User-friendly AI-based Decision Support Systems (DSSs)

Natalia Echeverry Giraldo
MSHCI Directed Final Project - Spring 2024
Rochester Institute of Technology
ne9320@rit.edu

INTRODUCTION

Public and private institutions increasingly rely on data science and visual analytics to transform vast amounts of raw data into decision support systems (DSSs) with practical insights for action. These DSSs usually take the form of AI-based data dashboards and digital assistants. Institutions wanting to adopt a more data-driven approach in their operations and services have evolved from building dashboards with descriptive and inferential statistics to incorporating predictive analytics models. Increasing our understanding of how users interact and interpret information from predictive analytics is essential for "complementary computing" [12], which is the notion that combining human and machine intelligence is more effective than relying on humans or AI alone.

DSSs with predictive analytics were commonly used by users with advanced quantitative and statistical skills. Nowadays, users with various levels of such skills are required to utilize AI-based DSSs to make decisions. Decisions initially made by human judgment alone now have an "algorithm-in-the-loop" [6].

ETHICAL CONSIDERATIONS OF USING AI-BASED DSSs

Combining human and artificial intelligence (AI) is more effective than relying on humans or AI alone. This has moved public and private institutions to incorporate AI-based DSSs to support workers in decision-making. Examples of this are the implementation of risk models in health care, social services [2], and the judicial system [23]. Predictive classification models can be inflexible systems that perpetuate racial biases and discrimination [3, 11]. Emphasizing the accuracy, reliability, fairness, and efficiency of such models is crucial. However, successful implementation of an AI-based DSS requires careful consideration of human factors that intervene when users interact with such systems [6, 17].

HUMAN FACTORS IN AI-BASED DSSs

AI-based DSSs demand different cognitive processes from users. Common visual analytics heuristics and DSS design patterns may not apply when the system includes predictive analytics. Nourani et al. found that generic representations of uncertainty can be counterproductive and possibly trigger misconceptions and cognitive biases in users as a coping mechanism [14]. Therefore, the design and implementation of AI-based DSSs require a renewed approach to the GUI design, risk communication strategies, and user training approaches [6, 7, 9, 19]. The search for risk communication heuristics and design patterns needs to consider findings such as Green et al., who found that AI explainability did not improve human performance [6] and that providing user feedback for human decision calibration decreased their accuracy [6].

Cognitive biases and interpretation errors

Cognitive biases are distorted perceptions and interpretations of objective information that can result in distortions or inaccuracies of judgment [7]. Cognitive biases that have been identified in research studies related to visual and predictive analytics are anchoring bias [22], automation bias [1, 18], confirmation bias [16], and interpretation error [6], among others.

Anchoring bias

Anchoring bias is the “anchoring effect” that happens when exposure to previous scores influences new estimates [2, 8, 22].

Automation bias

Automation bias is “the tendency of people to defer to automated technology when presented with conflicting information” [1]. Kawakami et. al. found that AFST users felt “discomfort around underlying assumptions of statistical predictions, such as case comparison and statistical notions of uncertainty.” [5] Their lack of understanding of the predictive model has led some of them to consistently defer to AFST’s risk scores [4, 15]. Reports of automation bias in COMPAS users have been made in the past and have resulted in the addition of a written disclaimer about the model's limitations and recommendation to exercise discretion when assessing a risk score concerning a defendant [6, 15, 22].

Confirmation bias

Dare et al. defined confirmation bias as the tendency to seek evidence that supports preferred views [24].

Interpretation errors

Contextual inquiries and experiments have shown that decision-makers often find it challenging to read risk predictions [4, 6]. This is usually because users do not have an accurate mental model of the underlying risk model of the AI-based DSS [4]. The mismatch between the user’s mental model and the AI-based DSS model can cause interpretation errors. Green et al. found that many participants in their study treated probability percentages on a scale of zero to hundred percent as binary; they shifted toward 0% or 100% [6]. Another example is the users' misunderstanding of Allegheny County’s AFST targets. AFST provides a risk score for the likelihood of an unfavorable outcome

happening in the next two years, but many users believed it predicts risk in the short term [4].

