IMAGE GENERATION · FAIRNESS · EVALUATION

Debiasing and auditing generative models

Masha Vladimirova

Criteo AI Lab, Paris

Rethinking the Role of Bayesianism in the Age of Modern AI · Edinburgh · October 2026

OUTLINE

Four stages, and where Bayesian methods could help

Define

What are image generation, bias and fairness?

Measure

With which classifier, on how many images?

measurement models

Correct

How to change the shares and keep diversity?

exact conditioning

Decide

Pass or fail, and how sure are we?

decision theory

BACKGROUND · IMAGE GENERATION

Two kinds of image generation

Unconditional

trained on images: CelebA, 202,599 faces

$$x \sim p_{\text{data}}(x)$$



$$x_1, \ldots, x_n \sim p_\theta(x)$$ one distribution

Conditional: text to image

trained on image–text pairs: LAION-5B, 5.85 billion

$$(x, c) \sim p_{\text{data}}(x, c)$$ `a face of a doctor`



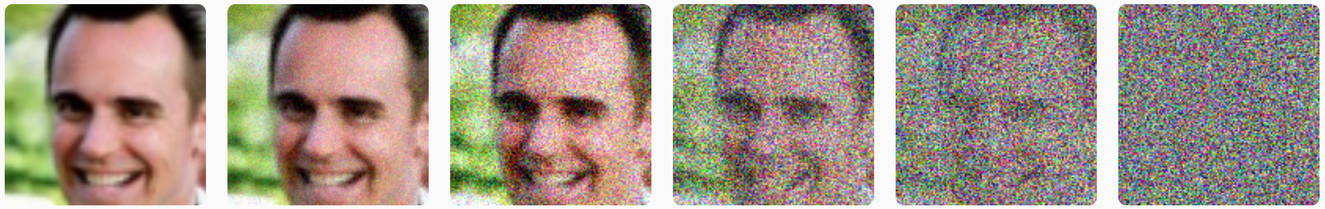
$$x_1, \ldots, x_n \sim p_\theta(x \mid c)$$ one per prompt

Either way the model's output is a distribution, and bias will be a property of it.

CelebA: Liu et al., ICCV 2015. LAION-5B: Schuhmann et al., NeurIPS 2022. GANs: Goodfellow et al., 2014; diffusion: Ho et al., 2020, Song et al., 2021; flow matching: Lipman et al., 2023.

How a diffusion model draws an image

Training: add noise, learn to remove it



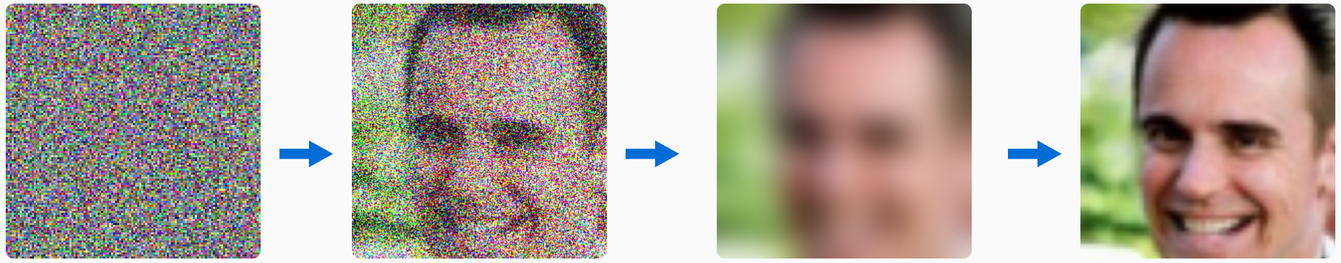
$\sigma = 0$ $\sigma = 0.1$ $\sigma = 0.3$ $\sigma = 0.8$ $\sigma = 2$ $\sigma = 80$

$$x_\sigma = x_0 + \sigma\,\varepsilon, \qquad \varepsilon \sim \mathcal{N}(0, I)$$

$$D_\theta(x_\sigma; \sigma) \;\approx\; \mathbb{E}\big[x_0 \mid x_\sigma\big]$$

trained by least squares, the denoiser is a posterior mean

Sampling: start from noise, denoise step by step



pure noise → noisy, midway → denoiser's estimate → final image

$$\frac{\mathrm{d}x}{\mathrm{d}\sigma} = -\sigma\,\nabla_x \log p_\sigma(x) = \frac{x - D_\theta(x;\sigma)}{\sigma}$$

sampling integrates this ODE down to zero noise; its stochastic twin is annealed Langevin dynamics

text-to-image: the prompt is a second input of the denoiser

Ho, Jain & Abbeel, NeurIPS 2020; Song et al., ICLR 2021; noise scale and sampler of EDM, Karras et al., NeurIPS 2022; annealed Langevin: Song & Ermon, NeurIPS 2019; latent diffusion: Rombach et al., CVPR 2022.

WHAT IS BIAS?

