Guidehouse
Outwit Complexity

Meta

VRS Compliance Metrics Verification

June 29, 2023

Table of Contents

Guidehouse
Outwit Complexity

I. Executive Summary

Guidehouse Inc. (Guidehouse or Reviewer) was proposed by Meta Platforms, Inc. (Meta) and had the consent of the United States Department of Justice (DOJ) to serve as the independent third-party Reviewer pursuant to ¶18 of the Settlement Agreement and Final Judgement entered in *United States v. Meta* Platforms, Inc., No. 22-Civ-5187 (S.D.N.Y.) on June 27, 2022, Dkt. No. 7 (Settlement Agreement).[1]

The Reviewer is an independent third-party and, pursuant to Settlement Agreement ¶17, will "review each Compliance Report and verify compliance with the VRS Compliance Metrics."[2]

Pursuant to Settlement Agreement ¶17 and the VRS Compliance Metrics Agreement, Guidehouse reviewed the Meta Compliance Report dated May 30, 2023 for the Reporting Period from January 10, 2023 to April 30, 2023 (First Reporting Period), and verified that Meta complied with the relevant VRS Compliance Metrics for both sex and estimated race / ethnicity for both Housing Advertisements with at least 300 Ad Impressions as well as Housing Advertisements with greater than 1,000 Ad Impressions.[3]

Meta represented to Guidehouse that certain parameters were set when working with DOJ to establish the VRS Compliance Metrics, including the inclusion of Differential Privacy (DP) in Meta's implementation of Bayesian Improved Surname Geocoding (BISG) as well as setting the BISG probability threshold at 50%.

For the First Reporting Period, Guidehouse verified compliance with the VRS Compliance Metrics by assessing Meta's implementation of BISG, aggregation of Potential Impressions and Actual Impressions, and computation of Variance for accuracy and robustness using synthetic data created by Guidehouse.[4] [5] [6] While Meta represented that certain parameters were set

---

[1] Capitalized terms are defined in Appendix – Definitions.

[2] United States v. Meta Platforms, Inc. f/k/a Facebook, Inc., 22 Civ. 5187 (JGK), Dkt. No. 7, Settlement Agreement ¶17. The Settlement Agreement is available at https://www.justice.gov/opa/pr/justice-department-secures-groundbreaking-settlement-agreement-meta-platforms-formerly-known.

[3] Meta Platforms, Inc. "VRS Compliance Metrics Agreement." 6 Jan. 2023.

[4] Potential Impressions and Actual Impressions are the field names in the First Reporting Period dataset provided by Meta that contain Ad Impressions associated with Eligible Audience and Actual Audience, respectively.

[5] Guidehouse created synthetic data to supplement analysis of the First Reporting Period data, as disaggregated data from the First Reporting Period is not available. Creation and analysis of the synthetic data is discussed further in Section V – Verification Methodology.

[6] Guidehouse's implementation of Earth Mover's Distance to calculate Variance is consistent with Meta's implementation, pursuant to the Settlement Agreement and the VRS Compliance Metrics Agreement.

when establishing the VRS Compliance Metrics, Guidehouse reviewed the impact of DP and BISG probability thresholds in its analysis of the synthetic data to understand the potential sensitivity of Variance and Coverage to such parameters.

Guidehouse also independently computed Variance, separately for sex and estimated race / ethnicity, for each Housing Advertisement in the First Reporting Period using aggregated data provided by Meta. Guidehouse used these Variances to calculate Coverage and compared such calculations to the VRS Compliance Metrics established in the January 6, 2023 VRS Compliance Metrics Agreement and Meta's reported Coverage for the First Reporting Period.

Guidehouse calculated a difference of zero percent in Coverage between Meta's Coverage reported in its Compliance Report compared to Guidehouse's independently calculated Coverage across all VRS Compliance Metrics, as shown in Table 1 and Table 2 below. As these values are higher than the required VRS Compliance Metrics, Guidehouse verified Meta's compliance with the VRS Compliance Metrics.

Table 1: Meta's Reported Coverage and Guidehouse's Calculated Coverage for Housing Advertisements with ≥ 300 Impressions

	Variance Threshold	VRS Compliance Metrics	Meta – Reported Coverage[7]	Guidehouse – Calculated Coverage[8]	Difference in Coverage
Sex	≤10%	80.6%	87.6%	87.6%	0.0%
	≤5%	68.5%	78.1%	78.1%	0.0%
Estimated Race / Ethnicity	≤10%	69.7%	75.6%	75.6%	0.0%
	≤5%	48.5%	52.1%	52.1%	0.0%

[7] Meta Coverage as reported in Compliance Report pursuant to *United States v. Meta Platforms, Inc.*, No. 22-Civ-5187 (S.D.N.Y.) for January 10- April 30, 2023.

[8] Guidehouse calculations use data aggregated at the Housing Advertisement level provided by Meta for the First Reporting Period.

Table 2: Meta's Reported Coverage and Guidehouse's Calculated Coverage for Housing Advertisements with >1,000 Impressions

	Variance Threshold	VRS Compliance Metrics	Meta – Reported Coverage[9]	Guidehouse – Calculated Coverage[10]	Difference in Coverage
Sex	≤10%	82.6%	89.7%	89.7%	0.0%
	≤5%	73.2%	81.6%	81.6%	0.0%
Estimated Race / Ethnicity	≤10%	72.2%	77.2%	77.2%	0.0%
	≤5%	54.3%	56.7%	56.7%	0.0%

Notwithstanding the verification of Meta's compliance with the VRS Compliance Metrics, Guidehouse had three observations as a result of its analysis of data.

