### Introduction

LIME collage placeholder

In previous work[1], we outlined that **the utility of a technique for explaining a decision** reached by an A.I. algorithm **depends on** the **nature of the task** being performed **and the** **role of the agent** consuming the explanation.

To explore this, **we have developed a framework of datasets, machine learning models and explanation techniques** which allows for the comparison between a range of explanation outputs in the context of different tasks.

Figure 2: Examples of a LIME explanation. Evidence towards the given classification (“Gun Wielder”) is highlighted in green, whilst evidence against is highlighted in red. Inconsistencies can be seen in the regions highlighted by the technique.

At the core of the framework are two objectives:

* **Simplify the building of intuition for current and future explanation techniques** allowing for innovations in their use and in the creation of novel techniques.
* **Facilitate producing data that can be used for developing metrics** for measuring the utility of explanation techniques and for performing experiments to measure their interpretability.

Figure 1: Using the interpretability framework to compare different explanation techniques generated relating to the same classification of an input image.

Explanation comparison placeholder

#### Current and Future Related Work

Using the framework, we demonstrated the debugging process of generating multiple explanations from the same technique (LIME, *Figure 2*) to measure and improve its stability (As explored in [2]).

We have observed that many existing explanation techniques are subject to post-hoc interpretation by the consuming agent. This can be greatly affected by the visualization and presentation of the explanation and can lead to the agent projecting their existing assumptions on to the explanation. This will be explored in future work.

In another piece of future work, we will be using the framework to consult with domain-expert decision makers to gain insights as to what constitutes a useful explanation for aiding in the performance of their tasks.

#### References

#### [1] Tomsett et al. “Interpretable to whom? A role-based model for analyzing interpretable machine learning systems” in 3rd Annual Workshop on Human Interpretability in Machine Learning (WHI 2018)

#### [2] Stiffler et al. “An Analysis of Reliability Using LIME with Deep Learning Models” in DAIS ITA AFM 2018