Towards Guidelines for Designing Human-in-the-Loop Machine Training Interfaces

Almar van der Stappen

Mathias Funk

almarvanderstappen@gmail.com

m.funk@tue.nl

Eindhoven University of Technology

Eindhoven, The Netherlands

ABSTRACT

Supervised machine learning approaches commonly require good availability and quality of training data. In applications that depend on human-labeled data, especially from experts, or that depend on contextual knowledge for training data sets, the human-in-the-loop presents a serious bottleneck to the scalability of training efforts. Even if human labeling is generally feasible, sustained human performance and high-quality labels in larger quantities are challenging. Interactive Machine Learning can help solve usability problems in traditional machine learning by giving users agency in deciding how systems learn from data. Yet, the field lacks clear design guidelines for such interfaces, specifically regarding the scaling of training processes. In this paper, we present results from a pilot study in which participants interacted with several interface variants of a recommender engine and evaluated them on interaction and efficiency parameters. Based on the performance of these different learning system implementations we propose design guidelines for the design of such systems and a score for comparative evaluation, in which we combine interaction experience and system learning efficiency into one relative scoring unit.

CCS CONCEPTS

• **Human-centered computing** → **Heuristic evaluations**; *Usability testing*; *User interface design*.

KEYWORDS

Machine Learning, Labeling Data, Algorithmic Training, Interaction, User Interface, UX Metric

ACM Reference Format:

Almar van der Stappen and Mathias Funk. 2021. Towards Guidelines for Designing Human-in-the-Loop Machine Training Interfaces. In *26th International Conference on Intelligent User Interfaces (IUI '21)*, April 14–17, 2021, College Station, TX, USA. ACM, New York, NY, USA, 6 pages. <https://doi.org/10.1145/3397481.3450668>

This work is licensed under a Creative Commons Attribution International 4.0 License. *IUI '21*, April 14–17, 2021, College Station, TX, USA
© 2021 Copyright held by the owner/author(s).
ACM ISBN 978-1-4503-8017-1/21/04.
<https://doi.org/10.1145/3397481.3450668>

1 INTRODUCTION

In recent years, we have become accustomed to machine learning (ML) integrated into products and services that we use on a daily basis. Increasingly, ML concepts influence the design of interfaces and how we interact with “smart” products and ML-supported services. End-user recommendations for e-commerce and media consumption are examples of providing personalized content [17] based on collaborative filtering and clustering. Such applications of machine learning require good availability and quality of training data. Especially when considering highly subjective and diverse user preferences and choices, training with large data sets is difficult and prone to introducing training bias. This process is often not transparent and cannot easily be debugged [14]. Correspondingly, the result can be irrelevant, biased or even offensive recommendations for end-users [16], because the learned “profile” over time does not account for users making mistakes, carrying out actions that are not actually indicative of regular behavior or producing the right results for the wrong reasons [12, 15, 19]. Instead of looking at traditional machine learning and increasingly larger data sets to solve these problems, we can also use interactive machine learning (IML) with a “human in the loop” [7, 9], who informs the training process through repeated interactions with the ML system. This can lead to better results and more perceived agency when, for instance, the user can convey to the learning system how they want it to interpret user data, if at all [10, 22].

Designers of such systems need to balance additional user agency in the training process with longer training times, user fatigue, and perception problems that might impact the quality of training data. This link between training quality and end-user workload has not been thoroughly explored from the viewpoint of the user training the system. In applications that depend on human-labeled data, especially from experts, or that depend on contextual knowledge for training data sets, the human-in-the-loop presents a serious bottleneck to the scalability of training efforts. This might be one of the reasons why human-in-the-loop machine training approaches are costly and seldom deployed in industry [21]. Even if human labeling can be facilitated, prolonged performance and the provision of high-quality labels in required quantities are problematic [3]. In this study, we explore the relationship between user experience and learning process agency for training by letting participants interact with four different interface variants of a recommender system. The task is to teach the recommender system to direct them towards a specific item in the database. After this task, participants review the four systems on both interaction speed and user experience. After

presenting the results, we end the paper with an exploration of a relative scoring unit for system training experience and conclude with an outlook to future work on design guidelines.

