Issues Affecting User Confidence in Explanation Systems

David A. Robb, Stefano Padilla, Thomas S. Methven, Yibo Liang, Pierre Le Bras, Tanya Howden, Azimeh Gharavi, Mike J. Chantler, Ioannis Chalkiadakis

This work is a joint contribution from all authors and names are listed in reverse alphabetical order.

Heriot-Watt University, EH14 4AS, Edinburgh, Scotland
S.Padilla@hw.ac.uk (corresponding author)

Abstract. Recent successes of artificial intelligence, machine learning, and deep learning have generated exciting challenges in the area of explainability. For societal, regulatory, and utility reasons, systems that exploit these technologies are increasingly being required to explain their outputs to users. In addition, appropriate and timely explanation can improve user experience, performance, and confidence. We have found that users are reluctant to use such systems if they lack the understanding and confidence to explain the underlying processes and reasoning behind the results. In this paper, we present a preliminary study by nine experts that identified research issues concerning explanation and user confidence. We used a three-session collaborative process to collect, aggregate, and generate joint reflections from the group. Using this process, we identified six areas of interest that we hope will serve as a catalyst for stimulating discussion.

Keywords: Explanations, Confidence, Decision Making, AI, Systems.

CEUR-WS.org/Vol-2151/Paper_S7.pdf

1 Introduction

Background and motivation. Our aim is to improve users' confidence in the use of AI systems. If users have confidence in (a) the inferences that these systems make, (b) the provenance of the data on which these inferences are made, and (c) in the explanation systems themselves then they will be more likely to employ these technologies [1]. This is particularly important in high-stakes scenarios where the risks of material and reputational damage are significant. This work therefore initially focused on a single question: *What issues do you think most effect user confidence in explanation systems?* However, during the study, the scope increased to encompass issues concerning confidence in the visualisation and data processing stages as well.

The ideation and reflection process. Nine researchers took part in the process (five PhD students, two senior postgraduate Research Associates and two Faculty members). All participants research explanation systems, user confidence, or both. The methodology used can be split into three phases:

i) Participants independently submitted responses to the following question: *What issues do you think most effect user confidence in explanation systems?*
ii) Participants sorted the ideas into groups using a distributed card sorting tool [2]. The grouping data was used to form six groups using a standard agglomerative

clustering algorithm. The six groups, which were assigned colours (Red, Blue, Green, Orange, Purple, and Yellow), together with visualisations of the card sorting results were supplied to the participants.

iii) A roundtable meeting was held in which each of the six groups was considered in turn. A simple 'round robin' protocol was used to ensure that all members of the study provided input and reflection. At the end of each discussion, a group name was agreed. Participants took individual notes which they used after the meeting to help them develop and add their reflections to a shared document.

Steps i) and ii) above are described in the document provided to participants [3], and the sections below describe the results from step iii) in more detail.

2 Reflections on User Confidence and Explanation Systems

Each of the following sections contain summaries of our expert participants' reflections on the various issues contained within each group, along with the agreed group title. In the discussion, each group was assigned a colour to allow for groups to be referenced before they were named. We include these as large coloured squares to allow the reader to quickly reference the groups in the original document [3] and to look up details of the ideas submitted. The sections are ordered arbitrarily and can be read independently.

2.1 Filtering for credibility

This group reflects on the issues concerning perceived credibility and filtering. It covers the need for users to find information at various stages of the system and perform checks on the information to reassure themselves of the system's performance and avoid broken expectations.

Filtering. This mechanism was seen as necessary for the credibility of the system as it allows users to look for specific, known information. It was believed that this mechanism should be 'deep', therefore not searching only the words on the screen but also the information behind it. It was argued that filtering rather that searching [4] can improve the credibility of the data provenance, and that users want to investigate both the input and outputs from the system. Finally, it was hypothesised that participants want to find known information in the system to perform a series of 'spot checks' allowing them to increase their confidence in the system's coverage and accuracy.

Missing information and broken expectations. Participants believed that users' expectations can be broken when information is missing (e.g. stop word removal) which can reduce confidence in the system. It was therefore suggested that there is a need to explicitly record these missing pieces of information and for design recommendation of how they can be shown to the user. Finally, it was argued that the models in a system need to relate to what users expect and be clearly explained or a system may not be perceived as credible.

2.2 Characteristics, compromises, and chunking

This group covers implications for visual design. Participants discussed the features of layouts [5] that aid explanation and understanding. Compromises between abstraction (that can aid cognitive chunking) and detail (that allows unusual events to be identified) together with the issue of outliers were debated.

Abstraction vs detail. This was seen as a key design issue by participants. It was thought that ideally it should adapt to user type (power to naïve) and depth of explanation. In addition, participants believed that expert users would desire more detail at more levels, and be able to process more complex visual presentations than naïve users. The representation of relationships was seen as an important issue; representing many possibly complicating the understanding of an overall mental model [6], while representing fewer might omit important detail. Participants considered ease of visual aggregation (to aid cognitive chunking [1]) and the minimisation of visual clutter as a critical design criterion for all classes of user.

Outliers and sparsity. Outliers and sparsity were seen as difficult issues, both from representational and explanation viewpoints. The explanation of outliers was thought to need particularly detailed, multi-level, data-driven explanation to enable data exceptions or unexpected inferencing to be identified and explained. It was recognised that for some applications it is the identification and characterisation of outliers that is the key task. Here it was thought that the ability of the user to adapt the specification of 'normal' and to further filter outliers was critical.

Compromise. Overall, participants recognised that the above issues presented difficult compromises for the designer. In particular, trading off the level of detail against the minimisation of visual clutter and possible cognitive overload, were seen to be challenging. The ideal solution was seen as being a user-controlled continuum, pre-set to a level suited to the naïve user.

