Position: We Can Measure XAI Explanations Better with Templates

Jonathan Dodge
dodgej@eecs.oregonstate.edu
Oregon State University

Margaret Burnett
burnett@eecs.oregonstate.edu
Oregon State University

ABSTRACT

This paper argues that the Explainable AI (XAI) research community needs to think harder about how to compare, measure, and describe the quality of XAI explanations. We conclude that one (or a few) *explanations* can be reasonably assessed with methods of the "Explanation Satisfaction" type, but that scaling up our ability to evaluate explanations requires more development of "Explanation Goodness" methods.

CCS CONCEPTS

• **Human-centered computing** → **User studies**.

KEYWORDS

Explainable AI, Evaluating XAI Explanations, Empirical Studies, Heuristic Evaluations

ACM Reference Format:
Jonathan Dodge and Margaret Burnett. 2020. Position: We Can Measure XAI Explanations Better with Templates. In *Proceedings of the IUI workshop on Explainable Smart Systems and Algorithmic Transparency in Emerging Technologies (ExSS-ATEC'20).* Cagliari, Italy, 5 pages.

1 INTRODUCTION

As AI plays an ever-increasing role in our lives, society needs a variety of tools to inspect them. Explanations have emerged to fill that role, but measuring their quality continues to prove challenging. Hoffman et al. [10] offers two terms to describe mechanisms for measuring the quality of an explanation, quoted at length here, with highlighting added to assist this paper's discussion.

> *"**Explanation Goodness**: Looking across the scholastic and research literatures on explanation, we find assertions about what makes for a good explanation, from the standpoint of statements as explanations. There is a general consensus on this; factors such as clarity and precision. Thus, one can look at a given explanation and make an a priori (or decontextualized) judgment as to whether or not it is "good." ... In a proper experiment, the researchers who complete the checklist [evaluation of the explanation] with reference to some particular AI-generated explanation, would not be the ones who created the XAI system under study."*
>
> *"**Explanation Satisfaction**: While an explanation might be deemed good in the manner described above, it may at the same time not be adequate or satisfying to users-in-context. Explanation Satisfaction is defined as the degree to which users feel that they understand the AI system or process being explained to them. Compared to goodness, satisfaction is a contextualized, a posteriori judgment of explanations."*

In this paper, we will use ExpG and ExpS to refer to Hoffman's concepts of Explanation Goodness and Satisfaction, respectively. Now, consider the three highlighted properties in each definition:

- **Contextualization:** ExpS is defined relative to a *task*, while ExpG is not.
- **Actor:** ExpS is measured from the perspective of a *user* performing a task, while ExpG is from the perspective of *researchers* (ideally dispassionate bystanders, but often the designers themselves).
- **Timing:** Because ExpS is defined relative to a task, it must be measured *after* the task is completed, while ExpG can be measured anytime.

The main thesis of this paper will be that we, as a research community, need to think harder about how to compare, measure, and describe the ExpG of explanation templates (clarified later in Section 5), as a potentially strong complement to ExpS.

To develop the main thesis, we attempt to argue several points:

- *Background:* Most current research has focused on ExpS.
- *Tasks:* ExpS is easier to operationalize, but incurs a great deal of experimental noise.
- *Benefits:* ExpS's usefulness is hampered by participants' limited exposure to the system.
- *Scope:* ExpG affords the opportunity to consider a wider range of behaviors, making ExpG mechanisms particularly well suited to reasoning about explanation templates.

Through these points, we hope to provoke thought about how explanation designers can better validate their design decisions via ExpG for explanation templates. This is of particular importance because a great many design decisions are *never* evaluated via ExpS mechanisms.

2 BACKGROUND: MOST CURRENT RESEARCH HAS FOCUSED ON EXPS

How have past researchers evaluated XAI design decisions? Most have used ExpS mechanisms, but a few have used ExpG mechanisms. For both types, an important criterion is rigor. The dangers of departure from rigorous processes for validating design decisions can be potentially severe if it devolves into "I methodology" [24], i.e., designers relying solely on their own views and assumptions about what their users will need and how they will use the functionalities the designers decide to provide.

