The Alan Turing Institute

Turing Commons

ACTIVITY 3

Selecting and Evaluating Claims

Responsible Research and Innovation in Data Science and AI

Summary

This activity is designed to help you (and your group) identify, understand, and evaluate claims made about a project, data, model or system in service of providing an explanation to some stakeholder or affected user. You will need to both identify specific types of claims, based on whether they support different objectives, such as project transparency, model interpretability or situated explanations, and also evaluate the quality of these claims, based on their ability to meet these objectives. Then, using a case study, you will need to develop a set of claims for the hypothetical project and evaluate whether they would be positively or negatively evaluated by the relevant stakeholders. The purpose of this task is to help you better understand the different facets of explainability and their relative strengths and weaknesses.

Learning Objectives

- Improve ability to identify the different facets of explainability based on example claims made about a project, its data, and the respective model or system
- Gain experience with developing claims that are well suited to different explainability objectives, such as project transparency, model interpretability or situated explanations
- Recognise how specific claims may be evaluated or judged by different stakeholders or affected users, and whether they are sufficient to meet their relative needs

Instructions

Pre-requisites

To carry out this activity, you will need the following:

- The list of sample claims provided in this document.
- A case study selected from our repository.
- If undertaking this activity as a group, you will need to form two (or more) breakout groups who will evaluate the claims developed by the other group(s).

Introduction

A request for an explanation is also a request for some reason or claim made about the target of the explanation. For instance, if you are asked why you were late for a meeting, you might respond as follows:

I was late because my train was delayed due to a signal failure.

This is a claim made about the reason for your lateness.

Claims differ in their quality and ability to meet the expectations of the person requesting the explanation. Again, using the example of being late for a meeting, you might respond as follows:

I was late because I spent too much time on social media before leaving the house.

While a valid explanation in terms of its truthfulness, it is unlikely to be well received by the person requesting the explanation.

In our explainability module, you were introduced to three different facets of explainability. These facets will be used in this activity to help you identify some pre-built claims and also develop and evaluate a set of claims made about a hypothetical case study. As a reminder, here are the three facets of explainability and a short summary about the types of claims that might be made to support them:

- Project Transparency: project transparency involves making clear the reasons or actions undertaken over the course of the project lifecycle, such as a design choice to engage a specific set of stakeholders or collect a specific type of data.
- Model Interpretability: model interpretability requires being able to employ the best method for interpreting the behaviour of a model, in order to then translate this information into clear and accessible explanations.
- Situated Explanations: situated explanations are reasons and claims that are grounded in an awareness of the sociocultural context in which a project is carried out, and with an understanding of how this context may have impacted the design, development, and deployment of the model or system.

Steps

If you are carrying out this activity as a group, these are the steps you should follow:

1 First, take a look at the sample claims provided in this supplementary document. Each claim is associated with one of the three facets of explainability and the set of claims will differ in terms of their quality and accessibility.

- You will need to first of all identify the facet of explainability that each claim supports.
- Next, you will need to rank the claims in terms of their quality.
- Finally, you will need to justify this ranking. For instance, why did you rank a particular claim as being of the highest quality? What makes it a good claim? Or, conversely, what makes a poor claim?

---

2 Once you have completed this task, you will need to form two (or more) breakout groups.

- In these breakout groups, using the same case study, you will need to develop three claims for each of the three facets of explainability and provide a ranking for each set.
- Record which facet of explainability each claim supports and why you have ranked them in the way you have. Keep this information hidden from the other group(s).
- Once you have completed this task, you will then swap your claims with another group.

---

3 Now, with the claims from one of the other groups you will have to a) identify which facet of explainability each claim supports and b) evaluate the quality of each claim.

- While carrying out this step, also reflect on which claims the other group(s) chose to develop and how this choice compares and contrasts to your own.

4 Finally, reconvene as a larger group and discuss the following questions:

- Did each group correctly identify the facet of explainability that each claim supported and the intended ranking?
- If not, why do you think this was the case? It could be for several reasons:
  - The claim was poorly written and did not clearly communicate how it supported the relevant explainability objective.
  - The claim supports more than one facet of explainability.
  - One (or more) of the teams has a gap in their understanding of the different facets of explainability.
  - The different sub-groups have different values for what constitutes a clear and accessible explanation.

Example

Here is a partial example of how the first step could be completed:

Three claims are selected, identified, and ranked as follows (best to worst):

1. "The goal of our project was to develop a decision support system for intelligence analysts. Their work is often carried out in safety critical environments and they are often required to make decisions in a time-constrained environment. As such, we prioritised a model with high intrinsic interpretability, which was further enhanced by visualisations that could further show information such as relative feature importance." (Situated Explanation)

2. "Our system uses a deep neural network to classify radiology images. As the model is not intrinsically interpretable we relied on visual attention maps to help us interpret the model's behaviour." (Model Interpretability)

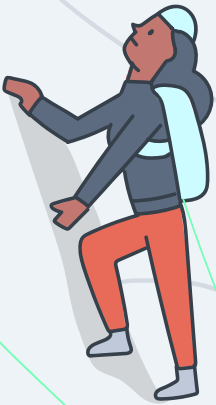
3. "We identified and engaged with a diverse set of stakeholders during project planning to ensure that design choices made about our system were inclusive and representative of their needs." (Project Transparency)

The justification for the ranking is as follows:

1. The claim demonstrates awareness of the context and environment in which the system is deployed (e.g. to support decision making in safety critical environments). Furthermore, the claim also helps explain why this contextual understanding shaped a decision about which model and interpretability method to use.

2. The claim is about the model interpretability method specifically, as it does not mention much about the context in which it is used (e.g. healthcare). It does not explain why visual attention maps were chosen over other methods.

3 The claim is about project governance and decision-making, but does not explain how stakeholder engagement was carried out to ensure inclusivity or representation.



Goals and Objectives

The primary goal for this activity is to help you deepen your understanding of the different facets of explainability and how they can be used to develop clear and accessible claims.

By the end of the activity you should have the following:

1 An ordered set of claims (with corresponding evaluations) from the supplementary document.

2 Either:

- An ordered set of claims based on the case study selected, and relevant feedback from discussion between groups.
- An unordered list of claims about the case study, which are linked to the relevant stakeholders or affected users.

Tips and Guidance

Remember that the case studies are hypothetical. As such, you can be quite flexible and creative with the claims that you develop for each of the sub-teams. If you need to add details to the project to support the explanation (e.g. pretending that a specific model interpretability method was used), then feel free to do so.

If you are struggling to develop claims, then take another look at the project lifecycle model from the earlier module. For instance, select a stage and consider what sort of task would be carried out at that stage. Then, consider how this task could a) need explaining to ensure project transparency, b) create some impact or constraint on the interpretability of the downstream model, or c) interact with the sociocultural context in which the project is being carried out.

Assets

- You will need the list of claims from this supplementary document.
- You will need a case study selected from our repository.
- You may also need a copy (printed or online) of the project lifecycle model.

In-Person

If you are conducting this activity in-person, you will need something to make group notes on (e.g. flip-chart, paper, sticky notes) and help keep track of the ordering and evaluation of the relevant claims.

Online

If carrying out this activity online as group you may wish to use an online whiteboard or collaborative note-taking tool such as Miro, Mural, HackMD. or Google Docs.