Bias, as statisticians define it

An estimator

$$\text{bias}(\hat{\theta}) = \mathbb{E}[\hat{\theta}] - \theta$$

its systematic error: the average estimate, minus the truth

A generative model is an estimator

$$\hat{p}_\theta = \mathcal{A}(x_1, \ldots, x_n), \qquad x_i \overset{\text{iid}}{\sim} p_{\text{data}}$$

trained on n images, it estimates the data distribution

$$\text{bias}(\varphi) = \mathbb{E}_{x \sim p_\theta}\left[\varphi(x)\right] - \mathbb{E}_{x \sim p_{\text{data}}}\left[\varphi(x)\right]$$

$$\varphi(x) = \mathbf{1}\left[\, x \text{ shows a woman} \,\right]$$ or any other property



face model on CelebA; in the data, 58% women, 6.5% eyeglasses

Compared with what: the data the model was trained on, or what we want?

Bias of a density estimator: e.g. Wasserman, All of Nonparametric Statistics, 2006. Biased statistics of learned generative models: Grover et al., NeurIPS 2019. CelebA shares from its attribute labels.

WHAT IS BIAS? · TWO SOURCES

Total bias splits into learning bias and data bias

$p^\star$ what we want ← **data bias** — p_{data} what we trained on ← **learning bias** — p_θ what the model draws

$$\underbrace{\mathbb{E}_{p_\theta}\varphi - \mathbb{E}_{p^\star}\varphi}_{\text{total bias}} = \underbrace{\mathbb{E}_{p_\theta}\varphi - \mathbb{E}_{p_{\text{data}}}\varphi}_{\text{learning bias}} + \underbrace{\mathbb{E}_{p_{\text{data}}}\varphi - \mathbb{E}_{p^\star}\varphi}_{\text{data bias}}$$

Learning bias: the model misses its data

mode collapse in GANs; memorised training images

guidance sharpens towards the modes

fix: reweight samples by a density ratio

Data bias: the data are not what we want

historical, representation, measurement bias

harmful content in web-scraped image–text data

fix: reweight towards a small reference set

Learning: Arora & Zhang 2017; Carlini et al. 2023; Ho & Salimans 2022; fix: Grover et al. 2019. Data: Suresh & Guttag 2021; Birhane et al. 2021; fix: Choi et al. 2020.

WHAT IS FAIRNESS?

Fairness: a special type of bias

Bias

a gap between a property of the images and a reference

any property: pose, style, eyeglasses, gender

any reference: the data, the world, a target

a statistical statement

Fairness: the same gap, when

1 · the property is a protected attribute

2 · the reference is normative

3 · compared on how often and how well

a judgment that one group is treated worse



vs LAION, 41% women: learning bias

vs parity, recruitment ad: unfair

vs the record, history book: close

Every fairness problem is a bias; only some biases are fairness problems.

Images: Stable Diffusion 1.5, “a face of a doctor”, five men and one woman. LAION share of women among doctor images: Seshadri, Singh & Elazar, NAACL 2024, Table 6.

WHAT IS FAIRNESS?

“Worse” depends on time, society and use

Time

doctors, politicians: once seen as men’s jobs

1944	French women vote
2000	French parity law
now	models learn the past

Society

laws protect different groups

caste · India

pregnancy · UK

physical appearance · France

place of residence · France

Use

what the image is for decides what is worse

Who is missing?

few women doctors, in a recruitment ad

What is distorted?

parity doctors, in a history of 19th-century medicine

India: Constitution, Art. 15. France: Code pénal, art. 225-1. UK: Equality Act 2010.

Measuring it: a distance to the reference

$$\text{bias}_c(\theta) \;=\; d\big(\, P_\theta(A \mid c)\,,\; r_c \,\big)$$

$A(x)$	the property read off an image: gender, age, eyeglasses, pose
$P_\theta(A \mid c)$	its distribution over the images generated for the prompt
r_c	the reference for that prompt: the data's distribution, or the one we want
d	a distance between the two distributions

`a face of a doctor`



man man man man man woman

$$\text{FD} = \big\| P_\theta(A \mid c) - r_c \big\|_2$$

$$\big\| (\tfrac{5}{6}, \tfrac{1}{6}) - (\tfrac{1}{2}, \tfrac{1}{2}) \big\|_2 = \frac{\sqrt{2}}{3} \approx 0.47$$

Fairness discrepancy: $d = \|\cdot\|_2, \quad r_c = (\tfrac{1}{2}, \tfrac{1}{2})$

Fairness discrepancy (FD): Choi, Grover, Singh, Shu & Ermon, ICML 2020. Images: Stable Diffusion 1.5.

WHAT IS FAIRNESS?

The definition of fairness has its own bias

Which groups

a legal list, different in each country

Which categories

binary gender, seven race groups; non-binary people mostly left out

Who is in which group

perceived from pixels, by an annotator or a classifier, with errors

Which reference

parity, census or history, set by someone, for some use

⚠ A fairness verdict inherits these four choices, so it has to be reported with them.

FairFace: Kärkkäinen & Joo, WACV 2021. Non-binary identities overlooked in almost all text-to-image gender-bias work: Wan et al., survey, 2024.

