Towards a Learner-Centered Explainable AI

Anna Kawakami[1]
akawakam@andrew.cmu.edu

Nikos Arechiga[2]
nikos.arechiga@tri.global

Luke Guerdan[1]
lguerdan@andrew.cmu.edu

Matthew Lee[2]
matt.lee@tri.global

Yang Cheng[1]
yanghuic@andrew.cmu.edu

Scott Carter[2]
scott.carter@tri.global

Anita Sun[1]
ningjins@andrew.cmu.edu

Haiyi Zhu[1*]
haiyiz@andrew.cmu.edu

Alison Hu[1]
ayhu@andrew.cmu.edu

Kenneth Holstein[1*]
kjholste@cs.cmu.edu

Kate Glazko[2]
kate.glazko.ctr@tri.global

1 Carnegie Mellon University
2 Toyota Research Institute
* Denotes equal contribution

CHI'22,, April 30–May 6, 2022, New Orleans, LA, USA
ACM 978-1-4503-6819-3/20/04.
https://doi.org/10.1145/3334480.XXXXXXX

Abstract

In this short paper, we argue for a refocusing of XAI around *human learning goals*. Drawing upon approaches and theories from the learning sciences, we propose a framework for the learner-centered design and evaluation of XAI systems. We illustrate our framework through an ongoing case study in the context of AI-augmented social work.

Introduction

Recent years have seen a surge of interest in the question of how AI systems can be made more "interpretable" or "explainable" to humans. Yet these terms are used in reference to many disparate goals within the literature [10, 17, 19]. For instance, work on interpretability has sometimes focused on enhancing humans' ability to mentally simulate and predict an AI system's behavior [16, 17, 22] or to evaluate counterfactuals [27]. Other work addresses ways to help humans decompose models, to understand their constituent parts (e.g., parameters) and how these parts fit together [17]. From a human-centered perspective, these design goals can be understood as supporting different **human capabilities**, each of which may be more or less useful in different real-world contexts. For example, decomposing a model may be useful when debugging an AI system. In a decision-making context, the ability to identify situations that could impact a model's reliability may be more helpful [11, 20].

In this paper, we argue that many, if not all, of the design goals in existing XAI research and practice can be produc-

tively reinterpreted as human **learning goals**. Much current XAI research focuses on designing ways to *make models explainable to humans*. By contrast, building upon recent arguments for centering human *understanding* in XAI research [19, 26], we focus on supporting humans in *learning* about particular AI systems and how to work with or around them. Whereas XAI research often aims at communicating information about an AI system instantaneously and with minimal effort on the part of a human recipient, some learning goals may best be met through longer learning engagements or through deliberate practice and feedback [3, 4, 15, 20].

Drawing lessons from the learning sciences—a scientific and design discipline dedicated to the study of human learning and ways to support it in real-world contexts—we explore the implications of adopting a learning-centered lens for the design and evaluation of human-centered XAI. We propose a framework for learner-centered XAI, which integrates and extends existing concepts from the learning sciences. Finally, we present an ongoing case study illustrating how this framework can be applied in practice.

A framework for learner-centered XAI

In this section, we propose a framework for the learner-centered design and evaluation of XAI. We describe how three concepts from the learning sciences—backward design [28], participatory design for learning [9], and "closing the loop" [7, 18]—can help to guide the design of XAI that positions humans as deliberate and continuous learners. The goals of this framework are to (1) offer a systematic process for designing XAI interfaces that target specific **learning outcomes**, (2) demonstrate how **context- and stakeholder-specific needs** can be surfaced and addressed during the design process, (3) combine **participatory and data-driven methods** to support more contextually-relevant XAI designs, and (4) provide a more rigorous approach for **evaluating the effectiveness** of XAI.

