The Problem of Explanations without User Feedback

Alison Smith
Decisive Analytics Corporation
Arlington, United States
alison.smith@dac.us

James J. Nolan
Decisive Analytics Corporation
Arlington, United States
jim.nolan@dac.us

ABSTRACT
Explanations are necessary for building users' understanding and trust in machine learning systems. However, users may abandon systems if these explanations demonstrate consistent errors and they cannot affect change in the systems' behavior in response. When user feedback is supported, then the utility of explanations is to not only promote understanding, but also enable users to help the machine learning system overcome errors. We suggest an experiment to examine how users react when a system makes explainable mistakes with varied support for user feedback.

Author Keywords
Explanations; user feedback; human-in-the-loop systems; human-machine interfaces

ACM Classification Keywords
H.1.2 User/Machine systems: Human Information Processing

INTRODUCTION
Analysts in domains such as military, intelligence, financial, and medical face the ever-growing problem of needing to perform multi-modal analysis on complex data sets. Machine learning techniques show promise for dramatically increasing the speed and effectiveness of analytic workflows for analyzing large amounts of data. However, because an analysts' credibility and reputation may be rated based on automated decisions, they hesitate if they do not have a full understanding of how the algorithm reached its final decision. To overcome this doubt, and to increase interpretability and trust of these systems, it is necessary to provide a transparent way to inspect, interrogate, and understand machine learning results.

While significant work explores this need for more explainable machine learning – whether by making the algorithms themselves more explainable [1,2] or by creating explanation interfaces to explain algorithm output [4], we argue that explanations alone are not always sufficient. Support for user feedback should be treated as an equally important component of an explainable machine learning system as providing an explanation without a method for feedback may lead to frustrated users and overall system disuse. For example, [5] find that users ignore explanation when the benefit of attending to them is unclear or if they are unable to successfully control predictions.

In this position paper, we discuss prior work on explainable systems that do and do not take user feedback into account and outline a study design to examine how users react when a system makes explainable mistakes with varied support for user feedback.

MOTIVATION
Imagine the following example (from a frequently-cited explainable machine learning paper [10]): an explainable system, specifically an image classification tool, makes an error, such as classifying an image of a husky incorrectly as a wolf. The user then requests an explanation of why the system produced this incorrect classification. Using *attention,*[1] the system can explain its mistake by displaying to the user that the presence of snow in the image led to the wolf misclassification. At this point, the user now understands why the system made this initial mistake, but what happens when the system makes the same or similar mistake again? Here we consider two possible outcomes. One possibility is that the user will be frustrated or choose not to use the system if they know that the it errs on certain types of data or problems, but they cannot do anything about it. Alternatively, the system may be deceiving to users who believe that the it can learn from mistakes (as a human who admits to making a mistake is expected to), but in fact it will continue to make the same mistake.

In fact, Ribeiro et al. [10] find that while 10 of 27 participants trust the model that misclassifies a husky for a wolf without any explanation, only three out of 27 participants trust the model when it explains the mistake. Thus, without a way to provide user feedback to improve

[1]In image classification, *attention* [8] can be used to determine the portion of an image that most affected the system's classification, or the part of the image that the system "attended to" the most when making a classification.

the system, explaining predictions is more likely to be utilized as a method for knowing when *not* to trust the system.

Alternatively, in our prior work [13], we developed a system for intelligence analysts that both provides evidence for its decisions and supports analyst feedback to improve the underlying model. This system automatically clusters entity mentions (people, places, and organizations) from large unstructured corpora to overarching entity clusters.[2] For example, clustering entity mentions throughout a large news corpus, such as *Mr. Obama*, *President Obama*, and *Barack*, into one entity cluster, President Barack Obama. The system provides as evidence the entity mention in context as well as the other entity mentions in the cluster. While this evidence may help the analyst to understand why certain mentions were incorrectly placed in clusters or other mentions are missing from a cluster, simply understanding the system's mistakes is not sufficient for supporting trust and utilization. To this end, the system supports interactive feedback mechanisms, such as accepting and rejecting mentions as well as merging clusters. While no formal user experiment has been performed with this system, we have received positive feedback from analysts regarding the interactive feedback mechanisms.

EXPERIMENT DESIGN

We outline an experiment design to examine how users react to an explainable system with varied support for user feedback. Specifically, we suggest two possible study methods: the first, which aims to explore user frustration or confusion that occurs when an explainable system (that does not update with user feedback) continues to make the same or similar mistakes, and the second, which compares users' reactions to versions of a system based on the amount of control given to the user.

Research Questions

The goal of this proposed experiment is to answer the following research questions:

Q1: Do users assume an explainable system learns from mistakes?

We hope to better understand what users expect when utilizing an explainable system. Whether or not users expect the system to continue making the same or similar mistakes impacts how negatively the users will be affected when it does. Furthermore, we would like to understand whether users' expectations change if we vary how explanations are attended to or the form of explanations.

Shneiderman [11] and Lanier [6] argue that systems (intelligent agents, in particular) should not have human-like characteristics as these lead users to believing that the system may act rationally or take some responsibility for its actions [3]. Therefore, we hypothesize that conversational (or apologetic) explanations may be more likely to lead users to thinking the system will learn from a mistake.

Similarly, whether users expect a system to improve may vary based on their interactions with explanations, such as simply clicking 'ok' to dismiss explanations as opposed to interactions such as 'accepting' or 'rejecting' classifications. The latter may lead users to believe they are correcting the system.

