Thanks to all reviewers for their comments and feedback. Please find below responses to the main points raised by each reviewer.

Review #1 **submitted on 23/Oct/2023**

Section 3.

1. Comment: As far as the objectives are concerned it is not clear (or not well motivated) why Explainability for AI experts and AI stakeholders have been splitted into two objectives. Clearly, the audiences are different and the generation of explanations for these categories will imply to adopt different techniques and methods, nonetheless, O2 and O3 seem to have a shared goal, that is the generation of intelligible explanations for human users.

   Response: O2 and O3 both refer to Explainability but the generated explanations are used to fuel two complementary aspects of the neuro-symbolic cycle: objective O2 focuses on the neuro-symbolic extraction-injection cycle whereby explanatory rules and relations provided to AI experts should be tuned and injected into an untrained network to improve/adjust learning; objective O3 focuses on the Explanation-Feedback-Control cycle whereby explanations provided to end-users and domain experts should be validated with respect to background and common-sense knowledge and used to augment cognitive reasoning for decision support. These two objectives have been rephrased and these aspects should be now clarified.

2. Comment: Perhaps a graphical representation of the neural-cognitive cycle would enhance the reading.

   Response: A graphical representation has been added.

3. Comment: Some parts of the text are a bit repetitive.

   Response: Text has been read again to reduce repetition as much as possible.

Section 4

4. Comment: It seems to me that H1, H2, and H3 more than hypotheses can be considered enabling techniques, methods and expertises that can be leveraged to support the neural-cognitive cycle discussed in Section 3.

   Response: in Section 4 we want to make sure our perspective is clear before the key elements of the cycle are outlined. We consider H1-2-3 hypothesis in the sense that we start from these ideas as assumptions that validate the suggested methods in 4.1, 4.2, 4.3. If se start from the assumption that H1. Graphs are a good representation formalism for our goals, H2. propagation of knowledge can produce better representations, and H3. human experts' role is key, then all the rest holds.

5. Comment: (Section 4.1) What is ILASP?

Response: Inductive Learning of Answer Set Programs (added to the text)

6. Comment: typos and references to be fixed

   Response to all minor points: all typos and references have been corrected

---

Review #2 **submitted on 16/Feb/2024**

7. Comment on Conceptual Framework and Methodology. The paper introduces a compelling conceptual framework for integrating neuro-symbolic approaches in explainable AI. However, it would benefit from a more detailed discussion on the methodologies for integrating these two paradigms. Specifically, how will the symbolic reasoning components interact with the neural network layers, and what are the anticipated challenges in this integration?

   Response: The conceptual framework has been clarified and represented in a diagram indicating the key synergies between the two paradigms.

8. Comment: A deeper exploration of the data types and structures that would be most suitable for this neuro-symbolic cycle would be beneficial. For instance, how will the proposed cycle handle different types of data (e.g., structured vs unstructured data)?

   Response: The general idea is that the framework would be applicable to feed-forward networks that can cater for different types of input data. The ability to handle data diversity relies on the neural element of the cycle. Despite the paper targets computer vision models, the approach can be extended to attention-based models and GNN as long as the learned representation is represented via knowledge extraction and mapping. In fact we among others have done work on extracting knowledge graph from transformer, CNN and more recently GNN.

9. Comment on Evaluation and Validation: The paper would benefit from an outline for the evaluation and validation of the proposed neuro-symbolic cycle. What metrics or criteria will be used to assess the effectiveness and accuracy of the explanations generated?

   Response: We acknowledge that there is still a gap in the field of XAI in terms of validation of generated explanations. We suggest the need to design novel evaluation metrics in terms of qualitative and quantitative measures, with a strong focus on human-centred evaluation. One way of quantifying some of these measurements would be to adapt methods such as the System Causability Scale (SCS) and Trustworthy Explainability Acceptance (TEA) as per text and relevant references included in the paper. However there are still opportunities and more to be done in this area.

