Comparing User Interface Designs for Explainable Artificial Intelligence

Ionut Danilescu & Chris Baber

University of Birmingham, United Kingdom

SUMMARY

A well-known approach to Explainable Artificial Intelligence (XAI) presents features from a dataset that are important to the AI system's recommendation. In this paper, we compare LIME (Local Interpretable Model-free Explanation), to display features from a classifier, with a radar plot, to show relations between these features. Comparative evaluation (with N = 20) shows LIME provides more correct answers, has a higher consistency in answers, and higher rating of satisfaction. However, LIME also showed lower sensitivity (using signal detection), a slightly more liberal response bias, and had a higher rating of subjective workload. Evaluating user interface designs for XAI needs to consider a combination of metrics, and it is time to question the benefit of relying only on features for XAI.

KEYWORDS

Explainable AI; LIME; radar plots; Signal Detection Theory

Introduction

In many applications Artificial Intelligence (AI) systems present themselves to users in the form of a 'black box' which means that the inner workings of AI systems are difficult to evaluate. Explainable AI (XAI) is intended to allow humans to comprehend the output of AI systems. Often approaches to XAI require the AI system to explain its results in terms of the features which may have led to a particular output (Kaur et al., 2022). It is an open question as to whether the visualisation of these features has an effect on user acceptance and interpretation of an explanation generated by an AI system. In this study, we are interested in whether a user interface that emphasises discrete features produces the same results as one that emphasises the relationship between features. Our assumption is that, when presented with discrete features, users will need to develop a model that can define relationships between these, which would lead to higher perceived workload than when they had a user interface that defined these relationships for them. What is less clear is whether this would also produce differences in decision performance.

Machine Learning and AI communities are developing three broad lines of XAI:

(i.) approaches which support simultability either because the human operator can calculate examples of these or because the models can be visualized in ways that make cause-effect relations easy to appreciate;
(ii.) approaches which present the goals or plans by which the agents are reasoning;
(iii.) approaches which allow induction of models, i.e., which can allow simplification of the underlying model or modification of weights in the model to reveal how features can be changed to apply different rules.

In the first approach, we only have the outputs to interpret and must either infer the processes that generated these outputs or be provided with some explanation of such processes. This could involve identifying one or two features in the data or output that can be used to infer the processes that generated the output (Gilpin et al., 2019; Baehrens et al., 2010). “Clinicians repeatedly identified that knowing the subset of features deriving the model outcome, is crucial. ”(Tonekaboni et al., 2019, p.6).

AI systems can master video games (such as ‘Pong!’ from Atari). As Greydanus et al. (2018) demonstrate, it is possible to record the fixations of the AI system as it plays the game and create saliency maps, i.e., images of highest fixation point (figure 1). From figure 1, you can *infer* what the AI system might have been doing. But, even with simple games such as Pong!, the strategy that the AI system used might be different to that used by a human. So, in Pong!, we might watch the ball as it leaves a bat and moves across the screen; the AI system calculates an end-point based on angle of incidence and impact force and thus ignores the path the ball takes as it flies across the field. While this is a simple example, it illustrates that how we might imagine an action being performed is likely to differ from how the AI system performed it.

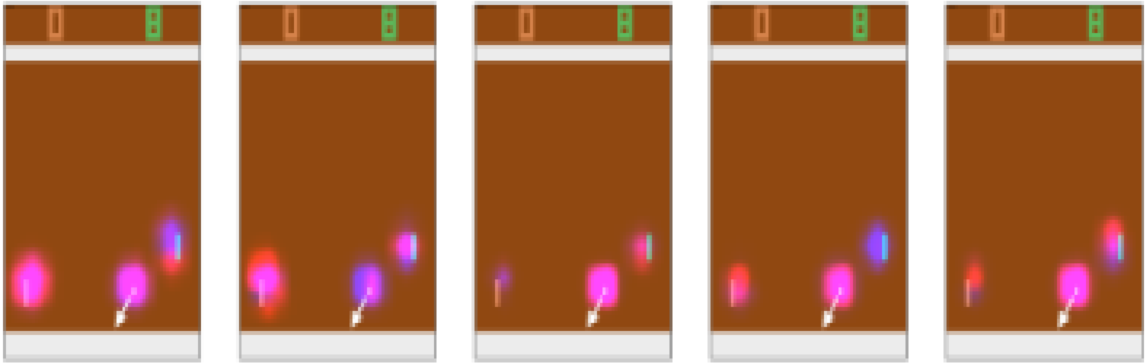
(c) Pong: learning a kill shot.