USERS' PERCEPTIONS OF THE AI-BASED DSS

In the same way that cognitive biases and interpretation errors can affect the quality of decisions, negative perceptions of the AI-based DSS in users can affect the quality of decisions as well. This issue has been described as users' confidence in the fairness of the risk predictions. The transparency of a system has been associated with its fairness [5, 25]. Explanatory approaches, communicating uncertainty, and model limitations are forms of transparency that may encourage users to trust the tool [5, 21, 26].

Awareness of AI-based DSS limitations and error rates can empower users to correct flawed predictions. De-Arteaga et al. analyzed users' responses to "erroneous algorithmic scores" provided by an early version of the AFST [15]. They found that users "are less likely to adhere to the machine's recommendation when the score displayed is an incorrect estimate of risk, even when overriding the recommendation requires supervisory approval." [15]. The assumption that users can identify errors and overriding recommendations [15] leads to the question: Can GUI design and the risk communication strategy make it easier for users to recognize prediction errors?

AI-BASED DECISION SUPPORT SYSTEMS DESIGN

Taking aside the fact that the underlying predictive models in AI-based DSSs can have many issues on their own [3, 11], it is important to recognize that while ensuring the accuracy, reliability, and fairness of such models is crucial, it is as important to consider practical aspects of the AI-based DSS design and implementation [6, 17]. The design and implementation of the graphical user interface (GUI) of AI-based DSSs differ from traditional DSSs because they communicate uncertainty as "statistical predictions" [5].

The use of predictive risk models that output a risk score has become prevalent in public and private institutions that have adopted AI-based DSS. Risk scores can be communicated numerically or verbally. As a probability percentage, ratio, single- or double-digit number, or as a text label. Whether the risk is communicated as a number or a text label, it could be supplemented by a graphical representation such as charts, linear and round gauge scales, and icon arrays, among others. Multiple combinations for the risk communication strategy are possible, however, ensuring the risk communication strategy use is appropriate for the user base of the AI-based DSS requires further experimentation [5].

BETWEEN-SUBJECTS STUDY

Hypotheses and variables

In this study, I aim to explore if the risk communication strategy makes any difference in the decisions users make and if there are there any risk communication strategies that contribute to mitigating cognitive biases and interpretation errors in users with low quantitative and statistical skills.
The GUI design of AI-based DSSs with underlying predictive classification models that output a risk score are the focus of this study. I will test if the risk communication strategy in an AI-based DSS has any effect on users' decisions. The null and alternative hypotheses are:

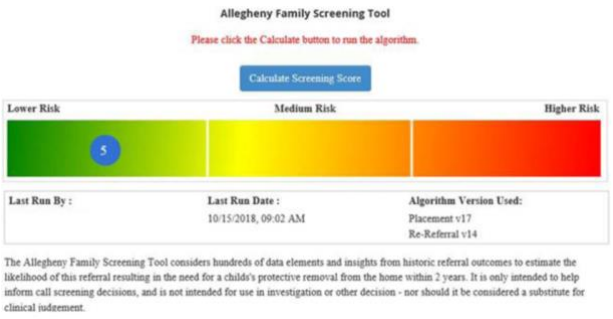
> **H0:** The risk communication strategy in the GUI of the AI-based DSS does not affect the user's decision.
> **H1:** The risk communication strategy in the GUI of the AI-based DSS affects the user's decision.

The dependent variable is the decision users make. The independent variable is the risk communication strategy used in the GUI design of the AI-based DSS. The risk communication strategies in this study include a risk score and its graphical representation.

"Real-world" AI-based DSS: The Allegheny Family Screening Tool (AFST)

The real-world AI-based DSS on which this study is based is the Allegheny Family Screening Tool (AFST) with an underlying predictive risk model (PMR) that outputs a risk score presented as a one- or two-digit number situated over a linear gauge scale with labels: "Lower Risk", "Medium Risk", and "Higher Risk" (Figure 1). I built four high-fidelity wireframes of an AI-based DSS modeled after the AFST [2]. Each of the four wireframes show a particular risk communication strategy (Table 1).

Figure 1. AFST V2 [2].



