Designing AI for Appropriation Will Calibrate Trust

ZELUN TONY ZHANG, fortiss GmbH, Research Institute of the Free State of Bavaria, Germany
YUANTING LIU, fortiss GmbH, Research Institute of the Free State of Bavaria, Germany
ANDREAS BUTZ, LMU Munich, Germany

Calibrating users' trust on AI to an appropriate level is widely considered one of the key mechanisms to manage brittle AI performance. However, trust calibration is hard to achieve, with numerous interacting factors that can tip trust into one direction or the other. In this position paper, we argue that instead of focusing on trust calibration to achieve resilient human-AI interactions, it might be helpful to design AI systems for appropriation first, i.e. allowing users to use an AI system according to their intention, beyond what was explicitly considered by the designer. We observe that rather than suggesting end results without human involvement, appropriable AI systems tend to offer users incremental support. Such systems do not eliminate the need for trust calibration, but we argue that they may calibrate users' trust as a side effect and thereby achieve an appropriate level of trust by design.

Additional Key Words and Phrases: appropriation, artificial intelligence, iterative problem solving, incremental support, trust calibration

ACM Reference Format:
Zelun Tony Zhang, Yuanting Liu, and Andreas Butz. 2023. Designing AI for Appropriation Will Calibrate Trust. In *CHI TRAIT '23: Workshop on Trust and Reliance in AI-Assisted Tasks at CHI 2023, April 23, 2023, Hamburg, Germany*. ACM, New York, NY, USA, 7 pages. https://doi.org/XXXXXXX.XXXXXXX

1 INTRODUCTION

AI systems are notoriously brittle, i.e. their performance deteriorates abruptly under conditions that fall outside of what was covered during their development [29]. One mechanism widely seen as key to managing the brittleness of AI is *trust calibration*: Humans should be able to judge when to trust and rely on AI and when not to. The focus on trust calibration is especially prevalent in human-AI decision-making [2, 27], but is also prominent in other AI applications, like autonomous driving [20], or applications of large language models such as code generation [26] or question answering [11].

However, trust calibration is a very delicate balancing act (Fig. 1), as countless factors can tip users' trust into one or the other direction. To start with, how well users can calibrate their trust depends on various user-specific factors, such as personality [21], domain [27] or AI expertise [25]. Further, trust can depend on model performance—both the stated performance and as experienced by users [30]. Users' first impression can also play a role, i.e. whether they experience good or bad model performance first [19]. Apart from model outputs, AI explanations can also influence trust calibration in many ways. Relevant factors include the type of explanation (feature-based, example-based, etc.) [27], the specific algorithm used for a particular explanation type [13], or the wording of explanations [31]. Furthermore, seemingly small details of the user's task can have an influence as well [1]. In fact, even the terminology used to introduce an AI system has an effect [18]. These are just some examples, many more factors have been investigated in the literature.

Fig. 1. Symbolic illustration of trust calibration, created with Stable Diffusion[1]. Not only can trust calibration be likened to carefully balancing the stones on each other; the image also illustrates how we cannot even be sure whether trust calibration is a viable objective at all, given that the pictured scene is not real.

Given this fragility of trust calibration, it appears ineffective to rely on it as primary mechanism for achieving resilient human-AI interactions. Studies with experts such as clinicians [12] or pilots [33] also show that users often do not want or do not have the capacity to engage in case-by-case trust calibration. But how else can we deal with AI brittleness? In this position paper, we argue that it might be helpful to design AI systems for *appropriation* first, i.e. allowing users to use an AI system according to their intention, beyond what was explicitly considered by the designer [8]. We discuss why and how to design AI for appropriation and then come back to how trust calibration fits into the picture.

2 THE KEY TO APPROPRIABLE AI: PATERNALISM VERSUS SUPPORT

AI systems are often envisioned to support complex cognitive tasks, like making complex decisions or writing sophisticated texts. Due to the complexity of these tasks, it is unlikely that designers of AI systems can foresee or even model every eventuality that could occur during usage [28]. The struggle of the autonomous driving industry to reach market maturity is maybe the most prominent case in point. It is therefore necessary to allow users to use AI flexibly to cope with conditions that are outside of what designers can foresee and include into AI models. This is what technology appropriation is about. In this section, we discuss what makes AI systems easy or difficult to appropriate.

