Unfair Unlearning? Accounting for Fairness in Machine Unlearning

George-Octavian Barbulescu
University of Warwick
United Kingdom
george-octavian.barbulescu@warwick.ac.uk

Francois Buet-Golfouse
Barclays AIML Global Markets
United Kingdom
ucahfbu@ucl.ac.uk

ABSTRACT

We study the intersection of machine unlearning and algorithmic fairness in high-stakes decision-making systems. While recent advances in unlearning focus on removing the influence of specific datapoints from trained models, little is known about how such procedures interact with fairness constraints—particularly when forgotten datapoints contribute to equitable outcomes. In this paper, we propose a principled variational framework for fair unlearning, which balances three competing objectives: (1) forgetting specific datapoints, (2) preserving fairness across protected groups, and (3) maintaining proximity to the original model via KL-regularization. We show that forgetting subgroups can reduce fairness, and visualize this trade-off via a Pareto frontier.

CCS CONCEPTS

• **Computing methodologies** → **Fairness in machine learning**; *Variational inference*; • **Security and privacy** → *Data privacy*; • **Applied computing** → Health informatics.

KEYWORDS

Machine unlearning, fairness, variational inference

ACM Reference Format:
George-Octavian Barbulescu and Francois Buet-Golfouse. 2018. Unfair Unlearning? Accounting for Fairness in Machine Unlearning . In *Proceedings of Make sure to enter the correct conference title from your rights confirmation email (KDD '25).* ACM, New York, NY, USA, 3 pages. https://doi.org/XXXXXXX.XXXXXXX

1 INTRODUCTION

As machine learning systems are increasingly deployed in high-stakes domains, two demands have emerged: the ability to *unlearn* specific datapoints upon request, and the enforcement of *algorithmic fairness* with respect to protected attributes. The right to be forgotten, as formalized in regulations like the GDPR [16], requires the removal of personal data from automated systems. Meanwhile, fairness concerns have led to a growing literature on mitigating bias in model predictions [1, 2, 5, 10]. These goals can be in tension: removing data points—particularly from marginalized groups—may undo fairness guarantees. Yet current unlearning methods [4, 8] largely ignore fairness, focusing instead on utility or privacy.

Our contributions are:

- We propose a variational unlearning framework that integrates fairness constraints during forgetting.
- We apply our method to the Adult Income dataset [6], focusing on the effect of forgetting datapoints from protected subgroups.
- Our findings highlight the importance of jointly optimizing unlearning and fairness to ensure responsible model behavior.

2 FAIR VARIATIONAL UNLEARNING

We consider a Bayesian setting where the original model is trained on dataset $\mathcal{D}$, and a subset $\mathcal{F} \subset \mathcal{D}$ must be unlearned. The goal is to construct a new posterior $q(\theta)$ that (i) minimizes the influence of $\mathcal{F}$, (ii) preserves fairness under a group-level criterion such as demographic parity, and (iii) stays close to the original posterior $q_0(\theta)$ for stability and certifiability.

2.1 Variational formulation

Let $\mathcal{L}_{\mathcal{F}}(\theta)$ denote the (negative) log-likelihood on the forget set, and let $\mathcal{F}_{\text{fair}}(\theta)$ be a differentiable penalty capturing fairness violations (e.g., squared demographic parity gap [9, 10]). Following prior work on variational unlearning [4, 8, 12], we define the fair unlearning posterior via a constrained optimization:

$$q^*(\theta) = \arg\min_{q \in \mathcal{Q}} \ \mathrm{KL}(q \,\|\, q_0) - \lambda_1 \, \mathbb{E}_q[\mathcal{L}_{\mathcal{F}}(\theta)] + \lambda_2 \, \mathbb{E}_q[\mathcal{F}_{\text{fair}}(\theta)], \quad (1)$$

where $\lambda_1, \lambda_2 > 0$ control the strength of forgetting and fairness constraints. This generalizes standard variational inference objectives [3, 17] by including a fairness penalty and negative influence term on $\mathcal{F}$.