ExSS-ATEC'20, March 2020, Cagliari, Italy

2.1 Research that uses ExpS mechanisms

Through extensive literature review, Hoffman et al. [10] identified a group of existing ExpS methods for mental model elicitation (their Table 4). Among them, many are essentially qualitative and focus on things people say (e.g. Think Aloud or Interview techniques). We felt the "Retrospection Task" [18] and the "Prediction Task" [20] looked to be the most well suited for quantitative study, and chose to use them for Anderson et al.'s empirical studies [4]. Approaching the problem from another angle, Dodge et al. [8] investigated several aspects of perceptions of fairness and explanations in a decision support setting.

Other researchers have used ExpS to understand a wide variety of effects in explanation. Providing explanation has been shown to improve mental models [15, 16]. Of particular importance to moderating the effects of explanation is the explanation's **soundness** and **completeness** [17]; most easily described with the phrase "the whole truth (completeness) and nothing but the truth (soundness)" about how the system is really working. Note that neither soundness nor completeness are binary properties, but a smooth continuum—with 100% soundness or completeness not always achievable. Explanation has also been shown to increase satisfaction (here we mean in the colloquial sense, the user's self-reported feeling) [2, 12], and understanding—particularly in low expertise observers [27]. Several different kinds of explanation have also been shown to improve user acceptance via setting appropriate expectations (e.g. by showing an accuracy gauge) [14]. There are many other researchers studying explanations using ExpS mechanisms, and we refer the reader to Abdul et al. [1] for a recent literature review.

2.2 Research that uses ExpG mechanisms

One XAI tool for ExpG is the checklist proposed by Hoffman et al. [10]'s Appendix A, composed of 8 yes/no questions that researchers and designers ask themselves (e.g. *"The explanation of the [software, algorithm, tool] is sufficiently **detailed**."*). Other approaches include Amershi et al. [3]'s guidelines for interactive AI, Kulesza et al. [15]'s design principles for explanatory debugging, and Wang et al. [25]'s guidelines to match human reasoning processes with XAI techniques. However, most XAI research is not explicit about *usage* of these or other ExpG mechanisms.

3 TASKS: EXPS IS "EASY" TO OPERATIONALIZE, BUT NOISY

ExpS sings a siren song of sorts; it appears simple to evaluate, one must simply define a task and criteria to measure performance at that task. Easy enough, right? Wrong.

We have been using one of the XAI tasks which we felt would be most well suited for quantitative study, the "Prediction Task", proposed by Muramatsu et al. [20]. However, we have run into a number of challenges using it in our XAI studies, some of which are apparent in Figure 1, taken from Anderson et al. [4].

First, participants' ability to perform the task (predict an AI's next actions) in a domain are moderated by a number of other things, such as their need for cognition, interest in the task, domain experience, etc. To illustrate the effect of variability in explanation consumers, imagine a scientist effectively describing quantum computing to another scientist. Then imagine the same person giving the same explanation to a child[1]. The explanation *itself* could be high quality, it just was not appropriate for that audience and needed reformulation. Thus, empirically measuring the explanation's quality is entangled with many factors beyond the explanation itself.

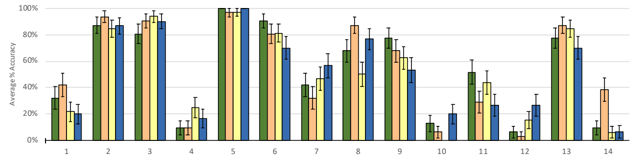


Figure 1: (Source: Anderson et al. [4]) Percentage of participants in four explanation treatments (4 colors) correctly predicting the AI's action at 14 decision points. This image shows: 1. No treatment was a clear winner. 2. Some decisions were easy enough that all participants predicted correctly—even those without explanations. 3. Conversely, others were hard enough that few participants predicted correctly, even with explanations. 4. There is no evident learning effect (decision points are shown sequentially over time).

Second, there is a great deal of variability in the state/action space (as we observed in [4]). This leads some choices to be easy, causing all treatments to have nearly 100% participant prediction accuracy—even those without explanation (e.g. the 5th decision point in Figure 1). In contrast, others are much harder, and all treatments had nearly 0% prediction accuracy (e.g. the 4th decision point in Figure 1). As a result of these floor and ceiling effects, some of the variation between treatments is obscured.

