THE BIAS OF HARMFUL LABEL ASSOCIATIONS IN VISION-LANGUAGE MODELS

Caner Hazirbas, Alicia Sun, Yonathan Efroni, Mark Ibrahim
Meta AI

ABSTRACT

Despite the remarkable performance of foundation vision-language models, the shared representation space for text and vision can also encode harmful label associations detrimental to fairness. While prior work has uncovered bias in vision-language models' (VLMs) classification performance across geography, work has been limited along the important axis of harmful label associations due to a lack of rich, labeled data. In this work, we investigate harmful label associations in the recently released Casual Conversations datasets containing more than 70,000 videos. We study bias in the frequency of harmful label associations across self-provided labels for age, gender, apparent skin tone, and physical adornments across several leading VLMs. We find that VLMs are $4-7$x more likely to harmfully classify individuals with darker skin tones. We also find scaling transformer encoder model size leads to higher confidence in harmful predictions. Finally, we find improvements on standard vision tasks across VLMs does not address disparities in harmful label associations.

1 INTRODUCTION

Vision-language models (VLMs) offer a new capability for zero-shot open-world object recognition: the ability to classify any object without additional training. Compared to standard supervised training where the set of classes is predefined, VLMs can classify any text-image pairs thanks to a shared embedding space for images and text. With this new capability however, comes the potential for VLMs to learn a multidude of harmful spurious correlations between text and images.

While prior work attempted to study such biases in the context of domain shifts (Kalluri et al., 2023), studies specifically tackling these sensitive axes have been limited due to dataset challenges (Richards et al., 2023). Few, if any datasets, contain large diverse visuals across these many axes of gender, geography etc. and when they do, they often lack ground truth labels. For example, DollarStreet contains 30k samples with ground truth labels only for income and geography. Here we study these critical axes using the recently released Casual Conversations datasets Hazirbas et al. (2022b); Porgali et al. (2023). Casual Conversations contains more than $70,000$ videos of $8,500$ individuals with self-provided age, gender, physical attributes & adornments and geo-location and apparent skin tone.

Our study focuses on vision-language model predictions associating a human with a harmful class label (on the standard 1k ImageNet classes)—in other words, a *harmful prediction*. We study several vision-language models pre-trained on web scale data that perform well on standard classification benchmarks, including CLIP Radford et al. (2021) of various transformer encoder sizes and BLIP-2 (Li et al., 2023) and other variants (Li et al., 2022). We measure the extent harmful predictions disproportionately affect some groups more than others across age, gender, and apparent skin tone as well as the effect of physical adornments such as glasses and hats. We find 1) VLMs are $4-7$x more likely to harmfully classify individuals with darker skin tones relative those with lighter skin tones, 2) larger transformer models are more confident in their harmful label associations, and 3) progress on standard vision tasks does not necessarily improve disparities in harmful label associations.

Recent work has shown that vision-language model could induce harms by amplifying societal biases (Agarwal et al., 2021; 2023; Zhang et al., 2022; Hirota et al., 2023), and causing performance disparities across demographic groups (Hall et al., 2023), yet these analysis were only conducted on specific demographic groups. In this work, we investigate harmful label associations across a more comprehensive set of axes.

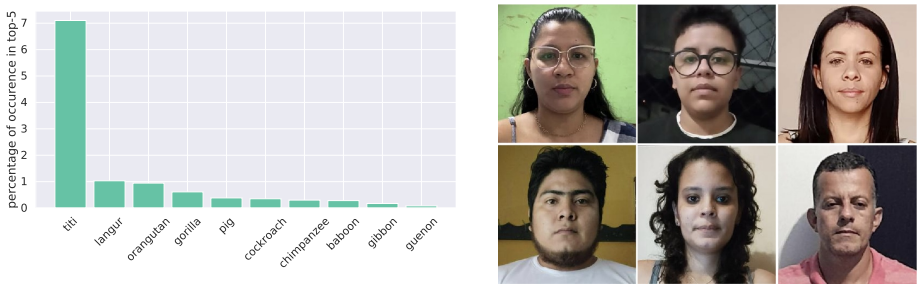

Figure 1: Most commonly selected harmful class labels for CLIP ViT-L14 (left) along with sample frames leading to harmful label associations from CCv2 (right).

