We thank all reviewers (**R1**, **R2**, **R3**, **R4**) for their detailed and encouraging comments, and we are pleased that the presentation was clear and the the work is overall well motivated. We tried to answer concerns raised by the reviewers below.

(*) The relation between the autoencoding and adversarial robustness is not clear(R1, R2), More intuition is needed about how the theory in 2. links with the proposed approach in 3. (R2), The results are empirical and disconnected from theory (R2). Learning a VAE is a highly over-complete and ill-posed problem. Our intuition is that we need regularization, but instead of modifying the decoder model, we constrain the encoder, by explicitly enforcing properties of the exact posterior. Our argument starts by proposing an extended model in 2.1 that admits VAE as a marginal. In 2.2, we say that the exact decoder model is by construction independent of the choice of the coupling strength ρ; 2.2 simply means that, if we had access to the exact decoder and if we 'could' do exact inference, any ρ will work, so the extended model is actually redundant, and in 2.3, we highlight the properties of an exact encoder. However, as we will be learning the decoder from data while doing only approximate amortized inference, we want to retain the properties of the exact posterior of the extended model (as it is more relevant for the representation learning Example 2.2), but how do we bake this in explicitly to the encoder? The answer comes in 3, where we propose AVAE, and in 3.1 we provide the justification that this choice coincides with the exact target conditional for $\rho = 1$, the representation learning case, when we encode and decode consistently.

How all this is related to adversarial robustness? The existence of 'surprising' adversarial examples, (where a pig is classified as a plane by slightly changing pixels), is typically the result of a problem with the smoothness of the representation (e.g. having a large Lipschitz constant), (see also example 3.1, VAE case). Authors in [4] attempted to fix this by data augmentation while training the encoder. We show here a more general framework, (where [4] is also a special case) and investigate an orthogonal choice that circumvents adversarial attacks. The data augmentation is achieved by using the learned generative model itself, as a component of the encoder. The AVAE objective ensures that samples that can be generated by the decoder in the vicinity of the representations corresponding to the training inputs are consistently encoded. In the experimental section we illustrate that this translates to a nontrivial adversarial robustness performance. We don't have formal guarantees, but in our opinion, learning an encoder that retains properties of an exact posterior is key in achieving adversarial robustness.

Discuss the effect of the architecture on the results, Report ELBO for comparison (R1) We find batch-norm useful for speeding up training while avoiding degenerate solutions, in especially VAEs that have powerful decoders. We also find that we need to choose decoders that are shallower than the encoders. Additionally in this paper, our focus was on representation learning, hence our evaluations were based on downstream tasks. We consider it a future work to investigate the issue of learning a good quality decoder (as measured by ELBO, MSE or the FID scores) using the AVAE objective alone and we conjecture that it is feasible to learn a better decoder while learning a robust encoder. Instead of the ELBO, we report the MSE as a proxy for the decoder quality as the AVAE or SE Elbo have additional terms that makes direct comparisons difficult.

More realistic data, such as CIFAR10 (R3) This is certainly a valid critique, and we agree that a 'wide domain' dataset such as CIFAR10 in contrast to 'limited domain' datasets MNIST (hand written digits) or CelebA (faces) are much more challenging. On the other had, our experience is that training VAE's for CIFAR10/CIFAR100 requires more advanced architectures choices, such as ResNets, or other improvements, such as VQ-VAE.

Title change (R4) This is a valid suggestion that we will consider; in fact our original title was 'Robust Representations with the Autoencoding Variational Autoencoder'.

Missing related work (R1, R2) The page limit has not allowed us to include a separate related work section but we will include more citations in the introduction and conclusion. Including Alain and Bengio (**R2**). Most work deals with modifying the generative model $\mathcal{P}$ but the spirit of our approach is regularization of the approximating distribution $\mathcal{Q}$ on an extended space.

Additional feedback ***(R2)*** **1)** *Figure 1 is not useful* The drift is a consequence of the inconsistency of the encoder and decoder, even if we choose $\rho = 1$, please see above (*) ***(R2)*** **4)** *Learning a natural ρ coupling parameter?* In the standard VAE data is assumed to be iid, and as Proposition 2.2 also shows, the marginal is independent of ρ. Hence this parameter is not identifiable from data, unless we assume additional relational structure, such as in a video where subsequent frames are closely related. But this requires observing at least pairs of data points which is actually not available in the standard benchmarks. ***(R2)*** **4.0)** *Why $\rho = 1$?* This is a hyper parameter that can be chosen freely (VAE implicitly has $\rho = 0$) Any choice close to one is reasonable in the context of representation learning; in lack of any other downstream task we would like to retain a representation that would enable us to reconstruct the same image. In our experiments, we tune this parameter and find $\rho = 0.95$ gives good results. ***(R2)*** **5)** *p_θ notation is unclear* Our wording seems to cause the misunderstanding here; the last sentence before 3.1 should have read 'The following proposition shows our justification for the choice of $\mathcal{Q}_{\text{AVAE}}$ distribution'. The justification for p_θ, does not follow from 3.1.