts are boxed. Key processes associated with managerial mediation (and possibly decoupling) or expert mediation are highlighted.

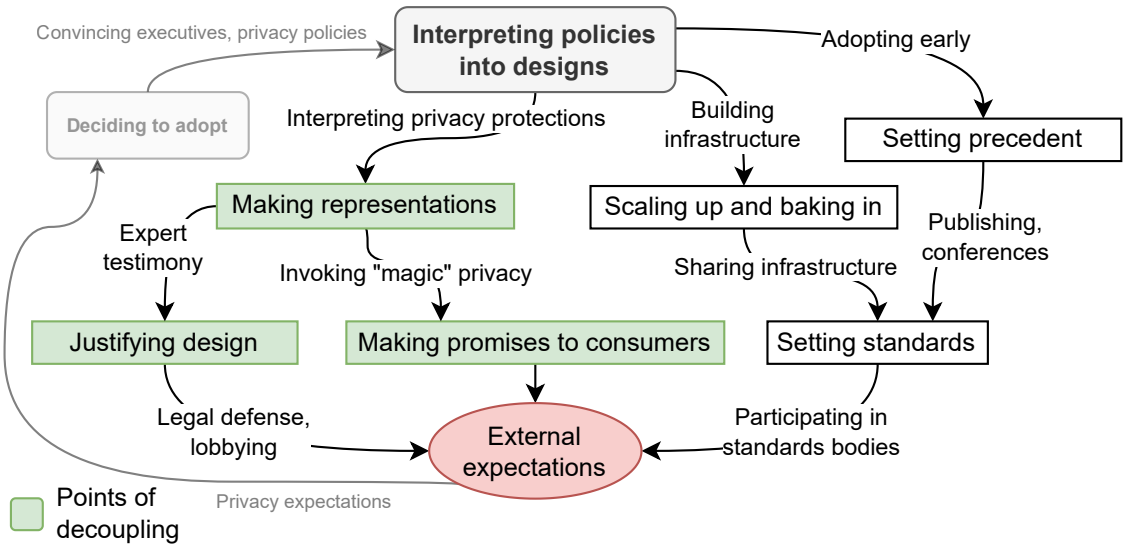


Figure C.4.3: Legitimating design. Second-order concepts are boxed. Key processes associated with managerial mediation (and possibly decoupling) or expert mediation are highlighted.

Appendix D

Measuring Social Biases in Unsupervised Image Generation

D.1 Attribute Words

We selected the following words for high/low valence and high imagery from the scores collected by Bellezza et al. (1986) in a laboratory experiment. A specific algorithm for systematically selecting words with high imagery and extreme valence is included in our code at github.com/ryansteed/ieat.

Positive words: baby, ocean, beach, butterfly, gold, rainbow, sunset, money, diamond, flower, sunrise
Negative words: devil, morgue, slum, corpse, coffin, jail, roach, funeral, prison, vomit, crash

D.2 Stimuli collection procedure

We collected n images for each verbal stimulus using the following procedure:

1. If there is a CIFAR-100 category corresponding to the stimulus, we selected a random sample of n images from that category in CIFAR (Krizhevsky, 2009).[1]

2. Otherwise, we searched for the verbal stimuli verbatim on Google Image Search in private Chrome window with SafeSearch off on September 5th, September 18th and October 1st, 2020. We accepted the first n results of the search meeting the following criteria:[2]

[1] Because the verbal stimuli are very specific, only 3 of over 105 IAT verbal stimuli appear in CIFAR-100; the rest were collected with Google Image Search.

[2] A few words were too abstract to be easily visualized. These words are listed in Appendix D.2 with a sample size of 0.

- Includes only the object, person, or scene specified by the stimulus.[3]
- For objects and people, has a plain background, to avoid including confounding scenes or objects.[4]
- Has no watermark or other text. Watermarks and text could confound the verbal stimulus being represented.
- Shows a real object, person, or scene - is not a cartoon or sketch. ImageNet does not include a great quantity of cartoons or sketches, so we do not expect our models to generalize well to these kinds of objects/scenes (Recht et al., 2019).

3. If no images in the first 50 results from the verbatim search met these criteria, we added a clarifying search term (e.g. "biology lab" instead of "biology").
4. Crop each image squarely (iGPT accepts only square images as input), centering the object or person of interest to ensure the entire object, person, or scene is included in the image.

For every verbal stimulus used to collect image stimuli for the verbal and mixed-mode IATs, we recorded the verbal stimulus (word or phrase), search terms used to collect images, and the number of images collected in a CSV file along with our code at github.com/ryansteed/ieat.

D.3 Disparate Bias Across Model Layers

Model design choices might also have an effect on how social bias is learned in visual embeddings. We find that embedded social biases vary not only between models pre-trained on the same data but also within layers of the same model. In addition to the high quality embeddings extracted from the middle of the model, we tested embeddings extracted at the next-pixel logistic prediction layer of iGPT. This logit layer, when taken as a set of probabilities with softmax or a similar function, is used to solve the next-pixel prediction task for unconditional image generation and conditional image completion (Chen, Radford, et al., 2020).

Table D.1 reports the iEAT tests results for these embeddings, which did not display the same correspondence with human bias as the embeddings for image classification. We found that unlike the high quality embeddings, next-pixel prediction embeddings do not exhibit the baseline Insect-Flower valence bias and only encode significant bias at the 10^{-1} level for the Gender-Science and Sexuality IATs.

To explain this difference in behavior, recall that the neural network used in iGPT learns different levels of abstraction at each layer; as an example, imagine that first layer encodes

[3]Some verbal stimuli (e.g. "salary") are difficult to express verbally without the use of symbols (e.g. a picture of cash). In these cases, we collected only the first image ($n = 1$) that meets the criteria, preferring image stimuli corresponding to other, more visual cues and representations.

[4]If no images with white or gray backgrounds appeared in the first 50 results, we searched for "[stimulus] + {white, plain} background."

lighting particularly well, while the second layer begins to encode curves. The contradiction between biases in the middle layers and biases in the projection head are consistent with two previous findings: 1) bias is encoded disparately across the layers of unsupervised pre-trained models, as Bommasani et al. (2020) show in the language domain; 2) in transformer models, the highest quality features for image classification, and possibly also social bias prediction, are found in the middle of the base network (Chen, Radford, et al., 2020). Evidently, bias depends not only on the training data but also on the choice of model.