As shown in Figure 1, our framework proposes that researchers should collaborate with relevant stakeholders in real-world human-AI interaction contexts, to iteratively co-design learning objectives, measures, activities, and evaluation approaches. Following a "backward design" approach, as discussed below, this collaboration should begin by specifying **learning objectives**: a set of specific capabilities that the learners should ideally have following a learning activity. Learners should then be involved in decisions about how to operationalize these learning objectives in the form of concrete **learning measures** which capture observable human behaviors as proxies for latent constructs such as "understanding" of a targeted concept [12]. For instance, researchers might engage learners in specifying *how they would know* whether a given intervention had succeeded in meeting one of their learning objectives: how would they behave differently, or what would they be able to do that they could not do previously? With these objectives and measures in mind, researchers can work with learners to co-design **learning activities** to try to help learners achieve their specified objectives. The measures specified previously can then be used to **evaluate learning outcomes**, to guide the iterative, data-driven refinement of learning activities. Below, we introduce three concepts from the learning sciences that inform this framework.

Wiggins and McTighe proposed **backward design** to address a longstanding challenge in instructional design: teachers and instructional designers often focus more on how to *teach* rather than on how to help students *learn* [2, 28]. Backward design is an approach that 'flips' the design process. Rather than starting with the design of instructional materials, designers are encouraged to first identify

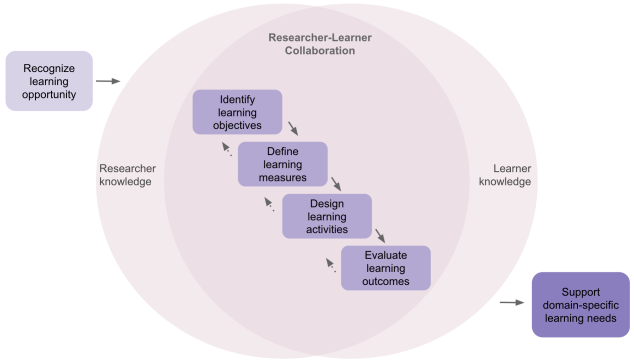

Figure 1: A framework for the design of learner-centered XAI.

desired learning *outcomes*, then to design *assessments* of those outcomes, and lastly to design *instruction* aimed at achieving those outcomes. These challenges in instructional design are echoed in the current XAI landscape: Even as research moves towards more human-centered XAI methods, it remains common to first propose an explainability technique, and then evaluate whether and how the technique is useful to users. In the learner-centered XAI framework, we propose a backward design process that **starts by identifying meaningful learning objectives for a given task context**, then operationalizes what it means to meet those learning objectives. Only after designing and operationalizing learning objectives that reflect stakeholder- and domain-specific needs are XAI designers prepared to design interfaces that meet these learning objectives.

The framework additionally draws from **participatory design practices** in the learning sciences. From a learning sciences perspective, participatory design is recast as an opportunity for relevant stakeholders and researchers to collaboratively learn new knowledge that can guide the design process, based on each others' complementary expertise [9]. Stakeholders with relevant lived experience are uniquely positioned to understand their own needs and desires. Meanwhile, researchers can bring unique scientific, design, and technical expertise that is critical to designing effective learning interventions. Moreover, as researchers and stakeholders' joint understanding of the problem space strengthens, the framework's emphasis on an **iterative design process** may encourage them to proactively reflect on their prior design decisions and refine them as needed. Empowering stakeholder participation earlier on in the design process, at the "defining learning objectives" stage, not just when evaluating the interfaces, may also open opportunities for different stakeholders in a given context to discuss any misalignments in their envisioned learning needs.

Finally, Figure 1 indicates that real-world evaluations of XAI techniques should inform the continuous process of iterative re-design. This aligns with the notion of **"closing the loop"** in the learning sciences, emphasizing the data-driven refinement of instructional materials based on analysis of data reflecting how people actually learn with them [7, 18]. This approach offers an opportunity to rigorously evaluate and iterate on co-designed learning objectives, measures, and interfaces, to address design misalignments, or to adapt to changes in stakeholder needs over time.

Case study: Using the framework to design training interfaces for AI-augmented social work

In this section, we illustrate how the learner-centered XAI framework can be used in practice, through an ongoing case study in the context of AI-augmented social work.

