Forward Reasoning Decision Support: Toward a More Complete View of the Human-AI Interaction Design Space

Zelun Tony Zhang
zhang@fortiss.org
fortiss GmbH, Research Institute of the Free State of Bavaria
Munich, Germany

Yuanting Liu
liu@fortiss.org
fortiss GmbH, Research Institute of the Free State of Bavaria
Munich, Germany

Heinrich Hußmann
hussmann@ifi.lmu.de
LMU Munich
Munich, Germany

ABSTRACT

Decision support systems based on AI are usually designed to generate complete outputs entirely automatically and to explain those to users. However, explanations, no matter how well designed, might not adequately address the *output uncertainty* of such systems in many applications. This is especially the case when the *human-out-of-the-loop* problem persists, which is a fundamental human limitation. There is no reason to limit decision support systems to such *backward reasoning designs*, though. We argue how more interactive *forward reasoning designs* where users are actively involved in the task can be effective in managing output uncertainty. We therefore call for a more complete view of the design space for decision support systems that includes both backward and forward reasoning designs. We argue that such a more complete view is necessary to overcome the barriers that hinder AI deployment especially in high-stakes applications.

CCS CONCEPTS

• **Human-centered computing → HCI theory, concepts and models**; **Interaction paradigms**.

KEYWORDS

decision support, output uncertainty, human-AI interaction, intelligent systems, transparency, explainability, user control, forward reasoning, backward reasoning

ACM Reference Format:

Zelun Tony Zhang, Yuanting Liu, and Heinrich Hußmann. 2021. Forward Reasoning Decision Support: Toward a More Complete View of the Human-AI Interaction Design Space. In *CHItaly 2021: 14th Biannual Conference of the Italian SIGCHI Chapter (CHItaly '21), July 11–13, 2021, Bolzano, Italy*. ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3464385.3464696

1 INTRODUCTION

In spite of the impressive progress in artificial intelligence (AI) in recent years, in many critical, high-stakes domains such as aviation, medical technology or law enforcement, AI is not yet widely deployable. Challenges like brittleness of the AI models [8] or algorithmic bias create significant barriers to practical usage in terms of safety, ethics or social justice. The *opaqueness* and *complexity* of modern AI technologies like deep learning are generally seen as the main issues in this regard, driving the rapidly growing interest in explainable AI (XAI). However, despite the increasingly active research in this field, results pertaining to how explanations benefit end users are ambiguous so far [22].

Much of the work on XAI is concerned with generating explanations for decision support systems that produce outputs automatically with no human intervention [4, 13, 21]. This one-sided focus ignores a large space of promising alternative designs to support human decision making. In this paper, we propose the notion of *forward reasoning decision support*, a design paradigm for decision support systems centered around human agency. We point out how such interactive designs can enable the deployment of AI to high-stakes applications and lay out directions for future work that is necessary to this end. With our proposal, we intend to provoke discussions that look beyond automatic, explanation-providing machines, and advocate for a more complete view of the design space of decision support systems and more generally of human-AI interaction.

2 BACKGROUND

2.1 The human-out-of-the-loop problem

The opaque nature particularly of deep neural networks hampers their development significantly. The current push for XAI was therefore mainly initiated by the desire of the AI community for a better understanding about the functioning of their models [1]. What makes sense in this technical context is not automatically guaranteed to translate well to end user interactions with AI systems. Nevertheless, the HCI community has adopted the focus on explanations [14, 16, 19], relegating the human to a passive supervisor in many applications. However, human factors researchers concerned with the design of automation have known for decades that humans are not good at passive supervision [2]. Without active engagement in the actual task, even a motivated person will struggle to notice erroneous behavior of an automated system that usually works reliably [2, 6]. This fundamental human limitation is known as the *out-of-the-loop* (OOTL) problem among human factors researchers and is still not satisfactorily addressed, even after decades of research [6]. By taking over from the AI community the paradigm of automated AI systems that explain themselves to the user, the HCI community is facing the very same barrier now.

CHItaly '21, July 11–13, 2021, Bolzano, Italy

ACM ISBN 978-1-4503-8977-8/21/06…$15.00
https://doi.org/10.1145/3464385.3464696

Considering the knowledge from automation research, it is questionable whether explanations alone can pave the way for AI to enter high-stakes applications, even if they were one day "perfectly" designed. After all, no matter how well explanations meet human information needs, there still remains the issue of the passive role of users in most decision support system designs.