Guidehouse identified two observations based on its analysis of synthetic data, which pertained to Meta's implementation of DP within BISG and Meta's selection of the BISG probability threshold.[11]

First, based on Guidehouse's analysis of the synthetic data, the noise added from DP impacted Meta's calculation of Variance and Coverage for the synthetic data. Meta explained that the effect of the DP noise, which is implemented as a privacy protecting measure, on calculated Variance is inversely related to the difference between the Potential Impression distribution and Actual Impression distribution. Meta also provided empirical evidence that DP noise increased the average Variance, and thus reduced Coverage, in 100 distinct implementations of DP for both synthetic data and Meta Housing Advertisement data. Based on the information provided and analyses performed by Meta, DP, on average, is not expected to result in an increase in Coverage.

Second, in the synthetic data, Guidehouse found Variance and Coverage to be sensitive to the probability threshold used in the implementation of BISG. As use of a 50% BISG probability threshold is consistent with academic, industry, and regulatory literature, and thus is reasonable, Guidehouse's verification of Meta's compliance with the VRS Compliance Metrics in the First Reporting Period is not impacted by this observation.

---

[9] Meta Coverage as reported in Compliance Report pursuant to *United States v. Meta Platforms, Inc.*, No. 22-Civ-5187 (S.D.N.Y.) for January 10- April 30, 2023.

[10] Guidehouse calculations use data aggregated at the Housing Advertisement level provided by Meta for the First Reporting Period.

[11] Meta uses a 50% BISG probability threshold, as discussed in their November 2021 white paper "How Meta is working to assess fairness in relation to race in the U.S. across its products and systems" found here: https://ai.facebook.com/research/publications/how-meta-is-working-to-assess-fairness-in-relation-to-race-in-the-us-across-its-products-and-systems.

Guidehouse's third observation is based on its analysis of the First Reporting Period data as it pertained to differences in Ad Impression counts for a given Housing Advertisement when Ad Impressions are counted across sex versus across estimated race / ethnicity. The discrepancies noted are due to Meta's treatment of unknown ZIP Codes or sex, ZIP Codes with populations too small for BISG to accurately estimate race / ethnicity, and Housing Advertisements with Eligible Audiences or Actual Audiences that are not large enough to implement DP, which may result in some Ad Impressions being omitted from the calculation of Variance and Coverage. The collective impact of these omissions was not large enough to affect Coverage in the First Reporting Period and, therefore, Guidehouse's verification of compliance with the VRS Compliance Metrics in the First Reporting Period is not impacted by this observation.

II. Verification of VRS Compliance Metrics

For the First Reporting Period, Guidehouse verified that Meta complied with the relevant VRS Compliance Metrics for both sex and estimated race / ethnicity for both Housing Advertisements with at least 300 Ad Impressions as well as Housing Advertisements with greater than 1,000 Ad Impressions, in accordance with the Settlement Agreement and the VRS Compliance Metrics Agreement.

In Table 3 and Table 4 below, Guidehouse has summarized the Target Coverage at the agreed upon Variance Thresholds for sex and estimated race / ethnicity for the First Reporting Period, along with Meta's Coverage reported in its Compliance Report compared to Guidehouse's independently calculated Coverage.[12] The difference in Coverage across all VRS Compliance Metrics was zero percent, and these figures were higher than the required VRS Compliance Metrics.

Table 3: Meta's Reported Coverage and Guidehouse's Calculated Coverage for Housing Advertisements with ≥ 300 Ad Impressions

	Variance Threshold	VRS Compliance Metrics	Meta – Reported Coverage[13]	Guidehouse – Calculated Coverage[14]	Difference in Coverage
Sex	≤10%	80.6%	87.6%	87.6%	0.0%
	≤5%	68.5%	78.1%	78.1%	0.0%
Estimated Race / Ethnicity	≤10%	69.7%	75.6%	75.6%	0.0%
	≤5%	48.5%	52.1%	52.1%	0.0%

[12] Compliance Report pursuant to *United States v. Meta Platforms, Inc.*, No. 22-Civ-5187 (S.D.N.Y.) for January 10- April 30, 2023.

[13] Meta Coverage as reported in Compliance Report pursuant to *United States v. Meta Platforms, Inc.*, No. 22-Civ-5187 (S.D.N.Y.) for January 10- April 30, 2023.

[14] Guidehouse calculations use data aggregated at the Housing Advertisement level provided by Meta for the First Reporting Period.

Table 4: Meta's Reported Coverage and Guidehouse's Calculated Coverage for Housing Ads with >1,000 Ad Impressions

	Variance Threshold	VRS Compliance Metrics	Meta – Reported Coverage[15]	Guidehouse – Calculated Coverage[16]	Difference in Coverage
Sex	≤10%	82.6%	89.7%	89.7%	0.0%
	≤5%	73.2%	81.6%	81.6%	0.0%
Estimated Race / Ethnicity	≤10%	72.2%	77.2%	77.2%	0.0%
	≤5%	54.3%	56.7%	56.7%	0.0%

[15] Meta Coverage as reported in Compliance Report pursuant to *United States v. Meta Platforms, Inc.*, No. 22-Civ-5187 (S.D.N.Y.) for January 10- April 30, 2023.

[16] Guidehouse calculations use data aggregated at the Housing Advertisement level provided by Meta for the First Reporting Period.

III. Observations

Through its verification of the VRS Compliance Metrics for the First Reporting Period, Guidehouse made three observations, two based on its analysis of the synthetic data and one based on its analysis of the First Reporting Period data.