2 RELATED WORK

Fails and Olsen [7] introduce interactive machine learning as a method to let users provide feedback improving a learner incrementally after evaluating its step-by-step generated results. As such, the users go through a fast-paced feedback cycle allowing for rapid ‘prototyping’ of the system. Yet, understanding and designing for the user’s involvement [22] in such interactions with a learning system requires more research. Amershi et al. [1] note that the explicit study of user interactions can challenge the assumptions made in traditional learning systems about their users and possibly better inform the design of interactive learning systems in the future. Consequently, the authors challenge the HCI community to look more deeply at interactions, and they stress the need for general design or interaction standards to appropriately evaluate interactive systems [1, 6]. Human-AI interaction remains to this day a tough design challenge, because of the large scope of unpredictable system behaviors it introduces, providing difficulty for both those training and designing the system [23]. In a 2015 study by Kulesza et al. [13], users were confronted with the practice of explanatory debugging. The participants received an explanation about where system results came from, which was followed by a substantial increase in their understanding as well as their ability to garner results. Kulesza et al. [15] already concluded in a prior study that one of the parameters that users find frustrating with learning systems is the lack of transparency, one of the main problems with engineering of intelligent algorithms in general [24]. The 2015 study supports that, while also showing that improved results can already be evident after limited training cycles. Interactive learning models need to work with a higher learning rate than traditional models, especially when there is a high number of features [7]. With interactive machine learning, we have to consider not only the rate at which the learning model operates but also the ‘human processing’ delay for every iteration. A second issue with interactive learning techniques is the trade-off between maintaining proper control, user satisfaction and getting results at a reasonable pace [11]. Labeling data, frequently a human task, is usually simple and easily understandable, but nonetheless time-consuming [5]. As a general principle, when presented with more choices, more effort is required from the human as they have to evaluate more data before they provide input. Fatigue is a second limiting factor to extended human labeling of data sets. Third, Stumpf et al. [22] report that the practice of labeling might be limited in its richness of interaction, thus negatively impacting the overall experience of training users. These three factors, fatigue by choice and task repetition, and limited natural expressivity, represent key challenges for the design of ML training interfaces. That is, to design future machine learning interactions that are sustainable due to a positive overall experience we need to carefully balance human effort and the quality of data that is required for the system to operate. To support the bringing forth of such designs, this study looks to explore the influence of user agency per training interaction on producing a satisfying, yet efficient experience overall.

3 METHOD

3.1 Design

A within-subject design was used to evaluate and compare different implementations of human-in-the-loop-learning interactions. Each participant interacted with the four different learning methods provided in a randomized order. Although not being strong enough to assuage it, this is meant to counteract a general learning effect as much as possible. The interactive learning system use case utilized to test the different interactions was that of a simple furniture recommender system. Participants have completed their task when they manage to train a system to recommend them a predefined item, a red chair, with no further specified style or attributes. Out of the data set of 40 furniture items, 2 items matched these criteria specifically. The features that were included differed per method, however at most it incorporated four features per item: type of item (e.g., chair, couch), main color, secondary color, and style (e.g., modern, classical). The test system was implemented as an interactive Processing sketch that is shared in an open-source repository¹.

3.2 Interfaces

Participants were asked to interact with four interface variants that provide different ways to train a recommender system. Each of them shows different aspects of data per interaction by affording a unique set of actions to the user.

The four different systems operate as follows: System 1 is a traditional method of machine learning, shown as a binary labeling task. Users are shown a furniture image and they classify the input either as “yes” or “no”. They train the recommender system towards the desired item as described in the previous section by liking products with similar features and disliking others. System 2, another binary classifier, has an additional component that allows the users to switch in and out of influencing certain feature classifiers. For example, if they were to turn off the color classifier, the next output would not change their predictions regarding the colors of the furniture shown in the previous output. However, if a classifier is left open, the system is free to change the represented feature in its next input to label. System 3 is a comparative learning system. For every output, this system shows three different items to the user, out of which they can then pick the item closest to the target product. System 4 is similar to system 1, except for the fact that every decision is met with a question. Whenever the user issues their choice, the system responds by asking them about what their motivation is. If a user would down-vote a green chair, the system would ask: “Did you dislike this item because it had the color green?”

