# Is Overreliance on AI Provoked by Study Design?

Zelun Tony Zhang[1,2][0000−0002−4544−7389], Sven Tong[2], Yuanting Liu[1][0000−0002−8651−6272], and Andreas Butz[2][0000−0002−9007−9888]

[1] fortiss GmbH, Research Institute of the Free State of Bavaria, Munich, Germany
{zhang,liu}@fortiss.org
[2] LMU Munich, Germany {sven.tong@campus,butz@ifi}.lmu.de

**Abstract.** Recent studies found that humans tend to overrely on AI when making decisions with AI support. AI explanations were often insufficient as mitigation, and sometimes even increased overreliance. However, typical AI-assisted decision-making studies consist of long series of decision tasks, potentially causing complacent behavior, and not properly reflecting many real-life scenarios. We therefore raise the question whether these findings might be favored by the design of these studies. In a first step to answer this question, we compared different study designs in an experiment and found indications that observations of overreliance might indeed be favored by common study designs. Further research is needed to clarify to what extent overreliance can be attributed to study designs rather than more fundamental human-AI interaction issues.

**Keywords:** human-AI interaction · AI-assisted decision-making · explainable AI · overreliance.

## 1 Introduction

Artificial intelligence (AI) is increasingly used to support human decisions, often in high-stakes domains such as healthcare, finance, or criminal justice (e.g. [2,6,18]). The hope is that AI complements human decision-making, given the supposedly complementary strengths and weaknesses of humans and machines [7,9]. However, recent studies repeatedly demonstrate that humans are prone to overrely on AI, i.e. they adopt AI outputs, even when they are flawed [1,2,3,6,8]. To address this issue, a common approach is to provide explanations of AI outputs. The reasoning is that by explaining how the AI comes to a result, humans should be able to better calibrate their reliance on the AI [17]. However, several studies show that in many cases, AI explanations may even increase blind trust in AI, rather than improve calibration [1,2,8,13]. Two recent studies [3,5] indicate that the reason for this effect is that people do not engage analytically with AI decision support, instead relying on fast but error-prone heuristic thinking [10].

Studies of AI-assisted decision-making typically involve a series of tasks which participants have to solve with the support of an AI model. Typically, these series are quite long, with up to 50 tasks being common, as shown in Table 1.

With such tiring task series lengths, one can reasonably suspect that participants become complacent over time, reducing their analytic engagement with the AI, and therefore increasing overreliance. At the same time, these long, intensive task series do not reflect well many real-world scenarios of AI-assisted decision-making. This raises the following question:

> **RQ**: Is overreliance indeed a fundamental issue of AI-assisted decision-making and explainable AI, or are the observations of overreliance in recent studies rather provoked by their long task series?

Table 1: Examples of studies that found overreliance in AI-assisted decision-making, along with the type and number of tasks participants had to solve.

| Publication | Study task | # Tasks |
|---|---|---|
| Bansal et al. [1] | Sentiment classification | 50 |
| | Law School Admission Test | 20 |
| Buçinca et al. [3] | Nutrition assessment | 26 |
| Green and Chen [6] | Recidivism risk assessment | 40 |
| | Loan risk assessment | 40 |
| Lai and Tan [11] | Deception detection | 20 |
| Liu et al. [12] | Recidivism prediction | 20 |
| | Profession classification | 20 |
| Schmidt and Biessmann [15] | Sentiment classification | 50 |
| Wang and Yin [17] | Recidivism prediction | 32 |
| | Forest cover prediction | 32 |

In this paper, we present a first attempt at answering this question. In particular, we approached the question through two novel study elements:

1. If long task series contribute to overreliance, the tendency to overrely should increase with the progression of the task session. We therefore employed a study design which allows us to measure how the tendency to overrely develops over the course of the task series (Section 2.2).
2. If long task series contribute to overreliance, the tendency to overrely should be less pronounced in shorter series. We therefore compare the common study design of a single, long task session to a design where participants solve the tasks in multiple short sessions (Section 2.3).