Participants

Call screeners of child protection hotlines usually have completed a bachelor's degree in social work or related fields and have some experience working with children and families. Because of their educational and professional background, it is uncommon for them to have substantial training in statistics. Given that the purpose of the study is to explore effective ways of communicating risk predictions to users with low and medium statistical skills, pair with the difficulty of working with call screeners directly, individuals older than 18 years old, that speak English, and have

completed some years of college or above are accepted as surrogate participants. A background in social work or a related field is preferred.

Online survey design and questionnaires

For this study, I administered an anonymous online survey with three main parts: (1) demographic, educational, and professional background questions, (2) a simulation of a call referral and case classification modeled after Allegheny County's child protection program, and (3) a set of system usability Likert-scale questions.

The first part of the survey asked participants for their age range, education level, and professional background. The second part of the survey, the simulation, included a wireframe of an AI-based DSS modeled after the Allegheny Family Screening Tool (AFST) (Figure 1), which I called the Family Screening Tool (FST) (Figure 2). Each participant was assigned a risk communication strategy randomly, using Qualtrics' built-in randomization function. There were four possible risk communication strategies created for this study: gauge scale with text label, gauge scale with percentage, icon array with text label, and icon array with percentage (Table 1).

Additional information such as participants' location (e.g., longitude and latitude, and IP address.), primarily language set up in their computer were collected automatically. This information was used in data cleaning to verify responses met the minimum acceptance criteria.

Simulation of a call referral and case classification with FST

After participants complete the first part of the survey, they are prompted with a short text describing a child protection program that requires call screeners to utilize an AI-based DSS when deciding how to classify a referral. The prediction target and error rate of the FST are communicated to participants. Participants are instructed to situate themselves in the role of a call screener.

> "The Department of Human Services manages a child protection hotline where people can report alleged cases of child abuse and neglect, also called referrals. The hotline is operated by call screeners who are required to collect information over the phone and utilize a risk assessment tool called the Family Screening Tool (FST). The FST is an AI-based tool that predicts the risk of an unfavorable outcome occurring to the child in the next two years. Call screeners are required to use FST and consider the risk level provided by the tool when deciding if there is a need for an intervention. The error rate measures the percentage of times the FST makes an incorrect risk prediction. The error rate of the Family Screening Tool is 5.5%. This survey will ask you to assume the role of a call screener in the following case scenario."

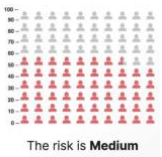
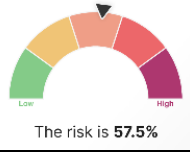
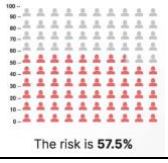
Then, they were prompted with the following scenario:

> "As a call screener, you receive a call referral. Dr. Patel, the pediatrician, informs you that Riley, the mother, brought her 13-year-old daughter, Remi, to the state hospital because she was refusing to eat more than a thousand calories daily. Dr. Patel examines Remi and finds that she is underweight. Additionally, Dr. Patel provides you with the parents' names, dates of birth, and state identification numbers. As the call screener, you enter this information into the Family Screening Tool and initiate the analysis. After completing the risk assessment, FST generates the following risk prediction. (Please refer to the screenshot below.)"

After participants had read the scenario, they were randomly assigned a wireframe of FST's with one of the risk communication strategies (Table 1). All risk communication strategies used a 57.5 risk score on a scale of 0 to 100. The case scenario does not present a clear-cut situation and the risk score is slightly above the midpoint. The error rate given was 5.5.%. These conditions aimed at encouraging participants to consider both the case and the risk score in their decision-making process. Then, they are asked to classify the case as low, medium, or high risk. This portion of the survey was presented as a Likert question. After participants selected a classification for the case, the survey asked them to type down the reason for the classification they chose.

Risk communication strategies: verbal versus numeric risk descriptions and graphical representations. Gresh et al. noted the limitation of relying upon a verbal description exclusively because of the difficulty of mapping it on a numeric scale [10]. Providing a graphical representation of the risk output in addition to a verbal or numeric risk description is considered a good practice [10]. However, graphical representations can be counterproductive if they are not selected according to the users' familiarity with them [17].

Table 1. Risk communication strategies.