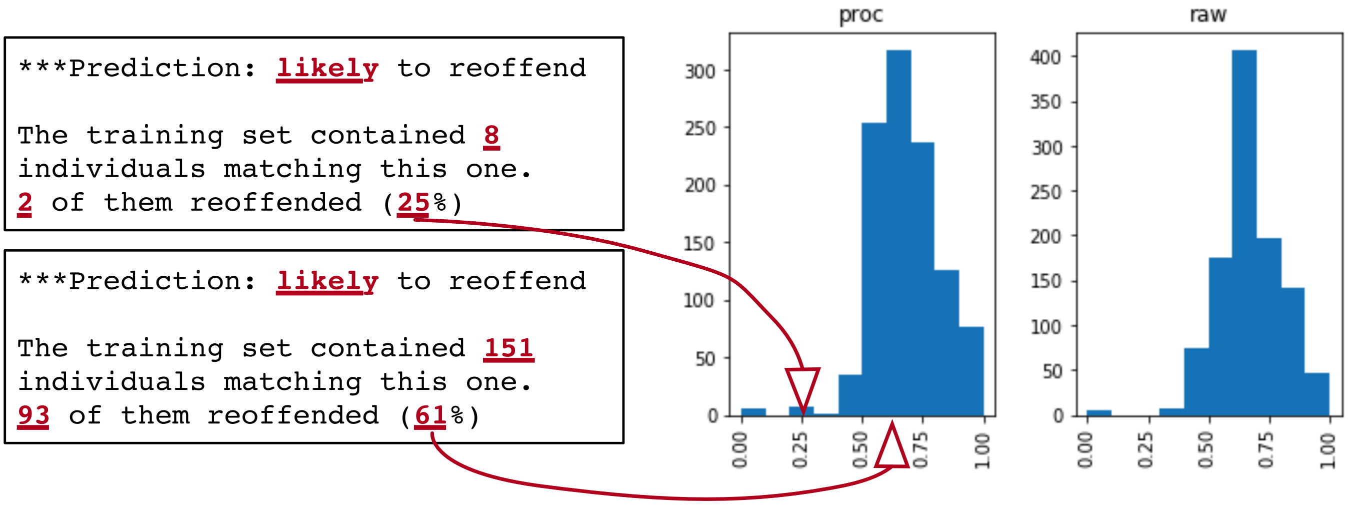
Third, it is difficult to assign "partial credit" for participants' predictions. In Figure 1's case, participants faced a choice of 4 options, leading random guessing to be right 25% of the time. However, AI is regularly used in domains with much larger action spaces, so the probability a participant picks right can be vanishingly small. As a result, it seems natural to think about which answers might be considered better than others. To do so, one might consider similarity in the action space (actions that *look* similar) or in the value space (actions that *produce similar consequences*), but either way it is a challenge to design rigorously.

4 BENEFITS: EXPS'S USEFULNESS IS HAMPERED BY LIMITED EXPOSURE

Consider that in-lab user studies are typically designed to be executed within a 2-hour window for a variety of reasons (e.g. reliability). As a result, the amount of participant exposure to the system is actually quite low. As an example, in Anderson et al.'s study [4] we showed 14 decision points to participants over the available 2 hours. In that paper, we point to this limited exposure as a possible reason that we did not observe any learning effect (evident in Figure 1).

Other challenges also surround exposing participants to the system when performing a ExpS evaluation. In particular, *which* decision points do we show to participants? Because this agent has

[1] https://www.youtube.com/watch?v=OWJCfOvochA conducts a similar exercise, though Dr. Gershon changes the explanation for 5 different audiences.



