Simplicity is Complexity Resolved: Considering Task Complexity in Empirical HC(X)AI Studies

SARA SALIMZADEH, Delft University of Technology, The Netherlands
UJWAL GADIRAJU, Delft University of Technology, The Netherlands

Increasingly, artificial intelligence is being used in decision-making contexts across a range of domains, including healthcare, finance, and education, thanks to advances in machine learning algorithms. By increasing the transparency of artificial intelligence systems or by providing explanations to aid human understanding, different research communities have attempted to optimize and evaluate human-AI team performance. Nevertheless, the variety of decision making tasks considered and their operationalization in prior empirical work has left it unclear how findings from one task or domain are transferred to another. Comparisons between decision tasks can not be made straightforwardly due to a lack of standardized task attributes. We argue that the lens of '*task complexity*' can be used to tackle this problem of under-specification and facilitate comparison across empirical research in this area. We found the absence of consideration of task complexity across various studies in this realm of research. Inspired by Robert Wood's seminal work on the construct, we operationalized task complexity concerning three dimensions (`component`, `coordinative`, and `dynamic`). We quantified the complexity of decision tasks in existing work accordingly. We then summarized current trends and proposed research directions for the future.

1 INTRODUCTION

As AI systems demonstrate considerable growth in predictive performance, their adoption in human-AI decision making has risen significantly in a wide variety of domains and applications [4, 13, 28]. To fully benefit from AI systems' capacities, many mechanisms have been introduced to assist human decision makers in effectively collaborating with AI systems and efficiently recognizing their weaknesses and strengths. Such methods, for instance, can improve AI systems transparency [20, 25] and explainability [17, 31] helping humans interpret AI decisions and functionality.

Prior studies have indicated that different design choices in empirical studies, such as the choice of decision tasks, affect human trust and reliance on AI systems and consequently their complementary performance [7, 23]. Although some work in the literature has incorporated multiple tasks with different characteristics in their experimental setups [24, 25, 27], there is a lack of a holistic view of how various task characteristics influence human-AI team performance. Concurrently, the lack of a standard framework for systematically comparing decision tasks can prevent the generalizability of research findings across studies that have been conducted. Such in-depth understanding can also guide us on how, why, and when human decision makers rely on AI systems and how they appeal for aid to facilitate their interaction with AI systems.

To provide a basis for comparison across decision tasks, we propose the yet-to-be-explored lens of **task complexity**. Task complexity is determined by the attributes it has that increase information load, diversity, or change of information. It's also possible to define task complexity independently of user capabilities or other factors that affect perceived complexity. [1, 18, 29]. Task complexity is a significant task attribute found to influence various factors in human-AI

This research has been supported by *ICAI AI for Fintech Research*.

Manuscript submitted to ACM

decision making, such as human-AI team performance [10, 29], trust and reliance on AI systems [14], error detection in AI decisions [2, 35].

In our paper, we first reviewed to what extent recent literature in human-AI decision making considers task complexity in their design choices. We limited our study to articles published in HCI conferences and journals over the recent four years. According to our retrospective study, we found limited support for considering task complexity in empirical study designs. We then proposed a framework to conceptualize task complexity, enabling comparison across decision tasks inspired by Robert Wood's seminal work [44]. We annotated the articles accordingly and highlighted current trends and possible future research directions.

We found that decision tasks were distributed in all levels of complexity, from low to high in each complexity dimension, *component*, **coordinative**, and **dynamic** complexity. We also reported that the majority of tasks included tasks with low or medium complexity. We also indicated that tasks with higher levels of complexity mainly resemble real-world scenarios requiring domain expertise and include higher risk. Although our proposed framework has limitations in incorporating all task attributes, we can assert task complexity as a dimension to compare decision tasks.

As the first to model task complexity in a human-AI decision making context, our paper advances the current conversation in this community. Further work is required to extend the operationalization of task complexity to incorporate other task characteristics and differentiate diverse methods of information visualization (e.g., plots, text, images) or task stakes. We hope to inspire future work in proposing methods to help inform and facilitate meaningful comparisons across empirical studies on human-AI decision making.