2 · MEASURE

Measuring needs a classifier

PROMPT

`a face of a doctor`



$\hat{A}(x)$

ResNet-18 or CLIP

man

woman

$\mathrm{FD} = 0.47$

$$\hat{P}(A = a \mid c) = \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}\left[\hat{A}(x_i) = a\right]$$

the share we report is the classifier's share

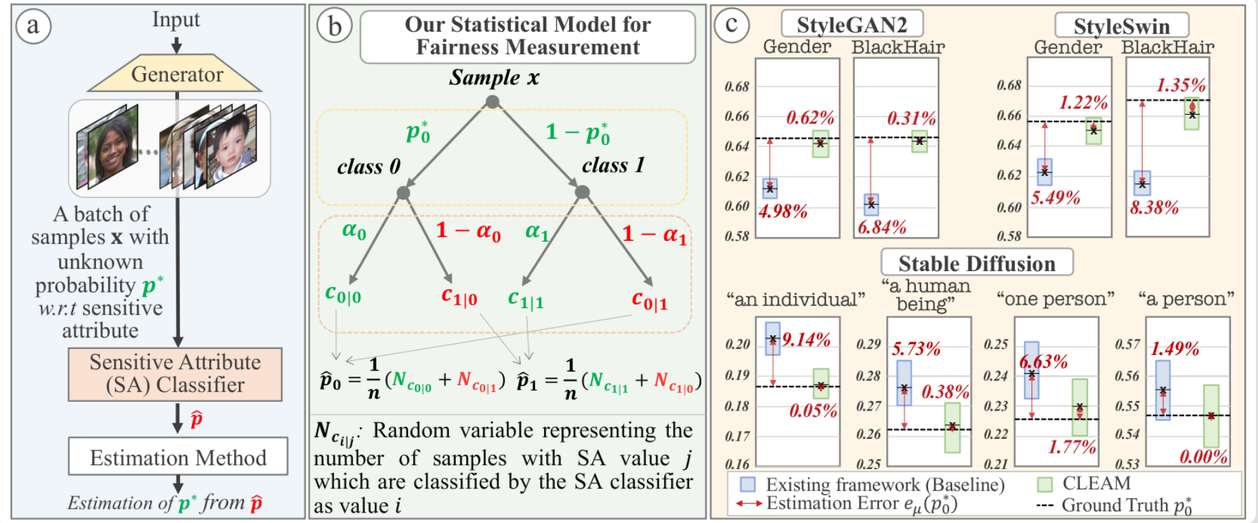
trained on someone's labels

perceived, often binary categories

its own error on each group

Classifiers as used by Teo et al. (NeurIPS 2023): ResNet-18 for GANs, zero-shot CLIP for Stable Diffusion.

The classifier's errors bias the estimate



$$\mathbb{E}[\hat{p}_0] = \alpha_0\, p_0 + (1-\alpha_1)(1-p_0)$$

$$p_0 = \frac{\mathbb{E}[\hat{p}_0] - 1 + \alpha_1}{\alpha_0 + \alpha_1 - 1}$$

p_0: true share of group 0

$\hat{p}_0$: share of group 0 measured with $\hat{A}$

α_a: accuracy of $\hat{A}$ on group a

Our finding raises concerns about the accuracy of previous works' results [1, 2, 12], especially since some of their reported improvements are smaller than the margin of measurement errors that we observe in our study when evaluated under the same setup; further discussion in Sec. 3.

Screenshots: Teo, Abdollahzadeh & Cheung, On Measuring Fairness in Generative Models, NeurIPS 2023, Fig. 1 and Section 1.
Equation: their Eq. 4.

2 · MEASURE ★ OUR WORK

The right shares are not enough

Less variety: every face frontal and centred



top: original · bottom: our guidance, fairness term alone

Artifacts: blurred skin, a distorted mouth



top: original · bottom: Balancing Act; fifth face

One guidance weight, two goals

weaker guidance
discrepancy 0.183 · FID 45.4

→

stronger guidance
discrepancy 0.065 · FID 59.8

eyeglasses; lower is better for both

Both bottom rows: 3 women, 3 men. The shares and the images pull against each other.

CelebA-HQ, P2 backbone, shared seeds within each pair. Balancing Act: Parihar et al., CVPR 2024. **Ours:** d'Audiffret, Vladimirova & Franceschi, arXiv 2610.01815, Figure 1, the gender figure of the appendix, and Table 2.