2 METHODOLOGY

We evaluate bias in vision-language models on geographically diverse video datasets of humans from the Casual Conversations v1 and v2 datasets. We compare harmful label associations of several leading vision-language models across self-identified attributes such as age, gender and apparent skin tone as well as physical adornments including face mask, beard/mustache, hair cover, etc.

In our experiments, we use the both versions of the Casual Conversations datasets Hazirbas et al. (2022b); Porgali et al. (2023). Both datasets provide self-identified age, gender and apparent skin tone per participant, while CCv2 is richer and more diverse in terms of participant and labels. CCv1 is composed of 3,011 participants collected across five cities in the US. CCv2 has 5,567 participants from seven different countries. We processed videos as described in A.1.

We focus this study on foundation models such as CLIP which encode both vision and text in the same representation space. We compare CLIP models with ViT, transformer encoders, of varying sizes including B16, B32, and L14. We also study a more recent vision-language model, BLIP2, trained with additional captioning and image-text matching objectives. To classify an image, we encode both the image and text prompts for each class label then predict the class label with the highest similarity to our image representation following the same procedure in Radford et al. (2021).

We focus our study of bias on harmful label associations: model label predictions that inappropriately classify a human as another harmful label, such as a primate. We use a standard zero-shot classification across the standard 1k ImageNet classes. We use the 80 text prompts [1] with simplified 1k ImageNet class labels [2] by extending with "people" and "face". Among these 1k labels, we mark labels as a harmful association using the taxonomy from Goyal et al. (2022), which designates classes such as gorilla as harmful label associations (see Appendix A.3 for details). To evaluate harmful labels we use the top-5 among a models' class predictions for all our analysis and consider a prediction harmful if the majority of labels in the top-5 constitute harmful label associations. We show in Figure 1 the most commonly predicted harmful labels by a CLIP ViT-L14 model. We find primates tend to be the most commonly predicted harmful label followed by "pig" and "cockroach".

Bias in harmful label associations. While we should strive to develop models that produce safe predictions (that is do not contain harmful labels) in this work we are particularly interested in the bias of harmful predictions across protected attributes such as age, gender, and apparent skin tone to understand whether harmful model predictions disproportionately affect some groups more than others. To do so, we examine the distribution of harmful labels across each group using the self-identified labels in the Casual Conversations datasets.

3 HARMFUL LABEL ASSOCIATION DISPARITIES

We first measure the percent of harmful label associations across several models. We then investigate disparities in harmful label associations across age, gender, and apparent skin tone. Next we account

[1] https://github.com/openai/CLIP/blob/main/notebooks/Prompt_Engineering_for_ImageNet.ipynb

[2] https://github.com/anishathalye/imagenet-simple-labels/blob/master/imagenet-simple-labels.json

for model confidence in our disparity measures to tease out whether some models predict harmful label associations with lower confidence. Finally, we study the effect of physical adornments on harmful predictions.

CLIP and BLIP2 exhibit reverse bias trends across gender and age. We measure the proportion of harmful label associations across gender, age, and apparent skin tones in Figure 2. For gender, we find transformer-based CLIP models exhibit a higher proportion of harmful label associations for women than for men. For example, CLIP ViT-B32 predicts harmful label associations for cis women at a rate of 57.5% compared to only 27.5% for cis men. The harmful association rate for non-binary individuals for CLIP models falls in-between that of cis women and cis men at 39.8%. On the other hand, the BLIP-2 model predicts harmful label associations much less frequently for cis women (33.6%) than cis men (45.7%) and similarly non-binary individuals (46.1%). We also find reversed bias trends for CLIP versus BLIP-2 for age. While CLIP models tend predict more harmful label associations for middle-aged and older adults relative to those for adult and young adults, BLIP-2's bias is reversed.

Harmful label associations are 4x more likely for darker skin tones. Finally, we find a stark difference in the percent of harmful label associations across apparent skin tones as show in Figure 2 with harmful predictions occurring nearly 4x more on average for darker skin tones (type vi Fitzpatrick) compared to lighter skin tones: 72.9% darker versus 21.6% lighter. The disparity is consistent across all models we evaluated with BLIP-2 exhibiting a disparity in harmful label associations of 7x across skin tones: 44.5% for darker versus just 6.7% for lighter.