Table D.1: iEAT tests for the association between target concepts X vs. Y (represented by n_t images each) and attributes A vs. B (represented by n_a images each) in embeddings for iGPT next-pixel prediction. Association effect sizes d, colored by conventional small (0.2), medium (0.5), and large (0.8) size are reported alongside permutation p-values.

	X	Y	A	B	n_t	n_a	d	p
Age†	Young	Old	Pleasant	Unpleasant	6	55	0.38	0.38
Arab-Muslim	Other	Arab-Muslim	Pleasant	Unpleasant	10	55	0.06	0.42
Asian§	European American	Asian American	American	Foreign	6	6	0.25	0.36
Disability†	Disabled	Abled	Pleasant	Unpleasant	4	55	-0.65	0.76
Gender-Career	Male	Female	Career	Family	40	21	0.04	0.44
Gender-Science	Male	Female	Science	Liberal Arts	40	21	0.37	0.06
Insect-Flower	Flower	Insect	Pleasant	Unpleasant	35	55	-0.32	0.91
Native§	European American	Native American	U.S.	World	8	5	0.32	0.26
Race†	European American	African American	Pleasant	Unpleasant	6	55	-0.17	0.62
Religion	Christianity	Judaism	Pleasant	Unpleasant	7	55	0.29	0.30
Sexuality	Gay	Straight	Pleasant	Unpleasant	9	55	0.69	0.08
Skin-Tone†	Light	Dark	Pleasant	Unpleasant	7	55	0.42	0.36
Weapon§	White	Black	Tool	Weapon	6	7	-1.64	1.00
Weapon (Modern)	White	Black	Tool	Weapon	6	9	-1.19	0.98
Weight†	Thin	Fat	Pleasant	Unpleasant	10	55	-0.84	0.97

§ Originally a picture-IAT (image-only stimuli). † Originally a mixed-mode IAT (image and verbal stimuli).

Appendix E

Gaps and Opportunities in AI Audit Tooling

E.1 Reflections

We consider our own cultural and professional perspectives throughout the interviews and our analysis (Charmaz, 2014). In particular, we view our position as both external to prominent AI developers and deployments, but still situated within the Western AI industry as a critical consideration. Our project was financially supported by a prominent U.S.-based foundation, and all the authors are either graduate students or graduates of well-funded universities in the U.S. and Europe. We thus recognize that our analysis is primarily scoped to the U.S. and the E.U. and may not be representative of the global AI auditing landscape or appropriate for or informed by other contexts. Because we drew our initial sources from our own fieldwork as audit practitioners and used English search engines for theoretical sampling, our dataset consists primarily of English language tools from from Western organizations, and our taxonomy reflects our own particular position.[1] Also, although we intentionally defined tools as resources more broadly and attempted to leave space for non-technical solutions in our analysis and discussion, we struggled to avoid a techno-solutionist framing in our conclusions (Wong et al., 2023). We hope that future work will expand and look outside of the mostly technical, Western perspective emphasized in this work.

E.2 Glossary

Definition 8. Accountability: Bovens (2007) defines accountability as "a relationship between an actor and a forum, in which the actor has an obligation to explain and to justify his or her conduct, the forum can pose questions and pass judgment, and the actor may face

[1]We also attempted to run translated queries in non-English search engines (e.g., Baidu) and to add regional keywords (e.g., African) to our searches, but we were unable to find any additional tools with initial tests of these methods. It thus seems likely that non-English, non-Western tools exist that our theoretical sampling was unable to identify.

consequences." We use the term "accountability" in this legal-political sense, meaning to face consequential judgment for system behaviors and impacts that do not align with articulated expectations and standards.

Definition 9. AI Audit: Birhane et al. (2024) define an audit as "any independent assessment of an identified audit target via an evaluation of articulated expectations with the implicit or explicit objective of accountability". Independence, as outlined by Birhane et al. (2024), ensures the auditors (which may or may not belong to the same organization as the developers) are operationally distinct from the team that engineered the examined AI system, maintaining separation from the engineering process. Identified audit targets refer to concrete, specific objects of examination, ideally tied to real-world AI deployments or widely-used open-source algorithms or datasets, which serve as relevant proxies. Implicit or explicit accountability refers to audits designed to inform consequential judgments, measuring the deployment's behaviors and impacts against clearly articulated expectations.

Definition 10. AI Auditor: An entity executing an AI audit, which may be viewed as either internal or external to the organization developing and/or operating the audited system. (Note that auditor independence is a nuanced spectrum, and there is not always a sharp divide between internal and external auditing.

- **Internal Auditor:** Raji, Xu, et al. (2022) define an internal auditor as an entity executing an audit or investigation with some contractual relationship with the audit target. They typically seek to minimize corporate liability and test for compliance with corporate or industry-wide expectations. Internal auditors are often hired voluntarily or to meet regulatory mandates.
- **External Auditor:** Raji, Xu, et al. (2022) define an external auditor as an entity executing an audit or investigation without any contractual relationship with the audit target. They typically execute audits voluntarily with a broader mandate of identifying and minimizing the harm impacting their constituents.

Definition 11. AI Audit Tool: We use the term AI audit tool to refer broadly to software, interfaces, code, benchmarks, frameworks, and other artifacts used by auditors in the AI audit process. Audit tools include resources that support algorithmic analysis and inspection (e.g., benchmarks/datasets, documentation templates) as well as resources that support the assessment of internal and external expectations for institutions across stages of design and development.

Definition 12. Abandonment & disgorgement: In the context of algorithmic systems and regulatory enforcement, the terms algorithm abandonment (Johnson et al., 2024) and algorithm disgorgement (Li, 2022) refer to the destruction or abandonment of a system deemed harmful, unethical, or in violation of societal or regulatory standards. Abandonment refers to "an organization's decision to stop designing, developing, or using an algorithmic system due to its (potential) harms" (Johnson et al., 2024), while disgorgement further requires the deletion of both improperly obtained data and any machine learning models, algorithms, or outputs derived from such data (Li, 2022).