1. Observations from review of synthetic data

a. DP adds noise that may potentially impact Variance and Coverage

User race / ethnicity is not self-reported information in the Meta user database. As such, Meta uses BISG to estimate user race / ethnicity. In its implementation of BISG, Meta applies DP "to prevent adversarial disclosure or re-identification by any party while still enabling aggregate analyses" by adding noise to the aggregated estimated race / ethnicity distributions produced by BISG.[17]

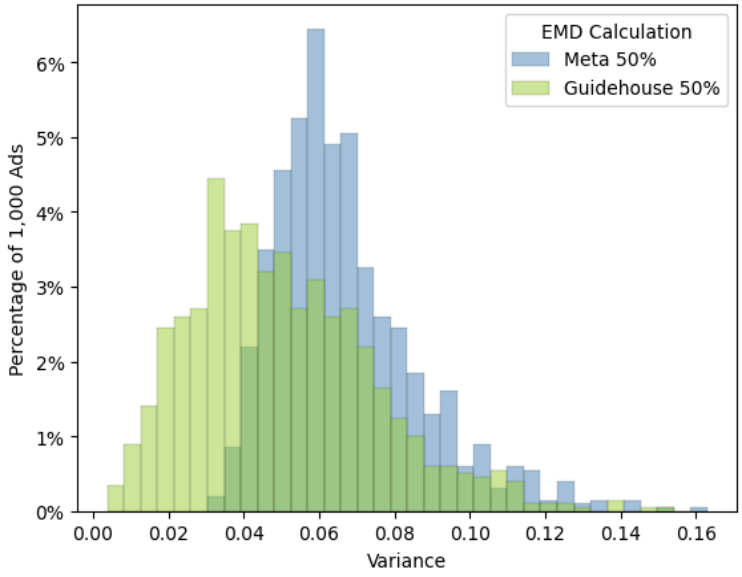
To evaluate the impact of DP on Variance and Coverage, Guidehouse generated synthetic user and Housing Advertisement data and compared the results of Meta's processing of the synthetic data, which included the addition of DP, to the results of Guidehouse's processing of the synthetic data, which did not include DP.[18] Meta processed the synthetic data 30 times, which produced 30 distinct sets of aggregated estimated race / ethnicity, Variance, and Coverage for the synthetic data. For the analysis, Guidehouse calculated the average Variance across Meta's 30 runs for each Housing Advertisement and assigned the average Variance to that Housing Advertisement.

Figure 1 below provides a comparison of the distribution of average Variance generated by Meta and the distribution of Variance generated by Guidehouse for all Housing Advertisements in the synthetic data.

[17] Meta's application of privacy enhancement is discussed further in its white papers available at https://ai.facebook.com/research/publications/how-meta-is-working-to-assess-fairness-in-relation-to-race-in-the-us-across-its-products-and-systems and https://about.fb.com/wp-content/uploads/2023/01/Toward_fairness_in_personalized_ads.pdf .

[18] Meta's and Guidehouse's implementation of BISG with a 50% probability threshold, aggregation of the data, and computation of Variance and Coverage were the same in this analysis to isolate the impact of DP.

Figure 1: Comparison of Meta's (with DP) and Guidehouse's (without DP) Variance Distribution



The average Variance computed by Meta across all advertisements in the synthetic data was 6.7%, versus an average Variance of 5.0% computed by Guidehouse. The minimum and maximum average Variance calculated by Meta was 3.3% and 16.3%, respectively, as compared to 0.4% and 15.2% computed by Guidehouse, respectively.

To provide further insight regarding the impact of DP on Variance, Guidehouse analyzed the fluctuation in the Variance computed by Meta for each Housing Advertisement across its 30 runs of BISG. Guidehouse observed the magnitude of the impact of DP on Variance differed across the 30 runs, despite consistent underlying impression data. The Housing Advertisement with the smallest observed fluctuation in Variance across the 30 runs had a minimum computed Variance of 1.6% and a maximum computed Variance of 6.2%, or a spread of 4.6%. The Housing Advertisement with the largest observed fluctuation in Variance had a minimum computed Variance of 2.9% and a maximum computed Variance of 26.3%, or a spread of 23.4%. These results indicate that the magnitude of the potential impact DP has on Variance may fluctuate.

In this analysis using synthetic data, the differences between Meta's and Guidehouse's Variance computations also resulted in discrepancies in the Coverage, as demonstrated in Table 5. Coverage is a measure of the portion of a population with Variance below a given threshold, so Coverage can be sensitive to changes in Variance, particularly when Variance in the population is clustered around the defined threshold. As discussed above, the average Variance in the synthetic data calculated by Guidehouse was 5.0%, with approximately 31% of Housing Advertisements with Variance within one percentage point of the 5% Variance threshold used to determine Coverage. Consequently, a very small amount of noise applied to the Variance for these Housing Advertisements could move them from one side of the 5% Variance threshold to the other, resulting in a large impact on Coverage at the 5% Variance threshold.

Table 5: Comparison of Meta's (with DP) and Guidehouse's (without DP) Coverage

	Meta*	Guidehouse	Difference
Coverage at Variance <= 5%	32.9%	54.2%	-21.3%
Coverage at Variance <= 10%	85.3%	96.0%	-10.7%

**Average across all of Meta's 30 runs*

As Table 5 shows, Guidehouse's computed Coverage with the synthetic data at the 5% Variance threshold was 54.2%, compared to Meta's average Coverage of 32.9% across 30 runs.[19] Therefore, the discrepancy in Variance due to DP caused a decrease of 21.3% in Coverage from Guidehouse's calculation to Meta's calculation at the 5% threshold. At the 10% Variance threshold, Guidehouse's computed Coverage was 96.0%, compared to Meta's average Coverage of 85.3% across 30 runs, resulting in a decrease of 10.7% in Coverage from Guidehouse's calculation to Meta's calculation.

Based on these results, Guidehouse observed that DP may potentially have an impact on the computed Variance and Coverage, and that the impact may fluctuate. To the extent that DP creates a bias in the distribution of impressions, the magnitude and direction of this bias may lead to changes in Coverage.