	
Gauge scale with text label (GL)	Icon array with text label (IL)
	
Gauge scale with percentage (GP)	Icon array with percentage (IP)

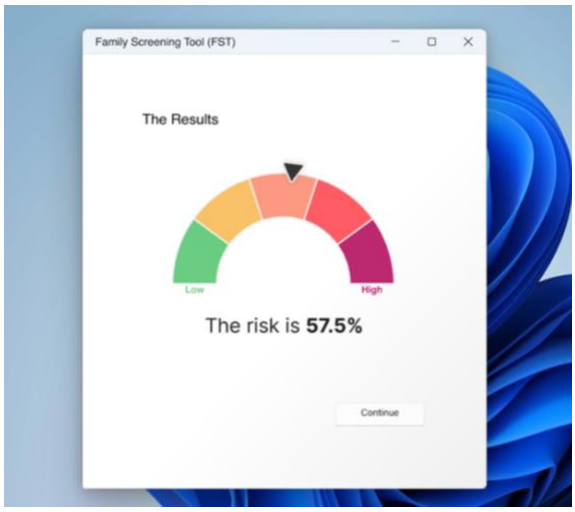
Gauge scale and icon array are the two graphical representations selected for this study. Gauge scales are pervasive in visual analytics, when they are used to communicate risk, they usually follow the speedometer metaphor that starts in low (green) and ends in high (red). The icon array is also used to represent risk graphically. Icon arrays are considered the best option for risk communication because they allow for a precise "discrete representation of risk" [13].

The risk communication strategies used for this study consist of a risk output communicated numerically or verbally and a graphical representation of it; a gauge scale or icon array (Table 1). I paired a verbal (i.e., text label) and numeric (i.e., percentage) description of the risk output with its graphical representation (i.e., gauge scale and icon array.) (Table 1).

Pilot study

Two users (T1 and T2) with research backgrounds and advanced numeracy skills tested the survey. They provided feedback on the settings and copy editing and FST image after completing the survey. They also shared screenshots of the FST image randomly assigned to them, which I used to test the survey randomization. T1 and T2 noted the risk communication strategy image was not sufficient to associate it with an AI-based DSS. I modified the images to simulate an FST window over a Windows virtual desktop (Figure 2).

Figure 2. FST wireframe.



Survey distribution and data cleaning.

A Qualtrics anonymous link was distributed through the subreddit r/SampleSize. An online Reddit community dedicated to research studies for school purposes where members can post links to surveys and polls. Participants were offered a $5 Amazon e-gift card as compensation for finishing the survey. The minimum acceptance criteria for receiving compensation were spending at least seven minutes completing the survey, providing a sensical answer to an open-ended question in the questionnaire, and passing Qualtrics' fraud detection filters [20]. Despite providing a clear statement of the minimum requirements for receiving compensation, many individuals encouraged by the compensation flooded the survey with fraudulent responses. Of the 493 responses received, 299 passed Qualtrics' quality response filters and language requirements [20]. However, 242 participants finished the survey while 57 did not. The minimum acceptance criteria for this study include a language requirement, a background in social work or a related field and having completed the survey. After data cleaning in RStudio, it was determined that a total of 225 responses met the minimum acceptance criteria.

DATA ANALYSIS AND SAMPLING

Participant demographics

The age range mode is 25-34 years old followed by 35-44 years old (Table 1). More than 60% of participants selected *Graduated 4-years College* as their highest level of education followed by *Postgraduate* with 20% and *Graduated 2-years College* with 15% (Table 2). All participants in this sample self-reported that they have formal training or experience in social work or a related field. These self-reported characteristics are within the acceptance criteria specified for this study.

Table 2. Participant demographics (n=225)

Age: What is your age range?	
18-24 years old	12
25-34 years old	119
35-44 years old	90
45-54 years old	4
Education: Highest grade or level of school completed.	
Some College	12
Graduated 2-year College	34
Graduated 4-year College	137
Postgraduate	42

Data sampling

Social work programs typically include statistical training in their curriculum. The extent of statistical training can vary depending on the program. College-level programs may include introductory courses on statistics and research methods, while post graduate programs typically have a stronger emphasis on statistics, including courses on descriptive and inferential statistics. Given that training in predictive statistics is not typically part of the curriculum even at the post-graduate level programs, participants in all education groups will be included in this study (Table 3).