Progress on standard vision tasks does not improve disparities in harmful label associations for apparent skin tones. Comparing harmful association disparities for BLIP-2 and CLIP models yields an important implication for the research community. While BLIP-2 achieves markedly better performance across a variety of vision tasks relative to CLIP, BLIP-2's disparities in harmful label associations across skin tones are more than 2x worse compared to those of CLIP: BLIP-2 is 7x more likely to harmfully classify individuals with apparent darker skin tones while CLIP's does so at 3x the rate. All the while, on standard vision benchmarks BLIP-2 achieves markedly better results. For example, BLIP-2 achieves zero-shot retrieval of 94.9% for Flick30k while CLIP only achieves 88% with BLIP-2 demonstrating comparable gains on other benchmarks (COCO zero-short retrieval 85.4% for BLIP-2 while only 77.0% for CLIP) (Li et al., 2023). This contrast between performance on standard benchmarks and disparities in harmful label associations for skin tones suggests *improving performance on standard vision benchmarks does necessarily not improve disparities in harmful label associations*.

Some individuals are consistently harmfully classified across all videos in the dataset We find for nearly 4.4% of individuals (245 out of 5566) the same individual is harmfully associated in model predictions across all videos. To control for confounding factors such as pose, resolution or lighting, we show a randomly selected set of frames for individuals in the right panel of Figure 1 as as well a random selection of frames of the same individuals across different videos who consistently harmfully associated by a CLIP ViT-L14 model in Appendix Figure 7.

3.1 Accounting for model confidence in harmful label associations

In addition to measuring the overall proportion of harmful label associations among a model's predictions, we account for a model's confidence in such predictions. To do so, we first compute the softmax normalized similarities between text label representations $t \in \mathbb{R}^{k \times d}$ and an image representation z as *normalized sim* = $\mathrm{softmax}(tz^\top)$. We then weigh each harmful prediction by its corresponding mean *normalized sim* in the top-5 predictions. We show the resulting confidence weighted percent of harmful label associations in Figure 4. We surface two noteworthy trends:

Larger ViT models are more confident in their harmful label associations. While model sizes across CLIP models showed no consistent trend with respect to harmful label associations when we measured the overall percent of harmful label associations, we find in contrast, CLIP models with larger encoders are much more confident in their harmful predictions relative to CLIP models

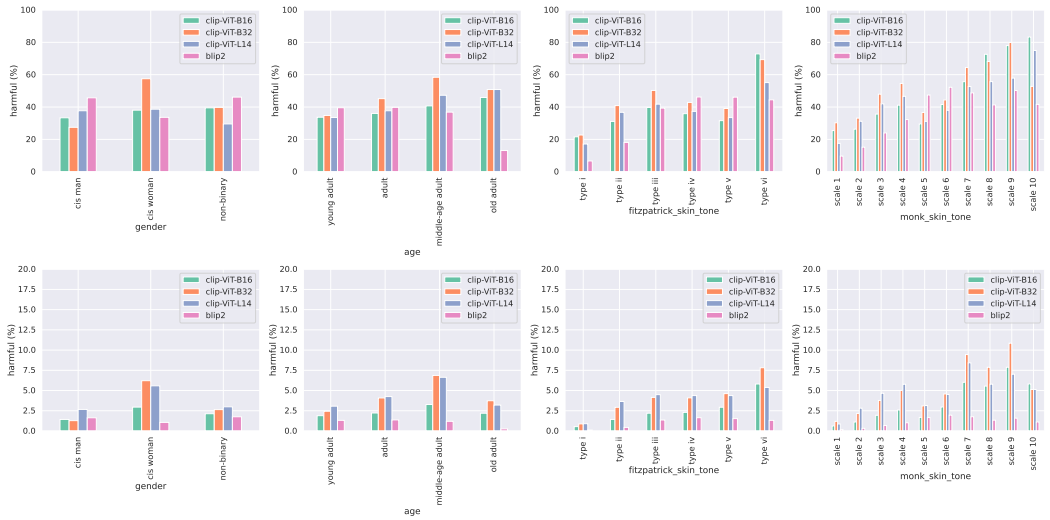

Figure 2: CCv2 breakdowns of harmful label associations. Second row shows confidence weighted scores.

with smaller encoders. This overconfidence of larger models was consistent across gender, age, and apparent skin tones.