2.2 Output uncertainty

The current focus on XAI is a natural answer to the opaqueness and complexity of many modern AI algorithms. However, while these are major issues when developing AI models, they might not actually be the core problems in end user interactions. As argued by [22], *output uncertainty* is what makes AI systems difficult to interact with for end users, i.e. the case-by-case uncertainty about whether the system output is a desired output. This output uncertainty is present whenever a system has less than 100% accuracy on its task. But even if entirely accurate AI models existed, many AI applications are inherently subject to output uncertainty as there are no definite answers to certain questions. Consider for example the paper review process: When is a paper worth publishing? Despite certain criteria, within the gray area between clear accepts and clear rejects, the answer to this question is up the reviewer's discretion, and different reviewers might judge differently. It is exactly such ambiguous problems where AI is often applied.

While opaqueness and complexity are frequent contributors to output uncertainty, they are not the same. A system could be arbitrarily opaque and complex without users ever caring, as long as they could be certain that the system behaved as expected. Vice versa, high output uncertainty could render interactions with the system problematic even if it was fully transparent and explainable.

3 XAI DOES NOT FULLY ADDRESS OUTPUT UNCERTAINTY

Extending the argumentation of [22], we suggest three features of designs that successfully manage output uncertainty. For the sake of brevity, we write about *AI errors* in the following, which also contains ambiguous instances where no clear "right" or "wrong" can be stated. In these cases, an AI error indicates an instance where users would disagree with the algorithmic result had they done the task themselves. The three features are:

- *Detectability*: Within the context of intended use, users can detect AI errors with an effort that is reasonable considering the task frequency and complexity. Infrequent and complex tasks might demand larger effort to examine the system output, while AI errors in frequent and simple tasks should be effortlessly noticeable.
- *Correctability*: Users can correct AI errors with an effort that is reasonable considering the error frequency and task complexity. Infrequent errors in complex tasks might demand larger effort to correct, while frequent errors in simple tasks should be effortless to correct. Ideally, the system can learn from the corrections about user intentions and/or task specifics.
- *Non-criticality*: Successful usage of the system does not hinge on the case-by-case correctness of the AI. In particular, an AI error should not hinder users in completing a task or lead to severe consequences if it goes unnoticed. Ideally, the system can provide value to users even if a result is not entirely correct.

Current XAI efforts only address the detectability of AI errors. However, even here, the OOTL problem might render AI errors hard to detect in certain use contexts, despite well-designed explanations. This could be true in setups where a system produces outputs fully automatically in high frequency, like a hiring manager using an automated recruiting tool to process large amounts of applications. Explanations might help to notice AI errors in an individual case. However, if users are asked to review a long stream of cases, they might miss AI errors due to the OOTL problem.

4 FORWARD REASONING DECISION SUPPORT

4.1 Automation plus XAI is not the only option

The predominant paradigm for decision support system design starts from a system that generates outputs automatically with no human intervention. XAI is then employed to help users to make sense of the automatic results, so that users can actually incorporate the system output into their decision making. Section 3 argues how this setup insufficiently manages the output uncertainty that plagues AI applications. Fortunately, there is no reason why decision support system designs should be limited to this paradigm.

Alternatively, one can start from the human decision maker and design a highly interactive support system around the user. The aim would be to help users make better decisions than either human or AI alone. At first glance, the difference seems to be one of automated and non-automated decision making. On closer inspection, this turns out to be a fallacy—at least if the final decision responsibility lies with a human, as is usually the case for high-stakes applications[1]. Since in that case users of an automatic AI system need to decide whether or not to follow the system output, they still have to go through a decision-making process on their own. Seeing the automatically generated result, users need to reason backward from the output and reconstruct the decision made by the system. There is little synergy between the human and the system as the decision is in effect made twice: once by the system, and a second time by the human in a backward reasoning process. In contrast, with the more interactive design, users reason forward from input to output, supported by an interactive AI system. Hence, the major difference between the two paradigms is whether users reason backward or forward during the decision-making process, as illustrated in Fig. 1. We therefore call the paradigm of automatic, explanation-providing systems *backward reasoning decision support*, while we name the more interactive alternative *forward reasoning decision support*.