2.3 User perception of coherence and consistency

The group reflected on the different aspects of how a system is perceived [7] and interacted with by a user and how that in itself can affect confidence. Issues such as ensuring results were consistent and could be quickly interpreted, as well as minimising user disagreement were discussed.

Consistency of outputs. It was conjectured that it is essential for users to consider a system's results as consistent; otherwise, a user's view of the credibility of said system will be adversely affected. This was seen as a different issue to users' mental models, due to it being more related to the data processes and visualisation itself (e.g. what if two similar items are categorised differently).

Credibility. The participants discussed whether it was possible to create one single, organised, and credible view. Furthermore, they suggested that systems should focus on what they are good at and minimise making inferences from incomplete or noisy data. It was considered that if there was substantial disagreement between users about

what a visualisation showed, or if it took a long time to interpret the results then the design might be inadequate or flawed. As a result, the participants believed the goal should not be to eliminate disagreement but to minimise it.

2.4 Algorithmic transparency

This group considers how one designs the pipeline for a system, in particular low-level issues including: transparency, uncertainty, algorithmic competency, and awareness of what processes were performed on the data.

Transparency: explanation and knowledge of competency. Participants reflected that visualisation creators have ethical and regulatory imperatives to ensure that algorithms are transparent. They discussed that the user should be able to determine (via a query, for instance) the reasons behind, and the competence of, any algorithmic inference. Competency could be communicated as an uncertainty measure on the outputs. Transparency was seen to have implications for user confidence and visualisation complexity. Finally, it was discussed whether carefully presented data would increase the transparency of the model, or increase user understanding of the data.

Uncertainty and ethical considerations. These issues were seen as being different to coherency as there might not be a correct platonic explanation or visualisation/layout. A question was left open as to how to visualise the uncertainty. One important issue discussed was that ethics in the process are essential. One cannot avoid the probabilistic nature of the models just because they are problematic. Data is not perfect, adding its own uncertainty, separate from the uncertainty caused by stochastic algorithms which produce different solutions on each run (e.g. topic modelling algorithms [6]). It was thought these issues could cause users to think a system is unreliable!

2.5 Data provenance and user bias

Compared to previous groups the ideas contained here were more diverse. Issues discussed in this group consisted of three main topics: the credibility and provenance of sources, users' own bias and mental models, and allowing users the opportunity to check the system.

Provenance. Participants discussed that it might be useful to distinguish between the provenance of the data and the provenance of the inferencing, as users may well need both. They hypothesised that provenance would increase confidence and debated how it can be visualised and proved. Proposed solutions included having multiple, reliable, and familiar sources of data, or sources that fit users' mental models.

User bias. What biases do users bring to a system? It was posited that bias could happen when users compared their mental model to what they are presented with. Participants suggested presenting an interactive explanation of the visualisation creation process [1], as well as showing the result. A link to confirmation bias was raised, along with the possibility of democratic data collection (e.g. from social media, or crowdsourced collections like Wikipedia). It was argued that users need to trust that

the data was collected representatively, fairly, and without bias. How do you show that reliable sources have been used and that the data is credible?

2.6 Language, culture user and user background

The discussion concerning the final group dealt broadly with the match (or mismatch) between the users' mental models, background and culture, and the models, conventions, and graphical representations of the application. Much of the debate reiterated issues covered extensively in the literature on user-centred design [8], including: the importance of aligning language, terminology, graphical representations, and underlying conceptual models with those of the user. Thus user characterisation, understanding existing user conventions and procedures, together with techniques such as iterative, in situ, development of design probes and prototypes were seen as key issues and particularly relevant for the development of explanation systems.

3 Conclusions

This paper presents the results of a three-stage ideation and consolidation process carried out by nine researchers on issues that affect user confidence in explanation systems. In summary the issues raised fell into six areas: 1) the use of filtering to help credibility assessment; 2) the trade-off of abstraction against detail; 3) users' perception of result consistency; 4) multilevel algorithmic transparency; 5) accessibility of data provenance; and 6) the importance of user-centred design. These, somewhat eclectic, incomplete and overlapping, sets of issues would benefit from extension and review. As such, we hope that this paper provides valuable input to the workshop and promotes vigorous discussion of this area.

References

1. Le Bras, P., et al. Improving User Confidence in Concept Maps: Exploring Data Driven Explanations. ACM CHI'18, USA (2018). doi: https://doi.org/10.1145/3173574.3173978
2. Methven, T.S., et al. Research strategy generation: avoiding academic 'animal farm'. ACM CSCW'14. USA (2014). doi: https://doi.org/10.1145/2556420.2556785
3. Chalkiadakis, I., et al. Confidence in Explanation System: Research Issues. Research Materials Well Sorted. UK (2018). doi: http://dx.doi.org/10.13140/RG.2.2.30733.03043
4. Wilson, M., Kules, B., Shneiderman, B. From keyword search to exploration: Designing future search interfaces for the web. Foundations and Trends in Web Science, (2010).
5. Ware, C. Information visualization: perception for design. 3rd Edition, Elsevier, (2012).
6. Padilla, S., et al. Understanding Concept Maps: A Closer Look at How People Organise Ideas. ACM CHI'17, 815-827, USA (2017). doi: https://doi.org/10.1145/3025453.3025977
7. Pieters, W. Explanation and trust: what to tell the user in security and AI? Ethics and information technology 13, 1, 53-64, (2011), 53–64.
8. Rogers, Y., Sharp, H., Preece, J. Interaction design: beyond human-computer interaction. John Wiley & Sons. (2011).