10. Comment: Including a discussion on potential benchmark datasets or experimental setups that could be used to validate the proposed approach would add considerable value to the paper.

Response: The ability to find suitable training data (such as radiology images) to learn deep representations for a classification task heavily depends on the specific application domain. For example, in the area of Cardiac MRI, the ACDC dataset (https://www.creatis.insa-lyon.fr/Challenge/acdc/) is a good starting point, but the list of required training data is heavily domain dependent. When it comes to clinical diagnostics and use of commonsense and clinical knowledge, we have suggested a list of publicly available datasets that can represent a starting point, and added them in a table in the paper

11. Comment on XAI: the paper could benefit from a deeper exploration of how neuro-symbolic integration can address specific challenges in explainability, such as dealing with complex, high-dimensional data.

    Response: As discussed already in the paper and in response to previous comments, the power of neuro-symbolic integration via the proposed cycle relies on the ability of learned neural representations to capture high-dimensional, complex data, while neuro-symbolic mapping facilitates transparency by unveiling deep representations in symbolic terms, and the neuro-symbolic cycle makes it possible to rely again onto neural deep representation learning with knowledge injection in order to learned better representations.

12. Comment: The paper would benefit from a more consistent formatting style for citations and footnotes.

    Response: Citation and footnote consistency addressed

Review #3 **submitted on 23/Dec/2023**

13. Comment on Introduction: There seems to be a missing link between argument against 'model-based interpretability' and argument for the proposed neuro-symbolic cycle. For example, how is the proposed neuro-symbolic cycle can overcome the limitations of incomplete and incorrect explanations? Will the proposed cycle able to produce both high accuracy preidctions and trustworthy explanations? In the given example (attribution maps), is it the case for all post-hoc and interpretable methods? As in do all XAI methods highlight a portions of an image that are responsbile for a predicted outcome without specifiying how the outcome was produced?

    Response: An additional paragraph aims to create this link. If we can extract a symbolic model from deep representations, this model can be used to generate explanations and also feed relevant knowledge back into the learning process and the cognitive reasoning process. This is better articulated in the paper but was missing in the introduction.

14. Comment on State of the Art: Perhaps some citations for for sentences from line 35-38? E.g., post-hoc approaches and attention-based approaches for medical image analysis.

Response: The main reference is provided as a survey: . Singh, S. Sengupta and V. Lakshminarayanan, Explainable Deep Learning Models in Medical Image Analysis, J. Imaging 6(6) (2020), 52. doi:10.3390/jimaging6060052.

15. Comment on Key Elements of the Extraction-Explanation-Injection cycle. H2: the assumption here would be that the knowledge extracted from the trained GNN or other networks will be treated as 'the absolute truth'. Although this extracted knowledge will be validated by the medical professionals, but how will this ensure the explainable model to be 'robust' and 'accurate' given that different optinions might be given for the validation process, especially in diagnosis?

    Response: H2 has been corrected as it refers to the injection of knowledge (independently of whether it is extracted by a CNN and used as is or corrected and revised by expert knowledge as we proposed in the neuro-symbolic cycle approach. The observation about subjectivity in the validation process is an interesting one. We have considered this aspect specifically in relation to the gender dimension of the research, and we advocate on the need to consider gender specific feedback.

16. Comment: Continue on from the previous point, what will happen if complete opposite opinions/knowledge are given by the experts during the validation process, will both be injected and propogated into the non-trained network and will both be given equal weights?

    Response: The personal bias generated by human involvement might be manifested in the evaluation, as perception can vary. The need to rely on a representative sample is important, as well as the need to mitigate bias with a process similar to inter-annotator agreement. When bias is generated by gender, it is important not only to make sure there is a representative sample of both genders in evaluators involved, but will also consider using different types of questionnaires and scales according to the gender of the evaluator. These considerations have been added to the discussion.

    Presentation comments have been addressed