Background

In an effort to augment social workers' abilities to efficiently process and prioritize among large volumes of child mal-

treatment referrals, child welfare agencies have begun to turn to new machine learning-based ADS tools [23, 6, 24, 29]. The Allegheny Family Screening Tool (AFST) has been in use in Allegheny County, Pennsylvania since 2016, where it assists child maltreatment hotline call screeners and supervisors in prioritizing among referred cases [21]. While the county has published public-facing reports discussing the ethics and validity of using such a tool [8], recent research raises new concerns around how effectively the tool has been integrated into the organizational and social context in which workers make day-to-day use of the tool. In particular, in a recent paper, we report findings from a series of interviews and contextual inquiries at this child welfare agency, to understand how workers currently make AI-assisted child maltreatment screening decisions. We found that workers had little to no opportunities to learn about the AI system they were using, nor about how to work with it effectively, limiting their ability to appropriately calibrate their reliance on the tool's predictions [13]. Moreover, we found that workers' decision-making objectives (focusing on short term risks to child safety) differed from the model's predictive targets (focused on *much longer-term predictions* of indirect *proxies* of risk). While the tool was intentionally designed to complement workers' focus on immediate outcomes, workers were unsure *how* exactly they were meant to integrate the tool's predictions of long-term risk with their own assessments of immediate safety.

Overall, these prior findings suggest a need to more broadly reconsider and reconceptualize what appropriate roles for ADS in social work might look like. This reconceptualization necessitates, at minimum, finding ways to understand, empower, and integrate worker perspectives in the design of ADS. As a first step towards this vision, we are currently exploring ways to address the gap between the current design of the AFST and workers' beliefs regarding what effective human-AI decision-making should look like, and how it should be measured. In this ongoing work, we engage workers in the design of training materials, as a means to identify and design worker-centered learning objectives, measures, and learning activities.

In this project, we do not plan to fully develop or deploy training materials for the AFST specifically. Indeed, based on our findings thus far, we expect that this co-design process will surface needs for fundamentally different kinds of ADS, not just building training interactions around the existing ADS. Rather, we view the AFST context as a rare opportunity to understand workers' learning goals and needs for support in a highly complex, social decision-making context where an ADS has already been in-use for many years (over half a decade). Beyond this context, we plan to explore the generalizability of our findings (e.g., regarding workers' learning goals) to other AI-augmented social decision-making contexts, such as AI-augmented content moderation.

Ongoing case study

Following the first step of the learner-centered XAI framework, we first **defined a set of fine-grained learning objectives**, such as "the ability to identify cases where a model may be more or less reliable", based on our design research with workers (see Appendix[1] for details). We plan to further explore, refine, and operationalize these learning objectives in collaboration with various stakeholders (e.g., workers, community members, and agency leadership).

Figure 2 shows two examples of initial training interface sketches that could address specific learning objectives in our taxonomy. In the first example, the learning objective is to improve workers' ability to appropriately rely on

[1] https://sites.google.com/andrew.cmu.edu/learner-centered-xai/home

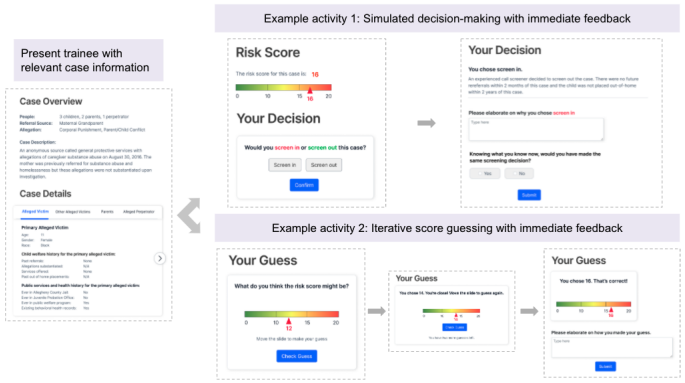

Figure 2: Example interfaces targeting different learning goals.