***Prediction: likely to reoffend

The training set contained 8 individuals matching this one. 2 of them reoffended (25%)

***Prediction: likely to reoffend

The training set contained 151 individuals matching this one. 93 of them reoffended (61%)

Figure 2: *Left:* Two example explanations provided by the system used by Dodge et al. [8]. The black text is static and the added red highlights show text that will be based on calculations about the input—intended to show how explanation templates get filled in. It demonstrates how the ExpS results vary based on the input (e.g. the top explanation is far less convincing). *Right:* Histogram of the matching percentages underlined in Figure 2, for the classifiers trained on raw and processed data. These histograms show how differently the two classifiers behaved, but also show an interesting result—namely how often case-based explanation self-refutes (by providing low %s), or does not substantiate any claim (by giving near 50%, in a binary classification setting). However, this insight might not have been observable for a user under the ExpS formulation, as users typically only work with a small number of explanations and self-refuting ones are rare.

been training for 30,000 episodes, scrutinizing the training data in its entirety would be a daunting task. After training is complete, one could imagine presenting test cases, some of which could be handcrafted. One way to select test cases to present to assessors is a recent approach to solving the problem of which decision points to show by Huang et al. [11]. Their approach measures the "criticality" of each state, and chooses the ones where the agent perceived its choice to matter the most.

Note again, the large variability in state/action space, which we commented on in Section 3. *When combined* with limited exposure, participants are essentially gazing into a vast expanse of behavior through a tiny peephole.

5 SCOPE: EXPG CAN CONSIDER A WIDER RANGE OF BEHAVIORS, VIA TEMPLATES

One great advantage of ExpG is, it supports explanation *templates*.

5.1 What is an explanation template?

The earliest evidence we could find for what we call "explanation templates" is from Khan et al. [13], and we think studying them can help address some of the problems discussed earlier in this paper. Explanation templates operate on a different granularity than an explanation. If an explanation describes or justifies an individual action, the explanation template is like the factory that creates the explanation.

We have built templates inspired by Binns et al. [5], who used a wizard of oz methodology to generate multiple types of explanation for a decision support setting. The decision they were trying to explain was an auto insurance quote ([5]'s Figure 2). One of their types, which they term "Case-based explanations", is demonstrated in the left side of Figure 2 as we used them in Dodge et al. [8] to explain an AI system's judicial sentencing recommendations. Note that the templates shown in this paper are for textual explanations, but the idea extends naturally to other types of explanations, like the visual explanations in Mai et al.'s Figures 1 and 2 [19].

5.2 Why consider explanation templates?

An explanation template can be combined with appropriate software infrastructure and a test set to generate a large set of explanation instances. We argue that examining the distribution of thousands of explanations generated in this way can be more illuminating than seeing individual ones (e.g., Figure 2, right).

Although ExpG mechanisms can be used on any explanation template one desires, consider using a case-based explanation template. One way to produce these explanations is by finding training examples "near" the input, then characterize how well the labels of the resultant set match the label of the input (Figure 2, left). If we run an explanation generator on the whole test set, we can create a histogram from those matching percentages, shown on the right side of Figure 2.

However, this introduces an issue with explanation soundness. To see why, note that many instances fall near 100%—these are the explanations a user might find "convincing". However, there is also a good chance the explanation finds that around 50% of the nearby training examples matched the input's label—these are the explanations that do not substantiate any claim (note that the classifier is binary). Even worse, there are instances that fall near

0%—these are the self-refuting explanations[2] along the lines of *"This instance was labelled an A because all the nearby training examples were B's."*.

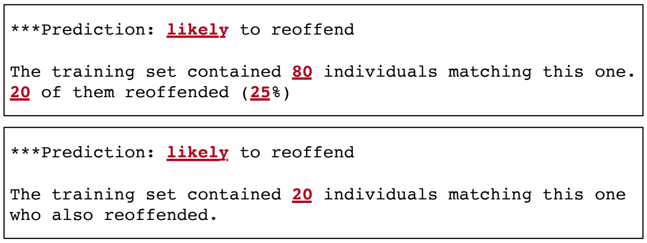
In this circumstance, the lack of soundness arises from the fact that the explanation uses nearest neighbors while the underlying classifier does *not*. Note also that the fix for these two problems is different. When the explanation lacks evidence, one should go find more evidence. But when the explanation self-refutes, one must figure out why the contradiction exists.