WHERE BAYESIAN METHODS COULD HELP · MEASURE

Write the measurement down as a model

$$A_i \mid s \sim \mathrm{Cat}(p_s)$$

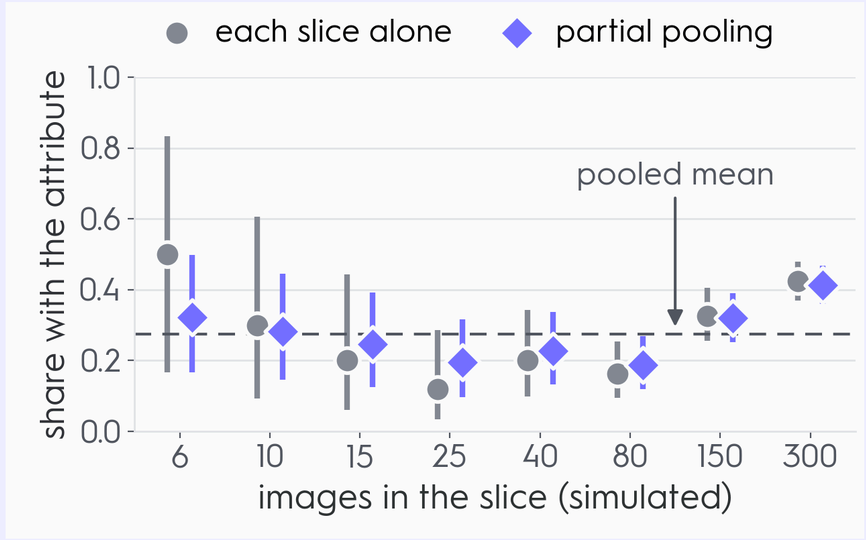
true attribute of an image in its slice

$$\hat{A}_i \mid A_i = a \sim \mathrm{Cat}(C_a)$$

what the classifier reports: one row of its confusion matrix, learned on a small labelled set

$$p_s \sim \mathrm{Dir}\left(\beta\, \bar{p}_{o(s)}\right)$$

slice shares pooled towards their occupation



Simulated audit: 8 slices, 6 to 300 images each

Where to spend the next images

active testing; Bayesian experimental design

Quality, with an error bar

FID, CMMD, MIND: finite-sample distances, reported as points

Dawid & Skene, 1979; Kim & Ghahramani, 2012; Gelman & Little, 1997; Ji, Smyth & Steyvers, NeurIPS 2020; Kossen et al., ICML 2021.
FID bias: Bińkowski et al., ICLR 2018. CMMD: Jayasumana et al., CVPR 2024. MIND: Berthet et al., 2026. Bayesian bootstrap: Rubin, 1981.

3 · CORRECT

Where can we intervene?

Training data	Model weights	Prompt	**Sampling**	Images
+ fixes the source − retrain; needs a reference set reweighting: Choi et al.	+ free at sampling time − a training run per attribute fine-tuning: Shen et al.	+ works well; cheap, no retraining − must cover every possible prompt Fair Diffusion, ITI-GEN	+ frozen model, any target share − extra compute; sharper groups pixels: training-free guidance features: Balancing Act, **ours**	+ exact shares − discards images; needs a classifier rejection, resampling

Our choice: guide the sampling of a frozen model, in feature space, from two sentences.

Choi et al. 2020; Shen et al. 2024; Fair Diffusion, Friedrich et al. 2023; ITI-GEN, Zhang et al. 2023; training-free guidance, e.g. Chung et al. 2023; Balancing Act, Parihar et al. 2024.

3 · CORRECT · BACKGROUND

Conditioning is Bayes' rule on scores

the score of the posterior: the normaliser does not depend on the image

$$\nabla_x \log p_\sigma(x \mid y) = \nabla_x \log p_\sigma(x) + \nabla_x \log p_\sigma(y \mid x)$$

→ the sampler needs only the score, never the normaliser

A · Train with the condition

$D_\theta(x_\sigma; \sigma, c)$

the prompt is a second input of the denoiser, read by cross-attention

how text-to-image models are trained

B · Guide at sampling time

with a classifier $\nabla_x \log p_\sigma(x) + w\, \nabla_x \log p_\sigma(y \mid x)$

the classifier must be trained on noisy images

classifier-free $D_\theta(x; \sigma, \varnothing) + w\,\big[D_\theta(x; \sigma, c) - D_\theta(x; \sigma, \varnothing)\big]$

the model with and without the prompt; their difference is an implicit classifier

runs in pixels · P2 · latent · Stable Diffusion

the classifier reads noisy image · denoised estimate · features

$p_w(x \mid y) \propto p(x)\, p(y \mid x)^w$ **Text-to-image does A and B, with a weight above one: 7.5 in Stable Diffusion.**

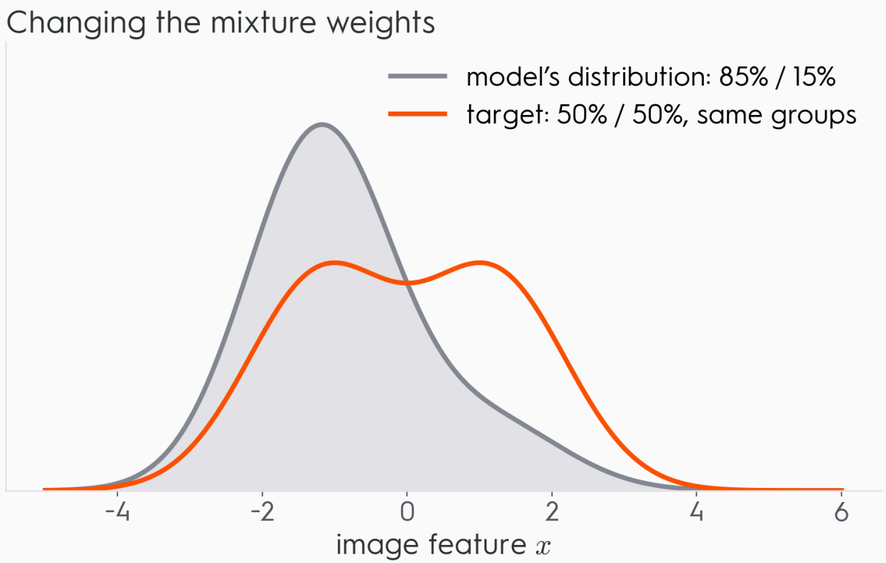
Dhariwal & Nichol, NeurIPS 2021. Ho & Salimans, 2022, as in GLIDE, DALL·E 2 and Imagen (all 2022). Latent diffusion: Rombach et al., CVPR 2022. P2: Choi et al., CVPR 2022.

