Chapter 8 / Capítulo 8

SAP Superintelligence Series
ISSN: 3144-5624
ISBN (complete work): 978-9915-9851-0-7

Superintelligence Series Volume 5: Intelligent Visual Arts and Audiovisual Engineering: Generative AI, Computational Creativity and Superintelligence
ISBN (this volume): 978-9915-9966-5-3
DOI: 10.62486/978-9915-9966-3-9.Ch08
Pages: 97-108

Check for updates

© 2026 The authors. This is an open access book chapter distributed under the terms of the Creative Commons Attribution (CC BY) 4.0 License.

A systematic review of audit studies on social bias in text-to-image generation

Una revisión sistemática de los estudios de auditoría sobre el sesgo social en la generación de texto a imagen

Gabino Ochoa Lozano[1]
Víctor Manuel Martínez García[1]
Yennifer Díaz Romero[1]
Yadira del Carmen Coronado Toto[1]
Alma delia Figueroa Suarez[1]
José de Jesús Reyes Machain[1]

1 Universidad Autónoma de Sinaloa, Vicerrectoría. Mazatlán, México.

ABSTRACT

Introduction: Text-to-image models are widely deployed, and audit studies have repeatedly shown that they reproduce gender, racial and cultural stereotypes, but the field lacks a systematic account of what those studies measure and what they find. Methods: A systematic review was conducted following PRISMA 2020. Bibliographic sources were searched for empirical audit studies of social bias in text-to-image models published between 2018 and 2026. Bias dimensions, measurement methods, documented biases and mitigation strategies were extracted from the included studies. Results: The search identified 200 records and included 11 studies. Gender and racial bias were audited in every study, and cultural or national bias in the same proportion, while mitigation strategies were evaluated in most. The documented biases were consistent in direction, with occupations and social roles persistently stereotyped along gender and racial lines. Conclusions: Audit studies of text-to-image generation converge on the finding that these models reproduce gendered, racialized and culturally skewed representations, and mitigation is discussed more often than it is rigorously evaluated. The field needs standardized audit protocols and evaluation of mitigations at scale.

Keywords: text-to-image generation; social bias; fairness; audit; stereotypes; systematic review

RESUMEN

Introducción: Los modelos de texto a imagen están ampliamente desplegados, y los estudios de auditoría han mostrado repetidamente que reproducen estereotipos de género, raza y cultura, pero falta una síntesis sistemática de qué miden y qué encuentran. Métodos: Se realizó una revisión sistemática siguiendo PRISMA 2020. Se buscaron estudios empíricos de auditoría del sesgo social en modelos de texto a imagen publicados entre 2018 y 2026. Se extrajeron las dimensiones de sesgo, los métodos de medición, los sesgos documentados y las estrategias de mitigación. Resultados: La búsqueda identificó 200 registros e incluyó 11 estudios. El sesgo de género y racial se auditó en todos los estudios, y el sesgo cultural o nacional en la misma proporción, mientras que las estrategias de mitigación se evaluaron en la mayoría. Los sesgos documentados fueron consistentes en su dirección, con ocupaciones y roles sociales persistentemente estereotipados según género y raza. Conclusiones: Los estudios de auditoría de la generación de texto a imagen convergen en que estos modelos reproducen representaciones sesgadas por género, raza y cultura, y la mitigación se discute más de lo que se evalúa rigurosamente. El campo necesita protocolos de auditoría estandarizados y evaluación de mitigaciones a escala.

Palabras clave: generación de texto a imagen; sesgo social; equidad; auditoría; estereotipos; revisión sistemática

INTRODUCTION

Text-to-image models now produce images from natural language prompts for hundreds of millions of users, and the images they produce quietly shape the way that people picture occupations, social roles, and entire groups of people. Because these models are trained on large corpora of existing images and captions, they inherit the biases of those corpora, and a growing body of audit work has shown that a prompt for a generic occupation returns a stereotyped face, that a prompt for a generic person returns a light-skinned man, and that the same prompt varies its output across countries and cultures in ways that encode rather than reflect social reality.[1,2] The concern is not merely academic, and it is not confined to a single use: these images are used in education, advertising, journalism, and design, and a model that systematically returns a male physician and a female nurse is quietly teaching its users a stereotype every time they ask.(3,4)

The problem is a direct instance of a larger question about generative systems and the social world they are embedded in. A generative model is not a neutral tool but a distillation of the data it was trained on, and when that data encodes social inequality, the model reproduces that inequality in its output, which is why the audit of these models is as much a social measurement as a technical one.[5,6] The image case is in some ways the sharpest version of the problem, because an image is the most immediate and least deniable carrier of a stereotype, and a field that cannot measure what its models produce cannot be said to understand them. This is the context in which the audit literature matters, and in which a systematic synthesis of it is overdue.