Note that if we were evaluating these explanations with ExpS, the result would depend strongly on whether the provided explanation was "convincing" or "self-refuting"—but the explanation *template* is the same in both cases, only the input changed. On the other hand, ExpG allows us to consider the wider scope—that the template is capable of generating explanations which *will occasionally refute itself*—and decide if that is acceptable.

To continue with the example of case-based self-refutation, suppose a member of the research team proposed an alternative explanation template that avoids refuting itself[3]. To do so, we adjust the static text and the calculation that fills in the variable parts, as illustrated in Figure 3. Note that the new proposal only highlights counts on things that match, which has the effect of essentially ignoring nearby training examples that do not match the input's predicted label.

The advantage of this proposal is that this alternative will not refute itself. But the advantage came at a cost: according to known taxonomies, it brings a decrease in completeness, as the explanation is telling less of the "whole truth."

So did the overall ExpG go up or down? We think most would argue down... but we cannot measure how much. This example exposes a critical weakness in the ExpG approach: the vocabulary and calculus currently available to us cannot adequately describe and measure the implications of a single design decision.



```
***Prediction: likely to reoffend

The training set contained 80 individuals matching this one.
20 of them reoffended (25%)
```

```
***Prediction: likely to reoffend

The training set contained 20 individuals matching this one
who also reoffended.
```

Figure 3: Two alternative designs for a case-based explanation template. The top shows one used in Dodge et al. [8], while the bottom illustrates strawman proposal for the discussion in this paper. Note how the alternative fails to acknowledge nearby training examples that did not match the input's predicted label. This will have the effect of decreasing completeness, but the alternative explanation will not self-refute (as the example shows).

5.3 Scalability: Many design decisions are *only* validated with ExpG

So where does this weakness leave us?

During a full design cycle, XAI designers face many design decisions, of which only a few can be evaluated with ExpS. To illustrate, consider that case-based explanation as originally proposed by Binns et al. [5] would be implemented by showing the single nearest neighbor. That approach could be extended by showing the k nearest neighbors for a number of different k. Or, it could be implemented by showing whatever neighbors lie within some feature space volume—which is the approach used by Dodge et al. [8] and illustrated in the left side of Figure 2. The space of these possible design decisions is large enough that we cannot hope to evaluate them all with ExpS mechanisms.

Given this, we suspect many XAI design decisions are *never* assessed, unless the XAI designers use ExpG—strictly due to the impracticality of doing ExpS at the scale needed for full coverage. In a sense, explanation *is* a user interface, and user interface designers have long used a wide variety of techniques relating to ExpG (e.g. design guidelines [21, 22], cognitive dimensions [9], cognitive walkthroughs [26], etc). Fortunately, there are some promising works that demonstrate the use of these approaches (e.g., [3, 15, 23]).

Hypothesis: Studying one (or a few) *explanations* is well-suited to ExpS oriented methods, but the *template* level may require ExpG methods.

This suggests that, to create XAI systems according to rigorous science—especially XAI systems that generate explanations with a template—we must develop improvements in the rigor and measurability of ExpG mechanisms. These will bring outsize benefit to designers and researchers as compared to improvements in ExpS.

6 ACKNOWLEDGMENTS

This work was supported by DARPA #N66001-17-2-4030. We would like to acknowledge all our co-authors on the work cited here, with specific highlights to Andrew Anderson, Alan Fern, Q. Vera Liao, Yunfeng Zhang, Rachel Bellamy, and Casey Dugan—research work does not happen in a vacuum, nor do ideas.

REFERENCES

[1] Ashraf Abdul, Jo Vermeulen, Danding Wang, Brian Y Lim, and Mohan Kankanhalli. 2018. Trends and trajectories for explainable, accountable and intelligible systems: An hci research agenda. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. ACM, New York, NY, USA, 582.

[2] S. Amershi, M. Cakmak, W. Knox, and T. Kulesza. 2014. Power to the people: The role of humans in interactive machine learning. *AI Magazine* 35, 4 (2014), 105–120.

