We would like to thank the reviewers for their constructive and thoughtful comments. They recognized that this is a *pressing issue* for the research community and that our work is potentially a *very important contribution* and *clearly presented*. We have made an effort to answer all their comments and will update our paper based on this rebuttal.

R1. We would like to highlight the contributions of this work. 1) We reveal new results, instead of simply confirming those of related work (e.g.; Buolamwini and Gebru, 2018). Prior work analyzed skin tone differences but not those across different regions. We report nuances between geographic regions including difference between South Asian and African regions, both of which feature people with darker skin tones. 2) While the individual components of our approach – cGAN, PG-GAN and Bayes sampling – already exist in the literature, we combine them in a novel way to characterize bias in classifiers. We show this approach allows for identification of 50% more failure cases than without Bayesian sampling and 16% more failure cases than without our conditional PG-GAN (see our response to R3 below). This leads to more efficient identification of bias. Furthermore, the synthesized images can then be used to improve the performance of a classifier (as described below). We will add the references suggested which are indeed relevant.

Per R1's suggestion, we performed additional experiments to show how the synthetic images can be used to improve weak spots of classifiers. Specifically. we trained three ResNet-50 gender classifications models. First, we sampled 14K images from our dataset (1K Black, 3K S Asian, 5K White and 5K NE Asian - this captures a typical model trained on unbalanced data). The classification accuracy was 81%. We then retrained this model, adding 800 synthesized failure cases discovered using our method, the retrained classifier achieved 87% accuracy. Compare this to a model trained on our balanced dataset for which the accuracy was 92%. Adding the synthesized images brings performance up, close to that of a model trained on a balanced dataset.

R2. 1) The classifier C in Eq (2) is independent of the classifier being tested. **2)** For the classifier being tested, the 0-1 classification loss is 0 if the classification of gender was correct or 1 if it was incorrect. We maximize this to find regions of the space in which the classifier fails the most. **3a)** In our preliminary tests, we varied the size of the set of previously found examples. We found the results were not sensitive to the size of this set and fixed it to 50 in our main experiments. **3b)** We ran the Bayesian Optimization for a fixed number of iterations (1000) in each trial. In all cases the percentage of failure cases had stabilized. As can be seen in Fig.4(a) the differences between the approaches typically become clear after several hundred iterations. **3c)** We will provide more details with the equations for Bayesian Optimization (BO). In summary, the BO routine aims to model the compositing function $L_c(\theta)$ via a Gaussian Process as $\sum_i \alpha_i$ * RBF(θ, θ_i). Here θ_i are training data points where $L_c(\theta_i)$ has already been evaluated. Since the goal of the BO is to find θ that maximizes the loss, we choose the next point to query via Expected Improvement (EI) as the one that promises the biggest gain in utility on average. A formal definition of EI is provided in Frazier et al. 2018[1] and we will summarize it in our paper. **3d)** The BO searches a continuous eight dimensional space and outputs a real-valued vector; we apply argmax and convert it into a one-hot vector. So the generator always receives a one-hot vector.

R3. Our reasons for using a GAN to synthesize images are two fold: 1) We can sample a larger number of faces with a greater variability than existed in the original photo datasets; 2) Synthesizing face images has an advantage of preserving the privacy of individuals, thus when interrogating a commercial classifier we do not need images of "real people". Per R3's suggestion, we have repeated our analysis by directly sampling from the real images used to train the PG-GAN. Using BO and the GAN images allowed us to discover gender detection failures at a higher rate (16% higher) than using BO and the real images (both with α=0.6), partially because the real image set is ultimately limited and in specific failure regions we eventually exhaust the images and have to sample elsewhere. This further supports the use of a GAN. To describe the quality for the different regions we have computed the *accuracy* of the model at producing images of the correct gender for the different regions and the *mean quality rating* (score from 0-5 in brackets): Black Men: 100% (3.31), Black Women: 98% (3.24), White Men: 100% (3.20), White Women: 100% (3.51), S. Asian Men: 100% (3.37), S. Asian Women: 100% (3.40), NE Asian Men: 88% (3.48), NE Asian Women: 100% (3.58).

To compute the FID scores for each region and gender, we synthesized 500 images from each region and gender (4K in total), using our conditional PG-GAN and StyleGAN (Karras et al., 2019). StyleGAN is the current state-of-the-art on face image synthesis (published after our paper submission) and serves as a good reference point. The FID scores were similar across all regions: Ours (Black: M 8.10, F 8.14; White: M 8.08 F 7.70; NE Asian: M 8.01 F 8.00; S Asian: M 8.06 F 8.10). StyleGAN (Black: M 7.70 F 7.92; White: M 7.68 F 7.8; NE Asian: M 7.76 F 7.80; S Asian: M 7.94 F 7.66). Our model produces comparable FID scores with the state-of-the-art results. Note that StyleGAN synthesizes new images by conditioning on a real image, while our model is just conditioned on labels. The results confirm that our dataset can be used to synthesize images across each gender and region with sufficient quality and diversity.

As this is the first work to use GANs to interrogate bias in facial classification systems we chose two particularly important dimensions to control: gender and race. The number of parameters could be extended with other critical dimensions, e.g., age; this would further accentuate the advantages of using generative models that we have highlighted.

[1] Frazier, P. I. (2018). A tutorial on Bayesian optimization. arXiv preprint arXiv:1807.02811