Figure 1: Saliency Maps for Atari game play [Greydanus et al., 2018]

The second approach is particularly well developed in robotics. In this field, explanation by robots focus on stating the purpose or goal that the robot is seeking to achieve (Dannenhauer et al. 2019). In order to do this, the robot needs to be able to state Beliefs (in terms of what information it has obtained from the environment and how it has interpreted this), Desires (in terms of the purpose or goal it is seeking to achieve), and Intentions (in terms of the plan that it will apply to achieve the goal). This leads to the concept of ‘explicable planning’: in terms of being able to generate plans that are amenable to human explanation; and explanation generation: the ability to tailor explanations to humans with given knowledge (Sreedharan et al. 2021).

The third approach is to produce a ‘surrogate model’, which works on a reduced version of the problem space and can highlight the important features that the AI uses. By reducing the data space over models an AI system operates, it is possible to highlight salient features for that model. These approaches, such as Local Interpretable Mode-agnostic Explanations (LIME) (Ribeiro et al., 2016). LIME (figure 2) is model-agnostic and concentrates on local fidelity, i.e., the relations within the vicinity of the specific instance. This could mean that the features selected might not always apply in other instances, i.e., that locally important features might not be globally important. A subset of features can allow the user to generate a ‘narrative’ as to why these (rather than other features) were used. This points to one of the problems with this approach: people can be very good at generating post-hoc rationalisation for *any* combination of features (particularly if the rationalisation fits a preferred narrative that they have imagined prior to the analysis). This could mean that, rather than objectively responding to the output of the AI system, people could use this to confirm their own biases and expectations. Even if we can be certain on the output rarely can we be certain how this output was achieved. This makes it difficult to defend the process by which the output was produced, especially under scrutiny or challenge.

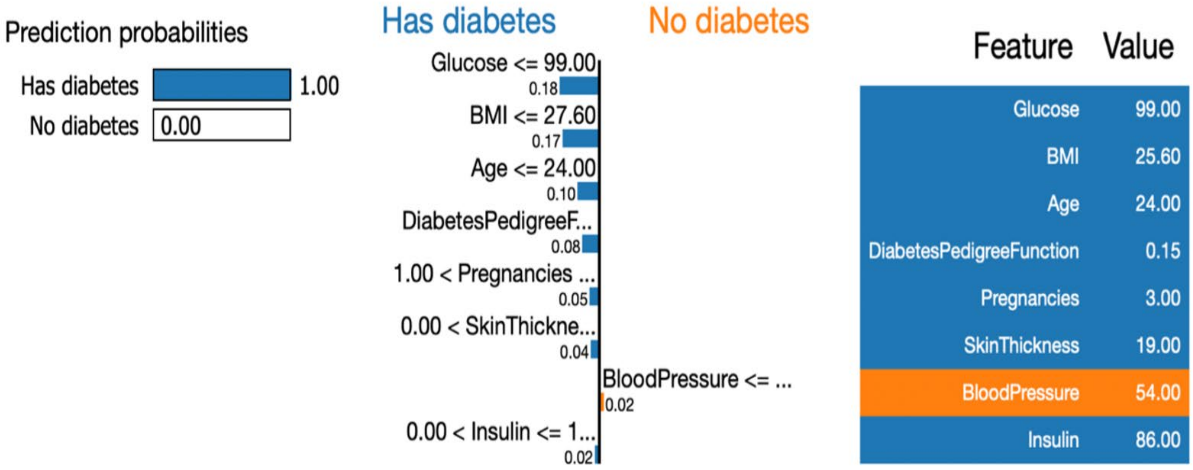

Figure 2: LIME

While LIME has proven to be a popular approach to visualising the features that an AI system uses to produce its output, there has been surprisingly little research evaluating its effectiveness for users. Dieber and Kirrane (2020) report a small-scale study in which six participants (3 with machine learning / AI experience) were asked to interpret LIME output. The authors believe that LIME presents "Too much information given in a badly structured way" (p8). Only 1/6 of the participants correctly identified why the model made its prediction, leading Dieber and Kirrane (2020) to "…conclude that while LIME helps to increase model interpretability, usability studies are needed in order to improve the user experience, and tools and techniques are needed in order to facilitate global comparisons." (p.12).