BLIP in contrast to CLIP is much less confident in its harmful label associations. Although BLIP-2 exhibited overall harmful association rates comparable or in some cases higher than those for CLIP, we find after accouting for confidence in these prediction BLIP-2 harmful prediction rates are consistently lower than those for CLIP across all model sizes. This suggests while BLIP-2 also produces harmful label associations, the model confidence is appropriately low for these harmful predictions. As shown in Appendix Figure 5, these trends hold both for the CCv2 and CCv1 datasets.

3.2 The effect of physical adornments on harmful label associations

Finally, given humans may appear with various physical adornments such as glasses, make up, face masks, and hair covers, we use the self-identified labels in CCv2 to study how the presence of such physical adornments affects models' harmful predictions. We study a collection of 8 physical adornments shown in Figure 3. We find the physical adornments that cover facial features such as facial masks, eye wear, bears or mustaches considerably decrease the percent of harmful label associations. For example, the presence of a face mask decreases the percent of harmful label associations drastically across all models as shown in the top left panel of Appendix Figure 3. This decrease is consistent across all models suggesting models may be relying on facial features obscured by face masks such as the mouth and nose in their harmful label associations. We find a similar trend when we account for model confidence of those prediction as shown in Appendix Figure 4.

4 Discussion

We investigated disparities in models' harmful label associations across age, gender, and apparent skin tone. We find models exhibit significant bias across groups within these important axes with the most alarming trend arising for apparent skin tone: CLIP and BLIP-2 are 4-7x more likely to harmfully associate individuals with darker skin than those with lighter skin. We also account for model confidence, finding larger models exhibit more confidence in harmful label associations suggesting scaling models, while helpful on standard benchmarks, can exacerbate harmful label associations. Finally, we find improved performance on standard vision tasks does not necessarily correspond to improvements in harmful association disparities, suggesting addressing such disparities requires concerted research efforts with this desideratum in mind.

REFERENCES

Sandhini Agarwal, Gretchen Krueger, Jack Clark, Alec Radford, Jong Wook Kim, and Miles Brundage. Evaluating clip: Towards characterization of broader capabilities and downstream implications. *ArXiv*, abs/2108.02818, 2021.

Sandhini Agarwal, Gretchen Krueger, Jack Clark, Alec Radford, Jong Wook Kim, and Miles Brundage. Identifying implicit social biases in vision-language models. *ICML 2023 Workshop on Challenges of Deploying Generative AI, Honolulu, Hawaii, USA.*, 2023.

Priya Goyal, Adriana Romero-Soriano, Caner Hazirbas, Levent Sagun, and Nicolas Usunier. Fairness indicators for systematic assessments of visual feature extractors. In *FAccT '22: 2022 ACM Conference on Fairness, Accountability, and Transparency, Seoul, Republic of Korea, June 21 - 24, 2022*, 2022.

Melissa Hall, Laura Gustafson, Aaron Adcock, Ishan Misra, and Candace Ross. Vision-language models performing zero-shot tasks exhibit disparities between gender groups. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*, pp. 2778–2785, October 2023.

Caner Hazirbas, Yejin Bang, Tiezheng Yu, Parisa Assar, Bilal Porgali, Vítor Albiero, Stefan Hermanek, Jacqueline Pan, Emily McReynolds, Miranda Bogen, Pascale Fung, and Cristian Canton Ferrer. Casual conversations v2: Designing a large consent-driven dataset to measure algorithmic bias and robustness. *arXiv*, 2022a.

Caner Hazirbas, Joanna Bitton, Brian Dolhansky, Jacqueline Pan, Albert Gordo, and Cristian Canton Ferrer. Towards measuring fairness in ai: The casual conversations dataset. *IEEE Transactions on Biometrics, Behavior, and Identity Science*, 2022b.