2.1 The problem with paternalistic AI systems

Many AI systems interact with their users in a "paternalistic" manner, i.e. they are designed to offer users complete solutions, without the need—or chance—for human involvement. When it comes to decision support for instance, AI decision support tools (DSTs) usually generate a ready-made assessment or decision recommendation [17]; the human decision maker can only evaluate the final result and takes no part in reaching that result. This approach to decision support and its limitations have been discussed under various terms, such as "backward reasoning decision support" by Zhang et al. [32], "end predictions" by Buçinca et al. [4], and "Oracle AI" by Cabitza et al. [6]. Such paternalistic patterns of human-AI interaction are effective as long as the AI output is exactly what users want, but unhelpful or even counterproductive otherwise.

The problem is that ready-made AI solutions are more often unhelpful than what high prediction accuracy or other model metrics would suggest. Speaking for the example of AI decision support again, a risk score or decision

[1] https://stablediffusionweb.com/

recommendation is often not very useful, since humans usually consider much more context than AI systems can: When screening child maltreatment cases, social workers might for instance know about the relationship between persons [14], which is unknown to the AI system. Clinicians might check on the general appearance of a patient ("How ill does the patient look?") [23], instead of only considering the data on which the DST recommendation is based. But not only do ready-made AI decision recommendations neglect that sort of context that is only accessible to humans; they also make it hard for human decision makers to combine their contextual knowledge with the evaluation provided by the AI. In the case of the social workers, they were mandated to use the DST, but considered it a "missed opportunity to effectively complement their own abilities" [15]. In another example, Blomberg et al. [3] report on a project to support a cloud services sales team with predictive models. The project failed despite the high accuracy and precision of the models because sellers were unable to incorporate the model predictions into their reasoning, which involved factors that were outside of the models.

These real-world cases suggest that what Dix has formulated for software systems in general appears to be true for AI systems as well: "*Designs that are closed are often more apparently sophisticated, because they may do more for the user, but ultimately do not allow the users to do more for themselves.*" [8] Apparently, by trying to directly solve a task for the user, paternalistic AI designs tend to be too closed and inflexible to be appropriated. As a result, they easily fail in practice when their output is imperfect.

2.2 Appropriable, incremental support enables co-decision and co-creation

But what are the alternatives? As Dix put it: "*Instead of designing a system to do the task you can instead design a system so that the task can be done.*" [8] For AI-based DSTs for instance, designers could turn their focus from providing ready-made decision recommendations to supporting decision makers' sensemaking [16], i.e. their process of building an understanding of the decision situation. Cai et al. [7] for example built a medical image retrieval system with control mechanisms that allow pathologists to specify which images they are looking for. The system supports in making diagnoses by helping pathologists to find similar cases for reference. Zhang et al. [33] designed a DST concept for pilots that continuously hints at possibly noteworthy properties of the surrounding airports. The purpose was to increase pilots' situation awareness—even during normal flight—so that they could always plan ahead, facilitating better decisions in case of an emergency.

Another noteworthy example is the academic research tool Elicit[2], which helps researchers find papers relevant to their research questions. One functionality of Elicit is to aid users in assessing the trustworthiness of a retrieved paper. Instead of displaying an aggregated trustworthiness score, Elicit considers which subquestions researchers might ask to assess the trustworthiness of a paper (e.g. "How many participants did the study have?", "Was the study pre-registered?", "Did the authors correct for multiple comparisons?", etc.). Elicit extracts the answers to these questions and links them to the paper, so users can easily check whether the system extracted the answers correctly. Users can further always formulate custom questions if the predefined ones are not sufficient to assess the trustworthiness of the paper. This way, users can decide themselves what is important for their assessment instead of relying on the signals that a trustworthiness prediction model would pick up.

In all of the above examples, there is no ready-made decision recommendation. In principle, one could surely add decision recommendations to each of them. In fact, for the aviation example, pilots were explicitly in favor of combining the continuous support with decision recommendations [33]. However, the key here is that decision makers