2.2 Stability-Fairness Guarantee

We provide a bound on how much fairness and loss metrics can change due to unlearning-induced posterior shift. This is based on transportation inequalities between KL divergence and Wasserstein distances [14, 15]:

Theorem 2.1. *Let $\mathcal{F}_{fair}$ and $\mathcal{L}_{test}$ be L_{fair} and L_{loss}-Lipschitz in θ, respectively. Suppose q satisfies $\mathrm{KL}(q \,\|\, q_0) \le \varepsilon$, and q_0 is strongly log-concave with variance proxy σ^2. Then:*

$$\left| \mathbb{E}_q[\mathcal{L}_{test}(\theta) + \lambda_2 \mathcal{F}_{fair}(\theta)] - \mathbb{E}_{q_0}[\cdot] \right| \le (L_{loss} + \lambda_2 L_{fair}) \cdot \sqrt{2\sigma^2 \varepsilon}. \quad (2)$$

KDD '25, June 03–05, 2018, Woodstock, NY

ACM ISBN 978-1-4503-XXXX-X/2018/06
https://doi.org/XXXXXXX.XXXXXXX

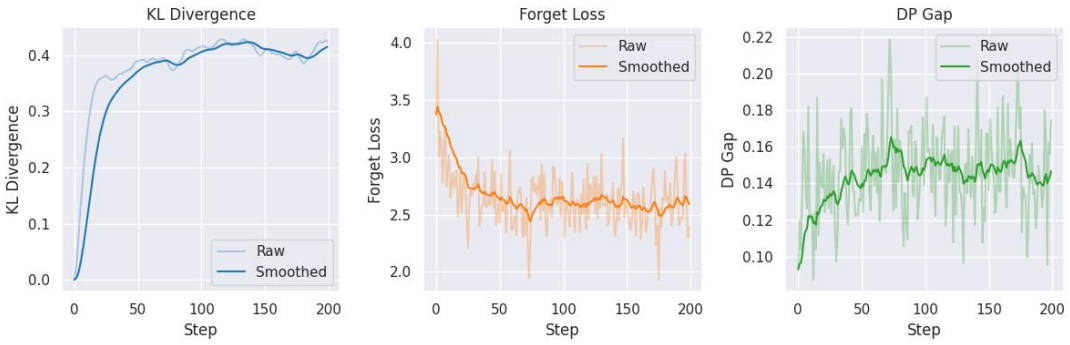

Figure 1: Evolution of losses when $\lambda_2 = 0$ (with $\lambda_1 = 1$ and $K = 10$). In particular, the DP gap *increases*.

This generalization bound ensures that as long as the new posterior q stays close to q_0, the model's performance and fairness will not degrade significantly. This aligns with recent concerns in unlearning research about certifiability and minimal retraining [7, 11].

3 ALGORITHM

We optimize the fair unlearning objective using stochastic variational inference, assuming a Gaussian posterior $q(\theta) = \mathcal{N}(\mu, \Sigma)$. The algorithm uses Monte Carlo sampling to estimate gradients:

Algorithm 1: Fair Variational Unlearning

Input: Original posterior q_0, forget set $\mathcal{F}$, retained set $\mathcal{R}$, fairness penalty $\mathcal{F}_{\text{fair}}$
Output: Updated posterior q

Initialize $\mu \leftarrow \mu_0$, $\log \sigma \leftarrow \log \sqrt{\text{diag}(\Sigma_0)}$
for $t = 1$ **to** T **do**
  Sample K points $\theta_k \sim \mathcal{N}(\mu, \sigma^2)$
  Compute forget loss: $\ell \leftarrow \frac{1}{K} \sum_k -\log p(\mathcal{F} \mid \theta_k)$
  Compute fairness penalty: $f \leftarrow \frac{1}{K} \sum_k \mathcal{F}_{\text{fair}}(\theta_k)$
  Compute KL divergence $\text{KL}(q \,\|\, q_0)$ in closed form
  Take gradient step on:

$$\mathcal{L} = \text{KL}(q \,\|\, q_0) + \lambda_1 \cdot \ell + \lambda_2 \cdot f$$

4 NUMERICAL RESULTS

We evaluate our method on the Adult Income dataset [6], using logistic regression with a Gaussian variational posterior. We define fairness via the demographic parity gap and conduct unlearning experiments where we remove 100 datapoints belonging to a protected group (e.g., females).

Demographic Parity Gap. We define the *demographic parity gap* as the absolute difference between group-wise positive prediction rates: $\text{DP Gap}(\theta) := |\mathbb{E}[f_\theta(X) \mid A = 0] - \mathbb{E}[f_\theta(X) \mid A = 1]|$. This serves as a fairness penalty $\mathcal{F}_{\text{fair}}(\theta)$ in our unlearning objective.

Trade-off Analysis. Figure 1 shows the trade-off between fairness and forgetting across different λ_1. As fairness constraints tighten (increasing λ_2), the model becomes more equitable but fits the forget set better—demonstrating a concrete trade-off between fairness and deletion effectiveness. We further plot the demographic parity gap as a 3D surface over the (λ_1, λ_2) grid (Figure 2), confirming that fairness loss increases predictably with stronger forgetting.