Meta provided a mathematical explanation of the behavior of DP noise, which posits that the effect of the noise on calculated Variance is inversely related to the difference between the Potential Impression distribution and Actual Impression distribution. Therefore, the DP noise is expected to be larger for smaller differences in the distributions, and smaller for larger differences. Meta also analyzed the impact of DP across 100 distinct implementations for both synthetic data and Meta Housing Advertisement data, which provided empirical evidence that the average noise resulting from DP increased Variance and thus reduced Coverage. Meta's analysis consisted of first adding DP noise to Potential Impression distributions and Actual Impression distributions for Advertisements in both the synthetic data and a sample of Housing Advertisements from Meta data and computing Variance for each Advertisement. Meta assumed this computed Variance to be the true value of Variance for each Advertisement. Meta then added DP noise one additional time to the assumed true value for each Advertisement and calculated the average difference in Variance between the second application of DP and the assumed true value for the 100 runs. Meta observed that the second application of DP resulted in an increase in Variance and decrease in Coverage, on average. Based on Meta's analyses, they asserted that DP, on average, is not expected to result in an increase in Coverage.

b. Variance and Coverage are sensitive to the BISG Probability threshold

BISG estimation assigns probabilities to each race / ethnicity bucket for a given ZIP Code / surname pair. To classify an individual as a single race / ethnicity, a probability threshold is defined. If the probability of an individual being a given race / ethnicity returned by BISG exceeds this probability threshold, the individual is assumed to be that race / ethnicity. There is

[19] Meta's Coverage across 30 runs ranged between 30.1% and 35.0%.

a tradeoff between the accuracy of the BISG estimation (i.e., a higher probability threshold) and the number of individuals whose race / ethnicity can be assigned by BISG.

Meta uses a 50% probability threshold in its implementation of BISG, as described in its November 2021 white paper "How Meta is working to assess fairness in relation to race in the U.S. across its products and systems."[20]

To assess Meta's implementation of BISG, Guidehouse used BISG with a 50% probability threshold to assign estimated race / ethnicities to the individuals in the synthetic data and compared the resulting output to the averages of outputs from Meta's 30 BISG synthetic data runs. In Table 6 below, the average count of individuals in each race / ethnicity bucket from the Meta runs is compared to the count of individuals in each race / ethnicity bucket per Guidehouse's implementation of BISG with a 50% probability threshold.

Table 6: Comparison of Synthetic Data Output of BISG with a 50% Probability Threshold

Estimated Ethnicity	Meta	Guidehouse	Difference
White	6,827.99	6,807	-20.99
Hispanic	1,491.07	1,473	-18.07
Black	686.01	703	16.99
Other	857.97	880	22.03
Unknown	134.26	134	-0.26
Total	9,997.30	9,997	-0.30

Meta represented that the differences in counts in each bucket were attributable to the impact of DP and the magnitude of the differences was in line with Meta expectations based on the parameters of DP described in the VRS Compliance Metrics Agreement.[21] As the results of Meta's and Guidehouse's implementation of BISG with a probability threshold at 50% were similar once DP is accounted for, Guidehouse concluded that Meta's implementation of BISG was consistent with Guidehouse's implementation.

[20] White paper is available at https://ai.facebook.com/research/publications/how-meta-is-working-to-assess-fairness-in-relation-to-race-in-the-us-across-its-products-and-systems/.

[21] Meta Platforms, Inc. "VRS Compliance Metrics Agreement." 6 Jan. 2023.

Academic, industry, and regulatory literature provide that BISG estimations can be implemented at various probability thresholds, and that higher thresholds produce better predictions.[22] However, a higher probability threshold decreases the number of individuals for whom race / ethnicity can be estimated using BISG. Because of this tradeoff between accuracy and identification, multiple probability thresholds can be considered when implementing BISG. The literature provides 50% - 60% as a range that strikes a good balance between accuracy and identification and is widely used as a best-practice in the financial services industry.[23]

To assess the sensitivity of Variance and Coverage to the BISG probability threshold across this probability threshold range, Guidehouse implemented BISG with a 60% probability threshold using the synthetic data and compared the Variance to that resulting from Meta's implementation of BISG using a 50% probability threshold.

When Guidehouse computed Variance for the synthetic data set using race / ethnicity estimated by BISG with a 60% probability threshold, Guidehouse observed an increase in the average Variance as compared to Meta's average computed Variance, which relies on race / ethnicity estimated by BISG with a 50% probability threshold.[24]

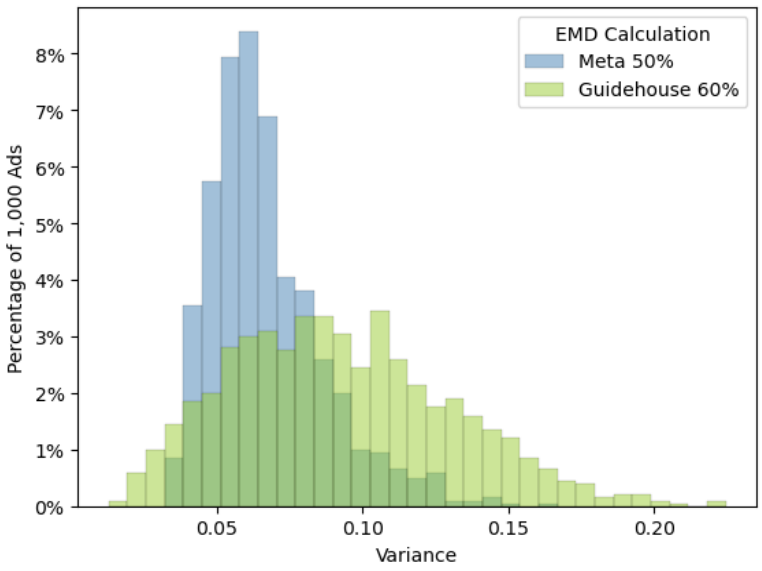
---

[22] Zhang (2018) cites research using a probability threshold no smaller than 50%, but also tests various thresholds and shows that choosing the maximum probability (BISG max) or 80% probability threshold produces more accurate estimates. Paper available at https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3169831. Additionally, Chen et al. (2018) shows that choosing the maximum probability over-weights the dominant class ("White" in this sample) in estimation. Jiahao Chen, Nathan Kallus, Xiaojie Mao, Geoffry Svacha, Madeleine Udell, 2018, "Fairness Under Unawareness: Assessing Disparity When Protected Class Is Unobserved" available at *https://arxiv.org/pdf/1811.11154.pdf*.