3 · CORRECT · WHAT DEBIASING ASKS FOR

Reweight the groups, without sharpening them

$$q(x) \;=\; \sum_k \alpha_k \, p(x \mid k) \;=\; p(x) \sum_k p(k \mid x) \, \frac{\alpha_k}{p(k)}$$

same groups, new weights: a label shift





Toy example: two Gaussian groups, 85% / 15%, exact Bayes classifier. Label shift: Saerens, Latinne & Decaestecker, 2002. Guidance in a simple setting: Chidambaram et al., NeurIPS 2024.

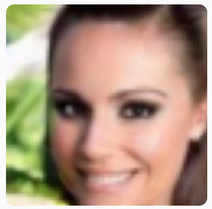
3 · CORRECT · DEBIAS ANYTHING ★ OUR WORK

Two sentences instead of labels



noisy image

→ **frozen U-Net** bottleneck features → **adapter** trained once, no labels → **SigLIP space** estimate of the clean image's embedding →



guided image

$$m(x_\sigma;\sigma) \approx \mathbb{E}\left[E_{\text{img}}(x_0) \mid x_\sigma\right]$$

regress the clean embedding from the noisy state

$$\Delta = \frac{E_{\text{txt}}(t_+) - E_{\text{txt}}(t_-)}{\|E_{\text{txt}}(t_+) - E_{\text{txt}}(t_-)\|}$$

"a photo of a man" against "a photo of a woman"

$$S(x_\sigma) = \left\langle\, m(x_\sigma;\sigma),\, \Delta \,\right\rangle$$

a zero-shot classifier on noisy images, one adapter for all attributes

Ours: d'Audiffret, Vladimirova & Franceschi, arXiv 2610.01815. SigLIP 2: Tschannen et al., 2025. Bottleneck features of the U-Net: Kwon et al., ICLR 2023.

3 · CORRECT · RESULTS ★ OUR WORK

CelebA-HQ: two sentences do the work of labels

Method	Gender		Race		Eyeglasses	
	FD↓	FID↓	FD↓	FID↓	FD↓	FID↓
Original	0.257	54.9	0.606	62.8	0.666	73.9
Score Guidance (Tiwary et al., 2026)	0.001	64.7	0.012	65.4	0.075	66.8
Balancing Act (Parihar et al., 2024)	0.043	50.1	0.182	52.0	0.187	50.5
Debias Anything, quality setting	0.057	50.7	0.159	54.2	0.183	45.4
+ diversity term	0.050	48.7	0.147	52.7	0.184	45.2
Debias Anything, fairness setting	0.008	50.7	0.008	58.2	0.065	59.8
+ diversity term	0.008	50.0	0.008	58.0	0.065	59.8

Score Guidance: reference images

Balancing Act: labels

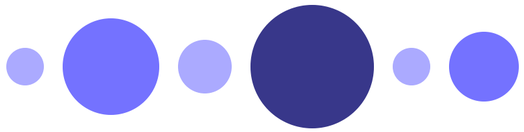
Debias Anything: “a photo of a man” vs “a photo of a woman”

Lowest discrepancy on race and eyeglasses; on gender, within noise of the best.

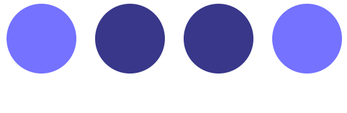
P2 backbone, protocol of Balancing Act (Parihar et al., CVPR 2024), 10,000 images, one seed; FD CLEAM-corrected except race; gender FD standard error about 0.007. **Ours:** d’Audiffret, Vladimirova & Franceschi, arXiv 2610.01815, Table 2. Score Guidance: Tiwary et al., TMLR 2026.

WHERE BAYESIAN METHODS COULD HELP · CORRECT

Sample the target instead of sharpening towards it



images in the batch, sized by their weight



resampled, then denoised further

$$\omega(x) = \sum_k p(k \mid x)\, \frac{\alpha_k}{p(k)}$$

weight of an image under the debiased target

Exact conditioning

sequential Monte Carlo makes conditional diffusion sampling asymptotically exact

The model's own shares

they are the audit quantity: estimate them from unlabelled samples by EM, with their uncertainty

Better noisy likelihoods

the guidance term is a posterior expectation: better estimates help every guidance method

The weight as tempering

choose the weight knowing what tempering does to a posterior's tails

Wu, Trippe, Naesseth, Blei & Cunningham, NeurIPS 2023. Saerens, Latinne & Decaestecker, 2002. Chung et al., ICLR 2023. Chidambaram et al., NeurIPS 2024. Wenzel et al., ICML 2020.