E.3 Additional Methods

E.3.1 Initial Sources

We drew our initial list of 143 tools from:

- tools mentioned by our interviewees;
- tools mentioned in previous surveys of fairness and other toolkits (Deng et al., 2022; Lee & Singh, 2021; Hickock, 2023)
- academic papers presenting new tools surfaced in a recent literature review of audit studies (Birhane et al., 2024); this literature review included audit studies from
  - the last five years of conference proceedings from: FAccT, AIES, EAAMO, AAAI, CVPR, ICWSM, WWW, WACV, EECV, and the ACL Anthology; also, the ACM Digital Library (including CHI and IC2S2) with the terms "audit", "accountability", "case study", "bias", "fairness", or "assurance" in the title, terms commonly used in AI studies published in computing venues;
  - reports from the government agencies ICO and NIST
- a convenience sample (accounting for 79 of the initial 143 tools) of other prominent audit tooling projects, academic papers, and tool-building organizations that we had encountered in our work as researchers and audit practitioners; most of these tools also appear in the sources above.

We included both tools designed specifically for AI auditing (such as AI fairness toolkits) and generic tools that have been used in AI audits (such as Selenium ("Selenium," 2023)). Our dataset is not an exhaustive list of all tools that have been or could be used in AI audit practice. Note that there are substitutes and competitors for many of the tools in our dataset (such as qualitative coding software), but we did not include them unless we observed them in one of the sources above or our subsequent searches (§ E.3.2). Conversely, inclusion of a tool does not indicate the authors' endorsement or preference.

E.3.2 Theoretical Sampling

Category descriptors from the initial taxonomy that were sourced either from our labels or from descriptors used by tools already collected were then searched in combination with general keywords we identified that were commonly used in the algorithmic auditing domain. These keywords include the following: "tool", "audit", "algorithm", "accountability", "responsible AI", "fairness", "discrimination", and "AI". For each of our categories, we conducted English Google searches using a selection of one or more relevant keywords combined with each of our initial category descriptors (e.g., "participatory audit", "participatory tool", etc.). In order to determine a point of saturation for each search, we examined each Google search page of 30 results at a time until two pages in a row contained no references to specific audit tools. It is important to note that we conducted searches for all categories in our initial taxonomy, but focused first on less saturated categories (for example, harms discovery and tools using participatory methods) in order to provide well-rounded definitions of categories that were

less common and/or visible. Categories with a more saturated set of examples were not given an equal amount of additional sourcing.

In addition to these targeted Google searches, we also added several new sources of tools to our initial list based on our initial taxonomy:

- News articles and reports by new organizations and civil society organizations including ProPublica, The Markup, the Pulitzer Center, the ACLU, and AlgorithmWatch;
- the Participatory ML Workshop at ICML (Kulynych et al., 2020);
- an additional Google search for startups working on "reg tech" (regulatory tech).

E.4 Interview Protocol

We used the following protocol in each of our interviews. Because the interviews were semi-structured, not all participants answered every question in the same order. Bolded questions, however, were prioritized—we asked all these questions of nearly all participants. The rest were optional follow-ups.

Thanks so much for taking the time to share your expertise. We really appreciate it and are looking forward to hearing your thoughts! [Briefly introduce the project.] [Confirm participant has completed consent form, remind participant of confidentiality, and confirm optional permissions.]

Background

- **How did you get involved in auditing to begin with?**
- **What do you hope to achieve?**
  - Would you describe yourself as an internal or external auditor?
  - What system was the target of your audit?
  - What was the motivation behind your audit work?
  - What are some notable successes?
  - Notable failures?
  - What were the most difficult aspects of the audit? How were those challenges overcome?
  - What were the easiest aspects of the audit?
- Who do you consider to be stakeholders and why?
- Tell me about the people who have a role in designing and executing the audits you're involved with.

Tool Usage and Development

- **Is there a specific tool or method (or set of tools and methods) that you employ? How do you choose these tools? What parts of the audits did they assist with?**
  - **What pain points did you encounter while using the tools?**
  - **Who made the decision to use this tool?\Why do you use this tool and not others?**
- What prompted you to use/develop a tool? What are the system behaviors that you worry about?
- **When do you know when to develop a tool vs. use an existing one?**
- Can you help me understand why these tools are helpful, from an ethical perspective?
- What is the intent of the tool/method used? Do you find that the way you've used the tools/methods aligns with those intents?
- Some people are trying to build more open-source audit methodologies and tools that are freely available to the community. Is this an important goal for your audit practice? Why or why not?

Exploring Gaps & Challenges in Tools

- **What common obstacles do you encounter while designing, building, performing, and communicating about audits/tooling? For tools, are there particular challenges (i.e., around adoption, maintenance, and distribution) that we should be aware of?**
- **Do you find that there are needs that are unmet with existing auditing tools? What are they? / Is there any tool you wish you had but didn't?** Have the tools/frameworks you've developed/used revealed any of the system behaviors that you are worried about? If yes, which ones (and to what extent)? If not, why do you think it did not uncover anything?
- In your experience, what are common properties that existing auditing tools/methods try to assess?
  - Do you think existing tools/methods are successful at measuring them?
  - Are there things it would be good to measure that current tools don't capture?
- To what degree do you find that existing auditing formats & methodologies are useful and impactful? Are there formats/methodologies that you would like to see or see more of?
- **We'd like to get a sense of how resource-intensive your tool(s)/methods are.** Would you be willing to talk about how much it cost to perform audits or develop tools? How many people were involved? And how long does it take? How hard was it to do the audit and how much did it cost you?

Wrap-up

- Is there anything else you'd like to talk about? Do you have any questions for me?
- [Confirm optional permissions again.]

E.5 Additional Landscape Analysis

To analyze the qualities of tools across our taxonomy, we manually labeled each tool with several tags describing the tool's documentation and function: license (open-source or proprietary); organization type (for-profit, non-profit, government, or academic); intended audit target (automated decision system, online platforms, large pre-trained online platforms autonomous vehicles, and/or other); intended user (internal and/or external); and format (e.g. API, software product, code/data repository, white paper, and/or other). One author created the labels and at least one other author reviewed each label for agreement.