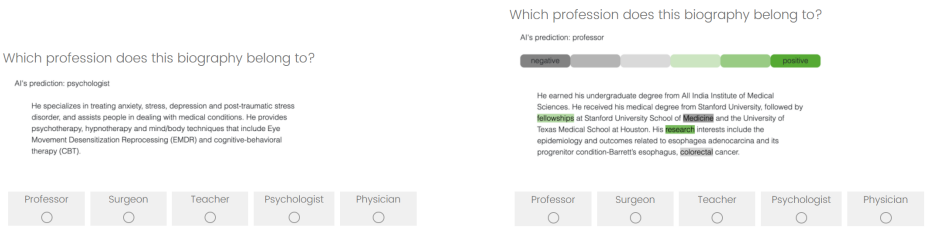
## 2 Experimental setup

### 2.1 Study task, apparatus and procedure

We followed Liu et al. [12] in choosing profession classification as the study task, since it does not require participants to have special knowledge or skills. The task

also bears some resemblance to AI applications in human resources, a domain where the stakes for the outcome of AI-assisted decision-making are high.

The task was based on a dataset by De-Arteaga et al. [4], consisting of short biographies scraped from the internet and each labeled with one of 29 occupations. The participants' task was to read a series of 50 biographies and to determine the occupations of the described persons, as shown in Fig. 1. To keep the task manageable, we limited the choice of occupations to the same five as in Liu et al. [12]. Above each biography, participants saw the prediction of a logistic regression model. Depending on the study condition (Section 2.3), participants also saw explanations for the predictions, generated with LIME [14] and visualized through color-coded text highlights, similar to Liu et al. [12].




(a) Without explanation. (b) With explanation.

Fig. 1: Examples of the occupation classification task presented to participants.

The study was set up as an online survey, distributed via university mailing lists and the online research platform SurveyCircle [16]. It started with a demographic questionnaire and an introduction to the task. After completing all tasks, participants answered an exit survey in which they could provide free text feedback. In addition to the 50 study tasks, we included two attention checks. Completing the study took 22.54 minutes on average. Each participant received a 10€ Amazon voucher as compensation. As an incentive to perform accurately, the most accurate participants could win an additional 5€ Amazon voucher.

### 2.2 Measures

We measured *agreement* with the AI and *overreliance* in two different ways: *per participant* and *per task*, i.e. as development over the course of the task series. Agreement per participant was measured as the share of all tasks in which a participant's answer was the same as the AI prediction. Overreliance was measured as the share of tasks in which a participant agreed with a wrong AI prediction. This conforms to measures commonly used in related work (e.g. [1,3,11,12,15,17]).

To measure the development throughout the task series, we divided the 50 tasks into ten blocks with five tasks each. Each block contained exactly one wrong AI prediction. This means that participants experienced an AI accuracy of 80%, which corresponds to the 86% test set accuracy of our logistic regression

model. The order of the blocks and the order of the questions within the blocks were both randomized. This yielded one measurement point for overreliance for each block of five tasks. For each block, we measured overreliance as the share of participants who agreed with the wrong AI prediction in that block. The agreement for each task was measured similarly as the share of participants whose answer was the same as the AI prediction. Lastly, we also recorded the *time* participants took for each task.

### 2.3 Study design and conditions

We employed a 2x2 between-subject design. The first factor was whether participants had to solve all 50 tasks in a single session (*single session group*—SSG) or in multiple short sessions (*multiple sessions group*—MSG). The SSG condition reflected the study design that is commonly used in related work and that we suspected to induce complacent behavior among participants. The MSG condition was meant to be less tiring for participants by keeping individual sessions short. In each session, participants would solve only one of the ten blocks of five tasks described in Section 2.2. After finishing one session, participants had to wait a minimum of one hour before the link to the next session was sent to them. Once they received the new link, participants were free to choose when to solve the task block. If a participant did not submit the current task block within 24 hours, they would receive a reminder message. We sent out session links and reminder messages via WhatsApp to make participation more convenient.

Previous studies showed that explanations sometimes increase participants' overreliance. We wanted to investigate whether this applied to participants in the MSG condition as well. If participants in the MSG condition were less complacent, they might engage with explanations more analytically, possibly leading to improved trust calibration. The second factor was therefore whether participants would see explanations for model outputs or not (see Fig. 1).