Q2: How is trust of and frustration with an explainable system affected by varied supports for user feedback?

Prior work implied, albeit without a formal experiment, that users may trust systems less when they explain their mistakes [10]. Similarly, Lim and Dey [7] find that users' impressions of a system are negatively impacted when systems are highly uncertain of their decisions (even when they behave appropriately). While supporting user feedback, particularly in cases of system error or high uncertainty, could mitigate these issue, the level of control given to the user may have varied effects on trust and frustration. For example, in prior work we discuss whether user feedback should be taken as a *command* or a *suggestion* for different types of interactive systems [12].

Method

To support examination of the identified research questions, we outline the following two-part study methodology.

The first part of the study will be performed as an interview study following a think-aloud methodology followed by a post-task survey. First, users will be shown an explainable system. When the system errs, it will provide an explanation. We will then ask users whether they believe the system will make the same or similar mistakes followed by measuring frustration and/or surprise when it does continue to do so. Frustration will be measured on the incident level and overall level following the methodology described by Bessier et al. [1]. For this part of the study we will vary what explanations look like and how users attend to explanations, as we hypothesize these will have an effect on whether users believe the system will learn from mistakes.

The second part of the study will be performed as a crowdsourced survey. In this case, we will incorporate user feedback into an explainable system. We will vary the system only in how it incorporates user feedback, representing the amount of control the user has over the system. We propose three system variants: one that ignores all user feedback, one that takes feedback into account as a *suggestion*, and one that takes feedback into account as a *command*. We will then measure how user trust, frustration, and other user reactions differ between these variants. Frustration will again be measured on the incident and overall level [1]. The users' impressions of the system, and in particular trust, will be measured by rating responses to relevant survey questions.

[2]This technique utilizes inter and intra-document entity co-reference, meaning it clusters entity mentions within and across documents.

CONCLUSION

In this position paper, we argue that while explainable systems are important, incorporating user feedback into these systems is equally important for supporting trust and continued use. And this goes both ways – systems that support user feedback must also ensure users understand how they work, such that they can give appropriate feedback. A truthful explanation into the system's black box improves users' understanding, which better prepares them for providing feedback to improve the system. We propose an experiment to provide additional evidence for this argument.

REFERENCES

1. Katie Bessiere, Irina Ceaparu, Jonathan Lazar, John Robinson, and Ben Shneiderman. 2003. Understanding Computer User Frustration: Measuring and Modeling the Disruption from Poor Designs. *Technical Reports from UMIACS*. Retrieved from http://drum.lib.umd.edu/handle/1903/1233%5Cnhttp://drum.lib.umd.edu//bitstream/1903/1233/1/CS-TR-4409.pdf
2. William Brendel and Sinisa Todorovic. 2011. Learning spatiotemporal graphs of human activities. In *Proceedings of the IEEE International Conference on Computer Vision*, 778–785. https://doi.org/10.1109/ICCV.2011.6126316
3. K. Höök. 2000. Steps to take before intelligent user interfaces become real. *Interacting with Computers* 12, 4: 409–426. https://doi.org/10.1016/S0953-5438(99)00006-5
4. Todd Kulesza, Margaret Burnett, Weng-Keen Wong, and Simone Stumpf. 2015. Principles of Explanatory Debugging to Personalize Interactive Machine Learning. In *Proceedings of the 20th International Conference on Intelligent User Interfaces - IUI '15*, 126–137. https://doi.org/10.1145/2678025.2701399
5. Todd Kulesza, Simone Stumpf, Margaret Burnett, Sherry Yang, Irwin Kwan, and Weng Keen Wong. 2013. Too much, too little, or just right? Ways explanations impact end users' mental models. In *Proceedings of IEEE Symposium on Visual Languages and Human-Centric Computing, VL/HCC*, 3–10. https://doi.org/10.1109/VLHCC.2013.6645235
6. Jaron Lanier. 1996. My Problems with Agents. *Wired*.
7. Brian Y. Lim and Anind K. Dey. 2011. Investigating intelligibility for uncertain context-aware applications. In *Proceedings of the 13th international conference on Ubiquitous computing - UbiComp '11*, 415. https://doi.org/10.1145/2030112.2030168
8. Volodymyr Mnih, Nicolas Heess, Alex Graves, and Koray Kavukcuoglu. 2014. Recurrent Models of Visual Attention. *Advances in Neural Information Processing Systems* 27: 1–9. https://doi.org/ng
9. Seyoung Park, Xiaohan Nie, and Song Chun Zhu. 2017. Attribute And-Or Grammar for Joint Parsing of Human Pose, Parts and Attributes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. https://doi.org/10.1109/TPAMI.2017.2731842
10. Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. Why Should I Trust You? Explaining the Predictions of Any Classifier. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '16* 39, 2011: 117831. https://doi.org/10.1145/2939672.2939778
11. Ben Shneiderman. 1997. Direct manipulation for comprehensible, predictable and controllable user interfaces. In *Proceedings of the 2nd international conference on Intelligent user interfaces - IUI '97*, 33–39. https://doi.org/10.1145/238218.238281
12. Alison Smith, Varun Kumar, Jordan Boyd-Graber, Kevin Seppi, and Leah Findlater. 2017. Accounting for Input Uncertainty in Human-in-the-loop Systems. In *Designing for Uncertainty Workshop at CHI 2017*.
13. Kevin Ward and Jack Davenport. 2017. Human-machine interaction to disambiguate entities in unstructured text and structured datasets. In *SPIE Conference on Next-Generation Analyst V*.