We also supplemented our dataset with data from Crunchbase ("Crunchbase," 2023) accessed in September 2023, a platform for tracking funding, employment, revenue and other data for technology ventures, and Github, a platform for hosting and developing software. From Github, we scraped repository activity—primarily the number of forks, stars, and issues—for the 98 tools with Github repositories in our dataset. Of the 347 organizations in our dataset, we were able to access Crunchbase records for 202 (accounting for 270 tools); 173 of these records include estimated employee counts. Of the 132 entries that are not for universities or government agencies, 91 include revenue estimates. We also collected total venture funding (adjusted to U.S. dollars) for 48 firms out of the 112 firms that are still private (i.e., have not undertaken an initial public offering). Additionally, we used the Google Scholar API to annotate each academic reference of an identified tool with the most recent available citation count in September 2023.

Figure E.5.1: Number of tools by taxonomy category, sorted by type of organization (our classification). Tools may be used in multiple stages.

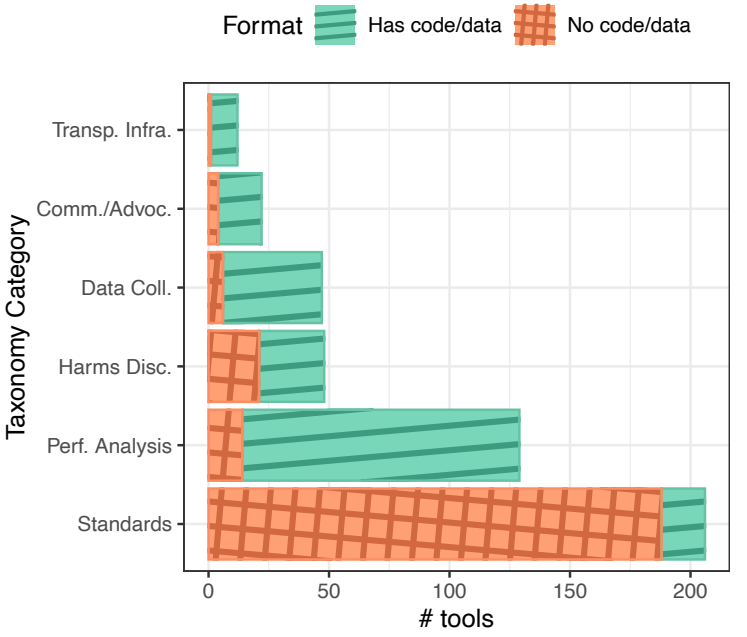


Figure E.5.2: Number of tools with code in each taxonomy stage. Tools may be used in multiple stages.

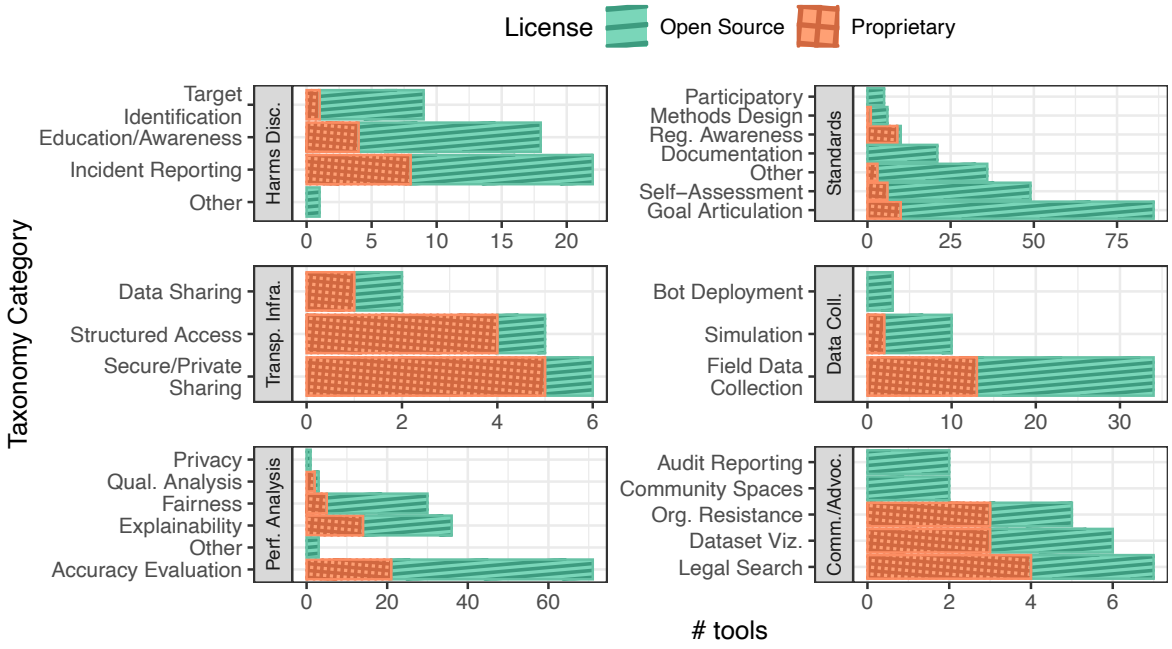


Figure E.5.3: Number of tools by taxonomy category sorted by license type. Tools may be used in multiple stages.

Figure E.5.4: Number of tools by taxonomy category sorted by audit target. Tools may be used in multiple stages.