3.3 Procedure and participants

Every participant was asked to go through a set of interactions after which they filled out a questionnaire. For each of these sessions, participants spent about 15 minutes in the test environment. They executed the task with the added information that the recommender systems featured chairs and couches as their main item pools. The procedure was explained as follows: (1) Launch program X (chosen at random). (2) Consider the item given by the system. (3) React

¹<https://github.com/Almarvds/Interactive-Machine-Learning-Methodologies>

Figure 1: Interface variants 1 and 2 (top left and right), 3 and 4 (bottom left and right)

to the item as is allowed by the interface. (4) Repeat steps 2 and 3 until you find the red chair. (5) Evaluate the system using the provided form and move to the next one. (6) After finishing all systems, identify your preferred interaction, and briefly motivate your decision. The study was conducted with 15 participants (7 female, age range 19-54, M 24, SD 8.5), without paid incentives. Requirements for participation were a good comprehension of English, no visual deficits, and no motor handicaps. All measurements were relative to within-subject counterparts. The study setting was a table with a laptop, on which participants would go through the interactions. The participants used a mouse, no keyboard actions were required. After completing the digital interaction sequences, the participants filled out the questionnaire.

3.4 Measurements

3.4.1 Efficiency. We consider the rate at which participants go through the set of tasks as efficiency. The combined statistic for efficiency is measured by two variables: (1) the number of interactions (number of times the system prompted user feedback and a new item was generated), and (2) the total task completion time starting with the first item and ending with the designated item as specified in the task (red chair).

3.4.2 Interaction Experience. The user experience is measured with a set of evaluating questions based on the NASA Task Load Index (TLX) [8]. The TLX consists of six scales: Mental demand, Physical demand, Temporal demand, Performance, Effort, and Frustration. These are combined into a single overall score for the perceived workload. This overall score is used to assess the effort required by the user to perform a task in the respective interface. For training learning systems, we hypothesize that effort determines comfort or discomfort. This method allows us to separate interaction parameters for every interface, provides us with a useful overall score, and is realistically implemented in the test setting. As an addition to the workload parameters, participants were given two additional scales on which they rated both the interactivity as well as the extent to which the interaction was perceived as being engaging. These featured the same scale with the same number of intervals. Besides that, users were asked for their favorite interface to interact with in terms of overall experience, that is, which human-in-the-loop interaction they perceived as most appropriate. Finally, participants were prompted to provide qualitative feedback.

4 RESULTS

To understand the performance of the different system variants we consider the total learning efficiency and experience of the trainer as the main determinants of quality. The data was checked for outliers (values that deviate more than 3 SD from the mean value), resulting in a single data point being removed for interface 4. By comparing the measured experience parameters across variants and viewing that in respect to what variants were experienced as being most pleasant, we can attempt to make a prediction on the parameters that are likely to dictate the overall experience of users. The way in which preference was expressed by users is also clearly distributed: users preferred interfaces 2 and 3 by a large margin (each 6 votes). The remaining votes went to interface 1 (2 votes) and 4 (1 vote). This hints at a correlation with the number of interactions (cf. Table 1), which were both lower on average for interfaces 2 and 3. Also, users rated their efforts lower for these two interfaces. Based on the replies given in the qualitative follow-up, users disliked the definitive nature of interfaces 1 and 4. A large factor of their frustration was that these two interfaces gave users the feeling that decisions were irreversible and may have unintended side-effects on learning, whereas interfaces 2 and 3 allowed users to steer more lightly. The participants indicated that their ability to complete the given task was better with interfaces 2 and 3 as well.

Especially for interface 4 participants reported that it appeared that they had to provide information more than once. This also coincided with higher frustration on average. Another cause for

Table 1: Mean results of participants after testing the different system interfaces. Standard deviation included in parenthesis.

Interface	Number of interactions	Effort 1-10	Interactivity 1-10	Fun and Engagement 1-10
Interface 1	6.0 (SD 0.7)	2.4 (SD 1.3)	4.5 (SD 2.5)	4.9 (SD 2.5)
Interface 2	5.0 (SD 1.2)	2.3 (SD 1.2)	6.6 (SD 2.0)	6.1 (SD 2.2)
Interface 3	5.0 (SD 1.2)	2.6 (SD 1.5)	6.0 (SD 2.1)	6.6 (SD 2.4)
Interface 4	6.0 (SD 0.9)	3.2 (SD 1.7)	5.7 (SD 1.9)	4.7 (SD 2.0)

frustration was recalling information that users were not particularly aware of. In many instances, participants made very quick decisions and when interface 4 would inquire about their motivations, they could not always recall them. Finally, the participants reported on the rather simplistic interface 1 that it provided them with too little room to explain their motivations.

