A Learning Approach for Increasing AI Literacy Via XAI in Informal Settings

Mira Sneirson, Josephine Chai, and Iris Howley[0000−0002−4694−9081]

Williams College, Williamstown, USA
{mls4, jhc4, ikh1}@williams.edu

Abstract. To achieve AI literacy, the AI community employs explainable AI (XAI), to increase AI literacy for those outside of formal educational settings. Designing and evaluating XAI remains an open question that can be guided by existing learning science research. When designers view their XAI through a learning lens, they may better define, assess, and compare explanation implementations. We surveyed and interviewed designers of interactive explanations for AI to identify how practitioners build their XAI and to better understand how a learning lens can be applied for explanations of complex AI concepts.

Keywords: AI literacy, Explainable artificial intelligence (XAI)

1 Introduction

As AI and ML algorithms significantly impact our daily lives, understanding the abilities of these algorithms, as well as their biases and flaws, is important. One solution to this need for AI literacy in the AI research community is explainable AI (XAI) [6,5] as it potentially increases AI literacy of current AI users through informal settings, rather than relying on formal educational contexts. In this article, we show that non-learning experts are already building AI explanations, often called *explainables*, and we ask how learning sciences can be applied to XAI, revealing ample opportunity for AI in Education researchers.

To do this, we created surveys customized to XAI design practitioners' products and analyzed responses and interviews for themes along goals designers hope their explainables enable users to achieve. Our findings show that XAI design practitioners struggle in evaluating the success of their artifacts and that using a learning objective framework is helpful scaffolding. Our investigation of cognitive learning objectives can be more widely applied and yield insight for designers creating complex interactive explanations.

1.1 Related Work

For the purposes of this article, we limit our scope to post-hoc XAI systems known as AI explainables [13]. These explainables often explain model predictions without specifying the underlying mechanisms by which they work [13].

Most explainables present as animated or interactive tutorials teaching users about a specific algorithm or application of an algorithm, such as those in Figure 1. To an AI in Education researcher, this interactive explanation may seemingly share much with intelligent tutoring systems. We expand upon this by examining how XAI practitioners design, compare, and evaluate their AI explanations, and providing a concrete framework for practitioner use.

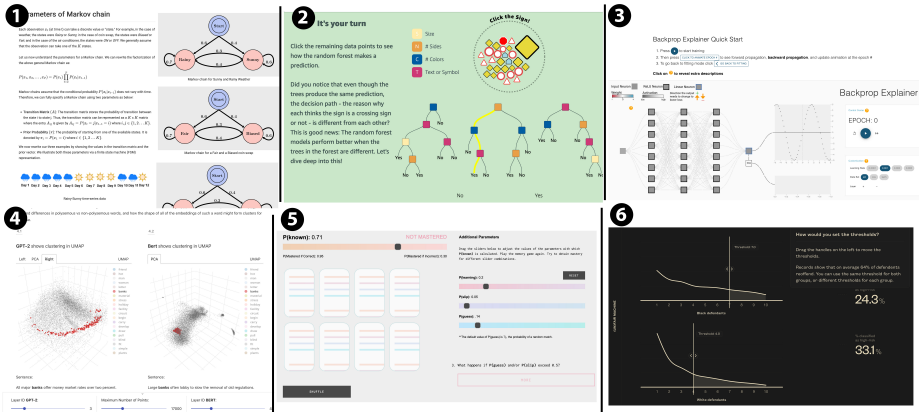

Fig. 1. Example screenshots from various XAI explainables: Exploring Hidden Markov Models (`https://nipunbatra.github.io/hmm/`), MLU-Explain The Random Forest Algorithm (`https://mlu-explain.github.io/random-forest/`), Backprop Explainer (`https://xnought.github.io/backprop-explainer/`), Demystifying the Embedding Space of Language Models (`https://bert-vs-gpt2.dbvis.de/`), Predicting What Students Know (`https://www.irishowley.com/res/bkt-esperanto/index.html`), (Un)Fair Machine (`https://unfair-machine.netlify.app/`).