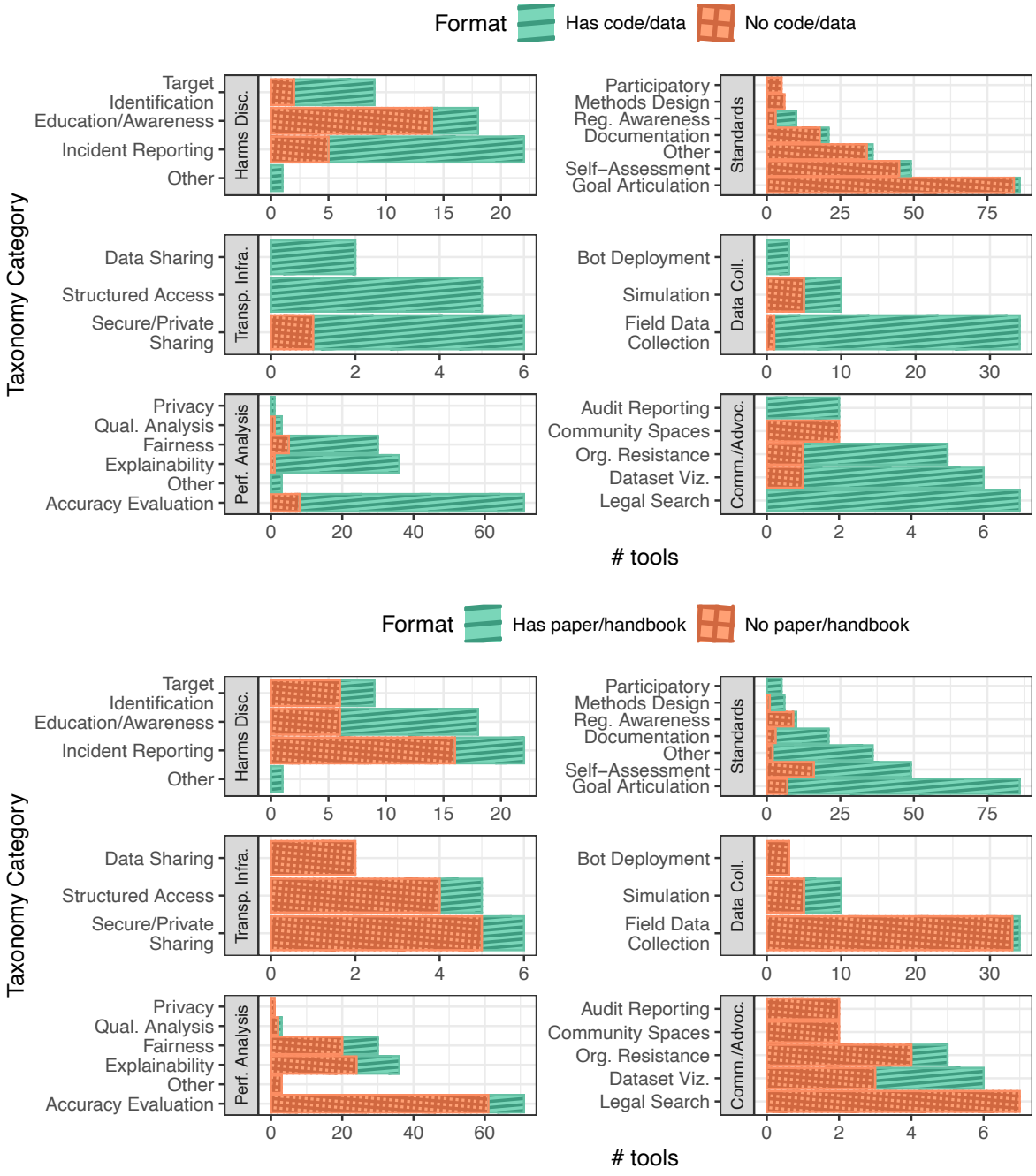


Figure E.5.5: Number of tools by taxonomy category sorted by format. Tools may be used in multiple stages.

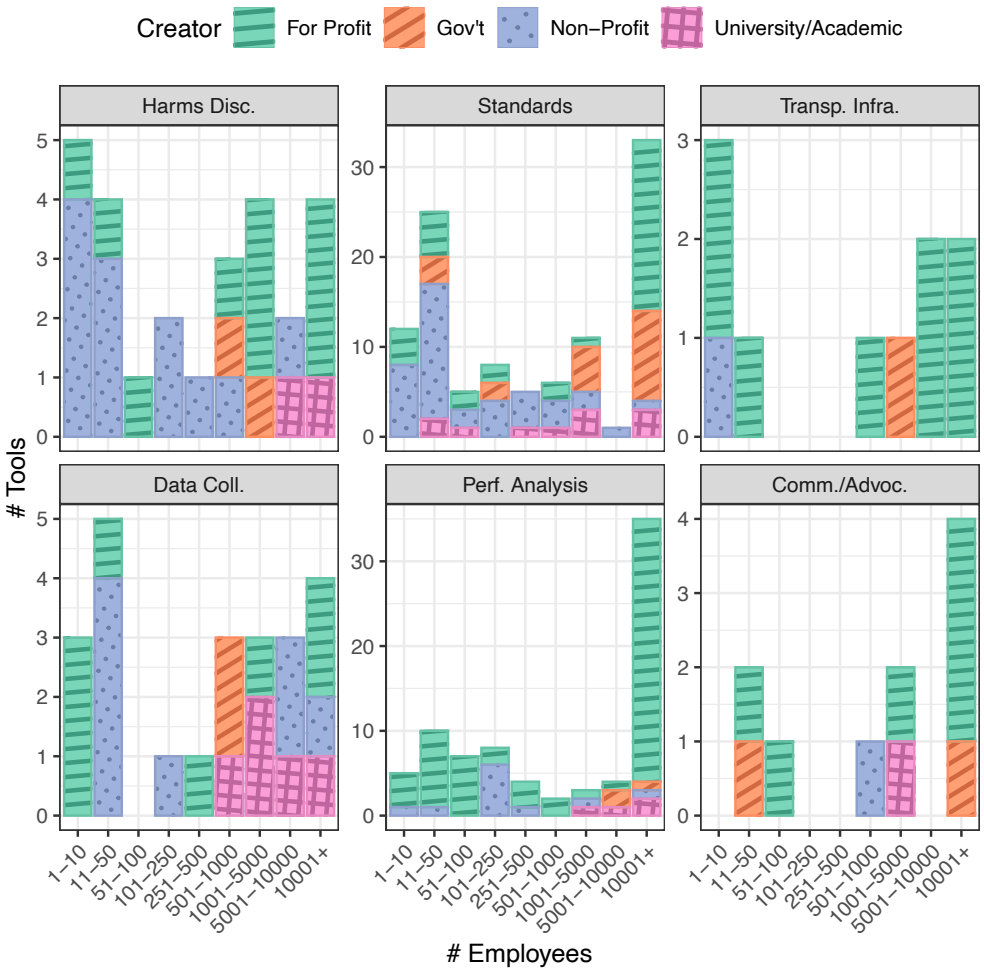


Figure E.5.6: Number of tools by taxonomy category. Tools may be used in multiple stages. Workforce size of creating organization sourced from Crunchbase ("Crunchbase," 2023). Sorted by type of organization (our classification). Tools from organizations without Crunchbase entries excluded.