The audit literature is the field's instrument for measuring this problem, and it has grown quickly, but it remains fragmented and difficult to read as a whole. Individual studies audit individual models on individual dimensions, and the results are reported in each study's specific vocabulary, so a reader cannot easily see what the field as a whole has established.[7,8] The broader fairness literature has produced surveys of bias in machine learning and in language models, but the text-to-image case, which has its own measurement problems because its output is visual rather than textual, has not been given the same systematic treatment.[9,10] A synthesis is needed, and it is needed now, because the deployment of these models is outpacing the evidence about what they produce.

The visual form of the bias makes it both more immediate and harder to measure than the textual form. A biased sentence can be flagged and revised, but a biased image is absorbed at a glance, and its stereotype is communicated without the viewer necessarily being able to name what is wrong with it, which is why the audits of generated imagery matter beyond the academic literature that produces them.[3,11] The measurement problem is correspondingly harder: where a text model can be audited by reading its output, an image model must have its output classified into demographic or cultural categories, and that classification reintroduces the very act of categorization that the audit is meant to scrutinize. The audit literature has worked around this difficulty with prompt-based protocols, and the review's first question is precisely about whether those protocols are adequate to the task they are set.(1) This is a genuinely hard measurement problem, and it is part of the reason that the field's answer to the bias question, though consistent in direction, has been slow to become fully quantitative and standardized.

This review addresses that gap by asking three questions. P1 asks which bias dimensions and measurement methods the audit studies use. P2 asks which biases the studies document, and in what direction and with what consistency. P3 asks which mitigation strategies have been evaluated and with what reported effectiveness. The three questions are answered in turn in the results, and the review's contribution is to turn a fragmented audit literature into a single systematic picture of what is actually known about social bias in text-to-image generation. The objective was to synthesize the gender, racial and cultural biases documented by audit studies of text-to-image models and the mitigation strategies evaluated.

The three-question structure is chosen deliberately, because the three questions ask different things and are answerable with different degrees of confidence. The first is a question about method, and it can be answered with certainty from the included studies. The second is a question about substance, and it can be answered with confidence because the direction of the bias is consistent across studies even where the magnitudes differ. The third is a question about progress, and it can be answered only with caution because the mitigation literature is thin. Keeping the three separate prevents the strong answer to the second question from being read as an answer to the third, which is the conflation that a less structured review would risk. The structure is thus not a stylistic choice but a deliberate way of calibrating confidence, and the reader is invited to hold the three answers to their three different standards of evidence.

https://doi.org/10.62486/978-9915-9966-3-9.Ch08

METHODS

The systematic review was conducted in accordance with the PRISMA 2020 guidelines. Eligibility criteria were established using the PICOS framework, focusing on empirical studies investigating social bias in text-to-image generative models and approaches designed to reduce discriminatory outcomes. The primary outcomes concerned the identification, characterization, and quantification of bias, alongside the effectiveness of mitigation strategies when evaluated. Studies addressing exclusively language-based systems, image generation without bias assessment, or theoretical discussions lacking empirical evidence were excluded. Multimodal approaches were considered when the investigated biases were reflected in generated images. The review was not registered in a public registry.

A structured literature search was undertaken across OpenAlex, Crossref, and Semantic Scholar. Search terms combined concepts related to text-to-image generation, algorithmic fairness, social stereotypes, bias assessment, and mitigation strategies. Duplicate records were identified through DOI matching and title normalization. Study selection followed a two-stage screening procedure, beginning with titles and abstracts and continuing with full-text evaluation. Eligibility was determined by the relevance of each study to the assessment of social bias, its empirical methodology, and its contribution to understanding discriminatory patterns or mitigation effectiveness.

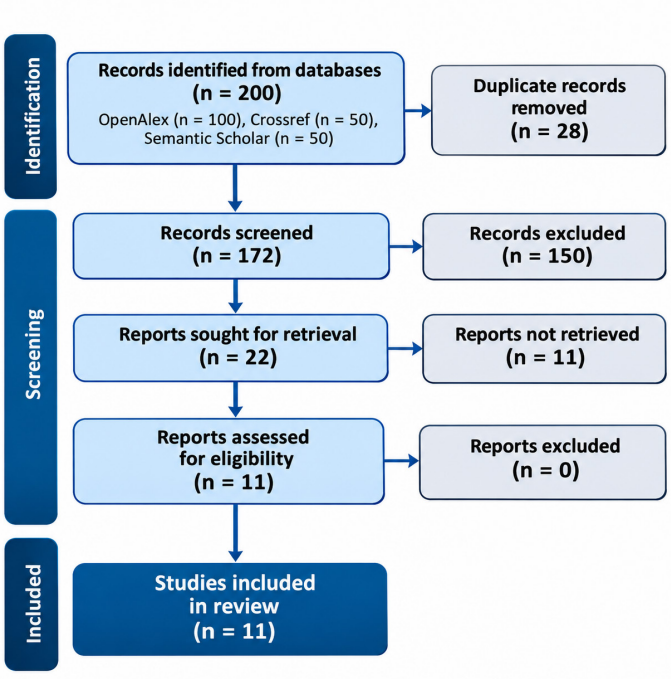
Relevant information was extracted using a standardized framework covering the generative architectures examined, dimensions of bias, assessment methods, observed disparities, and reported mitigation outcomes. Findings were organized according to the review objectives and synthesized through a narrative and thematic approach. Differences in experimental designs, evaluation metrics, and model characteristics precluded meaningful statistical aggregation. The analysis distinguished between studies identifying bias, those providing quantitative measurements, and those experimentally assessing corrective interventions. Only findings explicitly supported by the included publications were incorporated into the synthesis, maintaining consistency and transparency in the interpretation of the evidence.