## 3 Results

After filtering out drop-outs, submissions that failed the attention checks, and other invalid submissions, the number of participants was 47 (average age: 30.1 years, 20 female, 27 male). On average, participants' self-assessed English level was moderately high (3.78 on a five-point Likert scale with 1=basic, 5=native speaker). Participants' average self-assessed AI expertise was moderate (2.91 on a five-point Likert scale with 1=no expertise, 5=expert).

Fig. 2 shows the *agreement* with AI and *overreliance* in all four conditions per participant. While *agreement* was slightly higher with explanations, the main effect of explanations was not significant according to a two-way ANOVA test, $F(1, 43) = 1.547, p = 0.220$. Both the main effect of the factor *session*, $F(1, 43) = 0.150, p = 0.701$, as well as the interaction effect between *explanation* and *session*, $F(1, 43) = 0.253, p = 0.618$, were also not significant. For *overreliance*, there appeared to be a more pronounced interaction effect between the

factors *explanation* and *session*: Overreliance did not differ between the two SSG conditions, but was higher with explanations than without in the MSG conditions, which was against our expectation. However, this interaction effect was not significant according to a two-way ANOVA test, $F(1, 43) = 0.905, p = 0.347$. The main effects of *explanation*, $F(1, 43) = 0.498, p = 0.484$, and *session*, $F(1, 43) = 1.636, p = 0.208$, were also not significant.

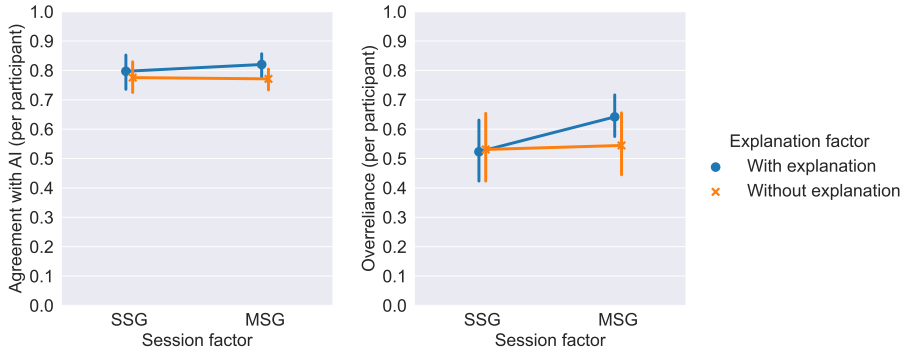



Fig. 2: Participants' agreement with AI predictions (left) and their overreliance (right) per participant. Error bars represent 95% confidence intervals.

Figures 3–5 show how participants' *overreliance*, their *agreement* with AI, and the *time* they took on average for a task developed throughout the task series in each of the four conditions. We analyzed these results using linear regression, as shown in Tables 2–4. We included the position of a task (or task block for overreliance), the session factor, and the explanation factor as main effects in each model. We further included interaction effects into the models where the data suggested the possible presence of interactions. While we performed linear regression with ordinary least squares for *agreement* and *time*, we resorted to robust linear regression using iterated re-weighted least squares (IRLS) and Huber's T norm for *overreliance* due to the inherently smaller number of measurement points and the resulting larger impact of outliers.

The most prominent observation is that in general, both *overreliance* and *agreement* significantly increased throughout the task series, while the *time to solve a task* decreased significantly. This suggests that participants indeed spent less effort on the tasks as the series progressed, as we expected. However, differently than expected, this observation applies to both the SSG and the MSG conditions. One notable exception is that *overreliance* (Fig. 3 and Table 2) slightly decreased throughout the task series in the MSG condition without explanations, while it significantly increased with explanations. While this interaction effect was significant, it has to be interpreted with caution due to the small number of participants and the resulting noise in the data. Also in contrast to our expec-

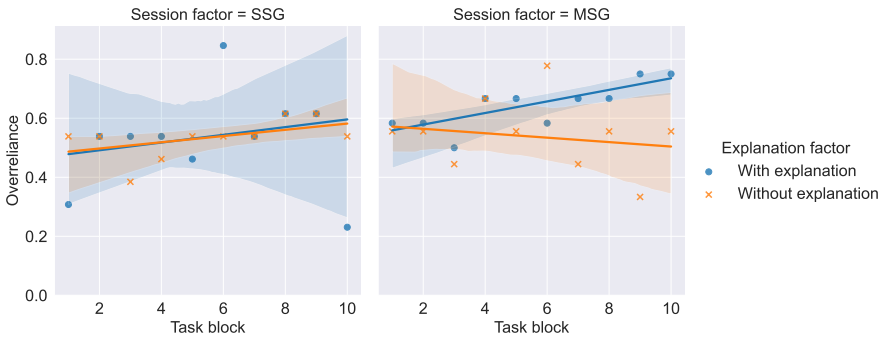



Fig. 3: Participants' overreliance over each task block. The lines represent linear regressions for each condition, the translucent bands around the lines represent 95% confidence intervals.