Yusuke Hirota, Yuta Nakashima, and Noa Garcia. Model-agnostic gender debiased image captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2023.

Tarun Kalluri, Wangdong Xu, and Manmohan Chandraker. Geonet: Benchmarking unsupervised adaptation across geographies, 2023.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation, 2022.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models, 2023.

Bilal Porgali, Vítor Albiero, Jordan Ryda, Cristian Canton Ferrer, and Caner Hazirbas. The casual conversations v2 dataset. In *CVPR*, 2023.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, 2021.

Megan Richards, Polina Kirichenko, Diane Bouchacourt, and Mark Ibrahim. Does progress on object recognition benchmarks improve real-world generalization?, 2023.

Yi Zhang, Junyang Wang, and Jitao Sang. Counterfactually measuring and eliminating social bias in vision-language pre-training models. In *Proceedings of the 30th ACM International Conference on Multimedia*, 2022.

Figure 3: CCv2 physical adornments breakdowns.

Figure 4: CCv2 physical adornments breakdowns (confidence weighted).

A Appendix

A.1 Datasets

For both datasets, we grouped the age into four buckets as "young adult" ($0-24$), "adult" ($25-40$), "middle-age adult" ($41-65$) and "old adult" ($66-100$) Hazirbas et al. (2022a). In CCv2, considering relatively lower number of participants, we combined "transgender" and "non-binary" as "non-binary". In order to run zero-shot classification, we processed the videos to crop faces as described in Hazirbas et al. (2022b), totalling of $45,116$ face crops/videos in CCv1 and $26,358$ face crops/videos in CCv2.

A.2 Harmful label associations for CCv1

We also measure harmful label associations and their bias across the same set of groups for the CCv1 dataset across the models we evaluated as shown in Figure 5.

A.3 Harmful label associations

Class label designations: 1) *harmful*: orangutan, gorilla, chimpanzee, gibbon, siamang, guenon, patas monkey, baboon, macaque, langur, black-and-white colobus, proboscis monkey, marmoset, white-headed capuchin, howler monkey, titi, Geoffroy's spider monkey, common squirrel monkey, cockroach, pig; 2) *possibly-human*: sunglass, sunglasses, sweatshirt, lipstick; 3) *human*: people and face. Although these lists are not the final, it allows us to measure the most prominent harmful label associations in VLMs. We show sample frames for individuals across self-labeled gender categories in Figure 6 all of which are harmfully associated by a CLIP ViT-L14 model. We also show sample frames for the same individuals who are consistently harmfully associated across all videos in Figure 7. In addition, Figure 8 demonstrates harmful associations on top-1 predictions.

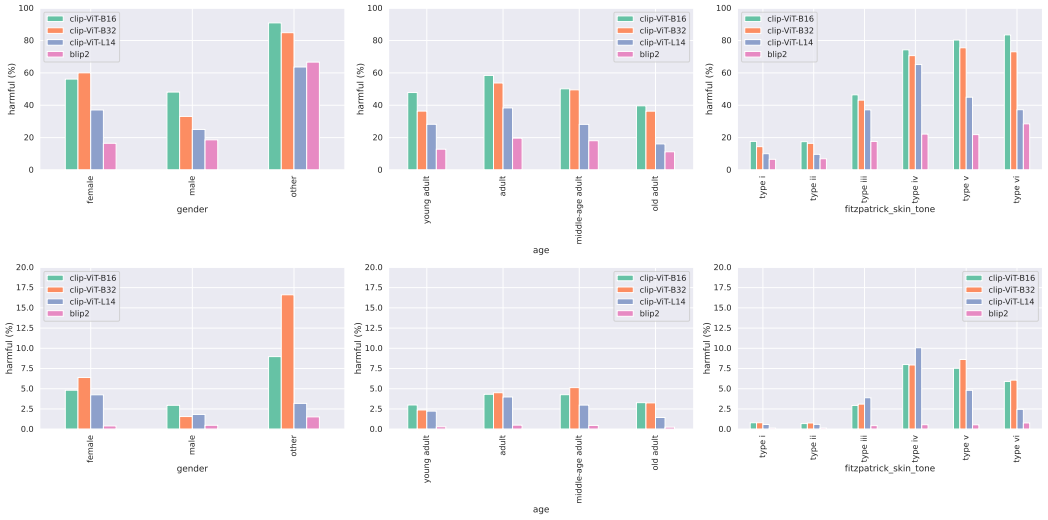

Figure 5: CCv1 breakdowns of harmful label associations. Second row shows confidence weighted scores.

