Sample Claims

This is a supplementary document to be used alongside the Selecting and Evaluating Claims activity.

Each claim in the following set is about a hypothetical project, model, or system. Various details have been omitted to focus on the claim itself. You should not need to know anything beyond what is expressed in the claim itself to complete the associated activity.

There are 15 claims in total. 5 claims have been designed as examples of 'project transparency' explanations, 5 claims have been designed as 'model interpretability' explanations, and 5 claims have been designed as examples of 'situated explanations'. However, the explanations also differ in terms of their quality, so some may be harder to identify than others.

The answers will be provided in a separate document to be reviewed after you have matched each explanatory claim to its corresponding facet of explainability.

Set of Example Claims

> ⓘ Important
>
> The following claims are not in any particular order.

1. "We chose to create a secure environment for performing data processing (e.g., extraction, analysis, and preprocessing) because our data is highly sensitive. In addition, this level of security prevents us from making information about our pipeline available to the public."
2. "We chose to use a neural network model because it gives us the highest levels of accuracy."
3. "Our stakeholder analysis process identified several groups of stakeholders, including the general public. To ensure that we met the different explainability requirements, we explored multiple processes for developing accessible explanations. For example, technical auditors have access to the results of complementary interpretability methods (i.e. Integrated Gradients and SHAP), while members of the public can access general information about how our system was built and implemented."
4. "Because of the context in which our system is intended to be used, there was a risk that our system could have lower levels of accuracy for specific sub-groups.

Therefore, we have ensured that the information about our training data and a variety of feature summary statistics are available to assist bias mitigation and assessment efforts."

5. "We used Gradient-weighted Class Activation Mapping (Grad-CAM)to help visualise the regions of an image that were most important for a particular prediction made by a neural network. We chose this technique because it is effective at generating visual explanations that are easy to understand and interpret, even for non-technical stakeholders."
6. "Our project has been carried out in accordance with best practices for transparency, interpretability, and explainability. Our AI system can provide clear and transparent insights into how it is making predictions that can be easily understood and trusted by stakeholders."
7. "Our system is for educational purposes only and will only be used by technical specialists. As such, no specific explanations about the system are needed, but our source code and notebooks are available on a public GitHub repository."
8. "Our random forest model outperforms other models we tested, and has also been designed with interpretability in mind. Firstly, we used permutation feature importance, partial dependence plots, and SHAP values to identify the most important features in the model. Then, we used supplementary techniques to visualize how the model responds to changes in individual features as well as to provide a more detailed understanding of the decision-making processes within the model. These methods and techniques allowed us to identify specific rules and patterns that are driving the model's predictions. We tested these outputs with a variety of our stakeholders to ensure they can be easily understood by people with a wide range of backgrounds and technical expertise."
9. "Our system is expected to be used by healthcare professionals in time sensitive situations. As such, simple explanations are required, even if this means some explanatory quality is lost."
10. "During our project, several models were extensively tested and evaluated using a range of interpretability methods, including sensitivity analysis and counterfactual analysis, to ensure that we trained and validated a robust and reliable model. The results of these tests are available in our documentation, which also provides information about why this approach was needed to build trust with stakeholders."
11. "To ensure regulators and auditors can fully assess the trustworthiness of our project and system, we have published a wide variety of documentation about our system, including details of our data collection and analysis process, as well as preliminary risk management process, such as an equality impact assessment carried out to identify risks of bias and discrimination."
12. "We developed a decision tree model to ensure it was intrinsically interpretable by a wide range of stakeholders. The important features and thresholds are clearly represented, and instructions for employing additional visualisation software (e.g. `dtreeviz`) are also available in our documentation."
13. "Once our problem was defined, we extracted the required data, analysed it, and processed it to get it ready for training our ML model. Due to the complexity of our

problem, we went through a rigorous process for choosing our model, training, and validating for accuracy, biases, and interpretability using multiple methods. This whole process was extensively documented, with the aim of making our project open and reproducible. Therefore, project documentation can be downloaded from our GitHub repository, which includes our data extraction and management processes, as well as model tests and results, but we also have a non-technical version available. The latter was designed with stakeholders from the general public in mind, as they too should have transparent and understandable access to the way the project was carried out."

14. "Because our project handles sensitive personal data and the decisions the AI model assists in making are consequential, we have used high security standards for data management."
15. "This project uses a logistic regression model, which has high levels of intrinsic interpretability, so therefore we did not do more to augment the models post hoc interpretability. However, given the nature of the project, we have developed detailed documentation to allow other research teams to understand our choice of model."

Answers

1. Project Transparency (Medium Quality)
2. Model Interpretability (Lower Quality)
3. Situated Explanation (Higher Quality)
4. Situated Explanation (Medium Quality)
5. Model Interpretability (Medium Quality)
6. Project Transparency (Lower Quality)
7. Situated Explanation (Medium Quality)
8. Model Interpretability (Higher Quality)
9. Situated Explanation (Lower Quality)
10. Project Transparency (Higher Quality)
11. Project Transparency (Medium Quality)
12. Model Interpretability (Higher Quality)
13. Project Transparency (Higher Quality)
14. Situated Explanation (Lower Quality)
15. Model Interpretability (Medium Quality)