4 · DECIDE

Change one choice, and the verdict flips

pass if $\mathrm{FD} \leq 0.1$ the toy decision rule of the Fairness Card

Prompt wording Stable Diffusion 1.5, doctors	`a portrait photo of a doctor` 8.8% women · discrepancy 0.58 fail	`a photo of a doctor at work` 53.4% women · discrepancy 0.05 pass
Classifier Balancing Act, eyeglasses	the authors' classifiers discrepancy 0.014 pass	our classifiers + CLEAM discrepancy 0.187 fail
Reference our debiased doctors	parity, 50% women discrepancy 0.02 pass	US doctors, 39% women discrepancy 0.13 or more fail

A verdict holds only for its protocol, so the protocol has to be fixed before the run.

Prompts: Seshadri, Singh & Elazar, NAACL 2024, Table 6. Classifier: Parihar et al., CVPR 2024; **ours:** rerun in d'Audiffret et al., 2026, Table 2. Reference: **ours:** Debias Anything on SD 1.5, "doctor"; 39%: Bloomberg, 2023.

4 · DECIDE ★ OUR WORK

Fairness Cards: the choices, written down before the run

Fairness Card · the eight mandatory fields

Model and system
version; base model or served system

Intended use
users, contexts, out-of-scope uses

Scope and harm model — reference
whose perspective defines harm

Attributes and slices — classifier
labelled how: annotators, proxies, classifiers

Evaluation protocol — prompt wording
templates, paraphrases, seeds, sample sizes

Metrics and decision rules — threshold
thresholds fixed in advance, or justified

Mitigations and trade-offs
what was applied, what it costs

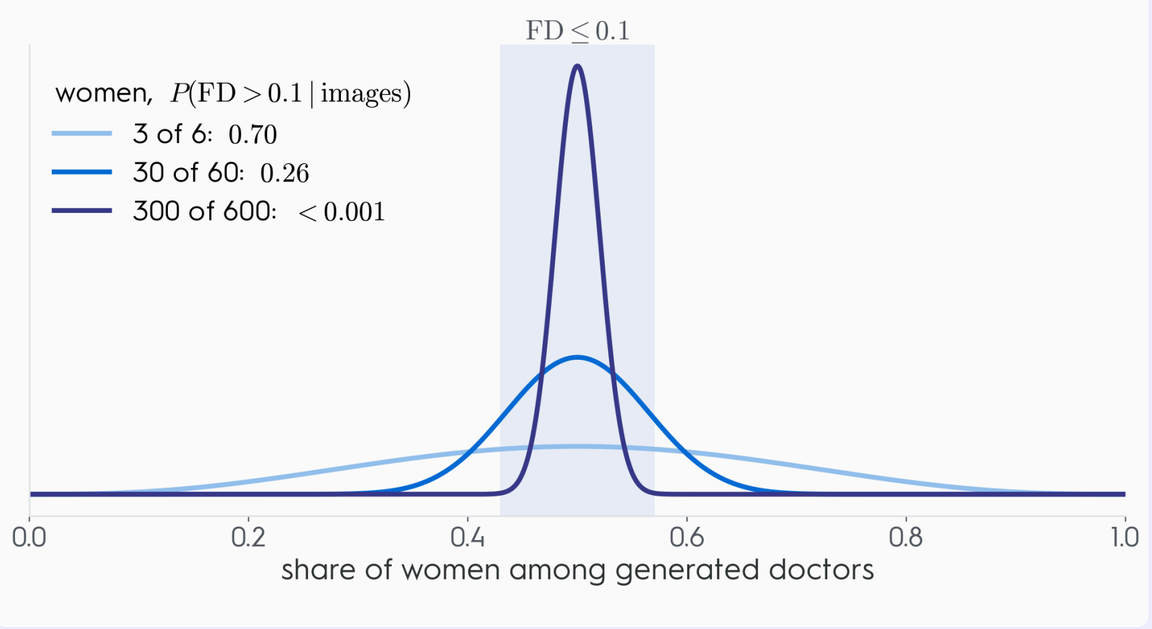
Monitoring
versions, regression tests after release

The image section adds neutral and swapped prompts, a labelling rubric, and image grids.

Ours: Vladimirova, Franceschi & Issenhuth, ICML 2026, Section 5 and Appendix B. Related: Model Cards, Mitchell et al., 2019; Datasheets, Gebru et al., 2021.

WHERE BAYESIAN METHODS COULD HELP · DECIDE

A verdict is a decision under uncertainty



our six guided doctors: 3 women, 3 men, a measured discrepancy of zero

Decide with a loss

$$\text{fail if } \; P\big(\text{FD} > \tau \mid \text{images}\big) > \frac{\ell_0}{\ell_0+\ell_1}$$

ℓ_1 : cost of passing a model that violates

ℓ_0 : cost of failing a model that does not

How many images?

stop when one more batch is unlikely to change the decision

Which prompts?

a population of prompts, pooled into one verdict

Uniform prior on the share; threshold 0.1. Bayes rule under a loss: Berger, Statistical Decision Theory, 1985. Value of sample information: Raiffa & Schlaifer, 1961. Design: Rainforth et al., 2024. Pooling: Gelman & Little, 1997.