In this paper we evaluate LIME by contrasting with an alternative form of visualisation. A challenge was to define a display that showed relations between variables. Multidimensional plots, such as radar, polar, polygon etc., are popular for multidimensional data (Draper et al., 2009), particularly when they only display a single data series. One explanation for the potential benefit of such displays is that they are 'Configural' (Carswell and Wickens, 1990), i.e., they offer the opportunity for relations between objects to 'emerge' such that the shape (or its distortion) of a polygon could provide high-level indication of warning (Greaney and MacRae, 1996). However, contemporary Ergonomics research shows that radar plots are much worse for response accuracy and decision time, compared with other formats (Abeynayake et al., 2023; Fischer et al., 2005; Holt et al., 2011). Given that LIME uses some of the formats that in its visualisation, we felt that a radar plot could give a reasonable baseline for comparison and allow us to hypothesise that LIME ought to demonstrate superior performance on decision tasks. On the other hand, from Dieber and Kirrane (2020) and our previous comments on relations between features, we might expect LIME to induce a higher workload because of the level of detail in the display.

Method

In order to compare the output of LIME with a radar plot, two user interfaces were created using data from Kaggle. The dataset concerned diabetes[1] and originates from the US National Institute of

[1] https://www.kaggle.com/datasets/iammustafatz/diabetes-prediction-dataset

Diabetes and Digestive and Kidney Disease. The dataset contains 786 instances of anonymised patient data (where the data have been collected from a specific population of female patients 21 years or older of Pima Indian heritage).

The LIME model was built using a worked example[2]. Figure shows the prediction probability of the patient having diabetes or not on the left. In the middle, the interface highlights which features (such as glucose or age) had the biggest impact in deciding if the patient is at risk. In this version, the feature is blue if the patient is predicted to have diabetes, and orange, if not. On the right, the feature tables show the features that contribute to the patient having diabetes are displayed in blue and the ones that contribute to the patient not having diabetes are displayed in orange.

For the radar plot, we used the plotly package in Python with visualise the same data (figure 3). The features are labelled as spokes and their contributions shown by the position on the scales of each spoke. The features are linked to form the blue polygon. An orange polygon (used to indicate average cases in which diabetes in absent) is presented as a reference. The reference polygon is the same in all cases and is intended allow the user to ask, ‘does the blue polygon extend beyond the limits of the reference?’