Table 3. Highest level of school completed.

Groups	**Less than 4-year of college**	**Postgraduate**
Gauge scale with text label	48	6

Gauge scale with percentage	44	9
Icon array with text label	50	14
Icon array with percentage	41	13

Risk communication strategies and groups.

Participants were randomly assigned to one of the four risk communication strategies: gauge scale with label, gauge scale with percentage, icon array with label, and icon array with percentage. The risk prediction in the FST was the same for all four groups, 57.5 on a scale from 0 to 100. The only difference between groups was the risk communication strategy used in the FST.

Summary of Risk Classifications

The distribution of the risk classification in all groups was the same or similar for all groups. The mean for gauge with label, gauge with percentage and icon array with percentage groups is 2.4 with a standard deviation of 0.5 (Table 4). The icon array with text label group had a slightly different mean of 2.3 and a standard deviation of 0.6 presumably because of a bigger sample size (Table 4).

Table 4. Group Designation and Risk Classification

Risk Communication Strategy	**LR (n)**	**MR (n)**	**HR (n)**	**M**	**Mdn**	**Sd**
Gauge scale with text label (n=54)	2	31	21	2.4	2	0.5
Gauge scale with percentage (n=53)	1	27	25	2.4	2	0.5
Icon array with text label (n=64)	4	37	23	2.3	2	0.6
Icon array with percentage (n=54)	2	30	22	2.4	2	0.5

One-way ANOVA of Risk Classification

I applied one-way ANOVA to test the null hypothesis that the mean risk classification of all four groups is the same. The test indicates that there is no statistical evidence to reject the null hypothesis (p-value=0.5). The risk communication strategy does not seem to influence the decision the user makes. The mean risk classification across groups is statistically the same (Figure 3). However, this may be due to a small sample size (n=225) and to data quality issues. Participants of the online anonymous survey may have had the incentive to self-report characteristics that are not accurate because of the offer to receive a reward.

Participants' perception of FST

The survey prompted participants with Likert-scale questions about their agreement with statements about the FST after they had classified the case and typed their responses. The one-way ANOVA tests for each statement did not find any statistically significant difference between groups except for the final statement: "Overall, I consider the FST to be useful for the evaluation of child referrals" (Figure 4) (Annex 1).

Participants in the *gauge with text label* and *icon array with text label* rated the FST lower in the Likert-scale with means of 3.8 and 3.9 while the *gauge with percentage* and *icon array with percentage* rated the FST higher both with a mean of 4.2. The p-value of the one-way ANOVA test is 0.04 which shows that there is statistical evidence to infer that the risk communication strategy may influence users' trust in the usefulness of the tool. A possible explanation for the differences between groups is that users who were administered the risk communication strategies with a percentage perceived the tool, as more specialized and precise.

Figure 3. One-way ANOVA (Low risk = 1, Medium risk = 2, High risk = 3)

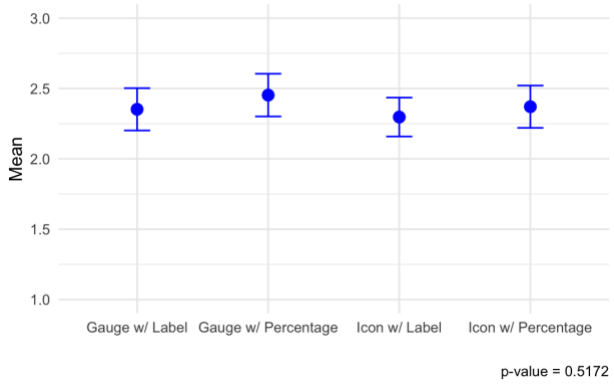
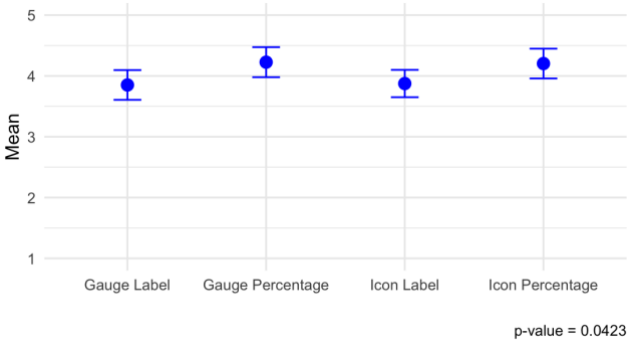