[3] Saleema Amershi, Dan Weld, Mihaela Vorvoreanu, Adam Fourney, Besmira Nushi, Penny Collisson, Jina Suh, Shamsi Iqbal, Paul N. Bennett, Kori Inkpen, Jaime Teevan, Ruth Kikin-Gil, and Eric Horvitz. 2019. Guidelines for Human-AI Interaction. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) *(CHI '19)*. ACM, New York, NY, USA, Article 3, 13 pages. https://doi.org/10.1145/3290605.3300233

[2] The original reason we generated the histogram on the right of Figure 2 was not to see if the explanation would self-refute, but to compare the two classifiers: one trained on raw data (raw) and another trained on the processed data (proc). "Processing" the data refers to the use of a preprocessor by Calmon et al. [6] intended to debias the data—perhaps inducing a classifier people consider more fair. In this effort, we looked to see what classifications were different, compared confidence score histograms, etc.

[3] Here we use a strawman explanation template that is known to be bad, in order to explore the extent of our ability to characterize how bad it is. Correll performed a similar exercise in the visualization community, proposing Ross-Chernoff glyphs as a strawman, *"...as a call to action that we need a better vocabulary and ontology of bad ideas in visualization. That is, we ought to be able to better identify ideas that seem prima facie bad for visualization, and better articulate and defend our judgments."* [7].

[4] Andrew Anderson, Jonathan Dodge, Amrita Sadarangani, Zoe Juozapaitis, Evan Newman, Jed Irvine, Souti Chattopadhyay, Alan Fern, and Margaret Burnett. 2019. Explaining Reinforcement Learning to Mere Mortals: An Empirical Study. In *Proceedings of the 28th International Joint Conference on Artificial Intelligence* (Macao, China) *(IJCAI'19)*. AAAI Press, Palo Alto, CA, USA, 1328–1334. http://dl.acm.org/citation.cfm?id=3367032.3367221

[5] Reuben Binns, Max Van Kleek, Michael Veale, Ulrik Lyngs, Jun Zhao, and Nigel Shadbolt. 2018. 'It's Reducing a Human Being to a Percentage': Perceptions of Justice in Algorithmic Decisions. In *2018 CHI Conference on Human Factors in Computing Systems* (Montreal QC, Canada) *(CHI '18)*. ACM, New York, NY, USA, Article 377, 14 pages.

[6] Flavio Calmon, Dennis Wei, Bhanukiran Vinzamuri, Karthikeyan Natesan Ramamurthy, and Kush R Varshney. 2017. Optimized Pre-Processing for Discrimination Prevention. In *Advances in Neural Information Processing Systems 30*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (Eds.). Curran Associates, Inc., Red Hook NY, 3992–4001. http://papers.nips.cc/paper/6988-optimized-pre-processing-for-discrimination-prevention.pdf

[7] Michael Correll. 2018. Ross-Chernoff Glyphs Or: How Do We Kill Bad Ideas in Visualization?. In *Extended Abstracts of the 2018 CHI Conference on Human Factors in Computing Systems* (Montreal QC, Canada) *(CHI EA '18)*. ACM, New York, NY, USA, Article alt05, 10 pages. https://doi.org/10.1145/3170427.3188398

[8] Jonathan Dodge, Q. Vera Liao, Yunfeng Zhang, Rachel K. E. Bellamy, and Casey Dugan. 2019. Explaining Models: An Empirical Study of How Explanations Impact Fairness Judgment. In *Proceedings of the 24th International Conference on Intelligent User Interfaces* (Marina del Ray, California) *(IUI '19)*. ACM, New York, NY, USA, 275–285. https://doi.org/10.1145/3301275.3302310

[9] Thomas R. G. Green and Marian Petre. 1996. Usability analysis of visual programming environments: a 'cognitive dimensions' framework. *Journal of Visual Languages & Computing* 7, 2 (1996), 131–174.

[10] Robert R. Hoffman, Shane T. Mueller, Gary Klein, and Jordan Litman. 2018. Metrics for Explainable AI: Challenges and Prospects. *CoRR* abs/1812.04608 (2018). arXiv:1812.04608 http://arxiv.org/abs/1812.04608

[11] Sandy H. Huang, Kush Bhatia, Pieter Abbeel, and Anca D. Dragan. 2018. Establishing Appropriate Trust via Critical States. *IROS* (Oct 2018). https://doi.org/10.1109/IROS.2018.8593649

[12] Ashish Kapoor, Bongshin Lee, Desney Tan, and Eric Horvitz. 2010. Interactive optimization for steering machine classification. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 1343–1352.