[2] https://elicit.org/

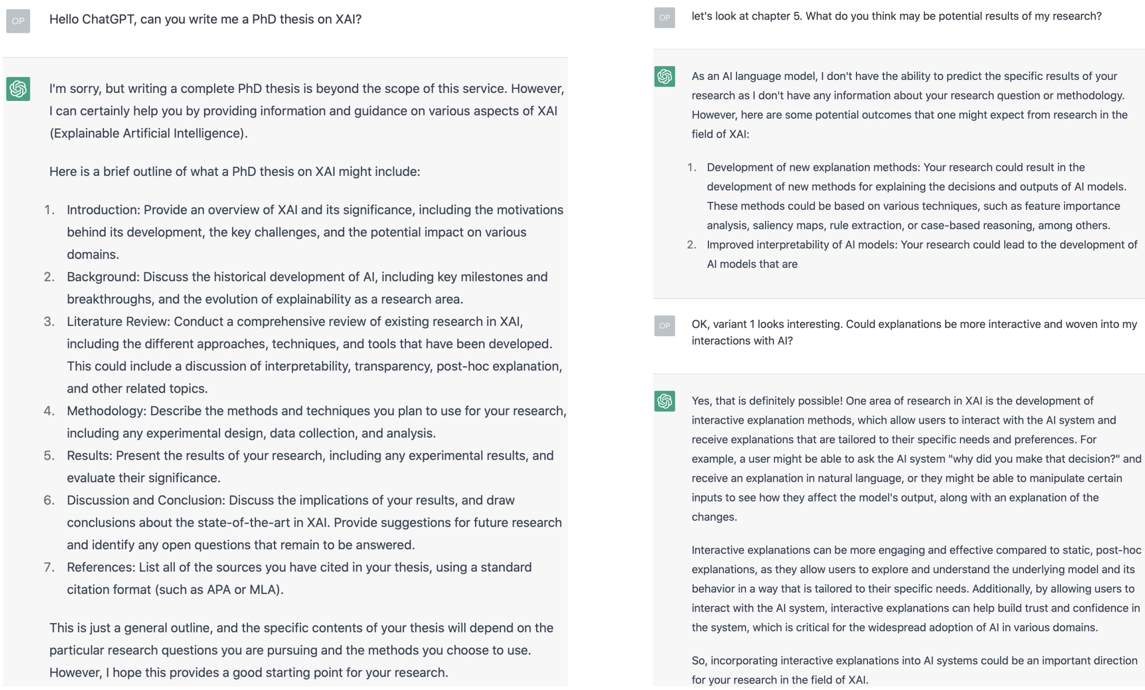

Fig. 2. Created with ChatGPT[3]. The AI refused to solve a complex problem at once, but helped structuring it. By a creative (appropriated) use of this tool, the user can still make progress toward eventually solving the original problem.

get incremental support that they can appropriate according to their own current sensemaking intention, enabling them to better combine human context information with AI support. This allows decision makers to benefit from the support provided even when it is imperfect, resulting in more resilient human-AI interactions.

Apart from decision support, using AI for creative purposes is another, maybe even more apparent area to discuss how appropriation of AI can look like and how it can be beneficial. Fig. 2 shows an exaggerated example for using ChatGPT to write a complex text. Similar to the examples in decision support, it demonstrates the principle of designing AI so that a task can be done, rather than designing to do the task. For a sufficiently complex task, ChatGPT's output will likely not live up to the user's intention. However, ChatGPT can provide small portions of text and also propose structures. A skillful user can take these intermediate outputs to iteratively develop ideas in dialogue and eventually co-create larger, more complex results. In the case of ChatGPT, the system is very general-purpose and its results largely depend on how it is being used. This does not only encourage, but actually requires appropriation.

2.3 Toward designing AI for appropriation

The above examples encourage thinking beyond paternalistic AI systems that try to solve tasks directly for users. A promising alternative role for AI is to incrementally support users to solve their tasks. However, paternalistic AI designs are arguably much easier to envision, given that AI research is largely driven by the desire to emulate human capabilities [22]. After all, what is more obvious than using these emulated capabilities to solve tasks for users? In

[3]https://chat.openai.com/chat

contrast, examples like those in Section 2.2 for incremental AI support are comparatively scarce, but a promising way toward more flexible, appropriable AI support tools is to learn from examples of how users appropriate AI, and then iterate these designs.

One example for AI appropriation is described by Ehsan et al. [9], where participants used and interpreted AI explanations in unanticipated ways based on their own intentions (either as affirmation for stable performance or diagnostic information for troubleshooting). Cai et al. [7] also observed appropriation with their medical image retrieval system mentioned in Section 2.2: Pathologists used the control tools provided to them in unexpected ways, e.g. to disambiguate whether surprising AI outputs were due to their own or due to the AI's mistake. Sivaraman et al. [23] found that human decision-making patterns are much more nuanced than typically assumed in human-AI decision-making experiments. In their study, many clinicians engaged with AI recommendations in a negotiation pattern by assessing the various components of a recommendation to determine which component can be accepted or needs adjustment. All of these examples give clues about how the respective AI system can be designed for more effective appropriation in a following iteration. They underline the importance of qualitatively investigating how people are actually using AI instead of only measuring quantitative outcomes.