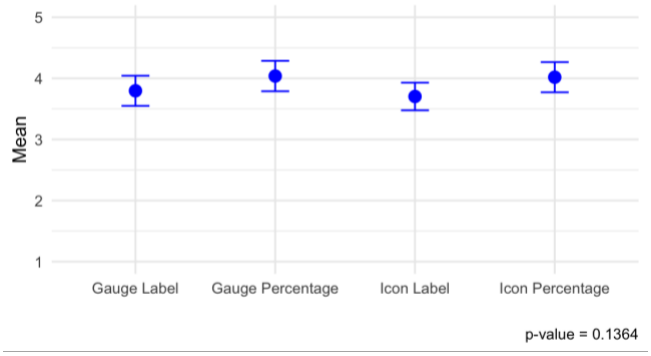
Figure 4. One-way ANOVA: "Overall, I consider the FST to be useful for the evaluation of child referrals." (Strongly agree = 5, Agree = 4, Neutral = 3, Disagree = 2, Strongly disagree 1)



A similar grouping appears in the one-way ANOVA test for the statement "All the information provided by the FST is useful and relevant." where groups *gauge scale with percentage* and *icon array with percentage* perceived the information in the FST as more useful and relevant compared to the other two groups (Figure 5). With a p-value of 0.1,

there is no statistical evidence to consider that the means are different. However, these results show similitudes between groups based on how the risk score is communicated. The graphical representation does not seem to have as much weight as the format of the risk score communication, whether is verbal or numeric.

Figure 5. One-way ANOVA: "All the information provided by the FST is useful and relevant." (Strongly agree = 5, Agree = 4, Neutral = 3, Disagree = 2, Strongly disagree 1)



Despite the lack of statistical significance, the *gauge scale with percentage* group had the highest mean rate compared to the other groups consistently in most Likert-scale statements such as: "The FST provided sufficient information", "The information in the FST is straightforward", and "I would not need the support of a supervisor or colleague to understand FST results". The "Overall, I consider the FST to be useful for the evaluation of child referrals" shared the mean of 4.2 with the icon array with percentage group. These statements have to do with users' trust in the statistical prediction given by the system and users' confidence in their capability of using FST to make decisions.

Participant's written responses

Participants in the icon array percentage share similarities with the gauge percentage group in terms of risk classification and explanations. P22 from the icon array percentage group, and P69 from the gauge percentage group classified the case as high risk, and wrote:

P22: "The risk rate is above average"
P69: "It's more than half"

CONCLUSIONS AND FUTURE WORK

Despite the result of the one-way ANOVA test does not show statistically significant results to reject the null hypothesis that the mean risk classification of all four groups is the same. I cannot necessarily conclude that risk communication strategy does not seem to influence the decision the user makes given the small sample size and to data quality issues. For future studies, pre-screening and eligibility surveys should not be neglected to ensure participants are representative of the user population of the study.

The ANOVA test results of the Likert-scale questions of the usability statements: "Overall, I consider the FST to be useful for the evaluation of child referrals" and "All the information provided by the FST is useful and relevant." Show similarity between groups based on how the risk score is communicated. A possible explanation is that users who were administered the risk communication strategies with a percentage perceived the tool, as more specialized and precise. Moreover, the significance of the one-way ANOVA tests for the "Overall, I consider the FST to be useful for the evaluation of child referrals" statement indicates that the risk communication strategy can influence users' perceptions of the usefulness of the AI-based DSS.

The risk communication strategies do not influence participants' decisions or case classification, but they do influence participants' perceptions of the usefulness of the system. Interestingly, participants who in the gauge scale with percentage and icon array with percentage rated FST agreed more with the usability statements. This indicates that participants may see numeric communication of a risk score as more reliable and transparent than the text label.