The goal of a forward reasoning design would be to enhance human decision-making abilities through synergies between human and AI, echoing the sentiment behind mixed-initiative user interfaces [10]. While AI, in particular machine learning (ML), is good at

[1] The European General Data Protection Regulation (GDPR) even prohibits by default fully automatic decision making in cases where the decisions might have significant impact on individuals. It is therefore a legal obligation to have a human make the final decision.

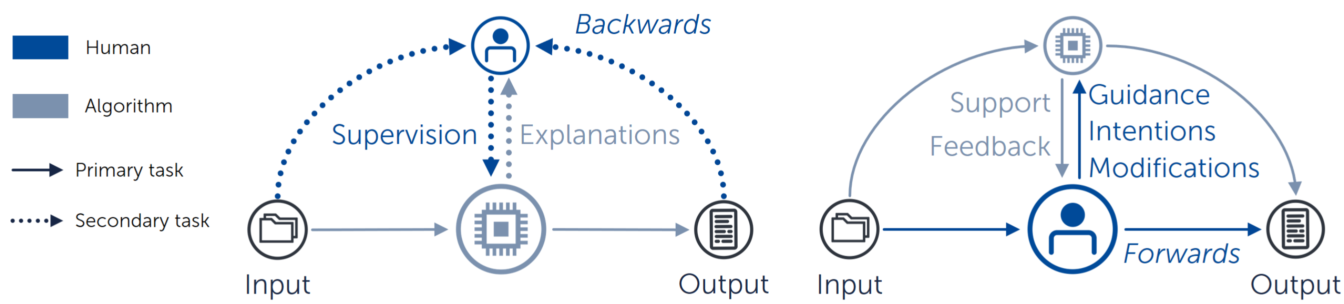

Figure 1: Schematic juxtaposition of *backward* (left) and *forward reasoning decision support* (right). With the former, the human is not involved in the primary task, but is occupied with the secondary supervision task. The system takes over the primary task completely, but also has to fulfill the secondary task of explaining to the human. In forward reasoning designs, both human and system are fully involved in the primary task. The backward reasoning paradigm starts with the automated system, and tries to get the human back into the loop. In the forward reasoning paradigm, the human is the starting point, and the aim is to integrate the AI into the loop.

recognizing patterns and correlations, current state-of-the-art algorithms do not understand causality and common sense. A forward reasoning design would thus contribute the pattern recognition power of AI/ML to the human decision-making process to enhance decisions driven by human judgement. Such forward reasoning decision support systems could for instance enhance the human ability to sift through large amounts of data, support users in discovering patterns and relationships, or mitigate human cognitive biases [20].

4.2 How forward reasoning decision support can address output uncertainty

Forward reasoning decision support relies heavily on techniques for rich interactions with AI systems, a selection of which is listed in Table 1. While these are not targeted at decision support, they can still serve as illustration for how forward reasoning designs could look like. Furthermore, these examples give a sense of how forward reasoning decision support addresses the issue of output uncertainty, based on the three features laid out in Section 3:

- *Detectability*: In all examples in Table 1, the intelligent components of the systems are closely coupled with users' current actions such that the AI increments on what users are currently doing. Users always have the full context and know what to expect since they themselves move the task completion forward. The OOTL problem is therefore not an issue and users easily notice when outputs deviate from what they expect. This suggests that explanations are not the only way to enable humans to detect AI errors.
- *Correctability*: *SMILY* and *PTM* provide examples for how to enable users to interactively refine AI-generated outputs. *SMILY* retrieves images from a database that are similar to a given query image. Its refinement tools help users to communicate their intentions to the system when it focuses on different aspects than what users are looking for. The editing capability of *PTM* on the other hand allows users to contribute their human knowledge on how to handle difficult nuances of language translation. The system can in turn learn from the user edits.
- *Non-criticality*: The systems in Table 1 collaboratively support users instead of automatically producing a complete end result. As such, their usefulness does not hinge on the correctness of a single AI inference. *Wrangler* and *RAVS* in particular are designed such that AI errors have minimal negative impact, if at all. With *Wrangler*, users can simply ignore unhelpful suggestions. Imperfect but close-enough suggestions can even be helpful, e.g. by triggering new ideas or because they can be adopted with minimal user edits. The AI model behind *RAVS* on the other hand is specifically tuned to avoid false negatives, while false positives are non-critical by design. This is because the system is not meant to make cancer predictions, but to suggest regions for physicians to review, eliminating the need to systematically search entire gigapixel-sized images for small and rare structures.

