Orchestrating a brighter world

NEC

A Human-Centric Assessment Framework for AI

Sascha Saralajew, Ammar Shaker, Zhao Xu, Kiril Gashteovski, Bhushan Kotnis, Wiem Ben Rim, Jürgen Quittek, Carolin Lawrence
NEC Laboratories Europe

Motivation

Given a dataset, we consider an AI system as a machine that can generate a solution for a given task. In our work, we tackle the question: How can we assess such systems in a **human-centric manner?**

Ground truth: The man at the bat readies to swing at the pitch while the umpire looks on.

Prediction: Two men playing baseball

Figure 1: If we have a system that generates captions for given images, how can we properly compare two captions?

Typically in order to evaluate a system solving a task, **we compare the generated solution with the ground truth**. Through this comparison, we can assess the AI system appropriately on a test set.

Issues with standard approaches: it assumes that the ground truth data can be generated or easily defined, which is not the case for many scenarios. For example, in order to assess an AI framework that generates explanations, we must first answer:

1. What is a good ground-truth explanation?
2. How do we compare explanations? The comparison is not straightforward.

Generation of ground truth data

- Subjective
- Ambiguous
- Difficult (or impossible)

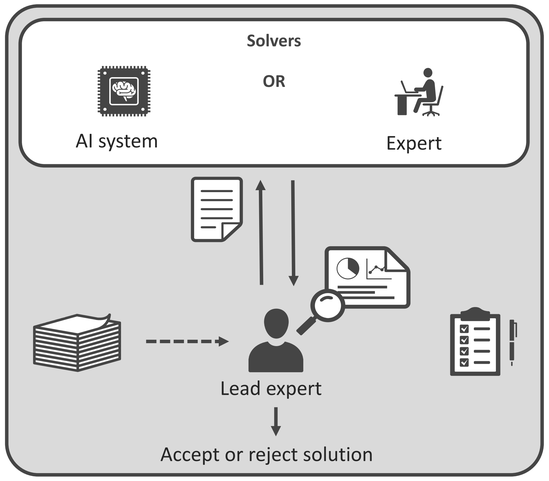
Comparison of generated data

- Subjective
- Ambiguous
- Difficult to describe mathematically

To assess the performance of two systems, we propose a blind experiment with two sides:

- A selected domain expert we call a **lead expert**
- **Solvers**: **AI system** or **another domain expert**

Proposed Framework



The framework consists of the following steps:

- The lead expert identifies a task to be solved by one of the solvers
- The task is assigned randomly to one of the solvers who tries to solve the task and returns a solution.
- The lead expert does not know who the solver is.
- Following approval guidelines, the lead expert evaluates the solution by accepting or rejecting it.

Figure 2: *By monitoring the experiment and by evaluating the acceptance rates for solvers, the AI system can be assessed in comparison to a human domain expert.*

Advantages:

- **Generic** and **human-centric** framework to assess AI systems
- **Blind assessment** of the AI system in comparison to a domain expert
- **Decoupling** of task solving and solution evaluation

Framework Instantiation

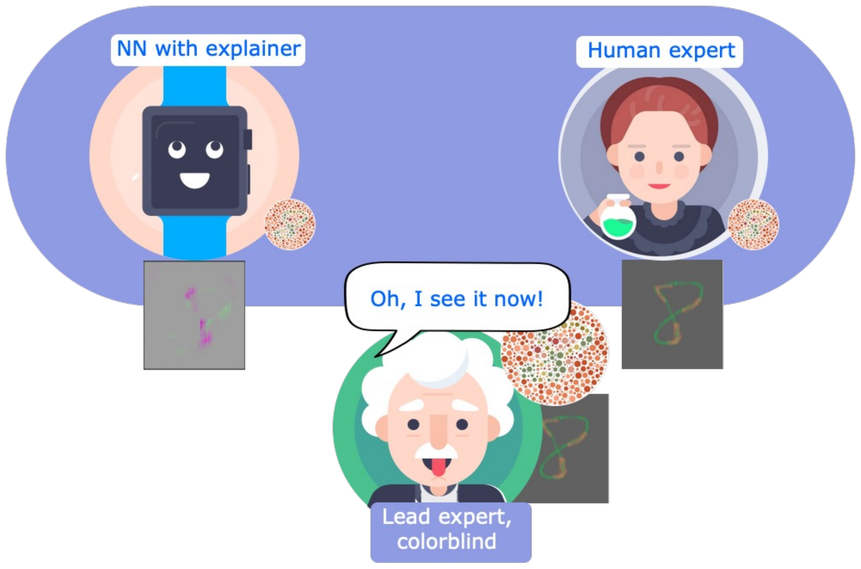


Figure 3*: Based on the framework we propose an instantiation to assess the usefulness of explanations by using Ishihara colourblind images.*

See our paper for framework instantiations:

- How it is related to assessment measures like classification accuracy.
- How the quality of explanations can be assessed on Ishihara's colorblindness test.

Conclusion: Machine explanations can be as good as human explanations, such that a colourblind lead expert would receive adequate assistance from an AI model to recognize colours.

Credit: Figure 3 illustrations by Laura Reen