3 APPROPRIATION WILL CALIBRATE TRUST AS A SIDE EFFECT

In many of the examples discussed above, it is still important that users recognize when to trust and rely on AI support and when not. However, we argue that trust calibration may not have to be a primary design goal when dealing with brittle AI performance. It may rather come as a side effect when the system supports users in achieving their goals or in their sensemaking, because users will encounter the imperfectness of AI at a much more granular level and are actively involved in shaping the end result.

The fragility of trust calibration as elaborated in Section 1 mainly stems from the fact that the corresponding systems provide end results without involving users. The consequence of this lack of involvement is that users do not engage purposefully with AI outputs and explanations [5, 10]. This could be addressed by letting users guide the interaction with a clear intention. A number of studies show that users do engage purposefully with AI when it does something surprising [7, 12, 23], but things can only be surprising when you have an expectation. Trust calibration will also likely become much easier when users engage with incremental AI support rather than checking complex end results (see the example of Elicit in Section 2.2). While seeking ways to appropriate the incremental AI support to solve their problems, users will likely learn about the AI's capabilities and weaknesses and thereby create an adequate level of trust as a by-product. In the case of the aviation DST mentioned in Section 2.2 for example, pilots themselves discussed how continuously supporting their situation awareness would help them build trust into the system [24].

4 CONCLUSION

In summary, we suggest that trust calibration is often too fragile to be the primary mechanism for managing brittle AI performance. We argue that it is important to design AI for appropriation first so that human-AI interactions can be resilient against conditions outside of the AI model or designers' expectations. We observe that systems that try to solve tasks directly without involving users in producing the end result are difficult to appropriate. Instead, AI systems should be designed to provide incremental support that is guided by users' intentions. This approach does not eliminate the need for trust calibration, but makes it potentially much easier, up to the point that appropriate trust may be established as a side effect as users engage actively with the incremental AI support. We have given a few successful literature and product examples for this strategy and propose to start from them and iterate over this class of designs, with a particular focus on qualitatively understanding how people use and appropriate AI. The goal is to eventually arrive at a general design strategy for intelligent systems that will incorporate trust calibration by design, instead of as an add-on.

ACKNOWLEDGMENTS

This work was supported by the German Federal Ministry for Economic Affairs and Energy (BMWi) under the LuFo VI-1 program, project KIEZ4-0.

REFERENCES

[1] Kasun Amarasinghe, Kit T. Rodolfa, Sérgio Jesus, Valerie Chen, Vladimir Balayan, Pedro Saleiro, Pedro Bizarro, Ameet Talwalkar, and Rayid Ghani. 2022. On the importance of application-grounded experimental design for evaluating explainable ML methods. http://arxiv.org/abs/2206.13503 .

[2] Gagan Bansal, Tongshuang Wu, Joyce Zhou, Raymond Fok, Besmira Nushi, Ece Kamar, Marco Tulio Ribeiro, and Daniel Weld. 2021. Does the whole exceed its parts? The effect of AI explanations on complementary team performance. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems (CHI '21)*. ACM, Yokohama, Japan, 81:1–81:16. https://doi.org/10.1145/3411764.3445717 .

[3] Jeanette Blomberg, Aly Megahed, and Ray Strong. 2018. Acting on analytics: accuracy, precision, interpretation, and performativity. *Ethnographic Praxis in Industry Conference Proceedings* 2018, 1 (2018), 281–300. https://doi.org/10.1111/1559-8918.2018.01208 .

[4] Zana Buçinca, Alexandra Chouldechova, Jennifer Wortman Vaughan, and Krzysztof Z. Gajos. 2022. Beyond end predictions: stop putting machine learning first and design human-centered AI for decision support. In *Virtual Workshop on Human-Centered AI Workshop at NeurIPS (HCAI @ NeurIPS '22)*. Virtual Event, USA, 1–4.

[5] Zana Buçinca, Maja Barbara Malaya, and Krzysztof Z. Gajos. 2021. To trust or to think: cognitive forcing functions can reduce overreliance on AI in AI-assisted decision-making. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW1 (April 2021), 188:1–188:21. https://doi.org/10.1145/3449287 .