2 BACKGROUND

Task complexity became a point of interest for over 50 years. Among all frameworks proposed in various domains, a seminal work by Wood [44] gained popularity with more than 2000 citations and became the basis of other frameworks. According to Wood [44], task complexity is defined in three dimensions: (i) component complexity indicating the number of distinct pieces of information required to accomplish the task, (ii) coordinative complexity showing the number of steps, and (iii) dynamic complexity specifying any changes in either component of coordinative complexity over time. Through adapting Wood's framework in human-AI decision making, we operationalized task complexity in our study and annotated decision tasks in these three dimensions.

3 METHOD

3.1 Scope of Literature Review

We followed a semi-systematic literature review which is adopted in prior studies [33, 36] through four stages, the definition of inclusion and exclusion criteria, search, selection, and analysis. We collected articles, including at least one quantitative empirical study pertaining to human-AI decision making. These articles also have been published in recent four years, from January 2019 to August 2022, as full papers in HCI venues CHI, CSCW, IUI, UMAP, FAccT, TOCHI, HCOMP, and UIST. Figure 1 depicts the steps followed in our study.

3.2 Task Complexity Modeling

After creating our corpus, we created a set of rubrics to model task complexity and calculate the corresponding three dimensions in decision tasks. Our rubrics specify the guideline to identify information cues and the number of required steps to perform the tasks. Accordingly, different experimental conditions of each study may have distinct complexity.

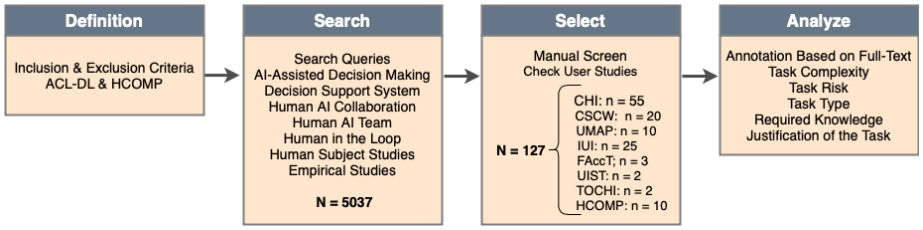


Fig. 1. A workflow diagram of the semi-systematic literature review process that we followed.

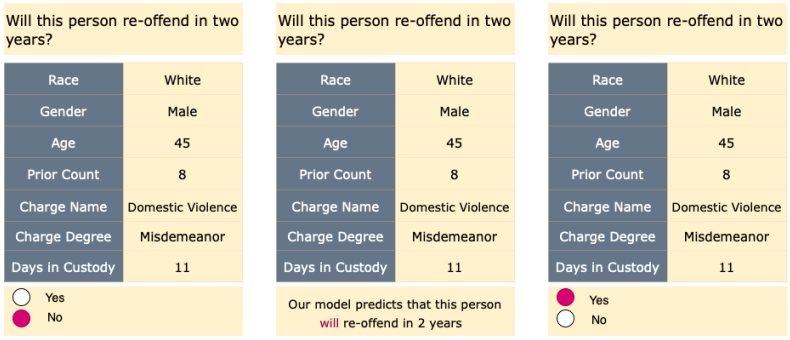


Fig. 2. A decision task study. The features of the defendant profile are recognized as information cues, each step is an act, and the final decision is considered the product.

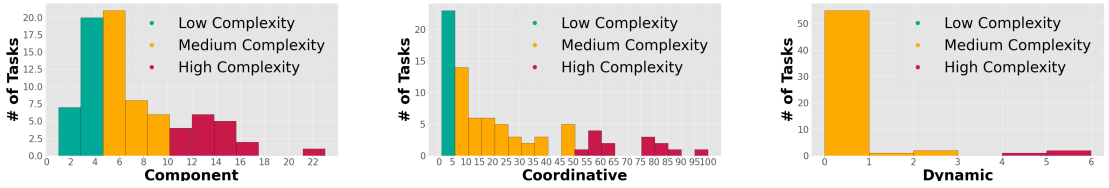
Additionally, we consider explanation methods as information cues and design a sub-set of rubrics to specify how they add component complexity per type. Figure 2 shows how complexity is measured per dimension, component, coordinative, and dynamic.