4.3 A more complete design space

We emphasize that our intention is *not* to pit forward against backward reasoning designs. However, we notice that the field is currently clearly focused on the latter and thereby neglects a large part of the design space. We argue that it is necessary to consider a wider range of solutions to break through the obstacles that hinder AI deployment in high-stakes applications. With their potential to address the problem of output uncertainty, forward reasoning designs constitute an important complement to the currently favored backward reasoning decision support systems.

Two streams of future work stand out to obtain a more complete view of the design space. For one, it is important to have a better understanding of when either backward or forward reasoning designs are more appropriate, as opposed to opting for the former per default. Our three features of successful output uncertainty management can serve as a possible starting point in this regard. Aspects like "reasonable effort" or the criticality of AI errors—and hence the appropriateness of either design paradigm to manage

Table 1: Literature examples of human-AI interaction techniques that are potentially useful for forward reasoning decision support.

System	Description	Interaction technique
SMILY [3]	Content-based image retrieval system for medical images.	Three ways to refine retrieved results: cropping regions of interest, pinning interesting results, sliders to (de-)emphasize certain concepts.
RAVS [15]	Visual search tool for cancer assessments.	Automated navigation to image regions that need to be assessed for tumor presence.
Wrangler [12]	Data cleansing tool with predictive data transformation suggestions.	"Automcomplete" suggestions for transformations based on users' interactions with a data table. Users can edit "close, but not perfect" suggestions.
PTM [7]	Automated language translation tool.	Users can edit machine-generated translations. The parts of the text affected by the user edits are automatically retranslated.

output uncertainty—depend highly on the application context. Combining elements of both paradigms also constitutes an interesting direction. For instance, Jacobs et al. found that clinicians seek explanations when the decision support system deviates from their expectation [11]. Thus, even when following a forward reasoning design paradigm, it might be useful at times to allow for backward reasoning.

Secondly, there needs to be a larger toolbox of interaction techniques like those in Table 1 to design effective forward reasoning decision support. Especially techniques like those in [3] that allow users to steer the outputs of complex deep learning models are scarce. Work on mixed-initiative user interfaces [10] can serve as inspiration on this front.

5 DISCUSSION: FORWARD REASONING DECISION SUPPORT AND RELATED WORK

Our notion of forward reasoning decision support bears a resemblance to the work of Wang et al. [20]. Based on the findings of a co-design exercise with clinicians, the authors recommend to support forward reasoning instead of the commonly triggered backward reasoning in decision support systems. However, their framework provides guidance on choosing XAI facilities to support human reasoning, and to help mitigate cognitive biases in particular. We take a broader view and aim to extend the design space for decision support systems to include more interactive and collaborative designs. This view is supported by the study of Cai et al. [3], who found that users employed their refinement tools to get a better understanding of the algorithm, suggesting that explanations are not the only means to this end.

While taking a different angle than the predominant lines of research, our proposal for forward reasoning decision support strikes a similar tone to some other voices in the field. Notably, Ben Shneiderman has recently laid out his vision of human-centered AI [17], essentially a reframing for AI of his well-known advocacy to think of computers as tools rather than agents [18]. He argues for shifting from the attempt to place humans into the loop around AI, to building AI-in-the-loop around humans [17]. By taking a wider view on the design space and considering both backward and forward reasoning designs, we adopt the human-centered conception of AI-in-the-loop. Our starting point is not the notion of an autonomous AI system, but the human and how to best support humans with AI. The solution *can* entail a backward reasoning design, but does not need to. We argue that it is often more appropriate to choose a more interactive, forward reasoning design to handle the output uncertainty in end user interactions with AI.

We further acknowledge the similarity of forward reasoning designs to human-in-the-loop or interactive machine learning (iML) [5], given their shared reliance on user-directed interactions. However, the goal of iML is usually to improve the ML model by integrating the user into the training process [9], while our focus is to enhance human decision making. We also do not consider forward reasoning decision support per se as our contribution, but rather want to direct attention towards the need for a more holistic view of the human-AI interaction design space. We therefore point out in Section 4.1 that the key difference between backward and forward reasoning designs is not whether users' decisions are supposedly automated, but rather the direction of users' reasoning process. We also suggest in Section 4.2 that interaction techniques like those in Table 1 can address the same issues typically approached with explanations.