Conducting online or in-person interviews and simulations with think aloud protocols would help understand better how users interact and interpret statistical predictions. Interviews would also give the opportunity for an in depth understanding of users' quantitative and statistical skills, and their perceptions of such systems. Online anonymous surveys without pre-screening have limitations around sampling and data quality.

REFERENCES

[1] Alan R. Wagner, Jason Borenstein, and Ayanna Howard. 2018. Overtrust in the robotic age. Commun. ACM 61, 9 (September 2018), 22–24. https://doi.org/10.1145/3241365

[2] Allegheny County DHS. 2023. The Allegheny Family Screening Tool. https://www.alleghenycounty.us/Human-Services/News-Events/Accomplishments/Allegheny-Family-Screening-Tool.aspx. Retrieved 2023-05-04.

[3] Allison Woodruff, Sarah E. Fox, Steven Rousso-Schindler, and Jeffrey Warshaw. 2018. A Qualitative Exploration of Perceptions of Algorithmic Fairness. In Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (CHI '18). Association for Computing Machinery, New York, NY, USA, Paper 656, 1–14. https://doi.org/10.1145/3173574.3174230

[4] 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 CHI Conference on Human Factors in Computing Systems (CHI ’22), April 29-May 5, 2022, New Orleans, LA, USA. ACM, New York, NY, USA, 18 pages. https://doi.org/10.1145/3491102.3517439

[5] 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 throughWorker-AI Interface Design Concepts. In DIS Conference on Designing Interactive Systems (DIS’22), June 13 - June 17, 2022. ACM, New York, NY, USA, 17 pages. https://doi.org/10.1145/3532106. 3533556

[6] Ben Green and Yiling Chen. 2019. The Principles and Limits of Algorithm-in-the-Loop Decision Making. Proc. ACM Hum.-Comput. Interact. 3, CSCW, Article 50 (November 2019), 24 pages. https://doi.org/10.1145/3359152

[7] Charvi Rastogi, Yunfeng Zhang, Dennis Wei, Kush R. Varshney, Amit Dhurandhar, and Richard Tomsett. 2022. Deciding Fast and Slow: The Role of Cognitive Biases in AI-assisted Decision-making. Proc. ACM Hum.-Comput. Interact. 6, CSCW1, Article 83 (April 2022), 22 pages. https://doi.org/10.1145/3512930

[8] Cho et al, "The anchoring effect in decision-making with visual analytics," in 2017. DOI: 10.1109/VAST.2017.8585665.

[9] Gagan Bansal, Tongshuang Wu, Joyce Zhou, Raymond Fok, Besmira Nushi, Ece Kamar, Marco Tulio Ribeiro, and Daniel S. Weld. 2021. Does the Whole Exceed its Parts? The Efect of AI Explanations on Complementary Team Performance. In CHI Conference on Human Factors in Computing Systems (CHI ’21), May 8–13, 2021, Yokohama, Japan. ACM, New York, NY, USA, 16 pages. https://doi.org/10.1145/3411764.3445717

[10] Gresh, D., Deleris, L. A., Gasparini, L., & Evans, D. (2011). Visualizing risk. in Proceedings of IEEE Information Visualization Conference 2011 (InfoVis 2011), Providence, RI, USA. IEEE computer society.

[11] Ihudiya Finda Ogbonnaya-Ogburu, Angela D.R. Smith, Alexandra To, and Kentaro Toyama. 2020. Critical Race Theory for HCI. In Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems (CHI '20). Association for Computing Machinery, New York, NY, USA, 1–16. https://doi.org/10.1145/3313831.3376392

[12] Joon Sung Park, Rick Barber, Alex Kirlik, and Karrie Karahalios. 2019. A Slow Algorithm Improves Users’ Assessments of the Algorithm’s Accuracy. Proc. ACM Hum.-Comput. Interact. 3, CSCW, Article 102 (November 2019), 15 pages. https://doi.org/10.1145/3359204

[13] Kurz-Milcke, E., Gigerenzer, G. and Martignon, L. (2008), Transparency in Risk Communication. Annals of the New York Academy of Sciences, 1128: 18-28. https://doi.org/10.1196/annals.1399.004