ADS outputs in specific cases. The sketch shows a simulated decision-making activity, which provides low-stakes opportunities for workers to practice integrating their own judgments with AI predictions on real historical data while receiving immediate feedback [1, 11, 14, 25]. The second example sketch focuses on honing workers' ability to mentally simulate the model's behavior through repeated practice opportunities on a score guessing exercise, with immediate feedback on the closeness of their guesses.

As next steps, we plan to **iteratively refine the learning objectives, measures, and activities** through co-design activities with social workers who use this ADS in their daily work, along with other stakeholder groups. Taking a **participatory design for learning** approach, we view the co-design of learning objectives and measures as an opportunity to surface and address value tensions across different stakeholder groups, regarding what human-AI decision-making in this context should look like in the first place [5, 13]. For example, while current worker-ADS decision-making performance measures are based on the ADS's predictive target, this assumes the workers should then learn to act like the system would. Our framework aims to involve workers in the design of improved learning measures, to offer alternative measures that counter these assumptions and align more closely with workers' own decision objectives or a mixture of workers' decision objectives and the systems' objectives, if that is believed to be desirable.

Open Questions

At the workshop, we hope to further explore several open questions. For example: How might learning objectives vary across different human-AI tasks (e.g., prediction, decision-making, or co-creation)? What are other implications of approaching XAI through a learning-centric lens?

REFERENCES

[1] Abdulmohsen H Al-Elq. 2010. Simulation-based medical teaching and learning. *Journal of Family and Community Medicine* 17, 1 (2010), 35.

[2] Vincent Aleven and Kenneth R Koedinger. 2013. Knowledge component (KC) approaches to learner modeling. *Design Recommendations for Intelligent Tutoring Systems* 1 (2013), 165–182.

[3] Gagan Bansal, Besmira Nushi, Ece Kamar, Walter S Lasecki, Daniel S Weld, and Eric Horvitz. 2019. Beyond accuracy: The role of mental models in human-AI team performance. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, Vol. 7. 2–11.

[4] Carrie J Cai, Samantha Winter, David Steiner, Lauren Wilcox, and Michael Terry. 2019. " Hello AI": Uncovering the onboarding needs of medical practitioners for human-AI collaborative

decision-making. *Proceedings of the ACM on Human-computer Interaction* 3, CSCW (2019), 1–24.

[5] Carrie J Cai, Samantha Winter, David Steiner, Lauren Wilcox, and Michael Terry. 2021. Onboarding Materials as Cross-functional Boundary Objects for Developing AI Assistants. In *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–7.

[6] Alexandra Chouldechova, Emily Putnam-Hornstein, Suzanne Dworak-Peck, Diana Benavides-Prado, Oleksandr Fialko, Rhema Vaithianathan, Sorelle A Friedler, and Christo Wilson. 2018. A case study of algorithm-assisted decision making in child maltreatment hotline screening decisions. *Proceedings of Machine Learning Research* 81 (2018), 1–15. http://proceedings.mlr.press/v81/chouldechova18a.html

[7] Doug Clow. 2012. The learning analytics cycle: closing the loop effectively. In *Proceedings of the 2nd International Conference on Learning Analytics and Knowledge*. 134–138.

[8] Allegheny County. 2017. *Ethical Analysis: Predictive Risk Models at Call Screening for Allegheny County*. Technical Report. https://www.alleghenycountyanalytics.us/wp-content/uploads/2019/05/Ethical-Analysis-16-ACDHS-26_PredictiveRisk_Package_050119_FINAL-2.pdf

[9] Betsy DiSalvo, Jason Yip, Elizabeth Bonsignore, and Carl DiSalvo. 2017. *Participatory Design for Learning*. Taylor & Francis.

[10] Finale Doshi-Velez and Been Kim. 2017. Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608* (2017).

[11] Kenneth Holstein, Vincent Aleven, and Nikol Rummel. 2020. A conceptual framework for human–AI hybrid adaptivity in education. In *International Conference on Artificial Intelligence in Education*. Springer, 240–254.

[12] Abigail Z Jacobs and Hanna Wallach. 2021. Measurement and fairness. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. 375–385.

