Questions:
Neural attention

a) Read the paper Attention is not Explanation, which argues against using attention weights as an explanation of model's predictions. Briefly summarize the key findings of the paper (no need to understand all details).

b) Attention is not not Explanation challenges the previous paper. What is the key point of this paper?

Answers:
a)
The paper “Attention is not Explanation” conducts comprehensive experiments across a wide range of NLP tasks to determine the extent to which attention weights offer valuable "explanations" for predictions. Here, they experimentally look at how attention weights, inputs, and outcomes relate to one another and aim to examine empirical properties of learned attention weights and to interrogate their interpretability and transparency.

This research and the conclusions that may be made from it have some significant limitations. They've discussed the moderately weak relationship between learned attention weights and many alternative measures of feature relevance, such as gradients. They don't mean to imply that such alternative measurements are always the best or that they need to be taken as "ground truth," either. Such measurements indeed have a distinct inherent (to the model) semantics, but nonlinear neural networks might make it challenging for humans to comprehend them. But practitioners should be concerned because attention regularly has weak correlations with several such measures.

In relation to these topics, their findings regarding attention weights in recurrent (BiLSTM) encoders are summarized as follows:
(1) Only weakly and inconsistently, and
(2) It is not frequently possible to design adversarial attention distributions that produce practically comparable predictions as when using the first generated attention weights, although paying to completely different input characteristics.

Even more startling, arbitrarily permuting attention weights frequently results in relatively little variations in output. In contrast, attention weights in basic, feedforward (weighted average) encoders perform better on these criteria.

Counterfactual attention experiments also show that there are other heatmaps that produce predictions that are equal, so one cannot infer that the model generated a certain prediction because it paid attention to inputs in a particular way. The attention module settings, however, may assess the hostile weights as improbable. Additionally, it's possible that different plausible explanations exist for a specific disposition, which might make interpretation more difficult. For these cases, the model should emphasize every feasible explanation; yet, one may consider a model to be credible if it offers a "sufficient" explanation.
Finally, they confined their evaluation to tasks with unstructured output spaces, excluding seq2seq tasks, which they decided they will investigate in the future and they feel that

interpretability is more typically a factor in categorization than in translation.

b)

For the paper “Attention is not not Explanation” the researchers proposed four alternative tests to determine when/if attention can be used as an explanation: a simple uniform-weights baseline, variance calibration based on multiple random seed runs, a diagnostic framework using frozen weights from pretrained models, and an end-to-end adversarial attention training protocol. Each enables meaningful understanding of RNN models' attention processes. They show that even when trustworthy adversarial distributions are established, they do not score well on the basic diagnostic, showing that paper “Attention is not Explanation” written by Jain and Wallace does not rule out the necessity of employment of attention processes for explainability.

While Jain and Wallace pose an important question and express genuine concern about the potential misuse of attention weights in explaining model decisions on English-language datasets, it was found that the key assumptions used in their experimental design leave an implausibly large amount of freedom in the setup, ultimately leaving practitioners without an applicable way to measure the utility of attention distributions in specific settings.

Jain and Wallace adjusted the attention distributions of trained models, referred to as base, in order to identify any other distributions that would result in the model producing almost equal prediction scores. These distributions may be found by first adversarially looking for maximally diverse distributions that nevertheless provide a prediction score inside the base distribution, and then by randomly permuting the base attention distributions on the test data during model inference. The idea that attention distributions cannot be explained because they are not exclusive is supported by these experimental findings. In order to do this, they adjust the attention distributions of trained models, which we will refer to as base from now on, in order to identify any other distributions that would result in the model producing almost equal prediction scores. These distributions may be found by first adversarially looking for maximally diverse distributions that nevertheless provide a prediction score inside the base distribution, and then by randomly permuting the base attention distributions on the test data during model inference. The idea that attention distributions cannot be explained because they are not exclusive is supported by these experimental findings.

In conclusion, according to “Attention is not not Explanation”, Jain and Wallace have not demonstrated the existence of an adversarial model that generates the stated adversarial distributions due to the per-instance nature of the presentation and the fact that model parameters have not been explicitly learnt or changed. Since they are not equally plausible or accurate explanations for model prediction, these cannot be considered as antagonistic attentions. Readers may wonder how antagonistic the discovered adversarial distributions are

as they haven't given a baseline for how much variance is often seen in learnt attention distributions.
Simultaneously, “Attention is not not Explanation” have offered a set of tests that researchers may utilize to decide for themselves how well their models' attention processes explain model predictions. When applied to a diagnostic MLP model, they have demonstrated that alternative attention distributions discovered through adversarial training approaches perform poorly in comparison to conventional attention mechanisms. These findings suggest that trained attention mechanisms in RNNs using our datasets do learn something valuable about the connection between tokens and prediction that cannot be simply "hacked" adversarially.