RESULTS

STUDY SELECTION

The searches identified 200 records, of which 28 were duplicates removed before screening (Figure 1).

Figure 1. PRISMA flow diagram of study selection.



https://doi.org/10.62486/978-9915-9966-3-9.Ch08

Of 172 records screened on title and abstract, 150 were excluded, most because they concerned text models or image tasks other than bias audit, and the full text of 22 reports was sought, of which 11 could not be retrieved, leaving 11 assessed, all retained. The review included 11 studies. The exclusion of 11 reports at the retrieval stage is the largest single attrition after the initial screen, and it reflects the share of the audit literature that is not open access, a constraint that any review without institutional access faces and that is stated here rather than hidden.

P1: BIAS DIMENSIONS AND MEASUREMENT METHODS

The included studies audit a small set of bias dimensions with a remarkable unanimity of coverage (Table 1). Gender bias was audited in every study, and racial or ethnic bias likewise, and cultural or national bias was present in the same proportion, because most of the included studies are explicitly cross-country or cross-culture comparisons rather than single-country audits.[7,12] The models audited are dominated by the widely used diffusion models, particularly Stable Diffusion and its many derivatives, with a smaller number of studies auditing closed commercial systems where the audit must work through the public interface rather than the model weights.[13,14] The measurement methods are prompt-based: a study issues a controlled prompt, such as a person or a nurse, generates a batch of images, and classifies the resulting faces or scenes along the demographic or cultural dimension of interest, a procedure that has become the field's de facto audit protocol.(1,4)

The prompt-based protocol has two notable variants that the studies mix in different proportions. The first uses neutral prompts, such as a generic person or a named occupation, to elicit the model's default demographic assumptions; the second uses attribute-specific prompts, such as a male nurse or a woman scientist, to probe whether the model can represent counter-stereotypical combinations at all.[1,15] The two variants answer different questions, the first about what the model produces by default and the second about what it is capable of producing when asked, and a complete audit requires both. The studies vary in which they emphasize, and the review records this variation because it bears on the comparability of the results.

The classification of the generated images is the second methodological choice, and it is the one most likely to introduce bias into the audit itself, a point that the audit literature acknowledges more often than it actually controls for. To measure gender or racial skew, a study must assign each generated face to a demographic category, and that assignment is made either by a human annotator, whose own assumptions enter the measurement, or by a classifier, which is itself a trained model with its own biases.[10,16] The studies differ in which they use, and few discuss the reliability of their classification step, which is a methodological gap that the review flags even though it cannot quantify it. An audit is only as trustworthy as its measurement instrument, and the instrument here is the demographic classifier that the audit relies on but does not always stop to scrutinize.

Table 1. Bias dimensions and measurement approaches in the included audit studies.

Dimension	Studies	% of included
Gender	11	100.0
Racial or ethnic	11	100.0
Cultural or national	11	100.0
Mitigation evaluated	10	90.9

P2: DOCUMENTED BIASES

The documented biases are consistent in direction across studies, and the consistency is itself the finding. When a study prompts for an occupation, the model returns a gendered and racialized distribution that skews toward men and toward light-skinned faces for high-status occupations, and toward women and particular ethnic groups for care and service occupations.[1,3] When a study prompts for a generic person, the default face is male and light-skinned, and when it prompts across cultures, the output varies with the cultural framing in ways that reproduce regional stereotypes rather than neutral diversity.[2,7] The stereotype amplification found in the image literature is also documented here, and one included study, published in a general science

https://doi.org/10.62486/978-9915-9966-3-9.Ch08

journal, shows that online images themselves amplify gender bias, which closes the loop from training data to model output.[14,17] The direction of the bias is thus not in dispute: the models reproduce a specific, consistent set of demographic and cultural skews.