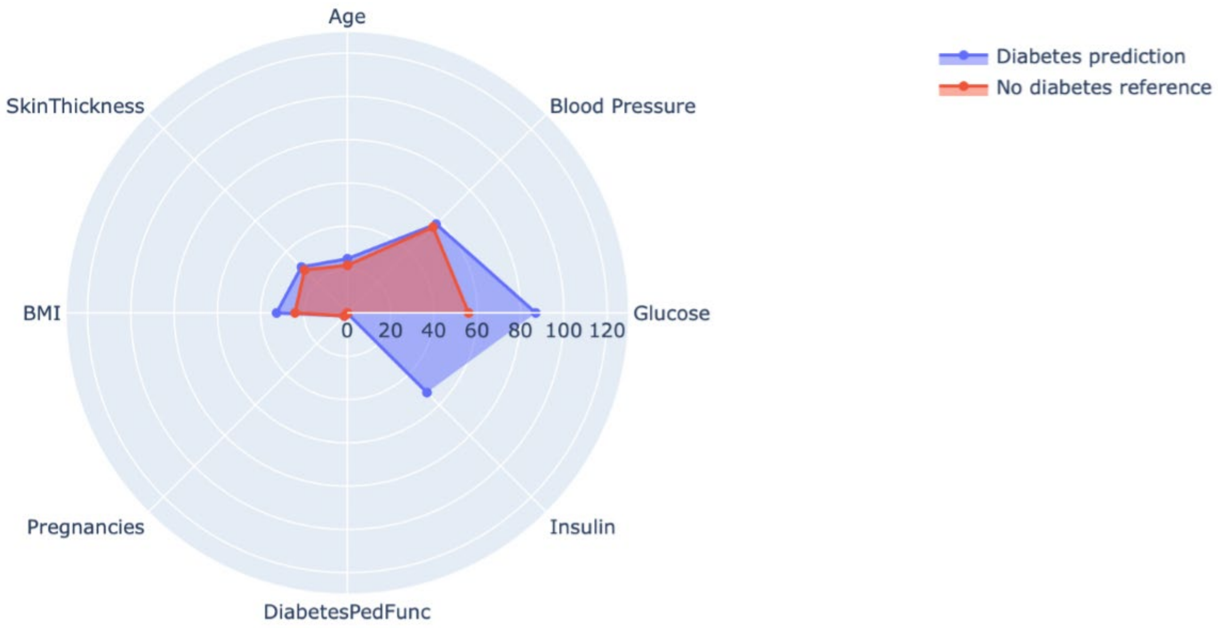

Figure 3: Radar plot

Participants

20 participants were recruited from University students in the School of Computer Science at the University of Birmingham. Ethics approval was provided by the School’s project board. Participants were selected because they reported some experience of diabetes (either as someone who was managing the disease or because they had family members who had lived with this). While we were using the user interfaces to make medical diagnoses, we felt that it was important for our participants to have some awareness of the symptoms.

[2] https://www.kaggle.com/code/jagannathrk/indian-diabetes-analysis-lime-shapley

Procedure

Each participant was presented with 20 questions of each user interface (counter-balanced across participants) and asked whether the data indicated if a patient was at risk of diabetes and which feature(s) contributed to their decision.

The questions were designed so that a user interface was presented with either diabetes present or absent, and a set of features that either agreed or disagreed with that diagnosis, to support the signal detection measure of sensitivity in terms of:

- Hit - if the participant thinks the patient is at risk and the correct answer is the same (the signal is present).
- Miss - if the participant thinks the patient is not at risk but the correct answer is the opposite (the signal is absent).
- False alarm - if the participant thinks the patient is at risk but the correct answer is the opposite (the signal is absent).
- Correct rejection - if the participant thinks the patient is not at risk and the correct answer is the same.

For each combination, we had 5 instances of user interface design.

Dependent Variables

The evaluation metrics used for this study are:

- Consistency of Response (Cronbach's alpha): whether participants agreed with each other.
- Explanation satisfaction: measuring participants' satisfaction (on a 5-point rating scale) with the explanations.
- Correctness: accuracy of the participants' answers relative to prior classification.
- SDT sensitivity: d' representing the sensitivity which is based on the signal and signal + noise values.
- SDT Bias:
    - Beta: the ratio of the normal density functions (we also report log(10) Beta);
    - C : the number of standard deviations.
- Subjective rating of mental workload: NASA task load index

Results

Consistency of Response: Cronbach's alpha showed 0.78 for the use of LIME and 0.54 for the use of the radar plot. A value of >0.7 is deemed 'acceptable' and <0.5 is 'poor'. This indicates that LIME had 'acceptable' consistency for the participants, and the radar plot had 'poor' consistency.

Explanation Satisfaction: 12/ 20 participants were satisfied with the LIME output, but only 5 / 20 were satisfied with the radar plot output, as explanations of the recommendation.

Participant Decision Accuracy: Figure 4 shows that participants were more likely to produce a correct answer when using LIME than with the Radar plot [t(19) = 17.1, p<0.0001].