- **Component Complexity**: The number of distinct information cues that need to be processed to complete a task. Our example uses the following information cues to identify the defendant's profile: race, gender, age, prior count, charge name, charge degree, days in custody, and model predictions. We have a score of 8 as component complexity with all these features.
- **Coordinative Complexity**: It's a set of sequences of actions that are required to complete the task. Figure 2 shows three steps.
- **Dynamic Complexity**: Changing the value of information cues or the number of actions leads to dynamic complexity. In dynamic complexity, we count the number of information cues with variable quantities or extra steps needed to finish the task. During the decision making process in our example, both component and coordinative complexity are static, so the dynamic complexity is zero.

Following that, we looked at different articles from different perspectives to get relevant information, like the type of decision tasks, their domain, component, coordination, dynamic complexities, whether domain expertise was required to accomplish a task, the stakes, and whether the task was proxy or actual [6].

4 RESULTS

There's no standard framework for conceptualizing task complexity among all the relevant articles we collected, as only a few have considered task complexity in the design of decision tasks. In our framework, we modeled complexity in three dimensions.



(a) Component (b) Coordinative (c) Dynamic

Fig. 3. Distribution of component, coordinative, and dynamic complexity in the decision tasks corresponding to our corpus.

4.1 Component Complexity

Component complexity was found to be within the range of 1 to 23 with the average of 6.9 (±4.3), Figure 3a. We categorized component complexity into three levels: low, medium, and high, according to a research led by Wood [44], which showed human short-term memory could hold 5 to 9 objects on average. Low-complexity tasks have fewer than 5 information cues (indicating component complexity). Those with 5-9 information cues have medium component complexity, and those with more than 9 have high component complexity. The decision tasks in recent literature have a medium level of component complexity (6.9) on average. There's a low level of component complexity at 33.7%, a medium level at 40%, and a high level in 26% of tasks. 12% of the tasks were outliers with a high level of complexity between 24 and 132. All the outliers used specific datasets, included many features, and employed sophisticated plots.

4.2 Coordinative Complexity

Coordinative complexity for tasks in our data lies between 1 and 100, with an average of 25.1±25.9, meaning participants have to do 25.1 steps to complete the task. Based on Figure 3b, 75.6% of the complexity of tasks are distributed between a range of 1 to 40. Using quartiles, we divided coordinative complexity into low, medium, and high levels, with low at the bottom, high at the top, and medium at the other two. Low-complexity tasks have a coordinative complexity below 5, medium-complexity tasks have a coordinative complexity between 5 and 50, and highly complex tasks have a coordinative complexity over 50. Overall, 25.9% of tasks were found to have low coordinative complexity, 56.8% have medium complexity, and 17.2% have high complexity. We also observed outlier articles with coordinative complexity of 130 to 420.

4.3 Dynamic Complexity

Our analysis revealed that dynamic complexity was distributed between 0 to 6 (cf. Figure 3c). Using the bottom and top quartiles, we classified the dynamic complexity level. Tasks with a dynamic complexity of 0 belong to the low dynamic complexity category; tasks with a dynamic complexity of 1 to 3 belong to the medium dynamic complexity category; and tasks with a dynamic complexity of 4 or more belong to the high dynamic complexity category. Also, 95% of decision tasks had dynamic complexity between 0 and 2. This shows that dynamic complexity isn't common in

empirical human-AI studies.The source of dynamic complexity is the non-stationary nature of coordinative complexity. Some studies let the participants choose when and if they want AI recommendations. By probing the inputs first, this approach made participants more cognitively involved in the decision-making process. In the end, there were different numbers of steps depending on the participants, which led to dynamic complexity. There were outliers with dynamic complexity scores ranging from 12 to 1890, with the source of dynamicity mostly coming from changes in component complexity.

4.4 Task Complexity as a Comparative Lens

There were some similarities between levels of complexity and across dimensions, such as task stakes, task expertise, and task type. However, there were also some differences within each dimension, which we discuss below.