4.1 Design guidelines

We can briefly summarize the qualitative study results in the following interpretative design guidelines:

- (1) An interactive learning interface likely scores higher when it features more opportunities for expression.
- (2) Opportunities for expression should be proactive rather than reactive, as recalling information can be frustrating and deliver inconsistent results.
- (3) When designing an interactive learning interface, one should avoid making the training process seem like there is the possibility of making irreversible mistakes that cannot easily be corrected for later on.
- (4) Quality of data in the feedback cycle can be increased without having a considerable negative impact on the total interaction time.

The first guideline corresponds to the intuitive notion that binary, course-grained decisions can be perceived as more difficult than more fine-grained choices, especially for items that are similar. Although more fine-grained choices might mean more effort in terms of interaction (navigating, selecting, clicking, sliding, etc.) learning algorithms often benefit from input data that is matched to individual features rather than entire instances. This is also reflected in other works [18]. The second guideline makes it so that there is not too much attention strain and thus fatigue on the individual teaching. Not having to recall past information can be seen as an extension of one of the human AI guidelines as laid out by Amershi et al., being that responses to information should be timely and recent interactions should remain accessible [2]. The third guideline clearly conflicts with of Shneiderman’s golden rules of interface design [20], i.e., permitting easy reversal of actions, which is often realized in a form of “undo” operation. Yet, it is important to distinguish a machine training interaction from conventional user interactions that aim to issue precise commands, settings, and data. Machine training is essentially an interaction with a layered statistical process and as such tolerant against outliers, that is, training actions that need correction. Rather than moving backwards through a sequence of choices to correct a past choice, it is far more efficient to continue and let the statistics “work out” the outlier. From a user point of view, this might change their

approach towards aiming for a good enough level of accuracy and moving through the training batch at a consistent pace, rather than perfecting choices in a lower number of training examples. The final guideline can be concluded from the data in table 1 which shows that the most engaging and quickest methods do not necessarily make up the ones with the lowest quality of data per interaction.

The four types of interactions differed at a fundamental level, as an attempt to capture different methods of providing training inputs. Interface 1 is a very common type of training interface with many applications in industry and the baseline in this study. The other interfaces add to the baseline in different ways: interface 2 adds dimensionality to the presented input and user choice, i.e., the user is prompted to consider different aspects of the input and make a more fine-grained choice. Interface 3 offers a comparative input, which also prompts a user to weigh different attributes of the presented input variants. Finally, interface 4 introduces a memory task; recalling, reflecting on, and justifying a choice with a given dimension. Given the possibly large design space, there is a need for rating, scoring, and benchmarking interfaces to learning systems. In the following section, we combine measurable aspects of interaction with learning user interfaces into a single score.

4.2 The IMLIQ score

With the results of how users rated their preference and the outcomes of the evaluation, we can make a prediction about how the individual factors presented in the evaluation influence overall experience. We propose an IMLIQ(Interactive Machine Learning Interaction Quality) score, which allows for the assessment of the relative performance of a human in-the-loop learning system in a controlled setting. As a method, that means it does not deliver universal results, but only results compared to systems tested in the same setting. After all, the score features subjective assessment, making it so that tests in different settings, with a different baseline, would result in disproportionate scoring. It should be noted that this is an example of how to construct such a performance assessment that takes into account user experience and interactions efficiency, rather than a fully empirically supported model. The exemplary IMLIQ scoring formula constructed with respect to this study can be defined as:

$$Q = I + Fe - E - 10 * (1 - Ni) - 10 * Cr \quad (1)$$

The IMLIQ score (Q) has a maximum of 20 points (matching the maximum score in the effort calculations) with a theoretical limit of $-\infty$ for unbounded interaction time. Interactivity (I) represents a subjective rating of the interface ranging from 1 to 10. A higher perceived interactivity thus leads to a higher perceived quality of