[13] Anna Kawakami, Venkat Sivaraman, Hao-Fei Cheng, Logan Stapleton, Yang Cheng, Diana Qing, Adam Perer, Steven Wu, Zhu Haiyi, and Kenneth Holstein. 2022. Improving human-AI partnerships in child welfare: Understanding worker practices, challenges, and desires for algorithmic decision support. In *Proceedings of the ACM CHI Conference on Human Factors in Computing Systems (CHI'22).*

[14] Kenneth R Koedinger, Albert T Corbett, and Charles Perfetti. 2012. The Knowledge-Learning-Instruction framework: Bridging the science-practice chasm to enhance robust student learning. *Cognitive Science* 36, 5 (2012), 757–798.

[15] Vivian Lai, Han Liu, and Chenhao Tan. 2020. " Why is 'Chicago' deceptive?" towards building model-driven tutorials for humans. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–13.

[16] Vivian Lai and Chenhao Tan. 2019. On human predictions with explanations and predictions of machine learning models: A case study on deception detection. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*. 29–38.

[17] Zachary C Lipton. 2018. The Mythos of Model Interpretability: In machine learning, the concept of interpretability is both important and slippery. *Queue* 16, 3 (2018), 31–57.

[18] Christopher J Maclellan, Erik Harpstead, Rony Patel, and Kenneth R Koedinger. 2016. The Apprentice Learner Architecture: Closing the loop between learning theory and educational data. *International Educational Data Mining Society* (2016).

[19] Tim Miller. 2019. Explanation in artificial intelligence: Insights from the social sciences. *Artificial intelligence* 267 (2019), 1–38.

[20] Hussein Mozannar, Arvind Satyanarayan, and David Sontag. 2021. Teaching humans when To defer to a classifier via examplars. *arXiv preprint arXiv:2111.11297* (2021).

[21] Allegheny County Department of Human Services. 2018. Allegheny Family Screening Tool, Frequently-Asked Questions. `https://www.alleghenycountyanalytics.us/wp-content/uploads/2018/10/17-ACDHS-11_AFST_102518.pdf.` (2018).

[22] Forough Poursabzi-Sangdeh, Daniel G Goldstein, Jake M Hofman, Jennifer Wortman Wortman Vaughan, and Hanna Wallach. 2021. Manipulating and measuring model interpretability. *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (2021), 1–52.

[23] Anjana Samant, Aaron Horowitz, Kath Xu, and Sophie Beiers. 2021. Family surveillance by algorithm: The rapidly spreading tools few have heard of. *American Civil Liberties Union (ACLU)* (2021). `https://www.aclu.org/sites/default/files/field_document/2021.09.28a_family_surveillance_by_algorithm.pdf`

[24] Devansh Saxena, Karla Badillo-Urquiola, Pamela J Wisniewski, and Shion Guha. 2020. A human-centered review of algorithms used within the US child welfare system. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–15.

[25] Randolph H Steadman, Wendy C Coates, Yue Ming Huang, Rima Matevosian, Baxter R Larmon, Lynne McCullough, and Danit Ariel. 2006. Simulation-based training is superior to problem-based learning for the acquisition of critical assessment and management skills. *Critical care medicine* 34, 1 (2006), 151–157.

[26] Jennifer Wortman Vaughan and Hanna Wallach. 2020. A human-centered agenda for intelligible machine learning. *Machines We Trust: Getting Along with Artificial Intelligence* (2020).

[27] Daniel S Weld and Gagan Bansal. 2019. The challenge of crafting intelligible intelligence. *Commun. ACM* 62, 6 (2019), 70–79.

[28] Grant P Wiggins and Jay McTighe. 2005. *Understanding by design*. Ascd.

[29] Alexandra Zytek, Dongyu Liu, Rhema Vaithianathan, and Kalyan Veeramachaneni. 2021. Sibyl: Understanding and addressing the usability challenges of machine learning in high-stakes decision making. *IEEE Transactions on Visualization and Computer Graphics* (2021).