The authors of [7] developed a model of the entire XAI process which includes relationships between the user, system, mental model, and task performance, with assessments for each, leading to appropriate trust and use. Considerations for measuring explanation effectiveness includes: user satisfaction, user mental model, user task performance, trust assessment, and (optionally) correctability. This XAI model includes several "tests" for evaluating XAI, which is expanded upon in the work in [12]. The authors created a literature-based taxonomy of XAI evaluation criteria which includes: faithfulness, completeness, stability, compactness, uncertainy communication, interactivity, translucence, comprehensibility, actionability, coherence, novelty, and personalization [12]. This more recent work fills in some specifics for the original XAI process model, illustrating how the evaluation criteria changes for specific XAI contexts, for specific user groups.

Much XAI research evaluating the goodness of explanations uses self-report or prediction tasks with questions like "What aspects of the news article contributed the most to this prediction?" [2]. However, examining the differences in human accuracy when assisted by decision-making systems with varying in-

terpretability allows researchers to draw conclusions about *how* interpretability impacts accuracy, but not *why* it does, which is where a learning science lens comes in. To understand XAI users' learning process, we must go beyond self-report questions to more accurate measures of understanding [15]. In order to do this, we must have specific, measurable goals against which to evaluate an XAI. Specificity in XAI goals or objectives enables finding the best strategy for a problem [3] and to better evaluate new approaches and designs. Within the learning sciences research, one common approach to creating measurable objectives is to adopt Bloom's Revised Taxonomy of Learning Objectives [3].

Similar questions of design intent and evaluation are explored in [1] where the authors propose a learning-based approach to better assess, and compare communicative visualizations [1]. The researchers analyzed participant-selected cognitive and non-cognitive learning objectives through surveys and interviews of visualization designers. Results suggested most designers' learning objectives were from the lowest cognitive level of "recall" and "fact" knowledge dimensions [1]. We adapted this approach for designer intent and evaluation of XAI systems.

Learning objectives describe what users will be able to do after the learning experience arranged into 3 domains: cognitive, affective, and psychomotor. [4]. The *cognitive* domain is knowledge-based learning, the *affective* domain is emotion-based learning, and the *psychomotor* domain is action-based learning [4]. Objectives within the cognitive domain combine a cognitive process (verb) with a piece of knowledge (noun). Bloom's Taxonomy provides an increasing-in-cognitive complexity set of categories, and sample verbs to define a cognitive objective for each category. Refer to the literature for the full listing [4]: **Remember**: recognize, recall. **Understand**: paraphrase, exemplify, categorize, generalize, extrapolate, compare, contrast, explain. **Apply**: execute, implement. **Analyze**: differentiate, organize, attribute. **Evaluate**: detect, critique. **Create**: hypothesize, plan, produce. Bloom's Taxonomy's knowledge dimension is a range increasing in abstractness [4]: **Factual** knowledge of specific details, such as "Recall the form and parameters of a Markov chain" from Example 1 in Figure 1. **Conceptual** knowledge integrates multiple facts. **Procedural** knowledge is skills or heuristics. **Metacognitive** knowledge is thinking about thinking.

2 Survey & Interview Studies

We still must determine if actual designers' intents also fit into these dimensions of the cognitive objective framework. To do so, we compiled a list of explainable designers from the 2018-2021 presenters from the IEEE VISxAI Workshop on Visualization for AI Explainability to participate in our IRB-approved study. According to the 5th Workshop on Visualization for AI Explainability (VISxAI) [1]: "*The goal of this workshop is to initiate a call for "explainables" / "explorables" that explain how AI techniques work using visualization.*"

[1] https://visxai.io/

Survey Study. There were 35 still accessible explainable artifacts with an average 2.4 authors per explainable. We identified 81 unique authors, 73 of which had accessible contact information. 33 unique authors completed the survey with participant backgrounds split between industry (61.8%) and academia (38.2%).

We modeled our survey items on the process in [1]. The survey asked respondents for the main goals of their explainable, to check which of a sample of cognitive objectives they had for users of their explainable, and any other learning objectives they planned. For each artifact, we provided *suggested learning objectives* that reflected our best inference of the intents of the designers based upon their artifact. Participants could also create their own learning objectives, which we refer to as *participant-suggested learning objectives*, and all did so.

32/33 respondents chose at least one of our suggested objectives as something they hoped their users would be able to do after their explainable. Labeling objectives with cognitive processes was mostly clear as the verbs in Bloom's Taxonomy could be consulted. For the knowledge dimension we developed a coding manual defining factual, conceptual, procedural, and metacognitive knowledge and had two coders label random samples until achieving a Cohen's κ coefficient of 0.71, which is considered substantial for labeling [9].