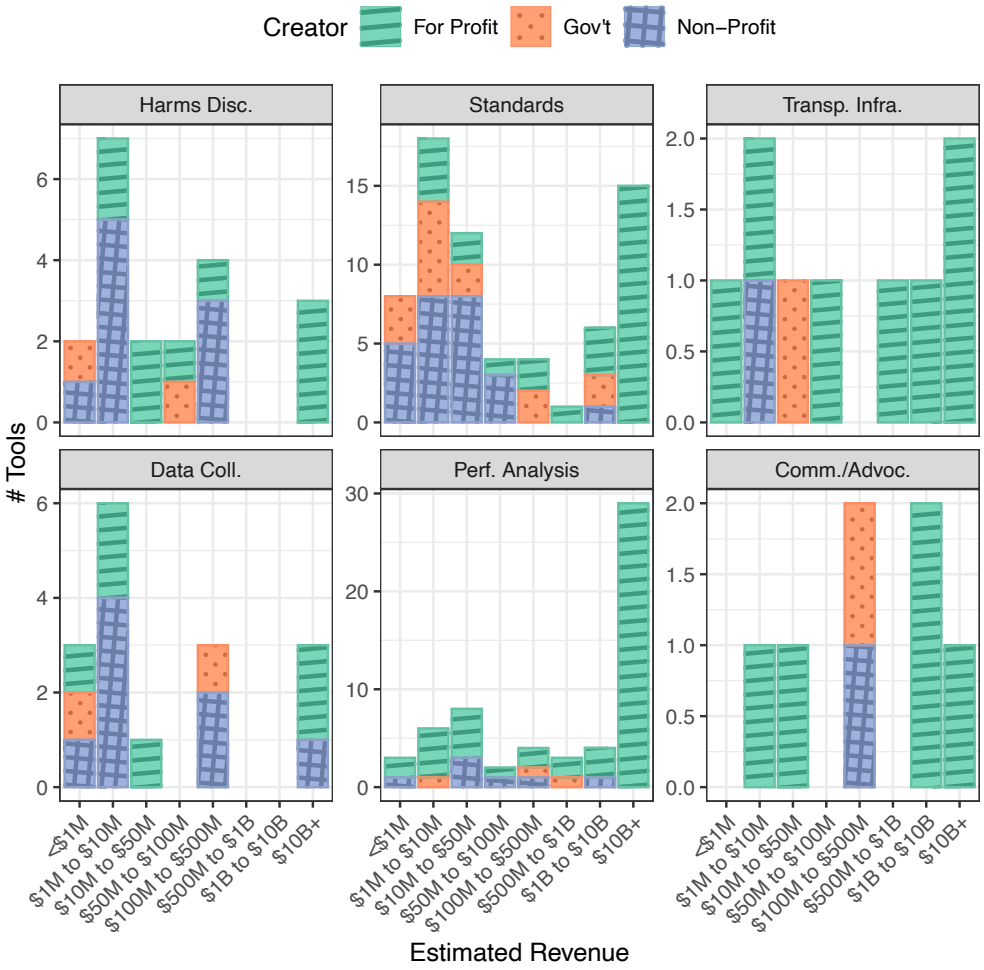


Figure E.5.7: Number of tools by taxonomy category. Tools may be used in multiple stages. Estimated revenue of creating organization sourced from Crunchbase ("Crunchbase," 2023). Tools from organizations without Crunchbase revenue estimates excluded.