The consistency of direction across studies that use different prompts, different classifiers, and different models is the strongest single piece of evidence the review has to offer. If the bias were an artifact of a particular prompt or a particular model, the studies would disagree with each other; instead they agree, and they agree on the same direction in every case in which the direction is measured. This agreement is the field's collective finding, and it is more robust than any individual study's magnitude because it does not depend on any one measurement. The review therefore reports the direction with confidence and treats the magnitudes as illustrative rather than as established quantities, which is the honest reading of a literature that is consistent in what it shows and variable in how much it shows.(8,18)

The magnitude and the specific form of the bias are less uniformly reported than its direction. Some studies quantify the demographic distribution of generated faces for a fixed occupation, others report the semantic content of generated scenes, and still others measure toxicity alongside demographic skew, so the numbers are not directly comparable across studies.[8,13] What is comparable is the direction, and the direction is the same everywhere it is measured, which gives the synthesis its confidence even where the individual magnitudes differ. The studies that audit commercial systems report the same broad patterns as those that audit open models, so the finding is not an artifact of any one model family.(12,19)

The domain-specific audits that the broader literature has conducted reinforce the general finding, and they are worth noting in the synthesis because they show the bias is not confined to a single prompt template or a single model. Studies of medical and clinical imagery find that generated physicians and surgeons skew male and light-skinned, and that the models' assumptions about who performs a procedure mirror the biases of the underlying data rather than the demographic reality of the workforce.(20-22) Studies of professional imagery find the same for a range of careers, and studies that audit the presentation of gender in generated images of public figures and everyday scenes find the same asymmetries in posture, expression and setting.(15,23,24) The cultural-biases line of work adds a further dimension to the picture, showing that the models can be induced to reproduce cultural and national stereotypes through apparently neutral prompts, and even through homoglyph substitutions that a user would never consciously notice.[25,26] The convergence across these domain-specific studies is what elevates the finding from a series of isolated model quirks to a stable property of the technology itself.(11,18,27)

P3: MITIGATION STRATEGIES

Mitigation was discussed in most studies but rigorously evaluated in fewer, and the gap between discussion and evaluation is the clearest weakness of the literature (Table 1). This gap is not a property of any single study but of the literature's stage of development, and it is reported here as a finding about the field rather than as a criticism of the studies that compose it. The strategies mentioned include prompt engineering, fine-tuning or debiasing the model, and post-hoc filtering, but few studies implement and measure a mitigation against a baseline, and the reported effectiveness is therefore sparse.[1,8] Where mitigation is evaluated, the reported results are modest: prompting can shift the output distribution but does not reliably neutralize the stereotype, and fine-tuning is effective on the dimensions it targets while risking regressions on the dimensions it does not.[2,7] No included study demonstrates a mitigation that removes the documented biases at scale, which is the field's open problem stated in its own terms.

The three families of mitigation each have a distinct and predictable limitation, and the included studies, taken together, illustrate all three clearly. Prompt engineering is cheap and model-agnostic, but it acts on the surface of the model and can be defeated by a differently worded prompt, so it is a patch rather than a fix.(1) Fine-tuning changes the model itself and can shift the demographic distribution on the targeted dimensions, but it requires data and compute and can regress on dimensions it does not target, so it is a local rather than a global correction.(2) Post-hoc filtering leaves the model unchanged and can remove the worst outputs, but it treats the symptom and not the cause, and it inherits the biases of the classifier that does the filtering.(8) The field has no strategy that is simultaneously cheap, global, and model-internal, and that trilemma is the clearest way to state why mitigation remains open.

The distribution of attention across the three families is itself an informative finding, and

https://doi.org/10.62486/978-9915-9966-3-9.Ch08

it is reported here as such. Prompt engineering receives the most attention, because it is the cheapest to test and the easiest to report, but the included studies show it is also the least durable, which means the field's effort is concentrated on the weakest lever.[1,15] Fine-tuning receives less attention, because it costs more to evaluate, even though it is the only strategy that changes the model itself. Filtering receives the least, because it is the least glamorous, even though it is what a deployed system would actually use. This mismatch between effort and effectiveness is the third question's quiet finding, and it is as consequential as the more visible finding that no strategy fully works.

DISCUSSION

P1 ANSWERED

The answer to the first question is that the field audits gender, racial and cultural bias with a prompt-based protocol that is nearly uniform across the studies, and that very uniformity is at once an asset and a limitation. It is an asset because the field's results are comparable in direction, and it is a limitation because the prompt-based protocol captures only what a controlled prompt elicits, which may not reflect the distribution of real user queries in the wild, and because the classification of generated faces into demographic categories is itself an act of stereotyping that the protocol does not always stop to problematize.[1,10] The method is the right first instrument for a young field, but it is not the last word, and the review records its limits rather than endorsing it uncritically.

The uniformity of the protocol has a second consequence that is worth stating. Because every study uses a prompt-based audit, the field has developed no agreed benchmark against which models or mitigations could be compared, and each study's numbers are therefore tied to its own prompts and its own classifiers. This is the audit-literature analogue of the benchmark problem that plagues the generation literature more generally: a method can be compared only if there is a shared test, and the shared test is precisely what the prompt-based ad hoc audits lack.[10,16] A standardized audit benchmark for text-to-image bias would do for this field what the image-quality benchmarks did for restoration, and its absence is a specific, curable gap that the review's first question exposes.