Table 2: Output of robust linear regression model for *overreliance*, using IRLS and Huber's T norm with median absolute deviation scaling. (*) indicates statistical significance at $\alpha = 0.05$.

| | Coefficient | Std. error | $z$ | $p$ |
|---|---|---|---|---|
| (Intercept) | 0.4537 | 0.032 | 14.328 | 0.000 |
| Session factor [MSG] | 0.1078 | 0.031 | 3.529 | 0.000* |
| Explanation factor [with] | 0.0026 | 0.031 | 0.087 | 0.931 |
| Task block | 0.0147 | 0.005 | 3.208 | 0.001* |
| Task block * Session factor [MSG] * Explanation factor [without] | -0.0194 | 0.007 | -2.931 | 0.003* |

tation, participants in the MSG conditions were on average significantly more overreliant than in the SSG conditions.

Some more subtle effects could be observed for the *agreement* with AI (Fig. 4 and Table 3). Explanations appeared to induce higher agreement with the AI for both SSG and MSG; however, the effect was not significant. There was also no statistically significant difference between SSG and MSG in terms of agreement with AI. The slight interaction between the explanation factor and task position in the SSG conditions visible in Fig. 4 was not significant either.

As for the *time to solve a task* (Fig. 5 and Table 4), MSG and SSG conditions differed significantly, with MSG participants taking significantly more time. The time participants took also decreased significantly faster in the MSG conditions. The reason for these differences is not clear, although we suspect that they might be the result of MSG participants using their smaller smartphone screens. The explanation factor had no significant effect on the time participants took, neither as a main effect, nor in interaction with the task position.

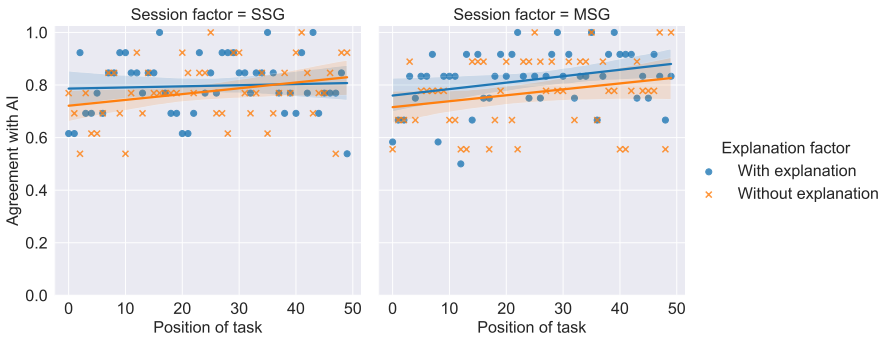



Fig. 4: Participants' agreement with AI predictions over each task. The lines represent linear regressions for each condition, the translucent bands around the lines represent 95% confidence intervals.

Table 3: Output of linear regression model for *agreement with AI*, $R^2 = 0.076, F(4, 195) = 4.033, p = 0.004 < 0.05$. (*) indicates statistical significance at $\alpha = 0.05$.

| | Coefficient | Std. error | $t$ | $p$ |
|---|---|---|---|---|
| (Intercept) | 0.7138 | 0.024 | 29.652 | 0.000 |
| Session factor [MSG] | 0.0094 | 0.016 | 0.578 | 0.564 |
| Explanation factor [with] | 0.0547 | 0.032 | 1.706 | 0.090 |
| Position of task | 0.0022 | 0.001 | 2.802 | 0.006* |
| Position of task * Explanation factor [with] | -0.0008 | 0.001 | -0.705 | 0.482 |

## 4 Discussion and outlook

Our results suggest that participants indeed spend significantly less effort as they progress through the task series of a typical AI-assisted decision-making study: Both participants' agreement with the AI and their overreliance increase throughout the task series, while the time they spend on a task decreases. Yet, by itself, this observation is not enough to conclude that overreliance in prior work was induced by the long task series in those studies. An alternative explanation could be, for instance, that people generally gain (potentially unjustified) trust into AI over time, irrespective of how tasks are presented to them. This would be a more fundamental issue for AI-assisted decision-making.