interaction, similar for the fun and engagement (Fe) component (1 - 10). The remaining measurements decrease the overall score: user effort (E) as calculated by TLX method [8] and the number of interactions (Ni) until task completion. The latter is a relative “handicap” measure based on the number of interactions to train the model divided by the minimal number of interactions recorded to train on the same data set in the same setting. This factor is multiplied by 10. Finally, the completion rate (Cr) represents the expected completion time in seconds. When testing multiple interfaces, this variable should remain a constant to properly show the relative individual performance of the different systems. Similar to the number of interactions, this variable is multiplied by 10. Note that the relative components limit the scoring model to comparative assessments. Applying the IMLIQ scoring model (with an expected testing time of 30 seconds) results in scores of 5.7, 7.9, 8.2, and -1.5 for interfaces 1 to 4 respectively.

4.3 Limitations and future work

The studied interface variants cannot represent all possibilities learning interfaces and further research is needed—also in different, more realistic training contexts. The way this study was performed felt reportedly close to a game and as such the baseline for its enjoyment may have been off compared to a real context. The study is limited in its predictive power by the number of participants, their background in higher-education and their interaction time. It could also be argued that the test results might be skewed towards a slightly better than average understanding of machine learning systems, a potential experimental artifact (cf. demand characteristics [4]). Still, the suggestions that are based on qualitative data form a basis for guiding future investigations into principles for interactive learning system design. The same holds for the proposed IMLIQ score, which is not fully empirically supported and requires further research, especially regarding the subjective component. The proposed scoring method shows how to approach performance assessment for training interfaces for machine learning systems, but its function is rather exemplary as the linear relationships within the model are hard to support with the lacking statistical power of the study. The proposed method of scoring would benefit from a larger study aiming to get a better grasp at the relationship between the parameters in the scoring model as well as comparing it to other assessment methods.

5 CONCLUSION

Training interfaces for machine learning systems can be designed in different ways, unintentionally limiting users in their interaction possibilities during system training. Literature indicates a need for better approaches to evaluate interactive machine learning systems and the involved interactions. This study evaluates four different interactive learning systems with varying levels and aspects of interactivity. Based on the resulting data we propose a scoring method that combines technical performance and user interaction into a single performance metric as a means of setting future guidelines. Second, the qualitative feedback provided in the study portrays a number of opportunities for potential guidelines to be more thoroughly explored. The findings portray different avenues for further exploration. As such, this study provides motivation as well as a

basis for future HCI and design research on novel training interface implementations for interactive human-in-the-loop learning systems.