4.4.1 Different Tasks Same Complexities. According to our analysis, there are different decision tasks that have similar complexity levels. For most component complexity score levels, at least two studies have the same score. Low-complex tasks are comparable in terms of stake, domain expertise, and task types. More than 93% are low-stake tasks that can be accomplished without domain knowledge. In general, these decision tasks are mainly artificial, without any explicit real-world applications. As the tasks are straightforward, they imply lower demand for humans to rely on [14].As we move up the complexity scale, we see more diversity among tasks in terms of stake and expertise. The number of tasks resembling real-world problems was higher in the medium-complexity bin than in low-complexity tasks. Around half of the tasks were still artificial [34, 49] or didn't necessarily require human input [37, 40], according to our study. Lastly, on the other side of the spectrum, we found that highly complex tasks tend to be high-stake tasks requiring domain knowledge to complete. Low-stakes tasks in this bin focus on recommendation systems, which include many features to capture human preferences. It's important to note that high-complex tasks are explicit examples of real-life problems. We found that the more complex the task, the more it resembles real-world use cases, the more domain knowledge it requires, and the bigger the stake.

Additionally, we found similar scores for coordinative complexity, which represents how many steps are required to accomplish a task. Low-complexity tasks are low stakes without the expertise needed. On the other hand, high-complexity tasks include high risk and require knowledge. It's interesting that these tasks also have a medium or high component complexity score. It could be because researchers are increasing the number of instances over time to examine human behavior. Although, with higher component complexity and stake levels, participants might feel mentally fatigued earlier [22]. Research might not take into account the fatigue effect in their studies. Comparing these user studies to actual scenarios, it's also rare for humans to make 100 decisions simultaneously. It's better to examine human behavior over time in different sessions instead of consequent cases all at once to minimize fatigue and model the real world.

4.4.2 Same Tasks Different Complexities. In contrast to studies with the same complexity score, we found some similar decision tasks with varied complexity scores presented in each level of component complexity, low to high. They incorporate different explanation methods, enriched with additional visualizations along with the AI decision to modify the component complexity of those tasks. Our survey showed that researchers could control component complexity by enhancing explanations, visualizations, and user preferences when making recommendations or changing domains.

5 DISCUSSION AND IMPLICATIONS

5.1 Potential Reasons Why Task Complexity Has been Overlooked in Study Design

Our reflections on highly complex tasks indicated that interest in promoting AI and opportunities to propose explanation methods could usually guide the design of such tasks. According to researchers, explanations can improve people's mental models and help them understand, especially for laypeople. [15, 16, 19, 41]. Additionally, to fill the knowledge gap between domain experts and laypeople or improve AI literacy, empirical studies engage with more explanations [11, 28, 38, 49]. Adding more user preferences can also make recommendations tasks more complex [21, 32, 48].

Increased task complexity could also be due to the need to study trust formation and AI reliance in such situations. [12, 30, 35, 46, 48]. A variety of features can ensure that human decision-makers have access to salient information needed to make better decisions, especially in high-stakes situations. [5, 9, 17, 26].

There's also a relationship between task complexity and the nature of the task. Tasks representing real-world cases, especially those with higher risk, tend to have more features and require more expertise to accomplish [28, 28, 42].

The use of cognitive forcing interventions is also being studied as a way to get humans to engage more thoughtfully with AI systems by: (I) asking humans to make decisions before seeing model predictions [43, 50], (II) varying AI systems response time [7, 34], and (III) providing feedback to humans [3, 17, 45, 47].

In terms of the arbitrary choice of task instances observed in many articles, researchers may include more instances to explore the impact of human-AI interaction over time. Having more time to collaborate with AI, human decision-makers familiarize themselves with AI systems, form their mental models, and calibrate their trust.

In contrast to orchestrating high task complexity, task complexity is mitigated in some studies. Due to the cost and limited accessibility to hire real-end users of AI systems, crowd workers simulate the decision making process. As crowd workers' knowledge is limited, decision tasks are either simplified, artificially created, or substitutes with common tasks that crowd workers have experience in are considered [6, 23, 40, 47]. Tasks with low complexity can help human decision makers have a better understanding of AI systems [8, 39].