6 CONCLUSION & FUTURE WORK

With this paper, we intend to provoke discussions to open up the currently narrow focus of the XAI field, which considers explanations as the only way to handle AI errors. This one-sided focus ignores a large part of the design space for decision support systems and does not fully address the issue of output uncertainty. This is especially the case in applications where the OOTL problem persists, which is a fundamental human limitation. Thus it is questionable whether explanations alone can be effective in such applications, no matter how well they are designed. We therefore advocate a more complete view of the design space and propose the notion of forward reasoning decision support to this end. While we put our focus on decision support systems here, we claim that the same is necessary for human-AI interaction in general. AI system design needs to look beyond fully automatic AI systems and needs to consider more interactive designs as well to pave the way for a wider deployment of AI, especially in high-stakes applications.

This paper also serves as motivation for our future work on forward reasoning decision support. We identified two broad directions for future work: 1) a better understanding of when either backward or forward reasoning designs are more appropriate, and 2) new human-AI interaction techniques for effective forward reasoning designs. As a first step, we plan to address the former by thoroughly analyzing the effectiveness of current XAI approaches. We further plan to perform an initial exploratory comparison between backward and forward reasoning decision support.

ACKNOWLEDGMENTS

This research was funded by the Bavarian Ministry of Economic Affairs, Regional Development and Energy.

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* (Montreal, Canada) *(CHI '18)*. ACM, 582:1–582:18. https://doi.org/10.1145/3173574.3174156

[2] Lisanne Bainbridge. 1983. Ironies of Automation. *Automatica* 19, 6 (Nov. 1983), 775–779. https://doi.org/10.1016/0005-1098(83)90046-8

[3] Carrie J. Cai, Martin C. Stumpe, Michael Terry, Emily Reif, Narayan Hegde, Jason Hipp, Been Kim, Daniel Smilkov, Martin Wattenberg, Fernanda Viegas, and Greg S. Corrado. 2019. Human-Centered Tools for Coping with Imperfect Algorithms during Medical Decision-Making. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland, UK) *(CHI '19)*. ACM, 4:1–4:14. https://doi.org/10.1145/3290605.3300234

[4] Hao-Fei Cheng, Ruotong Wang, Zheng Zhang, Fiona O'Connell, Terrance Gray, F. Maxwell Harper, and Haiyi Zhu. 2019. Explaining Decision-Making Algorithms through UI: Strategies to Help Non-Expert Stakeholders. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland, UK) *(CHI '19)*. ACM, 559:1–559:12. https://doi.org/10.1145/3290605.3300789

[5] John J. Dudley and Per Ola Kristensson. 2018. A Review of User Interface Design for Interactive Machine Learning. *ACM Transactions on Interactive Intelligent Systems* 8, 2 (July 2018), 8:1–8:37. https://doi.org/10.1145/3185517

[6] Mica R. Endsley. 2017. From Here to Autonomy: Lessons Learned from Human–Automation Research. *Human Factors: The Journal of the Human Factors and Ergonomics Society* 59, 1 (Feb. 2017), 5–27. https://doi.org/10.1177/0018720816681350

[7] Spence Green, Jason Chuang, Jeffrey Heer, and Christopher D. Manning. 2014. Predictive Translation Memory: A Mixed-Initiative System for Human Language Translation. In *Proceedings of the 27th Annual ACM Symposium on User Interface Software and Technology* (Honolulu, HI, USA) *(UIST '14)*. ACM, 177–187. https://doi.org/10.1145/2642918.2647408

[8] Douglas Heaven. 2019. Deep Trouble for Deep Learning. *Nature* 574 (Oct. 2019), 163–166. https://doi.org/10.1038/d41586-019-03013-5

[9] Andreas Holzinger. 2016. Interactive Machine Learning for Health Informatics: When Do We Need the Human-in-the-Loop? *Brain Informatics* 3, 2 (June 2016), 119–131. https://doi.org/10.1007/s40708-016-0042-6

[10] Eric Horvitz. 1999. Principles of Mixed-Initiative User Interfaces. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Pittsburgh, PA, USA) *(CHI '99)*. ACM, 159–166. https://doi.org/10.1145/302979.303030

