Figure 1: The 5 components with the highest reasoning probability for each type of reasoning for 'Newfoundland dog'.

Thanks to all reviewers for the effort and the patience leading to such interesting questions and very helpful comments to improve the quality of the contribution. As most reviewers highlighted similar questions and areas within the paper, we have chosen to discuss the comments in three main themes. But first we will give some general clarification.

Clarification: The essential new idea proposed in CBCs is the probabilistic constraint on both, the penultimate and the final layer of feedforward NNs for classification, as stated in the conclusion. Both support the interpretability of the network. The main complexity of CBCs originates in the feature extractor which is used to improve the component detection. In general, the CBC is not restricted to a CNN as feature extractor. The use of a Siamese architecture is also not a hard requirement. As it can be seen in the experiments on IMAGENET presented here and in the paper, a similar interpretability of the components can be achieved without a Siamese architecture. Please excuse that we have an incorrect translation in the current version of the supplementary that we will correct: *We define our probability space over a set of events related to a probability tree diagram and not to a decision tree.*

1) Clarity of the paper: To improve the clarity and self-containment of our paper, we propose to use the additional page of the final version to move the class-dependent probability tree diagram (Fig. 6 in the supplementary material), that models the class hypothesis probability $p_c(\mathbf{x})$, from the supplementary to the main part of the paper and provide a short mathematical derivation and a short explanation of the training procedure accordingly. This will provide clarity regarding the random variables (e.g. $A|I$) and improve the explanation of the visualizations in Sec. 4.1.2., as these can then be directly related to paths in the tree diagram. Including the probability tree diagram also clarifies how BIEDERMAN's RBC-theory is realized and the distinction between CBCs and the modeling of knowledge in general graph structures, e.g. in the mentioned ICML paper about the learning of trees. Opposed to general modeling of knowledge in graphs, CBCs only rely on the proposed kinds of reasoning and assume that components are stochastically independent. These assumptions are needed to keep the method simple. Moreover, we agree that our related work section has to be extended and that additional cross-references and citations have to be included to improve the paper.

2) Extended evaluation of complex datasets: We acknowledge that the experimental results presented in the paper are lacking complexity. This is partly due to the introductory nature of the paper. The datasets were chosen such that they reflect the functionality of different parts of the proposed approach. The experiments on MNIST show that the CBC architecture can fulfill the goals in principle. CIFAR-10 and GTSRB show that the model is capable of learning color components and IMAGENET shows the scalability of the approach. We have however neglected to show parts of the approach working on more complex datasets. To solve this, we plan to extend the evaluation on IMAGENET. To make space for this, we will move the experiments on CIFAR-10/GTSRB to the supplementary. In Fig. 1 the proposed extension of the experiments is shown. Using the discussed back projection method, we have computed a respective mapping of the $2 \times 2 \times 2048$ components used in the detection probability function to components that are part of the dataset. The proposed extension shows the five components with the highest reasoning probabilities for positive, negative, and indefinite reasoning for the class 'Newfoundland dog'. Investigating these components leads to a deeper understanding of the model's classification. For example the split into different types of dog snouts over positive, negative and indefinite components shows the importance of the type of dog snout for the classification. Notably, special breeds of the Newfoundland dog are known for their white marks. This characteristic is modeled by the first positive reasoning component. Considering that not all breeds of the Newfoundland dog have these marks, this might indicate a bias in the IMAGENET dataset.

We believe that this extension of the experiment on IMAGENET delivers further insight into the interpretability that the method provides, even without the usage of a Siamese architecture. Moreover, note that Fig. 1 is an example illustration. The figure for the final paper will also contain an input image of the class and make a comparison to components of other similar classes. If space allows, we will also move the learned MNIST patch components in Fig. 20 to the main part of the paper. These images highlight the decomposition of an image by the CBC.

3) Comparison to other methods: Thank you for the provided suggestion regarding additional comparisons. A first initial comparison to CNNs has already been presented in the supplementary during the ablation study. Moreover, we investigated a comparison of the CBC agreement heatmap visualizations to CAM methods. Disagreement cannot be compared to CAM methods as they cannot provide a similar visualization. However, such comparisons will be part of future contributions as well as a detailed study about adversarial robustness and outlier detection.