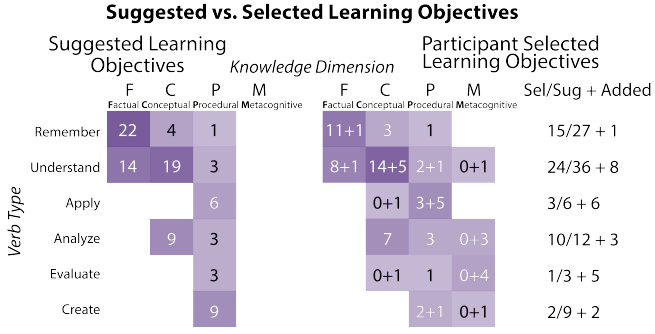

Fig. 2. Distribution of cognitive learning objectives in the surveys.

Figure 2 reflects data distribution of suggested, selected, and participant-created cognitive learning objectives. In surveys that received more than one response, we randomly selected one response to record in our data, resulting in 27 participant responses. In Figure 2, comparing the density map of our suggested learning objectives (left), to the density map of the participant selected and suggested objectives (right), we see that participants generally agreed with our distribution of suggested objectives, except in the case of the Metacognitive knowledge dimension. These goals are generally more difficult for an outside researcher to detect, as there may not be dedicated learning activities within the explainable for Metacognitive goals such as "Build upon our experiments and formulate their own hypotheses what this might be used for..."

The most frequent objectives are classified under the Remember process and the Factual knowledge dimension (22 items). This makes sense as both Remember and Facts are the lowest cognitively complex, and is also similar to results from related work [1]. Most participants agreed with our suggested cognitive objectives classified under the cognitive process, Analyze which was chosen 83.3%. Analyze was followed by Understand at 66.7%, Remember 55.6%, Apply 50%, Evaluate 33.3%, and Create 22.2%. Most of the suggested cognitive objectives were classified under the Understand verb type, though participants more often selected Analyze objectives. The most selected knowledge dimension objective was Conceptual at 75%, and then Factual at 52.8% and Procedural at 48%.

When suggesting new objectives, the distribution of targeted objectives differed. Out of the 25 participant-suggested cognitive objectives, most participants suggested Understand (8), followed by Apply (6), Evaluate (5), Analyze (3), Create (2) and then Remember (1). Out of the 4 knowledge dimensions, most participants suggested learning objectives classified as Procedural and Conceptual (7), followed by Metacognitive (9), and Factual (2). These results show that designers of explainables prioritize supporting their users in *doing*, not just *knowing*. The remaining objectives were categorized as 10 non-learning objectives, including 3 business goals ("Reach out to the team and ask for consulting services" [P52]), 1 affective learning objectives ("This article may ignite their interest in the topic and then they may explore..." [P41]), and 1 action-based psychomotor learning objective ("Do research on <algorithm>themselves" [P92]).

Interview Study. To gain further insight into the survey responses, we performed interviews with participants who agreed to do so. 9 explainable authors agreed to a 30-45 minute semi-structured interview. The structure of our interviews was heavily influenced by prior work focusing the interview on design process, design considerations, and audience [1]. To analyze our interview data, we followed a phenomological process applying guidelines from [8]. After performing phenomenological reduction, we coded them using an iterative process with structural coding, in vivo coding, and open coding methods [14]. Our initial pass through the data was done using in vivo coding, through which we created 132 unique codes. A subsequent pass was done using open coding to further elucidate themes present throughout the data and narrow down these 119 unique codes into 5 broader thematic categories encompassing the most common codes: structure planning, play/exploration, audience, accessibility, and future goals. Finally, we used structural coding to target our research questions and to analyze the data on demographics and past experience of the interviewee. We used 6 structural codes: prior experience in XAI, teaching experience, user goals, use of learning science, intended audience, and core concepts.

All 9 participants stated that learning goals played a role in their design process, commonly working backwards from these goals. Participants used words like "*building*" to envision goals in the early stages of their process, where they create "*a stack of concepts*" (P70) that "*progressively gets more and more complicated step by step until we can get to where we actually want it to be*" (P47).