[23] CFBP, 2014, "Using publicly available information to proxy for unidentified race and ethnicity" available at *https://files.consumerfinance.gov/f/201409_cfpb_report_proxy-methodology.pdf*.

[24] Meta computations include DP, which may also contribute to the disparities.

Figure 2: Comparison of Meta's (50% Probability Threshold and DP) and Guidehouse's (60% Probability Threshold) Variance Distribution



As Figure 2 shows, when a 60% probability threshold is applied to the BISG estimation in the synthetic data, the Variances, on average, increased. More specifically, Guidehouse's Variance estimates using a 60% BISG probability threshold are, on average, higher than those calculated by Meta using a 50% BISG probability threshold.

This may also translate into an impact to Coverage, as shown in Table 7.

Table 7: Comparison of Meta's (50% Probability Threshold with DP) and Guidehouse's (60% Probability Threshold) Variance Estimates and Coverage

	Meta	Guidehouse
Average Variance	6.7%	9.3%
Coverage at Variance <= 5%	32.9%	12.9%
Coverage at Variance <= 10%	85.3%	59.5%

In this case, the average Variance across all Housing Advertisements in the synthetic data computed by Guidehouse increased to 9.3% as compared to Meta's computed Variance of 6.7%, creating a 2.5% difference in the mean Variance. When evaluating at both 5% and 10% Variance thresholds, Guidehouse's computed Coverage was lower than the Coverage computed by Meta.

While the BISG probability threshold is a methodology decision that Guidehouse has observed may have an impact on Variance and Coverage, Meta's choice of 50% as the BISG probability threshold is consistent with academic, industry, and regulatory best practices.

2. Observations from review of First Reporting Period data

c. Meta's decisions related to the treatment of unknown ZIP Codes, ZIP Codes with low populations, Housing Advertisements with small daily Audiences, and unknown sex may result in a subset of Ad Impressions not being captured in VRS Compliance Metrics calculations

For less than 1% of the Housing Advertisements in the First Reporting Period data, there is a larger than 20% absolute difference in the sum of Potential Impressions across sex and estimated race / ethnicity. Similarly, for less than 1% of the Housing Ads in the First Reporting Period data, there is a larger than 20% absolute difference in sum of Actual Impressions across sex and estimated race / ethnicity.

As explained by Meta, these discrepancies in the sum of Ad Impressions can be attributed to one or more of the following factors:

1. When a user's ZIP Code is not known by Meta, their race / ethnicity is not estimated using BISG. Rather, they are assigned to the "Unknown" estimated race / ethnicity bucket.
2. When a Housing Advertisement is delivered to a user with a ZIP Code that does not have a sufficiently large total population, their race / ethnicity is not estimated using BISG. Rather, they are assigned to the "Unknown" estimated race / ethnicity bucket.
3. When a Housing Advertisement has an Eligible Audience or Actual Audience containing fewer than ten unique users for a given day, Meta does not run BISG on that subset of Ad Impressions, and the user race / ethnicity is not estimated. Rather, they are assigned to the "Unknown" estimated race / ethnicity bucket.
4. When a user does not self-report a sex of either male or female, their sex is considered "Unknown."

Any Ad Impressions delivered to users with "Unknown" estimated race / ethnicity are not counted in the VRS Compliance Metrics calculations for estimated race / ethnicity; however, they may be counted in the VRS Compliance Metrics calculations for sex. The converse is true in cases where sex is not known, but race / ethnicity is able to be estimated for an Ad Impression. Ad Impressions omitted for one of the reasons above could potentially impact Variance and Coverage.

However, the decisions enumerated above appear reasonable and their combined impact observed in the First Reporting Period data was not large enough to impact Coverage. Less than 1% of Housing Advertisements in the First Reporting Period had Ad Impression counts that deviated between sex and estimated race / ethnicity by more than 20% and the majority of Housing Advertisements with deviations greater than 20% had Variance exceeding both the 5% and 10% thresholds. As Meta-reported Coverage met the VRS Compliance Metrics by margins greater than 1%, these Housing Advertisements would not impact Meta's compliance with the VRS Compliance Metrics.

IV. Background - Settlement Agreement and Scope of Work

1. Settlement Agreement

On June 27, 2022, Meta entered into a settlement with DOJ.[25] DOJ filed the Settlement Agreement concurrently with a Complaint (Complaint) against Meta alleging violations of the Fair Housing Act (FHA) based on Meta's provision of Housing Advertisement targeting options on the basis of sex, race, and ethnicity and the placement of those Housing Advertisements. Meta denied liability and any and all wrongdoing related to these allegations.[26] DOJ designed the Settlement Agreement provisions to resolve the Complaint.

