It is important to understand the biases of the neural network, even if it is performing well on the training set. With AI models, explainability and methodology used in addition to purely performance are important. The paper shows that ImageNet trained CNNs have a texture bias, which is contrast to humans which have a shape bias in object recognition. Understanding how a pretrained ImageNet works is important when trying to apply it to different datasets with transfer learning. It now makes sense why the pretrained model may perform very poorly on datasets with images with just silhouettes, edges, or lack of textural cues that correctly align with the classification. I think that it is ideal for a model to have the same biases as humans, since we often use domain knowledge to help design experiments and the image augmentations for the experiment. The domain knowledge may be less helpful if the model is not understanding or utilizing the curated features in a similar way as the domain expert. In addition, shape based bias similar to humans has additional benefits such as better robustness against distortions even those it had never seen in training and better transfer learning in object detection.

Training on stylized images changes the biases of the network to shape based, because it is no longer able to make correct classifications solely by looking at local object related textural information. Therefore, the model is forced to learn by integrating and classifying global shapes of the object. This bias generalizes better to datasets with corruptions because it can still recognize the object shape information which is often still preserved while corruptions may degrade textural information through distortions like phase noise, contrast changes, or high/low pass filtering.