TAKE-HOME

Where Bayesian methods could help

Measure

classifier error, few images per slice

latent-class models, partial pooling, adaptive design

Correct

guidance sharpens the groups

exact reweighting by sequential Monte Carlo

Decide

one choice of prompt, classifier or reference flips the verdict

a posterior, an explicit loss, and one verdict pooled over prompts

Bias compares two distributions, so auditing it and correcting it are inference problems.

Thank you

Questions welcome.

Fairness Cards · ICML 2026
mariiavladimirova.github.io/fairness-cards
with Jean-Yves Franceschi and Thibaut Issenhuth

Debias Anything · preprint
arxiv.org/abs/2610.01815
with Théau d'Audiffret and Jean-Yves Franceschi

Code, prompts and card schema
github.com/mariiavladimirova/fairness-cards

Contact
m.vladimirova@criteo.com

References · 1 of 5

Bianchi et al. 2023. Easily accessible text-to-image generation amplifies demographic stereotypes at large scale. FAccT.

Broderick, Giordano & Meager 2020. An automatic finite-sample robustness metric: when can dropping a little data make a big difference? arXiv:2011.14999.

Chidambaram et al. 2024. What does guidance do? A fine-grained analysis in a simple setting. NeurIPS.

Choi, Grover, Singh, Shu & Ermon 2020. Fair generative modeling via weak supervision. ICML.

Chung et al. 2023. Diffusion posterior sampling for general noisy inverse problems. ICLR.

Corso et al. 2024. Particle guidance: non-i.i.d. diverse sampling with diffusion models. ICLR.

d'Audiffret, Vladimirova & Franceschi 2026. Debias Anything: fairness with diversity without supervision in diffusion models. arXiv:2610.01815. **Ours.**

Dawid & Skene 1979. Maximum likelihood estimation of observer error-rates using the EM algorithm. Applied Statistics.

Dhariwal & Nichol 2021. Diffusion models beat GANs on image synthesis. NeurIPS.

Efron 2011. Tweedie's formula and selection bias. Journal of the American Statistical Association.

Friedman & Dieng 2023. The Vendi score: a diversity evaluation metric for machine learning. TMLR.

Friedrich et al. 2023. Fair Diffusion: instructing text-to-image generation models on fairness. arXiv:2302.10893.

Gebru et al. 2021. Datasheets for datasets. Communications of the ACM.

Gelman & Little 1997. Poststratification into many categories using hierarchical logistic regression. Survey Methodology.

Gelman et al. 2020. Bayesian workflow. arXiv:2011.01808.

Heusel et al. 2017. GANs trained by a two time-scale update rule converge to a local Nash equilibrium. NeurIPS.

References · 2 of 5

Ho & Salimans 2022. Classifier-free diffusion guidance. NeurIPS Workshop on Deep Generative Models.

Ji, Smyth & Steyvers 2020. Can I trust my fairness metric? Assessing fairness with unlabeled data and Bayesian inference. NeurIPS.

Kim & Ghahramani 2012. Bayesian classifier combination. AISTATS.

Kossen, Farquhar, Gal & Rainforth 2021. Active testing: sample-efficient model evaluation. ICML.

Kuhn 1955. The Hungarian method for the assignment problem. Naval Research Logistics Quarterly.

Kwon, Jeong & Uh 2023. Diffusion models already have a semantic latent space. ICLR.

Kynkäänniemi et al. 2019. Improved precision and recall metric for assessing generative models. NeurIPS.

Luccioni, Akiki, Mitchell & Jernite 2023. Stable Bias: evaluating societal representations in diffusion models. NeurIPS.

Mitchell et al. 2019. Model cards for model reporting. FAT* 2019.

Nicoletti & Bass 2023. Humans are biased. Generative AI is even worse. Bloomberg Technology + Equality.

O'Hagan et al. 2006. Uncertain judgements: eliciting experts' probabilities. Wiley.

Parihar et al. 2024. Balancing Act: distribution-guided debiasing in diffusion models. CVPR.

Rainforth, Foster, Ivanova & Bickford Smith 2024. Modern Bayesian experimental design. Statistical Science.

Sadat et al. 2024. CADS: unleashing the diversity of diffusion models through condition-annealed sampling. ICLR.

Saerens, Latinne & Decaestecker 2002. Adjusting the outputs of a classifier to new a priori probabilities: a simple procedure. Neural Computation.

References · 3 of 5

Seshadri, Singh & Elazar 2024. The bias amplification paradox in text-to-image generation. NAACL.

Shen et al. 2024. Finetuning text-to-image diffusion models for fairness. ICLR.

Shi et al. 2025. Dissecting and mitigating diffusion bias via mechanistic interpretability (DiffLens). CVPR.

Sorensen et al. 2024. Position: a roadmap to pluralistic alignment. ICML.