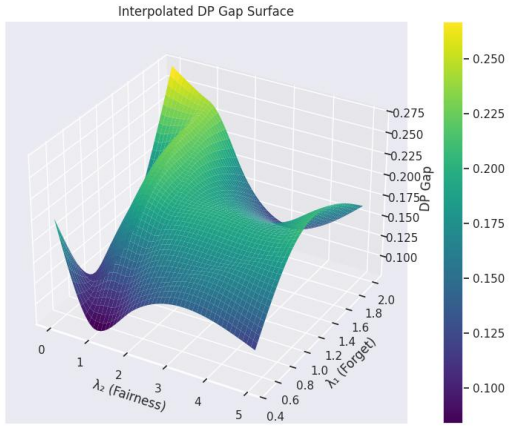

Figure 2: Pareto surface showing the trade-off between forgetting and fairness.

Key Insight. Removing fairness-supporting data can significantly increase demographic disparity.

5 CONCLUSION

We presented a variational framework for machine unlearning that incorporates fairness constraints. Our experiments show that forgetting subpopulations—especially those from protected groups—can significantly increase demographic disparities. By sweeping fairness and forgetting penalties, we revealed concrete trade-offs between equitable outcomes and deletion fidelity.

Our findings highlight the importance of integrating fairness considerations directly into unlearning objectives (and broader frameworks [13]). While our approach offers a principled starting point, there remains substantial work to be done: extending this framework to deep models, other fairness definitions, and real-world unlearning requests will be critical for deploying truly responsible ML systems.

REFERENCES

[1] Ashrya Agrawal, Florian Pfisterer, Bernd Bischl, Francois Buet-Golfouse, Srijan Sood, Jiahao Chen, Sameena Shah, and Sebastian Vollmer. 2021. Debiasing classifiers: is reality at variance with expectation? arXiv:2011.02407 [cs.LG] https://arxiv.org/abs/2011.02407

[2] Solon Barocas, Moritz Hardt, and Arvind Narayanan. 2019. *Fairness and Machine Learning*. http://fairmlbook.org.

[3] David M Blei, Alp Kucukelbir, and Jon D McAuliffe. 2017. Variational inference: A review for statisticians. *J. Amer. Statist. Assoc.* 112, 518 (2017), 859–877.

[4] Ludovic Bourtoule, Varun Chandrasekaran, Christopher A Choquette-Choo, et al. 2021. Machine unlearning. In *IEEE Symposium on Security and Privacy (S&P)*.

[5] Francois Buet-Golfouse and Islam Utyagulov. 2023. Fairness Trade-Offs and Partial Debiasing. In *Proceedings of The 14th Asian Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 189)*, Emtiyaz Khan and Mehmet Gonen (Eds.). PMLR, 112–136. https://proceedings.mlr.press/v189/buet-golfouse23b.html

[6] Dheeru Dua and Casey Graff. 2017. UCI Machine Learning Repository. https://archive.ics.uci.edu/ml.

[7] Aviad Ginart, Megan Guan, Gregory Valiant, and James Zou. 2019. Making AI forget you: Data deletion in machine learning. In *International Conference on Machine Learning (ICML)*. 3510–3519.

[8] Aditya Golatkar, Alessandro Achille, and Stefano Soatto. 2020. Eternal sunshine of the spotless net: Selective forgetting in deep networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 9304–9312.

[9] Moritz Hardt, Eric Price, and Nathan Srebro. 2016. Equality of opportunity in supervised learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, Vol. 29.

[10] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. 2021. A Survey on Bias and Fairness in Machine Learning. *ACM Computing Surveys (CSUR)* 54, 6 (2021), 1–35.

[11] Seth Neel and Aaron Roth. 2021. Desiderata for Machine Unlearning. arXiv:2107.00010 [cs.LG]

[12] Cuong V Nguyen, James Liu, and et al. 2022. Variational Machine Unlearning. arXiv:2209.10346 [cs.LG]

[13] Roberto Pagliari, Peter Hill, Po-Yu Chen, Maciej Dabrowny, Tingsheng Tan, and Francois Buet-Golfouse. 2024. A Comprehensive Sustainable Framework for Machine Learning and Artificial Intelligence. *arXiv preprint arXiv:2407.12445* (2024).

[14] Maxim Raginsky, Alexander Rakhlin, and Matus Telgarsky. 2017. Non-asymptotic analysis of stochastic gradient Langevin dynamics. *Conference on Learning Theory (COLT)* (2017), 1674–1703.

[15] Cédric Villani. 2008. *Optimal Transport: Old and New*. Springer.

[16] Paul Voigt and Axel von dem Bussche. 2017. *The EU General Data Protection Regulation (GDPR)*. Springer.

[17] Yang Zhang, Bo Dai, Yuchen Wang, and et al. 2021. Understanding posterior collapse in generative latent variable models. In *International Conference on Learning Representations (ICLR)*.