[13] Omar Zia Khan, Pascal Poupart, and James P. Black. 2009. Minimal Sufficient Explanations for Factored Markov Decision Processes. In *Proceedings of the Nineteenth International Conference on International Conference on Automated Planning and Scheduling* (Thessaloniki, Greece) *(ICAPS'09)*. AAAI Press, 194–200.

[14] Rafal Kocielnik, Saleema Amershi, and Paul N. Bennett. 2019. Will You Accept an Imperfect AI?: Exploring Designs for Adjusting End-user Expectations of AI Systems. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) *(CHI '19)*. ACM, New York, NY, USA, Article 411, 14 pages. https://doi.org/10.1145/3290605.3300641

[15] 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*. ACM, 126–137.

[16] Todd Kulesza, Simone Stumpf, Margaret Burnett, Weng-Keen Wong, Yann Riche, Travis Moore, Ian Oberst, Amber Shinsel, and Kevin McIntosh. 2010. Explanatory debugging: Supporting end-user debugging of machine-learned programs. In *Visual Languages and Human-Centric Computing (VL/HCC), 2010 IEEE Symposium on*. IEEE, 41–48.

[17] T. Kulesza, S. Stumpf, M. Burnett, S. Yang, I. Kwan, and W. K. Wong. 2013. Too much, too little, or just right? Ways explanations impact end users' mental models. In *2013 IEEE Symposium on Visual Languages and Human Centric Computing*. 3–10. https://doi.org/10.1109/VLHCC.2013.6645235

[18] Katherine Lippa, Helen Klein, and Valerie Shalin. 2008. Everyday expertise: cognitive demands in diabetes self-management. *Human Factors* 50, 1 (2008).

[19] Theresa Mai, Roli Khanna, Jonathan Dodge, Jed Irvine, Kin-Ho Lam, Zhengxian Lin, Nicholas Kiddle, Evan Newman, Sai Raja, Caleb Matthews, Christopher Perdriau, Margaret Burnett, and Alan Fern. 2020. Keeping It "Organized and Logical": After-Action Review for AI (AAR/AI). In *In 25th International Conference on Intelligent User Interfaces* (Cagliari, Italy) *(IUI '20)*. ACM, New York, NY, USA.

[20] Jack Muramatsu and Wanda Pratt. 2001. Transparent Queries: investigation users' mental models of search engines. In *Intl. ACM SIGIR Conf. on Research and Development in Info. Retrieval*. ACM.

[21] Jakob Nielsen. 2005. Ten usability heuristics. https://www.nngroup.com/articles/ten-usability-heuristics/

[22] Donald A Norman. 1983. Design principles for human-computer interfaces. In *Proceedings of the SIGCHI conference on Human Factors in Computing Systems*. 1–10.

[23] Oluwakemi Ola and Kamran Sedig. 2016. Beyond simple charts: Design of visualizations for big health data. *Online journal of public health informatics* 8 (28 12 2016). Issue 3. https://doi.org/10.5210/ojphi.v8i3.7100

[24] Nelly Oudshoorn and Trevor Pinch. 2003. *How Users Matter: The Co-Construction of Users and Technology (Inside Technology)*. The MIT Press, Cambridge, MA, USA.

[25] Danding Wang, Qian Yang, Ashraf Abdul, and Brian Y Lim. 2019. Designing Theory-Driven User-Centric Explainable AI. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. CHI*, Vol. 19.

[26] Cathleen Wharton, John Rieman, Clayton Lewis, and Peter Polson. 1994. The cognitive walkthrough method: A practitioner's guide. In *Usability inspection methods*. 105–140.

[27] Robert H Wortham, Andreas Theodorou, and Joanna J Bryson. 2017. Improving robot transparency:real-time visualisation of robot AI substantially improves understanding in naive observers, In IEEE RO-MAN 2017. *IEEE RO-MAN 2017*. http://opus.bath.ac.uk/55793/