REFERENCES

- [1] Saleema Amershi, Maya Cakmak, William Bradley Knox, and Todd Kulesza. 2014. Power to the people: The role of humans in interactive machine learning. *AI Magazine* 35, 4 (2014), 105–120.
- [2] 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). Association for Computing Machinery, New York, NY, USA, 1–13. <https://doi.org/10.1145/3290605.3300233>
- [3] Shubhomoy Das, Travis Moore, Weng-Keen Wong, Simone Stumpf, Ian Oberst, Kevin McIntosh, and Margaret Burnett. 2013. End-user feature labeling: Supervised and semi-supervised approaches based on locally-weighted logistic regression. *Artificial Intelligence* 204 (2013), 56–74.
- [4] Nicola Dell, Vidya Vaidyanathan, Indrani Medhi, Edward Cutrell, and William Thies. 2012. “Yours is better!” participant response bias in HCI. In *Proceedings of the sigchi conference on human factors in computing systems*. 1321–1330.
- [5] Mengnan Du, Ninghao Liu, and Xia Hu. 2019. Techniques for interpretable machine learning. *Commun. ACM* 63, 1 (2019), 68–77.
- [6] John J Dudley and Per Ola Kristensson. 2018. A review of user interface design for interactive machine learning. *ACM Transactions on Interactive Intelligent Systems (TiiS)* 8, 2 (2018), 1–37.
- [7] Jerry Alan Fails and Dan R Olsen Jr. 2003. Interactive machine learning. In *Proceedings of the 8th international conference on Intelligent user interfaces*. 39–45.
- [8] Sandra G Hart and Lowell E Staveland. 1988. Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. In *Advances in psychology*. Vol. 52. Elsevier, 139–183.
- [9] Andreas Holzinger. 2016. Interactive machine learning for health informatics: when do we need the human-in-the-loop? *Brain Informatics* 3, 2 (2016), 119–131.
- [10] Andreas Holzinger, Markus Plass, Katharina Holzinger, Gloria Cerasela Crisan, Camelia-M Pintea, and Vasile Palade. 2017. A glass-box interactive machine learning approach for solving NP-hard problems with the human-in-the-loop. *arXiv preprint arXiv:1708.01104* (2017).
- [11] Shervin Javdani, Siddhartha S Srinivasa, and J Andrew Bagnell. 2015. Shared autonomy via hindsight optimization. *Robotics science and systems: online proceedings* 2015 (2015).
- [12] Todd Kulesza, Margaret Burnett, Simone Stumpf, Weng-Keen Wong, Shubhomoy Das, Alex Groce, Amber Shinsell, Forrest Bice, and Kevin McIntosh. 2011. Where are my intelligent assistant’s mistakes? A systematic testing approach. In *International Symposium on End User Development*. Springer, 171–186.
- [13] 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*. 126–137.
- [14] Todd Kulesza, Simone Stumpf, Margaret Burnett, Weng-Keen Wong, Yann Riche, Travis Moore, Ian Oberst, Amber Shinsell, and Kevin McIntosh. 2010. Explanatory debugging: Supporting end-user debugging of machine-learned programs. In *2010 IEEE Symposium on Visual Languages and Human-Centric Computing*. IEEE, 41–48.
- [15] Todd Kulesza, Simone Stumpf, Weng-Keen Wong, Margaret M Burnett, Stephen Perona, Andrew Ko, and Ian Oberst. 2011. Why-oriented end-user debugging of naive Bayes text classification. *ACM Transactions on Interactive Intelligent Systems (TiiS)* 1, 1 (2011), 1–31.
- [16] Todd Kulesza, Weng-Keen Wong, Simone Stumpf, Stephen Perona, Rachel White, Margaret M Burnett, Ian Oberst, and Andrew J Ko. 2009. Fixing the program my computer learned: Barriers for end users, challenges for the machine. In *Proceedings of the 14th international conference on Intelligent user interfaces*. 187–196.
- [17] Ting-Peng Liang, Hung-Jen Lai, and Yi-Cheng Ku. 2006. Personalized content recommendation and user satisfaction: Theoretical synthesis and empirical findings. *Journal of Management Information Systems* 23, 3 (2006), 45–70.
- [18] Hema Raghavan, Omid Madani, and Rosie Jones. 2006. Active learning with feedback on features and instances. *Journal of Machine Learning Research* 7, Aug (2006), 1655–1686.
- [19] Patrick Schramowski, Wolfgang Stammer, Stefano Teso, Anna Brugger, Xiaoting Shao, Hans-Georg Luigs, Anne-Katrin Mahlein, and Kristian Kersting. 2020. Right for the wrong scientific reasons: Revising deep networks by interacting with their explanations. *arXiv preprint arXiv:2001.05371* (2020).
- [20] Ben Shneiderman, Catherine Plaisant, Maxine Cohen, Steven Jacobs, Niklas Elmquist, and Nicholas Diakopoulos. 2016. *Designing the user interface: strategies for effective human-computer interaction*. Pearson.

- [21] Patrice Y Simard, Saleema Amershi, David M Chickering, Alicia Edelman Pelton, Soroush Ghorashi, Christopher Meek, Gonzalo Ramos, Jina Suh, Johan Verwey, Mo Wang, et al. 2017. Machine teaching: A new paradigm for building machine learning systems. *arXiv preprint arXiv:1707.06742* (2017).
- [22] Simone Stumpf, Vidya Rajaram, Lida Li, Margaret Burnett, Thomas Dietterich, Erin Sullivan, Russell Drummond, and Jonathan Herlocker. 2007. Toward harnessing user feedback for machine learning. In *Proceedings of the 12th international conference on Intelligent user interfaces*. 82–91.
- [23] Qian Yang, Aaron Steinfeld, Carolyn Rosé, and John Zimmerman. 2020. Re-examining Whether, Why, and How Human-AI Interaction Is Uniquely Difficult to Design. In *Proceedings of the 2020 chi conference on human factors in computing systems*. 1–13.
- [24] Jürgen Ziegler. 2019. Challenges in User-Centered Engineering of AI-based Interactive Systems.. In *EICS Workshops*. 51–55.