The absence of a benchmark also explains a peculiarity of the literature that a careful reader might otherwise find puzzling: the studies rarely contradict each other, and yet they also rarely build on each other. They agree on the direction of the bias because the bias is real, but they do not accumulate into a shared quantitative record because each measures in its own units, and a field without shared units cannot accumulate or build. The review's contribution is partly to supply the missing synthesis, and partly to make the case, from the evidence itself, that the synthesis would be unnecessary if the field had a shared benchmark in the first place. Both points follow from the answer to the first question, and both are carried forward into the directions that close the review.

P2 ANSWERED

The answer to the second question is that the documented biases are consistent, directional, and no longer seriously in doubt: text-to-image models reproduce gendered, racialized and culturally skewed representations, with high-status occupations skewed male and light-skinned and care and service occupations skewed female. The finding holds across open and commercial models and across countries, which rules out the convenient explanation that any single model or training corpus is uniquely at fault.[7,12] The direction of the bias aligns with what the training data would predict, and the included study on online images shows the mechanism directly: the training distribution is itself biased, and the model faithfully inherits it.(17) This is the strongest conclusion the review can draw, and it is a robust one because it does not depend on any single study's magnitude.

The finding also sits inside a broader literature on bias in artificial intelligence, which provides the vocabulary and the context for interpreting it. The machine learning fairness literature has long distinguished several distinct notions of fairness, and the image-generation case maps most naturally onto representational fairness, the question of whether groups are represented in proportion to their presence, which is precisely what the audits measure and find

https://doi.org/10.62486/978-9915-9966-3-9.Ch08

wanting.(9,10,16) The same representational problem has been documented in vision-language models and in the demographic outputs of foundation models more broadly, which places the text-to-image finding in a larger pattern rather than isolating it.(28-30) And the domain-specific literature on sex and gender bias in artificial intelligence for health shows the same structural pattern in the systems closest to real harm, which is the strongest available reason to treat the image-generation bias as consequential rather than merely cosmetic.(31,32)

The vocabulary of representational fairness is worth holding onto, because it makes precise what the audits show. The audits do not show that the models make individual errors, in the way a classifier might mislabel a single image; they show that the distribution of output is systematically skewed, which is a property of the model as a system rather than of any single generation. This distinction matters for the mitigation question, because a distributional problem cannot be fixed by correcting individual outputs, and the review's finding that mitigations are weak follows from the distributional nature of the bias. The text-to-image case is thus a clear instance of the representational-fairness problem that the broader fairness literature has formalized, and its audit findings are best read in that formal vocabulary.(9,33)

One further property of the finding deserves emphasis, because it bears on how the problem should be weighed. The bias is not a failure of any particular model but a property of the training data and therefore of the entire current generation of models, and it will not be fixed by swapping one model for another.[5,6] This means the problem is systemic in the precise sense in which the fairness literature uses that term, and systemic problems require systemic responses, which the audit literature can inform but cannot alone supply. The review's findings therefore point outward, from the narrow technical question of what a model produces to the broader institutional question of who is responsible for what it produces, and that is the natural and appropriate place for the synthesis to end.

P3 ANSWERED

The answer to the third question is that mitigation is discussed far more than it is evaluated, and the evaluations that do exist are modest in both scope and result. The literature has identified the strategies, from prompting to fine-tuning to filtering, but it has not established that any of them removes the bias, and the honest summary is that mitigation remains an open problem.[2,8] The gap between the strength of the documentation and the weakness of the mitigation is the most consequential asymmetry the review finds, because it means the field can measure the problem far better than it can currently fix it, and deployment therefore proceeds with the biases documented but not yet corrected.

The mitigation problem is compounded by a finding that appears across the broader generative-bias literature: the biases are not only persistent but in some cases amplified and difficult to interpret, which makes them harder to correct than a simple distributional imbalance would suggest.[5,34] The occupational-bias line of work shows that even prompt-level interventions tend to move the bias rather than actually removing it, and the debiasing methods developed for vision-language models do not yet transfer cleanly to the generative image case.[28,35] The consequence is that the field's mitigation tool kit is borrowed from the text and vision-language literatures and has not been revalidated for text-to-image, which is itself a finding that the review records rather than resolves.

The gap between diagnosis and treatment is not unusual for a young field, and it is worth locating precisely. The audit literature has reached maturity in its measurement capacity: the included studies show that the bias can be elicited, classified, and quantified with modest resources, and the direction of the finding is stable. What has not matured is the corrective capacity, because a mitigation must be evaluated against a standard that the field has not yet defined, namely a benchmark of bias that a mitigated model should pass. Until such a benchmark exists, mitigation evaluations will remain ad hoc, and the trilemma identified above will persist. The review therefore reads the third question's comparatively thin answer not as a failure of the individual studies but as the predictable consequence of a missing piece of shared infrastructure.