[6] Federico Cabitza, Andrea Campagner, and Carla Simone. 2021. The need to move away from agential-AI: empirical investigations, useful concepts and open issues. *International Journal of Human-Computer Studies* 155 (Nov. 2021), 102696:1–102696:11. https://doi.org/10.1016/j.ijhcs.2021.102696 .

[7] 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 (CHI '19)*. ACM, Glasgow, Scotland, UK, 4:1–4:14. https://doi.org/10.1145/3290605.3300234 .

[8] Alan Dix. 2007. Designing for appropriation. In *Proceedings of the 21st British HCI Group Annual Conference on People and Computers (BCS-HCI '07, Vol. 2)*. BCS Learning & Development Ltd., Lancaster, UK, 27–30. https://doi.org/10.14236/ewic/HCI2007.53

[9] Upol Ehsan, Samir Passi, Q. Vera Liao, Larry Chan, I.-Hsiang Lee, Michael Muller, and Mark O. Riedl. 2021. The who in explainable AI: how AI background shapes perceptions of AI explanations. https://doi.org/10.48550/arXiv.2107.13509 .

[10] Krzysztof Z. Gajos and Lena Mamykina. 2022. Do people engage cognitively with AI? Impact of AI assistance on incidental learning. In *27th International Conference on Intelligent User Interfaces (IUI '22)*. ACM, Helsinki, Finland, 794–806. https://doi.org/10.1145/3490099.3511138 .

[11] Sean Hollister. 2023. *The new Microsoft Bing will sometimes misrepresent the info it finds.* The Verge. Retrieved 2023-02-15 from https://www.theverge.com/2023/2/7/23589536/microsoft-bing-ai-chat-inaccurate-results

[12] 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 (CHI '21)*. ACM, Yokohama, Japan, 659:1–659:14. https://doi.org/10.1145/3411764.3445385 .

[13] Sérgio Jesus, Catarina Belém, Vladimir Balayan, João Bento, Pedro Saleiro, Pedro Bizarro, and João Gama. 2021. How can I choose an explainer? An application-grounded evaluation of post-hoc explanations. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT '21)*. ACM, Virtual Event, Canada, 805–815. https://doi.org/10.1145/3442188.3445941

[14] Anna Kawakami, Venkatesh Sivaraman, Hao-Fei Cheng, Logan Stapleton, Yanghuidi Cheng, Diana Qing, Adam Perer, Zhiwei Steven Wu, Haiyi Zhu, 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 2022 CHI Conference on Human Factors in Computing Systems (CHI '22)*. ACM, New Orleans, LA, USA, 52:1–52:18. https://doi.org/10.1145/3491102.3517439 .

[15] Anna Kawakami, Venkatesh Sivaraman, Logan Stapleton, Hao-Fei Cheng, Adam Perer, Zhiwei Steven Wu, Haiyi Zhu, and Kenneth Holstein. 2022. "Why do I care what's similar?" Probing challenges in AI-assisted child welfare decision-making through worker-AI interface design concepts. In *Designing Interactive Systems Conference (DIS '22)*. ACM, Virtual Event Australia, 454–470. https://doi.org/10.1145/3532106.3533556

[16] G. Klein, B. Moon, and R.R. Hoffman. 2006. Making sense of sensemaking 1: alternative perspectives. *IEEE Intelligent Systems* 21, 4 (July 2006), 70–73. https://doi.org/10.1109/MIS.2006.75 .

[17] Vivian Lai, Chacha Chen, Q. Vera Liao, Alison Smith-Renner, and Chenhao Tan. 2021. Towards a science of human-AI decision making: a survey of empirical studies. https://doi.org/10.48550/arXiv.2112.11471 .

[18] Markus Langer, Tim Hunsicker, Tina Feldkamp, Cornelius J. König, and Nina Grgić-Hlača. 2022. "Look! It's a computer program! It's an algorithm! It's AI!": does terminology affect human perceptions and evaluations of algorithmic decision-making systems?. In *CHI Conference on Human Factors in Computing Systems (CHI '22)*. ACM, New Orleans, LA, USA, 581:1–581:28. https://doi.org/10.1145/3491102.3517527 .

[19] Mahsan Nourani, Chiradeep Roy, Jeremy E Block, Donald R Honeycutt, Tahrima Rahman, Eric Ragan, and Vibhav Gogate. 2021. Anchoring bias affects mental model formation and user reliance in explainable AI systems. In *Proceedings of the 26th International Conference on Intelligent User Interfaces (IUI '21)*. ACM, College Station, TX, USA, 340–350. https://doi.org/10.1145/3397481.3450639 .