[14] 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 26th International Conference on Intelligent User Interfaces (IUI '21). Association for Computing Machinery, New York, NY, USA, 340–350. https://doi.org/10.1145/3397481.3450639

[15] Maria De-Arteaga, Riccardo Fogliato, and Alexandra Chouldechova. 2020. A Case for Humans-in-the-Loop: Decisions in the Presence of Erroneous Algorithmic Scores. In Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems (CHI '20). Association for Computing Machinery, New York, NY, USA, 1–12. https://doi.org/10.1145/3313831.3376638

[16] Megha Srivastava, Hoda Heidari, and Andreas Krause. 2019. Mathematical Notions vs. Human Perception of Fairness: A Descriptive Approach to Fairness for Machine Learning. In The 25th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD ’19), August 4–8, 2019, Anchorage, AK, USA. ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3292500.3330664

[17] Naja Holten Møller, Irina Shklovski, and Thomas T. Hildebrandt. 2020. Shifting Concepts of Value: Designing Algorithmic Decision-Support Systems for Public Services. In Proceedings of the 11th Nordic Conference on Human- Computer Interaction: Shaping Experiences, Shaping Society (NordiCHI ’20), October 25–29, 2020, Tallinn, Estonia. ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3419249.3420149

[18] P. Robinette et al, "Overtrust of robots in emergency evacuation scenarios," in 2016, . DOI: 10.1109/HRI.2016.7451740.

[19] Qian Yang, Aaron Steinfeld, and John Zimmerman. 2019. 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 (CHI '19). Association for Computing Machinery, New York, NY, USA, Paper 238, 1–11. https://doi.org/10.1145/3290605.3300468

[20] Qualtrics: Fraud Detection. https://www.qualtrics.com/support/survey-platform/survey-module/survey-checker/fraud-detection/

[21] Ruotong Wang, F. Maxwell Harper, and Haiyi Zhu. 2020. Factors Influencing Perceived Fairness in Algorithmic Decision-Making: Algorithm Outcomes,

Development Procedures, and Individual Differences. In Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems (CHI '20). Association for Computing Machinery, New York, NY, USA, 1–14. https://doi.org/10.1145/3313831.3376813

[22] Spencer Soper, Fired by Bot at Amazon: 'It's You Against the Machine', 2021. https://www.bloomberg.com/news/features/2021-06-28/fired-by-bot-amazon-turns-to-machine-managers-and-workers-are-losing-out

[23] State of Wisconsin Department of Corrections. 2023. COMPAS. https://doc.wi.gov/Pages/AboutDOC/COMPAS.aspx. Retrieved 2023-05-04.

[24] Tim Dare and Eileen Gambrill. 2017. SECTION 2: Ethical Analysis: Predictive Risk Models at Call Screening for Allegheny County. https://www.alleghenycountyanalytics.us/wp-content/uploads/2019/05/Ethical-Analysis-16-ACDHS-26_PredictiveRisk_Package_050119_FINAL-2.pdf. Retrieved 2023-05-04.

[25] Umang Bhatt1,2, Javier Antorán2, Yunfeng Zhang3, Q. Vera Liao3, Prasanna Sattigeri3,, Riccardo Fogliato1,4, Gabrielle Melançon5, Ranganath Krishnan6, Jason Stanley5, Omesh Tickoo6, Lama Nachman6, Rumi Chunara7, Madhu Srikumar1, Adrian Weller2,8, Alice Xiang1,9 . 2021. Uncertainty as a Form of Transparency: Measuring, Communicating, and Using Uncertainty. In Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society (AIES '21), May 19–21, 2021, Virtual Event, USA. ACM, New York, NY, USA, 13 pages. https://doi.org/10.1145/3461702.3462571

[26] Valdemar Danry, Pat Pataranutaporn, Yaoli Mao, and Pattie Maes. 2023. Don't Just Tell Me, Ask Me: AI Systems that Intelligently Frame Explanations as Questions Improve Human Logical Discernment Accuracy over Causal AI explanations. In Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (CHI '23), April 23–28, 2023, Hamburg, Germany. ACM, New York, NY, USA, 13 pages. https://doi.org/10.1145/3544548.3580672