Comparing the single session with the multiple sessions study design was meant to enable a more conclusive interpretation of the above observation. However, there was no clear difference between the conditions. A possible reason could be that our setup did not have the intended effect. We aimed to make the multiple sessions conditions less tiring for participants by giving only five tasks at a time, by allowing them to start each session according to their own schedule,

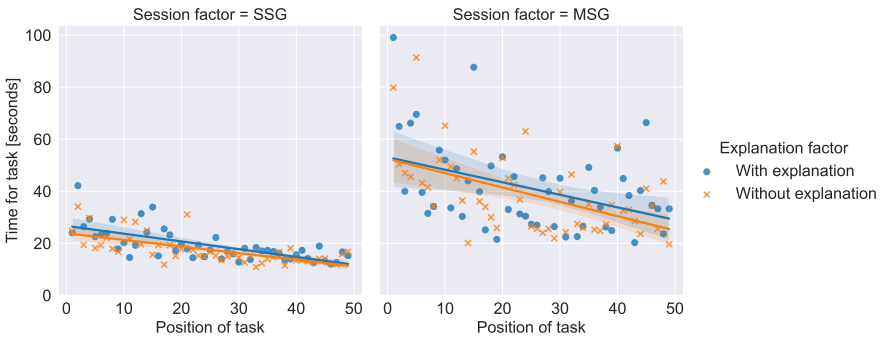



Fig. 5: Average time participants took to solve each task. The lines represent linear regressions for each condition, the translucent bands around the lines represent 95% confidence intervals.

Table 4: Output of linear regression model for *time for task*, $R^2 = 0.586, F(5, 190) = 53.78, p = 1.34\text{e}{-34} < 0.05$. (*) indicates statistical significance at $\alpha = 0.05$.

| | Coefficient | Std. error | $t$ | $p$ |
|---|---|---|---|---|
| (Intercept) | 24.4063 | 2.636 | 9.258 | 0.000 |
| Session factor [MSG] | 27.5137 | 3.044 | 9.038 | 0.000* |
| Explanation factor [with] | 1.7126 | 3.044 | 0.563 | 0.574 |
| Position of task | -0.2826 | 0.092 | -3.079 | 0.002* |
| Position of task * Session factor [MSG] | -0.2400 | 0.106 | -2.264 | 0.025* |
| Position of task * Explanation factor [with] | 0.0129 | 0.106 | 0.122 | 0.903 |

and by making access convenient via their smartphones. Still, free text feedback in the exit surveys reveals that participants were annoyed by the large number of sessions. Hence, we assume that the multiple sessions design did not differ enough from the single session design with regard to complacency. It therefore remains unclear whether the observed increase in overreliance is induced by typical study designs, or if it is a more fundamental issue of AI-assisted decision-making.

We think the question posed in this paper merits further investigation, since the answer would be crucial for the interpretation of previous results and the design of future studies. This work presents a first attempt with a small number of participants. Apart from recruiting more participants, a future follow-up study needs to devise a better way to administer the decision tasks in a less tiring manner. On the other hand, our task-block-based setup is a promising direction for future studies, enabling the measurement of overreliance over the course of the task series instead of merely aggregated over all tasks. This potentially provides more nuanced insights into how AI impacts human decision-making.