6 CONCLUSION AND FUTURE WORK

We reviewed the published literature on human-AI decision making tasks in the last four years. We found little evidence of its consideration as a design parameter. We then operationalized task complexity based on Robert Wood's seminal work. We analyzed different dimensions of task complexity and measured them using a set of well-defined rubrics. Our analysis found that tasks in the literature range in complexity across all levels and dimensions. Most of the tasks considered in empirical studies have low or medium complexity. The most complex tasks, which largely resemble real-world problems, were found to have higher risk levels, requiring domain expertise. Despite the limitations in our operationalization of task complexity — we did not account for other task characteristics that may affect task complexity — we found that it can still provide us with an axis for comparing decision tasks in human-AI studies. Such comparisons are particularly meaningful in tasks with lower levels of complexity. According to our analysis of empirical human-AI research, it's crucial to measure and report the different types of expertise or domain knowledge participants may have (numeracy, AI literacy, statistics, or visualization familiarity) in order to make meaningful comparisons across studies. A future study in this area could explicitly control how complex the task is across different conditions. Our goal is to expand our operationalization of task complexity to include other task features and develop a tool to help researchers design empirical human-AI studies by automatically measuring the complexity of tasks across Wood's three dimensions.

REFERENCES

[1] Abdullah Almaatouq, Mohammed Alsobay, Ming Yin, and Duncan J. Watts. 2021. Task complexity moderates group synergy. *Proceedings of the National Academy of Sciences* 118, 36 (2021), e2101062118. https://doi.org/10.1073/pnas.2101062118 arXiv:https://www.pnas.org/doi/pdf/10.1073/pnas.2101062118

[2] Gagan Bansal, Besmira Nushi, Ece Kamar, Walter S. Lasecki, Daniel S. Weld, and Eric Horvitz. 2019. Beyond Accuracy: The Role of Mental Models in Human-AI Team Performance. *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing* 7, 1 (Oct. 2019), 2–11. https://doi.org/10.1609/hcomp.v7i1.5285

[3] 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* (Yokohama, Japan) *(CHI '21)*. Association for Computing Machinery, New York, NY, USA, Article 81, 16 pages. https://doi.org/10.1145/3411764.3445717

[4] Reuben Binns, Max Van Kleek, Michael Veale, Ulrik Lyngs, Jun Zhao, and Nigel Shadbolt. 2018. 'It's Reducing a Human Being to a Percentage': Perceptions of Justice in Algorithmic Decisions. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (Montreal QC, Canada) *(CHI '18)*. Association for Computing Machinery, New York, NY, USA, 1–14. https://doi.org/10.1145/3173574.3173951

[5] Angie Boggust, Benjamin Hoover, Arvind Satyanarayan, and Hendrik Strobelt. 2022. Shared Interest: Measuring Human-AI Alignment to Identify Recurring Patterns in Model Behavior. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (New Orleans, LA, USA) *(CHI '22)*. Association for Computing Machinery, New York, NY, USA, Article 10, 17 pages. https://doi.org/10.1145/3491102.3501965

[6] Zana Buçinca, Phoebe Lin, Krzysztof Z. Gajos, and Elena L. Glassman. 2020. Proxy Tasks and Subjective Measures Can Be Misleading in Evaluating Explainable AI Systems. In *Proceedings of the 25th International Conference on Intelligent User Interfaces* (Cagliari, Italy) *(IUI '20)*. Association for Computing Machinery, New York, NY, USA, 454–464. https://doi.org/10.1145/3377325.3377498

[7] 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. *Proc. ACM Hum.-Comput. Interact.* 5, CSCW1, Article 188 (apr 2021), 21 pages. https://doi.org/10.1145/3449287

[8] Carrie J. Cai, Jonas Jongejan, and Jess Holbrook. 2019. The Effects of Example-Based Explanations in a Machine Learning Interface. In *Proceedings of the 24th International Conference on Intelligent User Interfaces* (Marina del Ray, California) *(IUI '19)*. Association for Computing Machinery, New York, NY, USA, 258–262. https://doi.org/10.1145/3301275.3302289

[9] Carrie J. Cai, Emily Reif, Narayan Hegde, Jason Hipp, Been Kim, Daniel Smilkov, Martin Wattenberg, Fernanda Viegas, Greg S. Corrado, Martin C. Stumpe, and Michael Terry. 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)*. Association for Computing Machinery, New York, NY, USA, 1–14. https://doi.org/10.1145/3290605.3300234

[10] Siew H. Chan, Qian Song, and Lee J. Yao. 2015. The moderating roles of subjective (perceived) and objective task complexity in system use and performance. *Computers in Human Behavior* 51 (2015), 393–402. https://doi.org/10.1016/j.chb.2015.04.059

[11] 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)*. Association for Computing Machinery, New York, NY, USA, 1–12. https://doi.org/10.1145/3290605.3300789