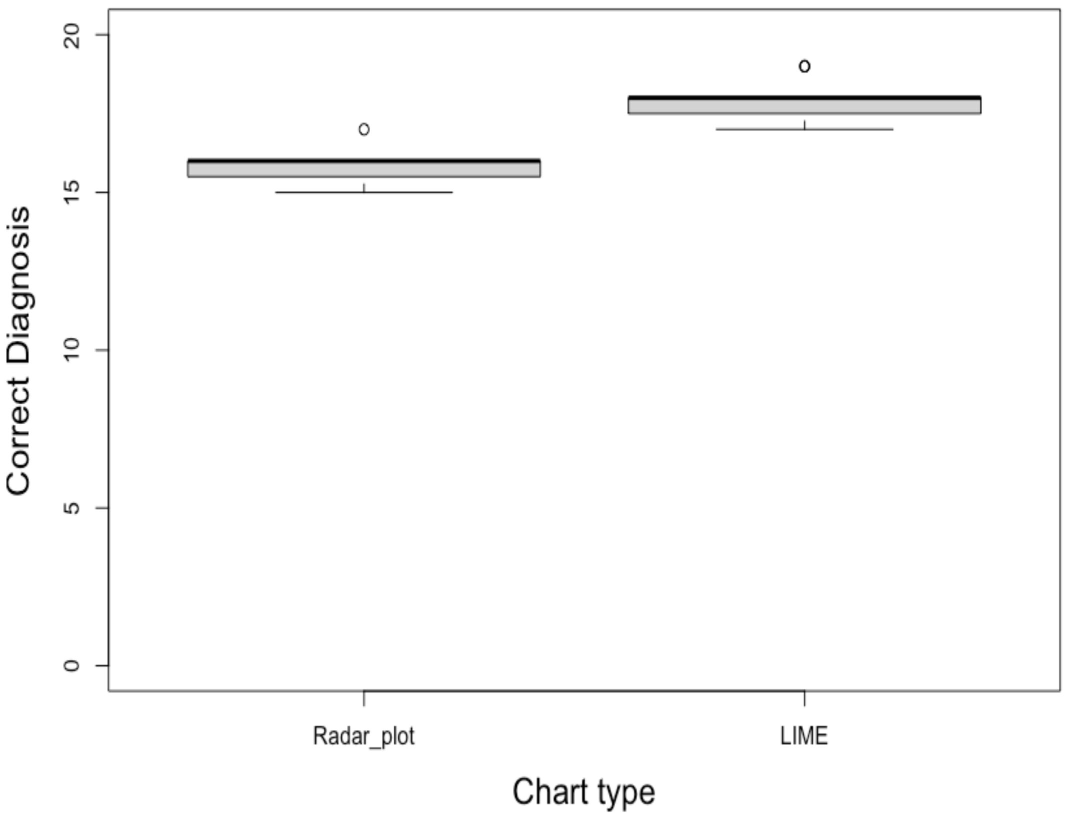

Figure 4: Comparison of Participants' ability to correctly diagnose a condition

Participants cited more factors in support of their decision when using the Radar Plot (3.24 vs. 2.13 for LIME) and a paired t-test showed that this difference was significant [t(19) = 44.5, p<0.0001].

Workload Ratings: Mental demand, as predicted, was higher with LIME (50 vs 23 for the radar plot) and own performance was rated as lower (50 vs 70).

Signal Detection Metrics: As table 1 shows, the radar plot had more correct hits, a higher d' and lower C, which suggests better sensitivity and lower bias.

Table 1: Signal Detection Measures for both User Interfaces

	SDT metrics					Number of Observations (% of total responses)			
	Beta	Beta (log10)	C	d'		Hit	FA	Miss	CR
LIME	1.61	0.21	-0.73	1.45		432 (54)	39 (0.05)	251 (0.31)	78 (0.1)
radar	4.0	0.6	-0.11	1.74		527 (66)	117 (0.15)	117 (0.15)	39 (0.05)

Participants seem to be slightly more liberal when using the LIME (as shown by the negative values of C, together with the higher value of Beta (log10) for the radar plot which puts the response criterion further to the right along the distribution). LIME has lower sensitivity, d', than the radar plot – although it should be noted that both are greater than 1 (we might expect values of d' to range from 0 to 2.5 on average, with higher values indicating better sensitivity). However, it should be noted that LIME had a much higher Miss rate than the radar plot and a low Hit rate.

Table 2 summarises the comparison between the two user interfaces. It is clear that there is not an obvious 'best' choice for this task and that the differences might reflect the ways in which participants were making use of the information available to them

Table 2: Results Summary (* indicates superior performance on a metric)

Metric	LIME	Radar plot
Consistency of response	0.78*	0.54
Satisfaction with explanation	0.6*	0.4
Correctness of diagnosis	0.99*	0.9
Number of features used	2.13	3.24*
Sensitivity (d')	1.45	1.74*
Response bias (Beta (log10))	0.21	0.6*
Subjective Workload	50	23*

Discussion

The study demonstrates an approach to evaluating user interfaces for XAI, using signal detection theory and shows how user interface designs differ in their impact on user decisions. While the task was not too demanding (the correctness scores for both user interfaces were close to 1 which suggests that we are close to a ceiling effect) and sensitivity for both user interfaces was greater than 1.0, there are some differences between the user interfaces. In line with prior work (Abeynayake et al., 2023; Fischer et al., 2005; Holt et al., 2011), the radar plot resulted in lower accuracy and consistency. This does not necessarily mean that LIME ought to be preferred because the Signal Detection measures (for sensitivity and response bias) were lower for LIME. One explanation of the potential for LIME to encourage a (slightly) more liberal response is that it presents the user with more information than the radar plot, which could lead to more demand on the user to filter this information (hence, the higher workload ratings). This seems to agree with the comments from Dieber and Kirrane (2020) concerning the problems that users have with the amount of information presented by LIME. Interestingly, in our study, the effort put into filtering the LIME output also resulted in fewer features being used (in comparison to decisions using the radar plot).