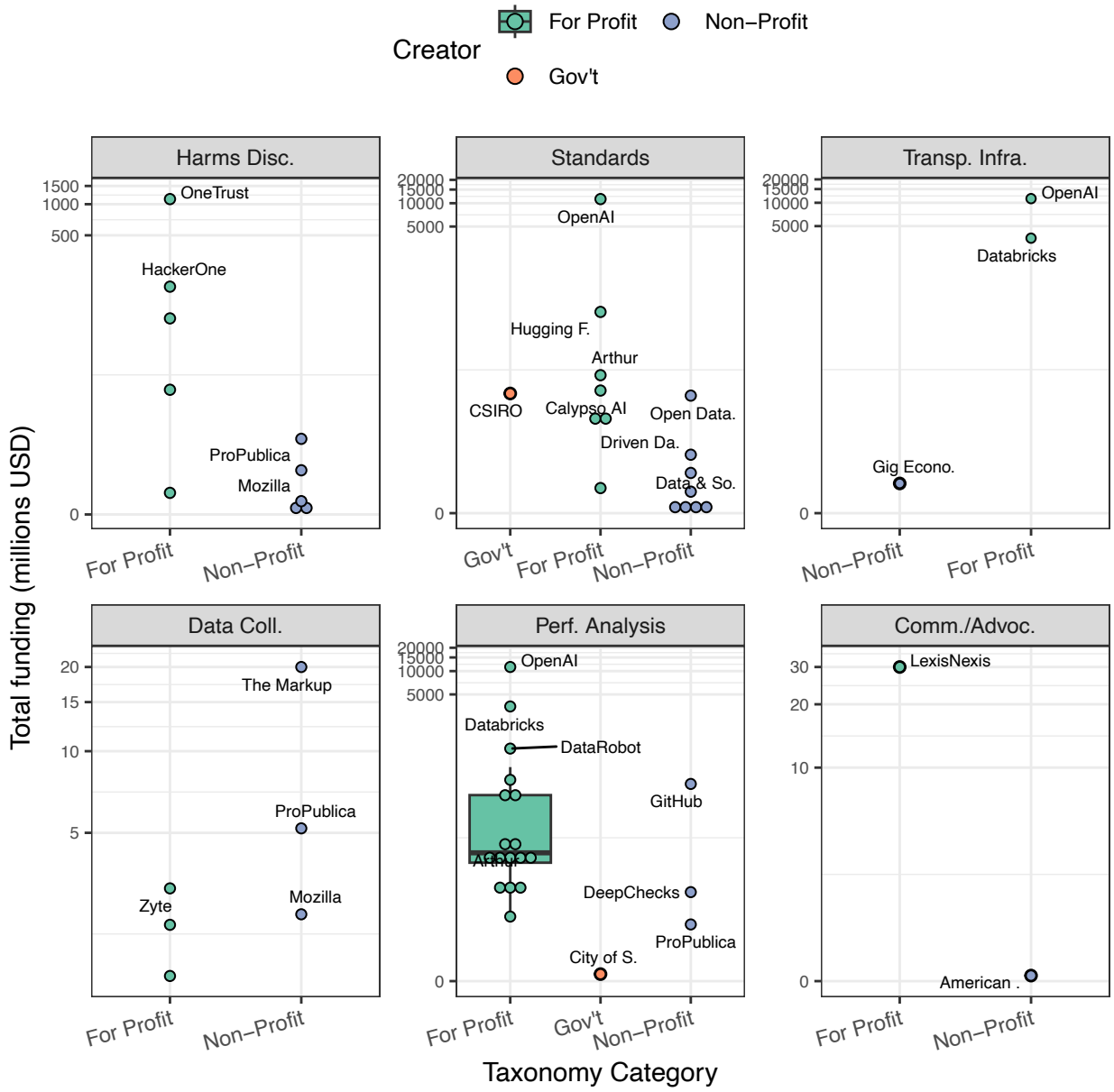


Figure E.5.8: Private (pre-IPO) organizations with Crunchbase entries (114/311 organizations). Total funding sourced from Crunchbase ("Crunchbase," 2023). Tools may be used in multiple stages.

Table E.1: Top-funded audit tool builders, per Crunchbase data ("Crunchbase," 2023). Includes only private (pre-IPO) organizations with Crunchbase entries.

Organization	Total funding (millions USD)	Estimated revenue	Employees	Stages
OpenAI	11303.12	$50M to $100M	501-1000	Perf. Analysis, Transp. Infra., Standards
Databricks	3497	$500M to $1B	5001-10000	Transp. Infra., Perf. Analysis
OneTrust	1120	$100M to $500M	1001-5000	Harms Disc.
DataRobot	1000.598	$100M to $500M	501-1000	Perf. Analysis
Hugging Face	395.2		101-250	Perf. Analysis, Standards
GitHub	350	$100M to $500M	1001-5000	Perf. Analysis
H2O.ai	251.099999	$10M to $50M	251-500	Perf. Analysis
Weights & Biases	250	$10M to $50M	251-500	Perf. Analysis
HackerOne	159.4	$10M to $50M	1001-5000	Harms Disc.
Bugcrowd	78.65	$1M to $10M	501-1000	Harms Disc.
Arthur	60.3		51-100	Standards, Perf. Analysis
Pymetrics	56.63	$1M to $10M	51-100	Perf. Analysis
Fiddler AI	45.2	$1M to $10M	11-50	Perf. Analysis
TruEra	42.284998		51-100	Perf. Analysis
CognitiveScale	40	$10M to $50M	51-100	Perf. Analysis
Calypso AI	38.2	$500M to $1B	11-50	Standards, Perf. Analysis
Seldon	33.691771	$1M to $10M	51-100	Perf. Analysis
Open Data Institute	32.835579	$1M to $10M	11-50	Standards
LexisNexis	30	$1B to $10B	10001+	Advocacy
The Markup	20	$1M to $10M	11-50	Data Coll.

Table E.2: 20 most popular Github repositories for tools in our database, sorted by number of forks.

Tool	Forks	Issues	Stars	Stages
Scrapy	10458	667	52280	Data Coll.
Selenium	8101	242	30129	Data Coll.
Appium	6052	137	18618	Data Coll.
CARLA	3562	1079	11075	Data Coll.
SHAP	3245	806	22426	Perf. Analysis
Evals	2568	130	14631	Perf. Analysis
LIME	1795	120	11492	Perf. Analysis
Language Model Evaluation Harness	1673	338	6312	Perf. Analysis
Adversarial Robustness Toolbox	1146	150	4738	Perf. Analysis
AI Fairness 360	827	199	2401	Perf. Analysis
Seldon Core	827	203	4334	Perf. Analysis
Interpret	726	105	6198	Perf. Analysis
Big Bench	582	107	2802	Perf. Analysis
Tensorflow Privacy	447	121	1915	Perf. Analysis
Foolbox	421	27	2713	Perf. Analysis
Fairlearn	410	163	1888	Perf. Analysis
Purple Llama	404	1	2476	Perf. Analysis
CodeSearchNet	384	14	2172	Perf. Analysis
Language Interpretability Tool	351	113	3453	Perf. Analysis
Error Analysis	344	86	1325	Perf. Analysis
Responsible AI Toolbox	344	86	1325	Perf. Analysis