[12] Chun-Wei Chiang and Ming Yin. 2022. Exploring the Effects of Machine Learning Literacy Interventions on Laypeople's Reliance on Machine Learning Models. In *27th International Conference on Intelligent User Interfaces* (Helsinki, Finland) *(IUI '22)*. Association for Computing Machinery, New York, NY, USA, 148–161. https://doi.org/10.1145/3490099.3511121

[13] Sam Corbett-Davies, Emma Pierson, Avi Feller, Sharad Goel, and Aziz Huq. 2017. Algorithmic Decision Making and the Cost of Fairness. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (Halifax, NS, Canada) *(KDD '17)*. Association for Computing Machinery, New York, NY, USA, 797–806. https://doi.org/10.1145/3097983.3098095

[14] Devleena Das and Sonia Chernova. 2020. Leveraging Rationales to Improve Human Task Performance. In *Proceedings of the 25th International Conference on Intelligent User Interfaces* (Cagliari, Italy) *(IUI '20)*. Association for Computing Machinery, New York, NY, USA, 510–518. https://doi.org/10.1145/3377325.3377512

[15] Oscar Gomez, Steffen Holter, Jun Yuan, and Enrico Bertini. 2020. ViCE: Visual Counterfactual Explanations for Machine Learning Models. In *Proceedings of the 25th International Conference on Intelligent User Interfaces* (Cagliari, Italy) *(IUI '20)*. Association for Computing Machinery, New York, NY, USA, 531–535. https://doi.org/10.1145/3377325.3377536

[16] Mitchell L. Gordon, Michelle S. Lam, Joon Sung Park, Kayur Patel, Jeff Hancock, Tatsunori Hashimoto, and Michael S. Bernstein. 2022. Jury Learning: Integrating Dissenting Voices into Machine Learning Models. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (New Orleans, LA, USA) *(CHI '22)*. Association for Computing Machinery, New York, NY, USA, Article 115, 19 pages. https://doi.org/10.1145/3491102.3502004

[17] Nina Grgić-Hlača, Christoph Engel, and Krishna P. Gummadi. 2019. Human Decision Making with Machine Assistance: An Experiment on Bailing and Jailing. *Proc. ACM Hum.-Comput. Interact.* 3, CSCW, Article 178 (nov 2019), 25 pages. https://doi.org/10.1145/3359280

[18] J. Richard Hackman. 1969. Toward understanding the role of tasks in behavioral research. *Acta Psychologica* 31 (1969), 97–128. https://doi.org/10.1016/0001-6918(69)90073-0

[19] Sophia Hadash, Martijn C. Willemsen, Chris Snijders, and Wijnand A. IJsselsteijn. 2022. Improving Understandability of Feature Contributions in Model-Agnostic Explainable AI Tools. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (New Orleans, LA, USA) *(CHI '22)*. Association for Computing Machinery, New York, NY, USA, Article 487, 9 pages. https://doi.org/10.1145/3491102.3517650

[20] Galen Harrison, Julia Hanson, Christine Jacinto, Julio Ramirez, and Blase Ur. 2020. An Empirical Study on the Perceived Fairness of Realistic, Imperfect Machine Learning Models. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* (Barcelona, Spain) *(FAT* '20)*. Association for Computing Machinery, New York, NY, USA, 392–402. https://doi.org/10.1145/3351095.3372831

[21] Daniel Herzog and Wolfgang Wörndl. 2019. A User Study on Groups Interacting with Tourist Trip Recommender Systems in Public Spaces. In *Proceedings of the 27th ACM Conference on User Modeling, Adaptation and Personalization* (Larnaca, Cyprus) *(UMAP '19)*. Association for Computing Machinery, New York, NY, USA, 130–138. https://doi.org/10.1145/3320435.3320449

[22] raymond Soames Job and James Dalziel. 2000. *Defining Fatigue as a Condition of the Organism and Distinguishing It From Habituation, Adaptation, and Boredom.* 466–476. https://doi.org/10.1201/b12791-3.2

[23] Johannes Kunkel, Tim Donkers, Lisa Michael, Catalin-Mihai Barbu, and Jürgen Ziegler. 2019. Let Me Explain: Impact of Personal and Impersonal Explanations on Trust in Recommender Systems. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) *(CHI '19)*. Association for Computing Machinery, New York, NY, USA, 1–12. https://doi.org/10.1145/3290605.3300717

[24] 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. *arXiv preprint arXiv:2112.11471* (2021).