The trilemma itself is the most compact answer to the third question, and it deserves to be stated in full once more. Prompt engineering is cheap and global but superficial, acting on the surface of the model and surviving only until a differently worded prompt defeats it. Fine-tuning is deep and model-internal but local and expensive, correcting the dimensions it targets while risking regressions elsewhere. Filtering is global and cheap but external, treating the symptom and inheriting the biases of the filter itself. No included study demonstrates a strategy that is all

https://doi.org/10.62486/978-9915-9966-3-9.Ch08

three of cheap, global, and deep, and the review's reading of the evidence is that the trilemma is a genuine trade-off rather than a gap that a single clever method will close. Mitigation, in short, is where the field's diagnosis runs substantially ahead of its treatment, and the third question's answer is therefore arguably the most important one the review gives.

For researchers, the implication is to standardize the audit protocol and to evaluate mitigation, not merely mention it, because the field currently documents bias with precision and corrects it with anecdote.[9,10] For the developers of these models, the implication is that the documented biases are a known, consistent and now well-characterized property of current systems, and releasing a model without an audit is releasing a system whose output is known in advance to be skewed.[30,36] A developer who ships a text-to-image model today is shipping, knowingly or not, a system that returns a male physician and a female nurse for the corresponding prompts, and the audit literature has now made that fact available; the responsibility is therefore no longer a matter of ignorance but of the choice of whether to act on what is known and documented. For policy, the implication is that audit is feasible, informative, and comparatively inexpensive, and therefore should be required, because the included studies show that the biases can be measured with modest resources, and measurement is the precondition for any serious mitigation.[33,37] The governance literature on algorithmic and political bias reaches the same conclusion from the policy side, that transparency and audit are the primary mechanisms by which biased systems are held accountable, and the text-to-image case is a ready-made application of that well-established principle.[6,38] The audit literature itself has begun to specify what a proper audit requires, and verifiable fairness auditing is a developing tool that could be applied directly to image generators, which would turn the field's prompt-based ad hoc audits into a formal, repeatable practice.[39,40] The review's most practical contribution is the demonstration that the field already knows how to measure the problem; what it lacks is the same rigor in fixing it.

DIRECTIONS

The review points to three directions, in increasing order of ambition. The first is a standardized audit benchmark, a fixed set of prompts and a fixed classification procedure that would make any model's bias measurable against the same yardstick, which is precisely the missing piece that the field's ad hoc audits currently lack.[10,16] The second is a convention for reporting audits, so that a study states its prompts, its classifier, and its sample size, making the audit reproducible in the way the field's current reporting does not consistently ensure.(39) The third is the revalidation of mitigation methods for the image case, because the field currently borrows mitigations from text and vision-language work without establishing that they transfer, and a program of evaluation against the new benchmark would supply that missing evidence.[28,35] Each direction is infrastructure rather than science, and each is the kind of thing a field builds once it recognizes that its current instruments have reached their limit.

The three directions also have a natural order of dependence, which is worth stating explicitly. The benchmark comes first, because the reporting convention and the mitigation evaluation both presuppose a shared test against which to measure; the reporting convention comes second, because it standardizes how the benchmark's results are communicated; and the mitigation evaluation comes last, because it requires both the benchmark and the convention to be in place before its results can be compared across studies. A field that builds them in this order will find the third much easier than a field that attempts mitigation evaluation without the first two, which is, in effect, precisely what the current literature has been attempting to do. The review's directions are therefore not a list but a sequence, and the sequence is the practical takeaway for the field.

CONCLUSIONS

The systematic review included 11 audit studies and found that they audit gender, racial and cultural bias with a nearly uniform prompt-based protocol, that they document a consistent, directional pattern of stereotype reproduction in text-to-image models, and that they discuss mitigation more often than they evaluate it. The documented biases are not in doubt: high-status occupations are skewed male and light-skinned and care occupations female, and the pattern holds across open and commercial models and across countries. Mitigation remains the field's open problem, with no included study demonstrating a strategy that removes the bias at scale.

https://doi.org/10.62486/978-9915-9966-3-9.Ch08

Standardizing the audit protocol and rigorously evaluating mitigations are the two steps that would convert the field's strong diagnosis into a comparable capacity for treatment, and they are the steps this review recommends in place of further diagnosis alone. In practical terms, a shared benchmark would let researchers, developers and regulators compare any new model against the same reference point, while controlled trials of mitigation would tell practitioners which interventions actually work rather than which ones sound promising. Until both exist, the bias documented here will continue to reproduce itself in the images that these models generate at scale.