Pursuant to the Settlement Agreement, Meta will:

1. Maintain publishing of active Housing Advertisements in the Ads Library, as required by the March 29, 2019 Settlement Agreement and Release (NFHA Settlement) between Meta and the National Fair Housing Alliance (NFHA), and take reasonable steps to notify users of Meta Platforms that active Housing Advertisements are available to search and view through the Ads Library, pursuant to Settlement Agreement ¶7;

2. Maintain Housing Advertisement identification processes established in the NFHA Settlement and, on the VRS Implementation date and every four months thereafter, submit a report to DOJ and the Reviewer with the number of Housing Advertisements sampled and the number of false positive and false negative Housing Advertisements identified in the reporting period, pursuant to Settlement Agreement ¶8;

3. Maintain limited Housing Advertisement targeting options made available to advertisers, pursuant to the NFHA Settlement. Any new targeting options added to the Housing Ad Flows in accordance with the standards set forth in Settlement Agreement ¶9.a must be shared DOJ, who will have thirty (30) days to review and notify Meta of any objections based on the standards set forth in Settlement Agreement ¶9.a prior to the option being added to Housing Ad Flows, pursuant to Settlement Agreement ¶9.b;

4. Stop delivery of Housing Advertisements targeted using the Special Ad Audience tool by December 31, 2022 and eliminate access to the Special Ad Audience tool and Lookalike Audience tool in Housing Ad Flows, pursuant to Settlement Agreement ¶9.c;

---

[25] United States v. Meta Platforms, Inc. f/k/a Facebook, Inc., 22 Civ. 5187 (JGK), Dkt. No. 7, Settlement Agreement.

[26] Pursuant to Settlement Agreement ¶5, the Extended Term of the Settlement Agreement will be four (4) years from the Effective Date of the Settlement Agreement. The term of the Settlement Agreement will be the Extended Term, ending on June 27, 2026. The Extended Term is defined in the Joint Letter filed by DOJ on behalf of both DOJ and Meta on January 9, 2023, Dkt. 12.

5. Develop a system, referred to as the VRS, to reduce the Variances in Ad Impressions between the Eligible Audience and Actual Audience for sex and estimated race / ethnicity, pursuant to Settlement Agreement ¶10;

6. Maintain the practice of requiring certification of compliance with anti-discrimination policies and applicable laws for all persons placing Housing Advertisements on Meta Platforms, pursuant to Settlement Agreement ¶11;

7. Maintain the practice of providing enhanced educational content on anti-discrimination policies and applicable laws to all persons placing Housing Advertisements on Meta Platforms, pursuant to Settlement Agreement ¶12;

8. Provide training on FHA to select Meta teams, pursuant to Settlement Agreement ¶13;

9. Make a statement on the Meta website about the Settlement Agreement, its obligations under the Settlement Agreement, and the importance of taking steps to prevent unlawful discrimination on internet platforms, pursuant to Settlement Agreement ¶14; and,

10. Prepare a Compliance Report every four (4) months during the term of the Settlement Agreement verifying compliance with the VRS Compliance Metrics, which will be shared with a third-party Reviewer, pursuant to Settlement Agreement ¶16.

2. Meta's VRS Compliance Metrics

The VRS Compliance Metrics are a measure of the effectiveness of VRS to reduce the Variances in Ad Impressions between the Eligible Audience and the Actual Audience for sex and estimated race / ethnicity, pursuant to Settlement Agreement ¶10, where:

1. Sex will be determined by information reported by users in their Meta profiles; [27]
2. Estimated race / ethnicity will be determined using privacy-enhanced BISG; [28] [29] and,
3. Each user in the Eligible Audience will be weighted by the total number of impressions for any Housing Advertisements displayed to the user on Meta Platforms in the prior thirty (30) days when measuring the Variance between Eligible and Actual Audiences. [30]

The VRS performance is measured using Earth Mover's Distance (EMD), also known as the Wasserstein Metric, and compliance will be determined based on VRS Compliance Metrics.

The VRS Compliance Metrics Agreement defines the "metrics for how much the VRS will reduce any Variances in Ad Impressions between Eligible Audiences and Actual Audiences for sex and estimated race / ethnicity" required by the Settlement Agreement ¶10(b). [31] On January 9, 2023, DOJ and Meta jointly filed a letter with the court advising that they had agreed to the VRS Compliance Metrics and setting forth those agreed-upon metrics. The court then adopted the parties' joint letter as an order. More specifically, VRS Compliance Metrics were set forth as shown in Table 8 and Table 9 below.[32]

---

[27] United States v. Meta Platforms, Inc. f/k/a Facebook, Inc., 22 Civ. 5187 (JGK), Dkt. No. 7, Settlement Agreement ¶10.a.v.

[28] Meta's BISG implementation process includes adaptations designed to preserve user privacy and prevent the creation of a durable records of user race / ethnicity, including obfuscating race / ethnicity buckets during BISG estimation and the addition of DP, or randomized noise, to the data to prevent reidentification of individual data from aggregate data. Meta's application of privacy enhancement is discuss further in white papers available at https://ai.facebook.com/research/publications/how-meta-is-working-to-assess-fairness-in-relation-to-race-in-the-us-across-its-products-and-systems and https://about.fb.com/wp-content/uploads/2023/01/Toward_fairness_in_personalized_ads.pdf.

[29] Ibid., ¶10.a.v.

[30] Ibid., ¶10.a.iv.

[31] Ibid., ¶10.b.

[32] United States v. Meta Platforms, Inc. f/k/a Facebook, Inc., 22 Civ. 5187 (JGK), Dkt. No. 7.

Table 8: VRS Compliance Metrics for Housing Advertisements with at least 300 Ad Impressions Delivered in the Reporting Period

	Variance	Coverage		
		By April 30, 2023	By August 31, 2023	By December 31, 2023
Sex	≤10%	80.6%	84.8%	90.2%
	≤5%	68.5%	73.4%	78.3%
Estimated Race / Ethnicity	≤10%	69.7%	74.0%	80.1%
	≤5%	48.5%	52.6%	56.8%

Table 9: VRS Compliance Metrics for Housing Advertisements with more than 1,000 Ad Impressions Delivered in the Reporting Period

	Variance	Coverage		
		By April 30, 2023	By August 31, 2023	By December 31, 2023
Sex	≤10%	82.6%	87.2%	91.7%
	≤5%	73.2%	79.1%	84.5%
Estimated Race / Ethnicity	≤10%	72.2%	76.1%	81.0%
	≤5%	54.3%	57.5%	61.0%

From December 31, 2023 through the end of the Extended Term of the Settlement Agreement, Meta agreed to reach the target Coverage ratios set forth under the December 31, 2023 columns in Table 8 and Table 9 above.