[25] Vivian Lai, Han Liu, and Chenhao Tan. 2020. "Why is 'Chicago' Deceptive?" Towards Building Model-Driven Tutorials for Humans. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–13. https://doi.org/10.1145/3313831.3376873

[26] Min Hun Lee, Daniel P. Siewiorek, Asim Smailagic, Alexandre Bernardino, and Sergi Bermúdez i Badia. 2021. A Human-AI Collaborative Approach for Clinical Decision Making on Rehabilitation Assessment. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI '21)*. Association for Computing Machinery, New York, NY, USA, Article 392, 14 pages. https://doi.org/10.1145/3411764.3445472

[27] Min Kyung Lee. 2018. Understanding perception of algorithmic decisions: Fairness, trust, and emotion in response to algorithmic management. *Big Data & Society* 5 (2018).

[28] Ariel Levy, Monica Agrawal, Arvind Satyanarayan, and David Sontag. 2021. Assessing the Impact of Automated Suggestions on Decision Making: Domain Experts Mediate Model Errors but Take Less Initiative. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI '21)*. Association for Computing Machinery, New York, NY, USA, Article 72, 13 pages. https://doi.org/10.1145/3411764.3445522

[29] Peng Liu and Zhizhong Li. 2012. Task complexity: A review and conceptualization framework. *International Journal of Industrial Ergonomics* 42, 6 (2012), 553–568. https://doi.org/10.1016/j.ergon.2012.09.001

[30] Zhuoran Lu and Ming Yin. 2021. Human Reliance on Machine Learning Models When Performance Feedback is Limited: Heuristics and Risks. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI '21)*. Association for Computing Machinery, New York, NY, USA, Article 78, 16 pages. https://doi.org/10.1145/3411764.3445562

[31] Ana Lucic, Hinda Haned, and Maarten de Rijke. 2020. Why Does My Model Fail? Contrastive Local Explanations for Retail Forecasting. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* (Barcelona, Spain) *(FAT* '20)*. Association for Computing Machinery, New York, NY, USA, 90–98. https://doi.org/10.1145/3351095.3372824

[32] Cataldo Musto, Alain D. Starke, Christoph Trattner, Amon Rapp, and Giovanni Semeraro. 2021. Exploring the Effects of Natural Language Justifications in Food Recommender Systems. In *Proceedings of the 29th ACM Conference on User Modeling, Adaptation and Personalization* (Utrecht, Netherlands) *(UMAP '21)*. Association for Computing Machinery, New York, NY, USA, 147–157. https://doi.org/10.1145/3450613.3456827

[33] Francisco Nunes, Nervo Verdezoto, Geraldine Fitzpatrick, Morten Kyng, Erik Grönvall, and Cristiano Storni. 2015. Self-Care Technologies in HCI: Trends, Tensions, and Opportunities. 22, 6, Article 33 (dec 2015), 45 pages. https://doi.org/10.1145/2803173

[34] 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 (nov 2019), 15 pages. https://doi.org/10.1145/3359204

[35] Forough Poursabzi-Sangdeh, Daniel G Goldstein, Jake M Hofman, Jennifer Wortman Wortman Vaughan, and Hanna Wallach. 2021. Manipulating and Measuring Model Interpretability. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI '21)*. Association for Computing Machinery, New York, NY, USA, Article 237, 52 pages. https://doi.org/10.1145/3411764.3445315

[36] Amon Rapp, Lorenzo Curti, and Arianna Boldi. 2021. The human side of human-chatbot interaction: A systematic literature review of ten years of research on text-based chatbots. *International Journal of Human-Computer Studies* 151 (2021), 102630.