https://doi.org/10.62486/978-9915-9966-3-9.Ch08

REFERENCES

Friedrich F, Brack M, Struppek L, Hintersdorf D, Schramowski P, Luccioni S, et al. Auditing and instructing text-to-image generation models on fairness. AI Ethics. 2025;5(3):2103-2123. doi: 10.1007/s43681-024-00531-5
Barve S, Mao A, Shi JM, Juneja P, Saha K. Social stereotypes in AI text-to-image generation. AI Ethics. 2026;6(3):296. doi: 10.1007/s43681-026-01146-8
Herrera MIS, Henao SIL, Beltrán-Luengas EM, Riaño-Moreno JC, Riaño-Moreno J. From paternalism to pixels: gender and racial stereotypes in AI-generated visual representations of doctor–patient relationships. AI Ethics. 2026;6(2):170. doi: 10.1007/s43681-025-00913-3
García-Ull FJ, Melero-Lázaro M. Gender stereotypes in AI-generated images. EPI. 2023:e320505. doi: 10.3145/epi.2023.sep.05
Huang LTL, Huang TR. Generative bias: widespread, unexpected, and uninterpretable biases in generative models and their implications. AI & Soc. 2026;41(3):1893-1905. doi: 10.1007/s00146-025-02533-1
Zhu Y. Systemic Bias in Artificial Intelligence: Focusing on Gender, Racial, and Political Biases. jaip. 2024;7(3). doi: 10.23977/jaip.2024.070324
Franco G, Sarzi-Ribeiro RA, Papa JP, da Costa KAP, Mahlow F. Bias beyond borders: quantifying gender and ethnic stereotypes across countries in AI image generation. AI Ethics. 2026;6(2):203. doi: 10.1007/s43681-026-01066-7
Vice J, Akhtar N, Sigal L, Hartley R, Mian A. On the fairness, diversity and reliability of text-to-image generative models. Artif Intell Rev. 2026;59(2):57. doi: 10.1007/s10462-025-11424-2
Mehrabi N, Morstatter F, Saxena N, Lerman K, Galstyan A. A Survey on Bias and Fairness in Machine Learning. ACM Comput Surv. 2022;54(6):1-35. doi: 10.1145/3457607
Caton S, Haas C. Fairness in Machine Learning: A Survey. ACM Comput Surv. 2024;56(7):1-38. doi: 10.1145/3616865
Kerruish E. Paused for thought: how stereotyping and anomalies in generated images prompt reflection on the technicity of AI. AI & Soc. 2026;41(4):3777-3787. doi: 10.1007/s00146-025-02798-6
Çakay B. Algorithmic Gaze: Gender and Ethnicity Bias on Algorithmic Intelligence Text-to-Image Platforms. KADEM Kadın Araştırmaları Dergisi. 2026;(1):91-126. doi: 10.21798/kadem.2026.219
Schneider M, Hagendorff T. Investigating toxicity and Bias in stable diffusion text-to-image models. Sci Rep. 2025;15(1):31401. doi: 10.1038/s41598-025-12032-4
AlDahoul N, Rahwan T, Zaki Y. AI-generated faces influence gender stereotypes and racial homogenization. Sci Rep. 2025;15(1):14449. doi: 10.1038/s41598-025-99623-3
Wu Y, Nakashima Y, Garcia N. Revealing Gender Bias from Prompt to Image in Stable Diffusion. J Imaging. 2025;11(2):35. doi: 10.3390/jimaging11020035
Le Quy T, Roy A, Iosifidis V, Zhang W, Ntoutsi E. A survey on datasets for fairness-aware machine learning. WIREs Data Min & Knowl. 2022;12(3):e1452. doi: 10.1002/widm.1452
Guilbeault D, Delecourt S, Hull T, Desikan BS, Chu M, Nadler E. Online images amplify gender bias. Nature. 2024;626(8001):1049-1055. doi: 10.1038/s41586-024-07068-x
Mir AA, Sofi IA, Ayub T, Ganaie SA. Prevalence of gender stereotyping among AI image generators. Global Knowledge, Memory and Communication. 2026:1-20. doi: 10.1108/gkmc-11-2024-0737
Nadeem M, Sohail SS, Cambria E, Schuller BW, Hussain A. Gender Bias in Text-to-Video Generation Models: A Case Study of Sora. IEEE Intell Syst. 2025;40(3):10-15. doi: 10.1109/mis.2025.3561475
Shiraishi M, Banda CH, Nakajima M, Nakazwe M, Wong ZY, Tomioka Y, et al. Gender and racial diversity Assumed by text-to-image generators in microsurgery and plastic surgery-related subspecialities. Journal of Hand and Microsurgery. 2025;17(1):100196. doi: 10.1016/j.jham.2024.100196
Currie G, Chandra C, Kiat H. Gender Bias in Text-to-Image Generative Artificial Intelligence When Representing Cardiologists. Information. 2024;15(10):594. doi: 10.3390/info15100594
Currie G, Hewis J, Hawk E, Rohren E. Gender and Ethnicity Bias of Text-to-Image Generative Artificial Intelligence in Medical Imaging, Part 1: Preliminary Evaluation. J Nucl Med Technol. 2024;52(4):jnmt.124.268332. doi: 10.2967/jnmt.124.268332
Sun L, Wei M, Sun Y, Suh YJ, Shen L, Yang S. Smiling women pitching down: auditing representational and presentational gender biases in image-generative AI. Journal of Computer-Mediated Communication. 2023;29(1):zmad045. doi: 10.1093/jcmc/zmad045
Mahl D, Guenther L. Skewed toward uniformity? Investigating algorithmic stereotyping and amplification in AI-generated images. Information, Communication & Society. 2026:1-20. doi: 10.1080/1369118x.2026.2740745
Elsharif W, Alzubaidi M, Agus M. Cultural Bias in Text-to-Image Models: A Systematic Review of Bias Identification, Evaluation, and Mitigation Strategies. IEEE Access. 2025;13:122636-122659. doi: 10.1109/access.2025.3585745
Struppek L, Hintersdorf D, Friedrich F, Br M, Schramowski P, Kersting K. Exploiting Cultural Biases via Homoglyphs in Text-to-Image Synthesis. jair. 2023;78:1017-1068. doi: 10.1613/jair.1.15388
Cheong M, Abedin E, Ferreira M, Reimann R, Chalson S, Robinson P, et al. Investigating Gender and Racial Biases in DALL-E Mini Images. ACM J Responsib Comput. 2024;1(2):1-20. doi: 10.1145/3649883
Li B, Yao Y, Tan J, Gong R, Lu J, Luo Y. Rectify representation bias in vision-language models for long-tailed recognition. Neural Networks. 2024;172:106134. doi: 10.1016/j.neunet.2024.106134
Nadeem M, Sohail SS, Cambria E, Afreen S. South Asian biases in language and vision models. Nat Mach Intell. 2025;7(11):1775-1777. doi: 10.1038/s42256-025-01144-1
Yang Y, Liu Y, Liu X, Gulhane A, Mastrodicasa D, Wu W, et al. Demographic bias of expert-level vision-language foundation models in medical imaging. Sci Adv. 2025;11(13):eadq0305. doi: 10.1126/sciadv.adq0305