Per the VRS Compliance Metrics Agreement, for the three reporting periods in 2023, Meta agreed to include in the VRS Compliance Metrics Housing Advertisements that both begin and

end delivery of Ad Impressions during the given four-month reporting period. For reporting periods beginning in 2024, Meta intends to include in the VRS Compliance Metrics Housing Advertisements that have ended delivery of Ad Impressions during the given four-month reporting period, regardless of the impression delivery start date.

3. Reviewer's Role and Scope

Guidehouse was proposed by Meta had the consent of DOJ to serve as the independent third-party Reviewer, pursuant to ¶18 of the Settlement Agreement. The Reviewer is an independent third-party and pursuant to Settlement Agreement ¶17 will "review each Compliance Report and verify compliance with the VRS Compliance Metrics."[33]

For the First Reporting Period, Guidehouse verified compliance with the VRS Compliance Metrics by:

1. Assessing the following components of the Meta VRS Compliance Metrics calculation process for accuracy and robustness, using synthetic data created by Guidehouse:[34]
    a. BISG implementation; and,
    b. Aggregation of Eligible Audience and Actual Audience impressions and the subsequent computation of Variance through EMD; and,
2. Confirming that the Variance and Coverage metric calculations for sex and estimated race / ethnicity performed by Meta are accurate, using actual aggregated data provided by Meta to Guidehouse for the First Reporting Period.

[33] United States v. Meta Platforms, Inc. f/k/a Facebook, Inc., 22 Civ. 5187 (JGK), Dkt. No. 7, Settlement Agreement ¶17.

[34] Disaggregated impression data for the First Reporting Period is not available, so synthetic data is used for evaluation of processes requiring individual user- or impression-level data.

V. Verification Methodology

Guidehouse adopted a two-step verification approach, where the first step assessed components of the VRS Compliance Metric calculation process using synthetic data, and the second verified the Meta-reported Coverage by independently replicating the calculation steps using aggregated impression data for Housing Advertisements subject to the VRS Compliance Metrics in the First Reporting Period.

1. Step 1: Assessment of VRS Compliance Metrics Calculation Process

Guidehouse assessed the following components of the VRS Compliance Metrics calculation process:

1. Meta's implementation of BISG to estimate race / ethnicity; and,
2. Meta's aggregation of Potential Impressions and Actual Impressions and the subsequent computation of the Variance.

To assess these processes, Guidehouse generated two sets of synthetic data.[35]

The first dataset, used to assess Meta's implementation of BISG, contained ZIP Code / surname pairs generated from 2010 U.S. Census data. As Guidehouse derived the synthetic data from the U.S. Census data, these ZIP Code / surname combinations are similar to what one could observe in a sample of Meta users.

To generate a list of ZIP Codes, Guidehouse used the 2010 Census data, which provides all ZIP Codes and their associated populations. Guidehouse sampled 6,719 ZIP Codes from the Census data, with weights for population. Therefore, ZIP Codes with greater populations had a greater probability of being sampled than ZIP Codes with lower populations.

Guidehouse followed the same process to generate the list of surnames to be implemented in BISG, using the frequently occurring surnames Census data. To generate this list, Guidehouse created weights for each surname based on the number of individuals with a particular surname in the Census data. 10,000 surnames were then sampled with replacement from the population of surnames. Then, ZIP Code and surname combinations were generated by pairing each ZIP Code with a corresponding surname based on logic that produced a representative sample of persons from the U.S. population.

The synthetic dataset was used by both Meta and Guidehouse to estimate the races / ethnicities of the synthetic users with BISG. Guidehouse then compared aggregated results and evaluated any differences in Variance and Coverage.

---

[35] As the processes being evaluated are agnostic to distributional properties, Guidehouse can evaluate them using a data set with any distribution. As such, the synthetic data will not be dependent on the Meta user base.

To test the aggregation of the impressions across synthetic users and the computation of Variance, Guidehouse generated a second dataset. To create this dataset, the surname / ZIP Code combinations from the first data set were mapped to Housing Advertisement IDs, which were randomly generated strings and not related to actual Meta Housing Advertisements. Guidehouse also randomly assigned a User ID and sex to the synthetic persons in the first dataset, and generated actual impression counts for each person / Housing Advertisement combination in a random fashion.

Meta and Guidehouse independently aggregated the impressions across the Housing Advertisements to compute the Variance and Guidehouse compared the results of the independent analyses.

2. Step 2: Verification of VRS Compliance Metrics for the First Reporting Period

Guidehouse used data compiled by Meta to compute the Variance and Coverage and compared the calculated Coverage to the VRS Compliance Metrics for the First Reporting Period. Meta provided the data for the First Reporting Period in the schema in Figure 3 below.

Figure 3: Meta VRS Compliance Metrics Reporting Schema

						Inputs to Calculate Variance													
						Potential Impressions						Actual Impressions							
		Ad	Ad	Impression Bucket		Sex		Estimated Race/Ethnicity				Sex		Estimated Race/Ethnicity					
#	Hashed Ad ID	Start Date	End Date	300-1000	>1000	Male	Female	White	Hispanic	African American	Other	Male	Female	White	Hispanic	African American	Other	Variance (Sex)	Variance (Estimated Race / Ethnicity)
1																			
2																			
3																			
...																			
n																			

To compute Variance, Guidehouse calculated the proportion of Potential Impressions and Actual Impressions in Meta's data for each sex and race / ethnicity bucket for a given Housing Advertisement, where the buckets for sex are "Male" and "Female" and for race / ethnicity are "White," "Hispanic," "African American," and "Other," pursuant to the VRS Compliance Metrics Agreement.[36] To calculate the proportion, Guidehouse took the Potential Impression count and Actual Impression count in each sex and race / ethnicity bucket for a given Housing Advertisement and divided them by the total Potential Impression count and total Actual Impression count for that Housing Advertisement, respectively. For example, if there are 600 and 400 potential Impressions for male and female, respectively, the ratios would be 60% (600/1,000) and 40% (400/1,000).