## References


1. Bansal, G., Wu, T., Zhou, J., Fok, R., Nushi, B., Kamar, E., Ribeiro, M.T., Weld, D.: 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. pp. 81:1–81:16. CHI '21, ACM, Yokohama, Japan (May 2021). https://doi.org/10.1145/3411764.3445717
2. Bussone, A., Stumpf, S., O'Sullivan, D.: The role of explanations on trust and reliance in clinical decision support systems. In: Proceedings of the 2015 International Conference on Healthcare Informatics. pp. 160–169. ICHI 2015, IEEE, Dallas, TX, USA (Oct 2015). https://doi.org/10.1109/ICHI.2015.26
3. Buçinca, Z., Malaya, M.B., Gajos, K.Z.: 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), 188:1–188:21 (Apr 2021). https://doi.org/10.1145/3449287
4. De-Arteaga, M., Romanov, A., Wallach, H., Chayes, J., Borgs, C., Chouldechova, A., Geyik, S., Kenthapadi, K., Kalai, A.T.: Bias in bios: a case study of semantic representation bias in a high-stakes setting. In: Proceedings of the Conference on Fairness, Accountability, and Transparency. pp. 120–128. FAT* '19, ACM, Atlanta, GA, USA (Jan 2019). https://doi.org/10.1145/3287560.3287572
5. Gajos, K.Z., Mamykina, L.: Do people engage cognitively with AI? Impact of AI assistance on incidental learning. In: 27th International Conference on Intelligent User Interfaces. pp. 794–806. IUI '22, ACM, Helsinki, Finland (Mar 2022). https://doi.org/10.1145/3490099.3511138
6. Green, B., Chen, Y.: The principles and limits of algorithm-in-the-loop decision making. Proceedings of the ACM on Human-Computer Interaction **3**(CSCW), 50:1–50:24 (Nov 2019). https://doi.org/10.1145/3359152
7. Guszcza, J.: Smarter together: why artificial intelligence needs human-centered design. Deloitte Review (22), 15 (Jan 2018)
8. Jacobs, M., Pradier, M.F., McCoy, T.H., Perlis, R.H., Doshi-Velez, F., Gajos, K.Z.: How machine-learning recommendations influence clinician treatment selections: the example of antidepressant selection. Translational Psychiatry **11**(1), 108:1–108:9 (Jun 2021). https://doi.org/10.1038/s41398-021-01224-x
9. Jarrahi, M.H.: Artificial intelligence and the future of work: human-AI symbiosis in organizational decision making. Business Horizons **61**(4), 577–586 (Jul 2018). https://doi.org/10.1016/j.bushor.2018.03.007
10. Kahneman, D.: Thinking, fast and slow. Farrar, Straus and Giroux, New York (2011)
11. Lai, V., Tan, C.: 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. pp. 29–38. FAT* '19, ACM, Atlanta, GA, USA (Jan 2019). https://doi.org/10.1145/3287560.3287590
12. Liu, H., Lai, V., Tan, C.: Understanding the effect of out-of-distribution examples and interactive explanations on human-AI decision making. Proceedings of the ACM on Human-Computer Interaction **5**(CSCW2), 408:1–408:45 (Oct 2021). https://doi.org/10.1145/3479552
13. Poursabzi-Sangdeh, F., Goldstein, D.G., Hofman, J.M., Vaughan, J.W., Wallach, H.: Manipulating and measuring model interpretability. In: Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems. pp. 237:1–237:52. CHI '21, ACM, Yokohama, Japan (May 2021). https://doi.org/10.1145/3411764.3445315

14. Ribeiro, M.T., Singh, S., Guestrin, C.: "Why should I trust you?": explaining the predictions of any classifier. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. pp. 1135–1144. KDD '16, ACM, San Francisco, CA, USA (Aug 2016). https://doi.org/10.1145/2939672.2939778
15. Schmidt, P., Biessmann, F.: Calibrating human-AI collaboration: impact of risk, ambiguity and transparency on algorithmic bias. In: Machine Learning and Knowledge Extraction. pp. 431–449. CD-MAKE 2020, Springer International Publishing, Dublin, Ireland (Aug 2020). https://doi.org/10.1007/978-3-030-57321-$8_24$
16. SurveyCircle: Research website SurveyCircle. Published 2016. (2022), `https://www.surveycircle.com`
17. Wang, X., Yin, M.: 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. p. 11. IUI '21, ACM, College Station, TX, USA (Apr 2021). https://doi.org/10.1145/3397481.3450650
18. Yang, Q., Steinfeld, A., Zimmerman, J.: Unremarkable AI: fitting intelligent decision support into critical, clinical decision-making processes. In: Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems. pp. 238:1–238:11. CHI '19, ACM, Glasgow, Scotland, UK (May 2019). https://doi.org/10.1145/3290605.3300468