From table 1, it is apparent that the radar plot generates more False Alarms (false positives) whereas LIME generates many more Misses. Knowing the prevalence of Misses and False Alarms might influence one's choice on which user interface to recommend, i.e., in some applications one might need to reduce Misses to as low a value as possible, and in other applications one might wish to minimise False Alarms. We also note that neither user interface reduced False Alarms or Misses to zero, and it would be interesting to determine what type of user interface (for XAI) could help in minimising these values.

Given the Misses and False Alarms (even in a task which we claim is relatively undemanding, and with reasonable accuracy and sensitivity) we argue that users will need to bring additional experience and knowledge to support their interpretation. To date, very few approaches to XAI consider the need for human-centred approaches (Hoffman et al., 2018).

Our aim in this study was not to define a 'best' user interface for XAI but to consider the pros and cons of two possible alternatives. The differences that we have identified across the various metrics suggest that evaluation of XAI must include multiple measures of performance and that there will be differences in choice of user interface depending on a host of task-related factors.

References

Abeynayake, H.I.M.M., Goonetilleke, R.S., Wijeweera, A. and Reischl, U., 2023. Efficacy of information extraction from bar, line, circular, bubble and radar graphs. *Applied Ergonomics*, *109*, p.103996.

Carswell, C.M. and Wickens, C.D., 1990. The perceptual interaction of graphical attributes: Configurality, stimulus homogeneity, and object integration. Perception & Psychophysics, 47, pp.157-168.

Dannenhauer, Z., Molineaux, M. and Cox, M.T., 2019. Explanation-based goal monitors for autonomous agents. Advances in Cognitive Systems, pp.1-6.

Dieber, J. and Kirrane, S., 2020. Why model why? Assessing the strengths and limitations of LIME. *arXiv preprint arXiv:2012.00093*.

Draper, G.M., Livnat, Y. and Riesenfeld, R.F., 2009. A survey of radial methods for information visualization. *IEEE transactions on visualization and computer graphics*, *15*(5), pp.759-776.

Fischer, M.H., Dewulf, N. and Hill, R.L., 2005. Designing bar graphs: Orientation matters. *Applied Cognitive Psychology: The Official Journal of the Society for Applied Research in Memory and Cognition*, *19*(7), pp.953-962.

Greaney, J. and MacRae, A.N., 1996. Diagnosis of fault location using polygon displays, *Ergonomics*, *39*(3), pp.400-411.

Greydanus, S., Koul, A., Dodge, J. and Fern, A. (2018) Visualizing and understanding Atari agents, arXiv:1711.00138v5

Hoffman, R.R., Klein, G. and Mueller, S.T. (2018) Explaining explanation for "explainable AI", *Proceedings of the human factors and ergonomics society annual meeting*, 62, Los Angeles, CA: Sage, 197-201.

Holt, J., Bennett, K. and Flach, J., 2011, September. Ambiguity and content mapping among display types. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*(Vol. 55, No. 1, pp. 390-393). Sage CA: Los Angeles, CA: SAGE Publications.,

Kaur, H., Nori, H., Jenkins, S., Caruana, R., Wallach, H. and Wortman Vaughan, J. (2020) Interpreting interpretability: understanding data scientists' use of interpretability tools for machine learning, *CHI'20,* New York: ACM, 1-14.

Li, D., Liu, Y., Huang, J. and Wang, Z., 2023. A Trustworthy View on Explainable Artificial Intelligence Method Evaluation. *Computer*, *56*(4), pp.50-60.

Pionek, J., Heidenreich, M., Shanahan, M. and Schoepke, D., 1997, October. Factors affecting user response to polygon displays. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* (Vol. 41, No. 2, pp. 1052-1055). Sage CA: Los Angeles, CA: SAGE Publications.

Sreedharan, S., Chakraborti, T. and Kambhampati, S., 2021. Foundations of explanations as model reconciliation. Artificial Intelligence, 301, p.103558.