[36] "VRS Compliance Metrics Agreement." 6 Jan. 2023.

Using these ratios, Guidehouse summed the absolute differences in ratios between Potential and Actual Impressions separately for sex and estimated race / ethnicity, and divided this sum by two to calculate Variance:

Variance (Sex) = $(|Ratio_{p,m} - Ratio_{e,m}| + |Ratio_{p,f} - Ratio_{e,f}|) \div 2$, and

Variance (Estimated Race / Ethnicity) = $(|Ratio_{p,w} - Ratio_{e,w}|+|Ratio_{p,h} - Ratio_{e,h}| + |Ratio_{p,a} - Ratio_{e,a}| + |Ratio_{p,o} - Ratio_{e,o}|) \div 2,$

where $p\ and\ e$ denote "Potential Impressions" and "Actual Impressions," $m\ and\ f$ denote "male" and "female," and $w, h, a, and\ o$ denote "White," "Hispanic," "African American," and "Other," respectively.

Finally, Coverage was computed by finding the percentage of Housing Advertisements with calculated Variance below the 5% and 10% Variance thresholds defined in the VRS Compliance Metrics Agreement.[37]

[37] United States v. Meta Platforms, Inc. f/k/a Facebook, Inc., 22 Civ. 5187 (JGK), Dkt. No. 12.

Appendix – Definitions

The capitalized terms listed below will have the following meaning, consistent with their definitions in the Settlement Agreement ¶3, 9, 10, 16, and 17 and the January 6, 2023 VRS Compliance Metrics Agreement, unless otherwise noted: [38] [39]

Actual Audience: All users in an Eligible Audience to whom at least one Impression of a Housing Advertisement is displayed.

Ad Impressions or Impressions: Display of ads on Meta Platforms, or any potential or synthetic ads not displayed on Meta Platforms.[40]

Ads Library: An interface that allows users to search and view active Housing Advertisements by advertiser or by location targeting options selected by advertisers.

Coverage: The percentage of Housing Advertisements where the Variance is less than or equal to the prescribed Variance threshold.

Compliance Report: Meta-prepared report confirming that it has met the VRS Compliance Metrics for the previous four-month reporting period.

Differential Privacy: A privacy-enhancing technology that protects against re-identification of individuals within aggregated data sets by adding randomized noise.[41]

Effective Date: The Effective Date of the Settlement Agreement, or the date upon which the Settlement Agreement is entered by the Court or an application to enter the Settlement Agreement is granted, whichever occurs first, as recorded on the Court's docket.

---

[38] United States v. Meta Platforms, Inc. f/k/a Facebook, Inc., 22 Civ. 5187 (JGK), Dkt. No. 7, Settlement Agreement ¶3, 9, 10, 16, 17.

[39] "VRS Compliance Metrics Agreement" 6 Jan. 2023.

[40] Definition of term expanded beyond that of the Settlement Agreement for the purposes of discussing Potential Impressions not displayed to Meta Platforms' users or synthetic Impressions in Guidehouse-generated synthetic data.

[41] Meta's discussion of Differential Privacy is available at privacytech.fb.com/differential-privacy/ and in white papers available at https://ai.facebook.com/research/publications/how-meta-is-working-to-assess-fairness-in-relation-to-race-in-the-us-across-its-products-and-systems and https://about.fb.com/wp-content/uploads/2023/01/Toward_fairness_in_personalized_ads.pdf.

Eligible Audience: All users who (1) fit targeting options selected by an advertiser for an ad, and (2) received one or more Impressions of any type of ad on Meta Platforms during the last thirty days.

FHA-Protected Classes: Race, color, religion, sex, disability, familial status, and national origin within the meaning of the FHA.

Lookalike Tool: Legacy tool available to advertisers on Meta platforms to create audiences, now replaced by the Special Ad Audience tool.

Meta Platforms: Facebook, Instagram, and Messenger.

Housing Advertisement: An advertisement offering a specific opportunity to rent, lease, sell, hold, convey, transfer, or buy a residential dwelling, and / or offering a specific real-estate related transaction such as residential mortgage, homeowner's insurance, or home appraisal services within the meaning of FHA.

Housing Ad Flows: Interfaces that advertisers use to create Housing Advertisements for publication on Meta Platforms.

Special Ad Audience: A tool in Housing Ad Flows that allows advertisers to create audiences with commonalities to a group of users, such as the advertisers' current customer, visitors to their websites, or people who like their Facebook page.

Reviewer: An independent third-party responsible for reviewing each Compliance Report and verifying compliance with the VRS Compliance Metrics.

Variance: The distance between the potential Impression distribution for the Housing Advertisement and the actual Impression distribution for the Housing Advertisement, for both sex (Male, Female) and estimated race / ethnicity (White, Hispanic, African American, and Other) separately, measured using Earth Mover's Distance.

Variance Reduction System (VRS): A Meta-developed system designed to reduce the Variance in Ad Impressions between Eligible Audiences and Actual Audiences for sex and estimated race / ethnicity.

VRS Compliance Metrics: Metrics agreed upon by DOJ and Meta and filed with the Court on how much the VRS will reduce any Variances in Ad Impressions between Eligible Audiences and Actual Audiences for sex and estimated race / ethnicity.