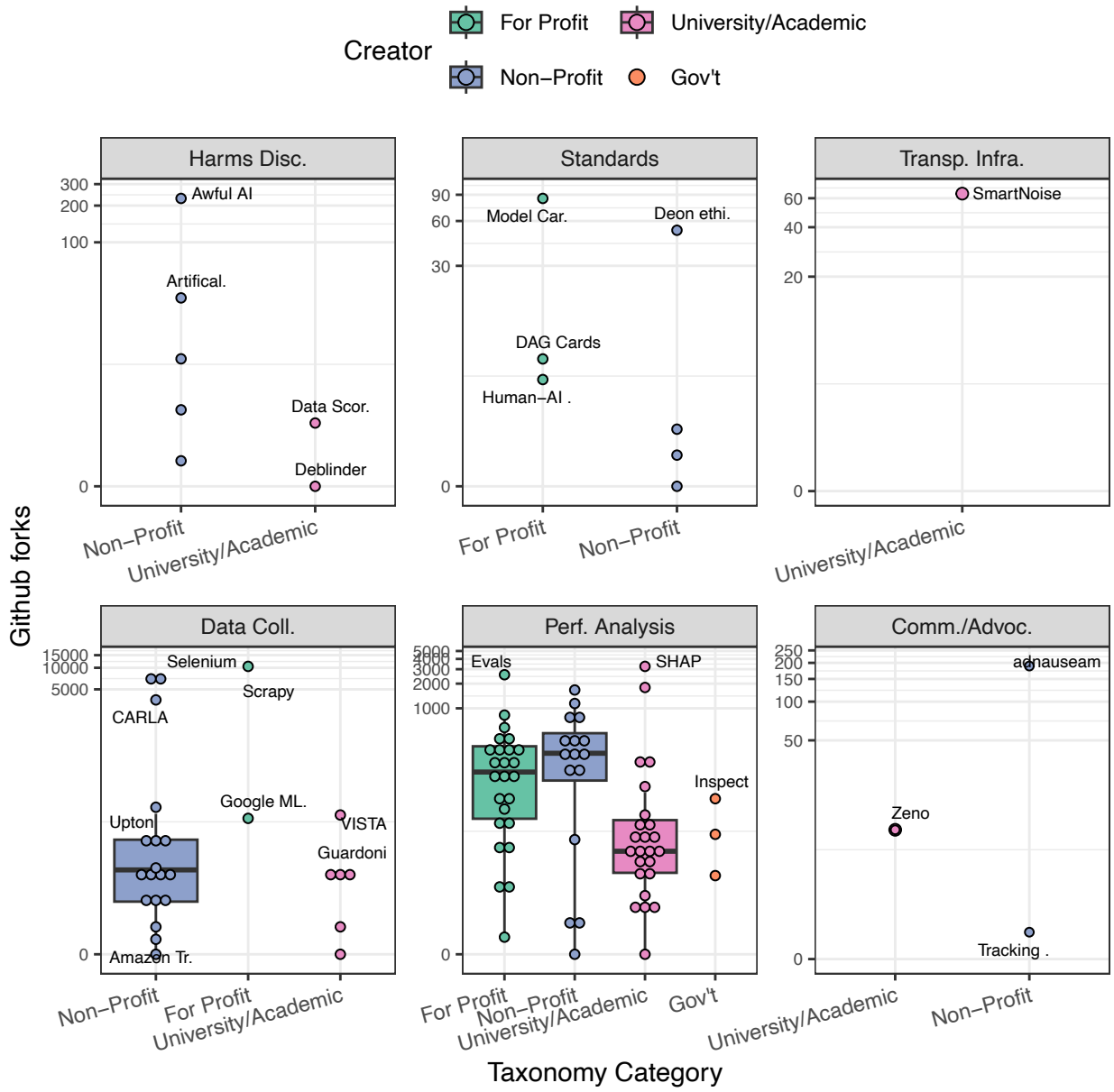


Figure E.5.9: Github forks by taxonomy category (for tools with Github repositories), sorted by type of organization (our classification). Tools may be used in multiple stages. Box-and-whisker plots included for categories with more than 10 points.

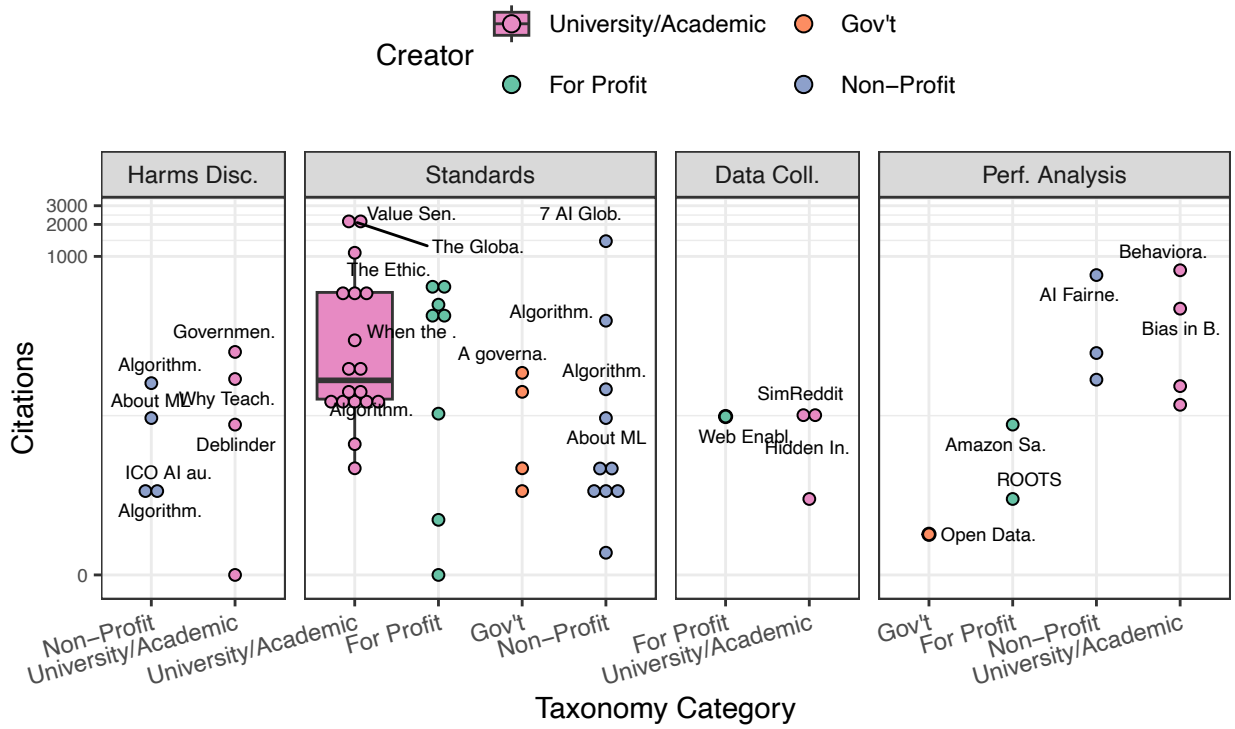


Figure E.5.10: Citations by taxonomy category (for tools with papers), sorted by type of organization (our classification). Tools may be used in multiple stages. Box-and-whisker plots included for categories with more than 10 points.