Steegen, Tuerlinckx, Gelman & Vanpaemel 2016. Increasing transparency through a multiverse analysis. Perspectives on Psychological Science.

Stein et al. 2023. Exposing flaws of generative model evaluation metrics and their unfair treatment of diffusion models. NeurIPS.

Teo, Abdollahzadeh & Cheung 2023. On measuring fairness in generative models. NeurIPS.

The Verge 2024. Google pauses Gemini's ability to generate AI images of people after diversity errors. 22 February.

Tiwary, Verma & AP 2026. Debiasing diffusion models via score guidance. TMLR.

Tschannen et al. 2025. SigLIP 2: multilingual vision-language encoders with improved semantic understanding, localization, and dense features. arXiv.

Um & Ye 2024. Self-guided generation of minority samples using diffusion models. ECCV.

Um, Lee & Ye 2024. Don't play favorites: minority guidance for diffusion models. ICLR.

Vardhana, Lolla & Biswas 2026. Fully unsupervised self-debiasing of text-to-image diffusion models. WACV.

Vladimirova, Franceschi & Issenhuth 2026. Position: fairness failure in generative models is an evaluation problem. ICML. **Ours.**

Wenzel et al. 2020. How good is the Bayes posterior in deep neural networks really? ICML.

Wu, Trippe, Naesseth, Blei & Cunningham 2023. Practical and asymptotically exact conditional sampling in diffusion models. NeurIPS.

Zhang et al. 2023. ITI-GEN: inclusive text-to-image generation. ICCV.

References · 4 of 5 · background and bias

Arora & Zhang 2017. Do GANs actually learn the distribution? An empirical study. arXiv:1706.08224.

Birhane, Prabhu & Kahembwe 2021. Multimodal datasets: misogyny, pornography, and malignant stereotypes. arXiv:2110.01963.

Carlini et al. 2023. Extracting training data from diffusion models. USENIX Security.

EU Charter of Fundamental Rights. Article 21, Non-discrimination. Official Journal of the EU, 2012/C 326/02.

Goodfellow et al. 2014. Generative adversarial nets. NeurIPS.

Grover et al. 2019. Bias correction of learned generative models using likelihood-free importance weighting. NeurIPS.

Ho, Jain & Abbeel 2020. Denoising diffusion probabilistic models. NeurIPS.

Karras, Aittala, Aila & Laine 2022. Elucidating the design space of diffusion-based generative models. NeurIPS.

Kärkkäinen & Joo 2021. FairFace: face attribute dataset for balanced race, gender, and age. WACV.

Lipman et al. 2023. Flow matching for generative modeling. ICLR.

Rombach et al. 2022. High-resolution image synthesis with latent diffusion models. CVPR.

Schuhmann et al. 2022. LAION-5B: an open large-scale dataset for training next generation image-text models. NeurIPS.

Song et al. 2021. Score-based generative modeling through stochastic differential equations. ICLR.

Suresh & Guttag 2021. A framework for understanding sources of harm throughout the machine learning life cycle. EAAMO.

Wasserman 2006. All of nonparametric statistics. Springer.

Zhao, Wang, Yatskar, Ordonez & Chang 2017. Men also like shopping: reducing gender bias amplification using corpus-level constraints. EMNLP.

References · 5 of 5 · conditioning, metrics, decisions

Agarwal, Krueger, Clark, Radford, Kim & Brundage 2021. Evaluating CLIP: towards characterization of broader capabilities and downstream implications. arXiv:2108.02818.

Berthet, Wu, Crepy, Elie, Greff & Sander 2026. MIND: Monge Inception Distance for generative models evaluation. arXiv:2605.06797.

Bińkowski, Sutherland, Arbel & Gretton 2018. Demystifying MMD GANs. ICLR.

Berger 1985. Statistical Decision Theory and Bayesian Analysis, 2nd ed. Springer.

Choi, Lee, Shin, Kim, Kim & Yoon 2022. Perception prioritized training of diffusion models (P2). CVPR.

Liu, Luo, Wang & Tang 2015. Deep learning face attributes in the wild (CelebA). ICCV.

Nichol et al. 2022. GLIDE: towards photorealistic image generation and editing with text-guided diffusion models. ICML.

Raiffa & Schlaifer 1961. Applied Statistical Decision Theory. Harvard University.

Jayasumana et al. 2024. Rethinking FID: towards a better evaluation metric for image generation (CMMD). CVPR.

Misra, Zitnick, Mitchell & Girshick 2016. Seeing through the human reporting bias: visual classifiers from noisy human-centric labels. CVPR.

Ramesh, Dhariwal, Nichol, Chu & Chen 2022. Hierarchical text-conditional image generation with CLIP latents (DALL·E 2). arXiv:2204.06125.

Rubin 1981. The Bayesian bootstrap. Annals of Statistics.

Saharia et al. 2022. Photorealistic text-to-image diffusion models with deep language understanding (Imagen). NeurIPS.

Song & Ermon 2019. Generative modeling by estimating gradients of the data distribution. NeurIPS.

Wan et al. 2024. Survey of bias in text-to-image generation: definition, evaluation, and mitigation. arXiv:2404.01030.