Several participants noted that they had not initially considered their goals as learning objectives. While one stated that they "*didn't write them down as learning goals*" at the time, but identifying the "*broad points I wanted people to walk away with*" (P79) was not only a key part of their planning process, but also one that they would consider to be making use of learning objectives despite not using that exact term previously. One participant proposed a solution: hiring a collaborator with a background in education, expressing "*if we now also had a set of only a publicity person, we would also have an educator person.*" (P37).

A prominent theme through the interviews was learning objectives' compatability with the explorative nature of explainables. When asked whether or not they thought that further use of learning objectives would be helpful, they clarified that they "*stand by the idea of intentionally leaving it a little bit open, in terms of allowing the user to make their own conclusion*" (P9). However, these open-ended goals could be specified with higher cognitively complex learning objectives. The importance of exploration in explainables was mentioned by 7 of 9 of our participants, desiring that the user "*develop or extract a non-trivial insight playing with the material that is presented*" (P70). One participant stated having the user "*really to get a deeper sense of understanding for these individual parts by playing around with them*" (P37) was their main goal. To reckon with this, several participants balanced objectives and open-ended discovery. One designer described this balance as a "*back and forth between allowing the user to sort of interact and learn things themselves. And me sort of explaining what there is to do with it, rather than hand-holding them the entire way*" (P47).

As in the survey analysis, we also categorized each objective mentioned into the cognitive, affective, and psychomotor domains in Bloom's Revised Taxonomy [4], identifying 27 cognitive learning objectives, and 2 affective objectives. Out of these 27 objectives, the most suggested verb type was Understand (12 times by 10 unique participants). A common subtype of the Understand objective was a Summarize objective (3 times). Participants described this particular objective as being able to relay the gist of a topic "*if he's asked about [the topic] at a party*" (P103). Understand was followed by Remember (7 times), Create (3 times), Analyze (3 times), and Apply & Evaluate (1 time).

Beyond the learning objectives identified from the interviews, we observed additional non-learning objectives designers named as goals during their processes. These objectives included 1 business objective (to "*get a lot of attention*" (P80) with the result of their work), as well as desires to achieve replication of the designer's own past learning experience, and a design goal of accessibility both in terms of level of prior knowledge and of mobile/low-speed internet compatibility.

XAI designers focused on both learning and non-learning measures when evaluating their explainables. A common metric among explainables (mentioned by 6 participants) was the praise and public response generated upon release or publication. Many participants took this to be a reliable indicator of success, stating that "*other people have cited it and other people have looked at it and other people have messaged me about it...So obviously it must provide some value*" (P47). Multiple participants clarified that this metric was used due to the

absence of other more formal measures: "*We're not sure about how to measure the success of these, but people like them on Twitter and that seems good*" (P79). Additional evaluation strengths authors identified included meeting their defined user goals and serving as a learning experience for the author themselves. Participants also identified 6 areas in which they felt their work fell short, including: using overly complex material, needing more time, not meeting all goals, not enough interactive components, and lacking technical robustness.

3 Conclusion

Learning objectives should lead to better XAI by providing a means to compare XAI design decisions and more rigorously evaluate the success of the XAI systems [1]. The contributions of this article to the AIED community include: investigation of the current state of informal AI literacy support, providing a broader perspective of designers' explainable design process, identifying that XAI designers struggle assessing their artifacts, exemplifying the use of learning objectives for designing XAI as helpful scaffolding, and discovering support of metacognitive goals as an important focus of explainable designer intent.

One of the most striking results is how a large number of metacognitive goals were suggested. Goals like "I hope that users will more effectively question what different algorithms they interact with are doing" (P44) are often more difficult for an outsider to infer, but these metacognitive goals are a critical piece of the rhetoric around the need for interpretable AI [13] and AI literacy. Therefore, focusing on the development and assessment of these metacognitive goals is a top priority and a fruitful avenue for AIED researchers to contribute to XAI.

Beyond cognitive objectives, many of the participant-suggested learning objectives from the survey included phrases like: "*...ignite their interest...*", "*They might want to refer to our article whenever they are dealing with <concept>*", or "*Reach out to the team and ask for consulting services*". These affective, reference, and business objectives (respectively) cannot be encapsulated by cognitive learning objectives, and remain important goals of XAI. The revised Bloom's Taxonomy does not include an affective dimension, and there does not appear to yet be a widely accepted holistic model that does so [4]. These affective goals are also missing from question banks for XAI, such as [11]. Communicative visualization researchers are developing affective objectives for their domain [10], and it is possible some of the affective verbs and nouns from this work could be adapted to XAI contexts. Reference objectives reflect longer term goals and also cannot be well encapsulated by XAI question banks, similar to affective objectives. Business goals made a small appearance in our results through mentions of measuring success through public use, and are also an important consideration for designers building AI explainables to increase AI literacy in informal settings.