(a) Cis Woman (b) Cis Man (c) Transgender/Non-binary

Figure 6: Sample face crops. All videos of these participants are harmfully associated by the CLIP ViT-L14 model.

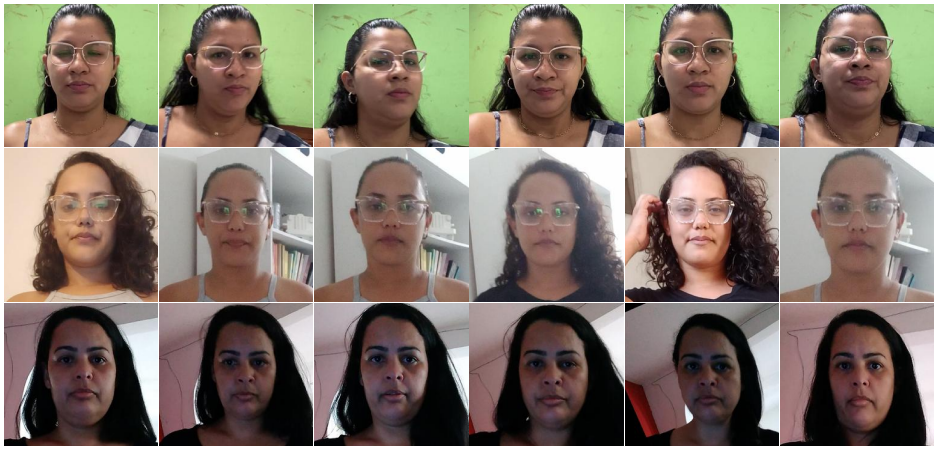

Figure 7: All of the videos of these participants classified harmful by CLIP ViT-L14.

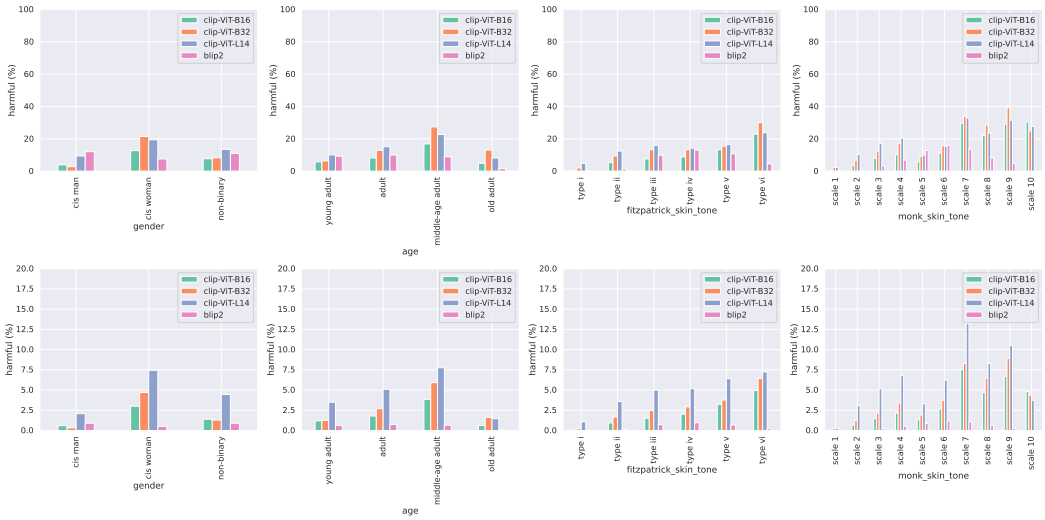

Figure 8: CCv2 breakdowns of harmful label associations on top-1. Second row shows confidence weighted scores.