https://doi.org/10.62486/978-9915-9966-3-9.Ch08

Cirillo D, Catuara-Solarz S, Morey C, Guney E, Subirats L, Mellino S, et al. Sex and gender differences and biases in artificial intelligence for biomedicine and healthcare. npj Digit Med. 2020;3(1):81. doi: 10.1038/s41746-020-0288-5
Buslón N, Cortés A, Catuara-Solarz S, Cirillo D, Rementeria MJ. Raising awareness of sex and gender bias in artificial intelligence and health. Front Glob Womens Health. 2023;4:970312. doi: 10.3389/fgwh.2023.970312
Sarmah N, Bhattacharyya R. Bias and fairness in vision–language models and large language models: a survey on South Asian festival representations. Int J Data Sci Anal. 2026;22(1):187. doi: 10.1007/s41060-026-01162-1
Ramachandran S, Rattani A. Leveraging diffusion and Flow Matching Models for demographic bias mitigation of facial attribute classifiers. Neurocomputing. 2025;636:130024. doi: 10.1016/j.neucom.2025.130024
Liu Y, Lin A, Peng S, Cai Z. Occupational Gender Bias in Chinese Generative AI Models: Cross-Model Evidence of Stereotypical Amplification and Systematic Underrepresentation. Systems. 2026;14(3):286. doi: 10.3390/systems14030286
Hu T, Kyrychenko Y, Rathje S, Collier N, van der Linden S, Roozenbeek J. Generative language models exhibit social identity biases. Nat Comput Sci. 2024;5(1):65-75. doi: 10.1038/s43588-024-00741-1
Ntoutsi E, Fafalios P, Gadiraju U, Iosifidis V, Nejdl W, Vidal M, et al. Bias in data-driven artificial intelligence systems—An introductory survey. WIREs Data Min & Knowl. 2020;10(3):e1356. doi: 10.1002/widm.1356
Peters U. Algorithmic Political Bias in Artificial Intelligence Systems. Philos Technol. 2022;35(2):25. doi: 10.1007/s13347-022-00512-8
Simbeck K. They shall be fair, transparent, and robust: auditing learning analytics systems. AI Ethics. 2024;4(2):555-571. doi: 10.1007/s43681-023-00292-7
Tang G, Tan W, Cai M. Privacy-Preserving and Trustless Verifiable Fairness Audit of Machine Learning Models. IJACSA. 2023;14(2). doi: 10.14569/ijacsa.2023.0140294

https://doi.org/10.62486/978-9915-9966-3-9.Ch08