Our studies show XAI designers are unsure how to evaluate their artifacts, most relying on their explainables' popularity on social media as a proxy for success. There is a very clear gap in how practitioners build and evaluate their post-hoc XAI, and how the research community views evaluation of post-hoc XAI.

Our work opens many research opportunities for the AIED community, including: supporting XAI practitioners in developing metacognitive goals, creating an affective & reference objectives taxonomy, and investigating the relationship between exploration, spontaneity, and prescriptive XAI learning goals.

This article provides insights into how designers create XAI, focusing on their intention to change the viewer through increased AI literacy. Presenting the information in an explainable, counter-factual, or even transparent algorithm is not enough to ensure that the user correctly understands the AI model. Without that understanding, it is difficult to achieve the goals of trust and fairness for which XAI designers aim. By framing XAI goals as learning objectives, designers can evaluate whether their design was successful. In surveying and interviewing XAI designers, we demonstrated that their goals and intentions can be mapped to learning objectives. In doing so, we also discovered additional dimensions that could be added to the framework to more holistically design and evaluate XAI.

References

1. Adar, E., Lee, E.: Communicative visualizations as a learning problem. IEEE Transactions on Visualization and Computer Graphics **27**(2), 946–956 (2020)
2. Alvarez-Melis, D., Kaur, H., Daumé III, H., Wallach, H., Vaughan, J.W.: From human explanation to model interpretability: A framework based on weight of evidence. In: Proc of the AAAI Conf on HCOMP). vol. 8, p. 3. AAAI, USA (2021)
3. Ambrose, S., Bridges, M., DiPietro, M., Lovett, M., Norman, M.K.: How learning works: 7 research-based principles for smart teaching. John Wiley & Sons (2010)
4. Anderson, L.W., Krathwohl, D.R. (eds.): A Taxonomy for Learning, Teaching, and Assessing. A Revision of Bloom's Taxonomy of Educational Objectives. Allyn & Bacon, New York, NY USA (December 2001)
5. Anik, A.I., Bunt, A.: Data-centric explanations: Explaining training data of ml systems to promote transparency. In: Proc of the CHI Conf. pp. 1–13. ACM (2021)
6. Doshi-Velez, F., Kim, B.: Towards a rigorous science of interpretable machine learning (2017), `https://arxiv.org/abs/1702.08608`
7. Gunning, D., Aha, D.: Darpa's explainable artificial intelligence (xai) program. AI Magazine **40**(2), 44–58 (2019)
8. Hycner, R.H.: Some guidelines for the phenomenological analysis of interview data. Human studies **8**(3), 279–303 (1985)
9. Landis, J.R., Koch, G.G.: The measurement of observer agreement for categorical data. Biometrics **33**(1), 159–174 (1977)
10. Lee-Robbins, E., Adar, E.: Affective learning objectives for communicative visualizations. IEEE Trans. on Visualization & Computer Graphics **29**(1), 1–11 (2023)
11. Liao, Q.V., Gruen, D., Miller, S.: Questioning the ai: informing design practices for explainable ai user experiences. In: Proc of the CHI Conf. pp. 1–15. ACM (2020)
12. Liao, Q.V., Zhang, Y., Luss, R., Doshi-Velez, F., Dhurandhar, A.: Connecting algorithmic research and usage contexts. In: Proc of the AAAI Conference on Human Computation and Crowdsourcing. vol. 10, pp. 147–159. AAAI, USA (2022)
13. Lipton, Z.C.: The mythos of model interpretability. Queue **16**(3), 31–57 (2018)
14. Saldaña, J.: The coding manual for qualitative researchers. sage, USA (2021)
15. Sitzmann, T., Ely, K., Brown, K.G., Bauer, K.N.: Self-assessment of knowledge: A cognitive learning or affective measure? Academy of Management Learning & Education **9**(2), 169–191 (2010)