[20] Daniel Omeiza, Helena Webb, Marina Jirotka, and Lars Kunze. 2022. Explanations in autonomous driving: a survey. *IEEE Transactions on Intelligent Transportation Systems* 23, 8 (Aug. 2022), 10142–10162. https://doi.org/10.1109/TITS.2021.3122865 .

[21] Philipp Schmidt and Felix Biessmann. 2020. Calibrating human-AI collaboration: impact of risk, ambiguity and transparency on algorithmic bias. In *Machine Learning and Knowledge Extraction (CD-MAKE 2020)*. Springer International Publishing, Dublin, Ireland, 431–449. https://doi.org/10.1007/978-3-030-57321-8_24 .

[22] Ben Shneiderman. 2020. Design lessons from AI's two grand goals: human emulation and useful applications. *IEEE Transactions on Technology and Society* 1, 2 (June 2020), 73–82. https://doi.org/10.1109/TTS.2020.2992669 .

[23] Venkatesh Sivaraman, Leigh A. Bukowski, Joel Levin, Jeremy M. Kahn, and Adam Perer. 2023. Ignore, trust, or negotiate: understanding clinician acceptance of AI-based treatment recommendations in health care. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (CHI '23)*. ACM, Hamburg, Germany, 1–18. https://doi.org/10.1145/3544548.3581075

[24] Cara Storath, Zelun Tony Zhang, Yuanting Liu, and Heinrich Hussmann. 2022. Building trust by supporting situation awareness: exploring pilots' design requirements for decision support tools. In *CHI TRAIT '22: Workshop on Trust and Reliance in Human-AI Teams at CHI 2022*. New Orleans, LA, USA, 1–12.

[25] Maxwell Szymanski, Martijn Millecamp, and Katrien Verbert. 2021. Visual, textual or hybrid: the effect of user expertise on different explanations. In *Proceedings of the 26th International Conference on Intelligent User Interfaces (IUI '21)*. ACM, College Station, TX, USA, 109–119. https://doi.org/10.1145/3397481.3450662 .

[26] Helena Vasconcelos, Gagan Bansal, Adam Fourney, Q. Vera Liao, and Jennifer Wortman Vaughan. 2022. Generation probabilities are not enough: improving error highlighting for AI code suggestions. In *Virtual Workshop on Human-Centered AI Workshop at NeurIPS (HCAI @ NeurIPS '22)*. Virtual Event, USA, 1–4.

[27] Xinru Wang and Ming Yin. 2021. Are explanations helpful? A comparative study of the effects of explanations in AI-assisted decision-making. In *Proceedings of the 26th International Conference on Intelligent User Interfaces (IUI '21)*. ACM, College Station, TX, USA, 318–328. https://doi.org/10.1145/3397481.3450650 .

[28] David D. Woods. 2016. The risks of autonomy: Doyle's Catch. *Journal of Cognitive Engineering and Decision Making* 10, 2 (June 2016), 131–133. https://doi.org/10.1177/1555343416653562 .

[29] David D. Woods. 2018. The theory of graceful extensibility: basic rules that govern adaptive systems. *Environment Systems and Decisions* 38, 4 (Dec. 2018), 433–457. https://doi.org/10.1007/s10669-018-9708-3 .

[30] Ming Yin, Jennifer Wortman Vaughan, and Hanna Wallach. 2019. Understanding the effect of accuracy on trust in machine learning models. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (CHI '19)*. ACM, Glasgow, Scotland, UK, 1–12. https://doi.org/10.1145/3290605.3300509 .

[31] Qiaoning Zhang, Matthew L Lee, and Scott Carter. 2022. You complete me: human-AI teams and complementary expertise. In *CHI Conference on Human Factors in Computing Systems (CHI '22)*. ACM, New Orleans, LA, USA, 114:1–114:28. https://doi.org/10.1145/3491102.3517791 .

[32] 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)*. ACM, Bolzano, Italy, 18:1–18:5. https://doi.org/10.1145/3464385.3464696

[33] Zelun Tony Zhang, Cara Storath, Yuanting Liu, and Andreas Butz. 2023. Resilience through appropriation: pilots' view on complex decision support. In *Proceedings of the 28th International Conference on Intelligent User Interfaces (IUI '23)*. ACM, Sydney, NSW, Australia, 1–13. https://doi.org/10.1145/3581641.3584056