[11] Maia Jacobs, Jeffrey He, Melanie F Pradier, Barbara Lam, Andrew C Ahn, Thomas H McCoy, Roy H Perlis, Finale Doshi-Velez, and Krzysztof Z Gajos. 2021. Designing AI for Trust and Collaboration in Time-Constrained Medical Decisions: A Sociotechnical Lens. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan). ACM, 1–14. https://doi.org/10.1145/3411764.3445385

[12] Sean Kandel, Andreas Paepcke, Joseph Hellerstein, and Jeffrey Heer. 2011. Wrangler: Interactive Visual Specification of Data Transformation Scripts. In *Proceedings of the 2011 Annual Conference on Human Factors in Computing Systems* (Vancouver, BC, Canada) *(CHI '11)*. ACM, 3363–3372. https://doi.org/10.1145/1978942.1979444

[13] 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* (Atlanta, GA, USA) *(FAT* '19)*. ACM, 29–38. https://doi.org/10.1145/3287560.3287590

[14] Brian Lim, Advait Sarkar, Alison Smith-Renner, and Simone Stumpf. 2019. ExSS: Explainable Smart Systems 2019. In *Proceedings of the 24th International Conference on Intelligent User Interfaces: Companion* (Marina del Ray, CA, USA) *(IUI '19)*. ACM, 125–126. https://doi.org/10.1145/3308557.3313112

[15] Martin Lindvall, Claes Lundström, and Jonas Löwgren. 2021. Rapid Assisted Visual Search: Supporting Digital Pathologists with Imperfect AI. In *Proceedings of the 26th International Conference on Intelligent User Interfaces* (College Station, TX, USA) *(IUI '21)*. ACM, 504–513. https://doi.org/10.1145/3397481.3450681

[16] Gonzalo Ramos, Jina Suh, Soroush Ghorashi, Christopher Meek, Richard Banks, Saleema Amershi, Rebecca Fiebrink, Alison Smith-Renner, and Gagan Bansal. 2019. Emerging Perspectives in Human-Centered Machine Learning. In *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland, UK) *(CHI EA '19)*. ACM, W11:1–W11:8. https://doi.org/10.1145/3290607.3299014

[17] Ben Shneiderman. 2020. Human-Centered Artificial Intelligence: Three Fresh Ideas. *AIS Transactions on Human-Computer Interaction* 12, 3 (2020), 109–124. https://doi.org/10.17705/1thci.00131

[18] Ben Shneiderman and Pattie Maes. 1997. Direct Manipulation vs. Interface Agents. *Interactions* 4, 6 (Nov. 1997), 42–61. https://doi.org/10.1145/267505.267514

[19] Alison Smith-Renner, Styliani Kleanthous, Brian Lim, Tsvi Kuflik, Simone Stumpf, Jahna Otterbacher, Advait Sarkar, Casey Dugan, and Avital Shulner. 2020. ExSS-ATEC: Explainable Smart Systems for Algorithmic Transparency in Emerging Technologies 2020. In *Proceedings of the 25th International Conference on Intelligent User Interfaces: Companion* (Cagliari, Italy) *(IUI '20)*. ACM, 7–8. https://doi.org/10.1145/3379336.3379361

[20] Danding Wang, Qian Yang, Ashraf Abdul, and Brian Y. Lim. 2019. Designing Theory-Driven User-Centric Explainable AI. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland, UK) *(CHI '19)*. ACM, 601:1–601:15. https://doi.org/10.1145/3290605.3300831

[21] Fumeng Yang, Zhuanyi Huang, Jean Scholtz, and Dustin L. Arendt. 2020. How Do Visual Explanations Foster End Users' Appropriate Trust in Machine Learning?. In *Proceedings of the 25th International Conference on Intelligent User Interfaces* (Cagliari, Italy) *(IUI '20)*. ACM, 189–201. https://doi.org/10.1145/3377325.3377480

[22] Zelun Tony Zhang and Heinrich Hußmann. 2021. How to Manage Output Uncertainty: Targeting the Actual End User Problem in Interactions with AI. In *Joint Proceedings of the ACM IUI 2021 Workshops co-located with the 26th ACM Conference on Intelligent User Interfaces, to appear* (College Station, TX, USA) *(IUI '21)*.