[37] Maria Riveiro and Serge Thill. 2022. The Challenges of Providing Explanations of AI Systems When They Do Not Behave like Users Expect. In *Proceedings of the 30th ACM Conference on User Modeling, Adaptation and Personalization* (Barcelona, Spain) *(UMAP '22)*. Association for Computing Machinery, New York, NY, USA, 110–120. https://doi.org/10.1145/3503252.3531306

[38] James Schaffer, John O'Donovan, James Michaelis, Adrienne Raglin, and Tobias Höllerer. 2019. I Can Do Better than Your AI: Expertise and Explanations. In *Proceedings of the 24th International Conference on Intelligent User Interfaces* (Marina del Ray, California) *(IUI '19)*. Association for Computing Machinery, New York, NY, USA, 240–251. https://doi.org/10.1145/3301275.3302308

[39] Alison Smith-Renner, Ron Fan, Melissa Birchfield, Tongshuang Wu, Jordan Boyd-Graber, Daniel S. Weld, and Leah Findlater. 2020. No Explainability without Accountability: An Empirical Study of Explanations and Feedback in Interactive ML. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–13. https:

//doi.org/10.1145/3313831.3376624

[40] Aaron Springer and Steve Whittaker. 2019. Progressive Disclosure: Empirically Motivated Approaches to Designing Effective Transparency. In *Proceedings of the 24th International Conference on Intelligent User Interfaces* (Marina del Ray, California) *(IUI '19)*. Association for Computing Machinery, New York, NY, USA, 107–120. https://doi.org/10.1145/3301275.3302322

[41] Maxwell Szymanski, Martijn Millecamp, and Katrien Verbert. 2021. Visual, Textual or Hybrid: The Effect of User Expertise on Different Explanations. In *26th International Conference on Intelligent User Interfaces* (College Station, TX, USA) *(IUI '21)*. Association for Computing Machinery, New York, NY, USA, 109–119. https://doi.org/10.1145/3397481.3450662

[42] Niels van Berkel, Jorge Goncalves, Daniel Russo, Simo Hosio, and Mikael B. Skov. 2021. Effect of Information Presentation on Fairness Perceptions of Machine Learning Predictors. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI '21)*. Association for Computing Machinery, New York, NY, USA, Article 245, 13 pages. https://doi.org/10.1145/3411764.3445365

[43] Xinru Wang and Ming Yin. 2021. Are Explanations Helpful? A Comparative Study of the Effects of Explanations in AI-Assisted Decision-Making. In *26th International Conference on Intelligent User Interfaces* (College Station, TX, USA) *(IUI '21)*. Association for Computing Machinery, New York, NY, USA, 318–328. https://doi.org/10.1145/3397481.3450650

[44] Robert Wood. 1986. Task complexity: Definition of the construct. *Organizational Behavior and Human Decision Processes* 37 (02 1986), 60–82. https://doi.org/10.1016/0749-5978(86)90044-0

[45] 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)*. Association for Computing Machinery, New York, NY, USA, 189–201. https://doi.org/10.1145/3377325.3377480

[46] 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* (Glasgow, Scotland Uk) *(CHI '19)*. Association for Computing Machinery, New York, NY, USA, 1–12. https://doi.org/10.1145/3290605.3300509

[47] Kun Yu, Shlomo Berkovsky, Ronnie Taib, Jianlong Zhou, and Fang Chen. 2019. Do I Trust My Machine Teammate? An Investigation from Perception to Decision. In *Proceedings of the 24th International Conference on Intelligent User Interfaces* (Marina del Ray, California) *(IUI '19)*. Association for Computing Machinery, New York, NY, USA, 460–468. https://doi.org/10.1145/3301275.3302277

[48] Rachael Zehrung, Astha Singhal, Michael Correll, and Leilani Battle. 2021. Vis Ex Machina: An Analysis of Trust in Human versus Algorithmically Generated Visualization Recommendations. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI '21)*. Association for Computing Machinery, New York, NY, USA, Article 602, 12 pages. https://doi.org/10.1145/3411764.3445195

[49] Qiaoning Zhang, Matthew L Lee, and Scott Carter. 2022. You Complete Me: Human-AI Teams and Complementary Expertise. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (New Orleans, LA, USA) *(CHI '22)*. Association for Computing Machinery, New York, NY, USA, Article 114, 28 pages. https://doi.org/10.1145/3491102.3517791

[50] Yunfeng Zhang, Q. Vera Liao, and Rachel K. E. Bellamy. 2020. Effect of Confidence and Explanation on Accuracy and Trust Calibration in AI-Assisted Decision Making. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* (Barcelona, Spain) *(FAT* '20)*. Association for Computing Machinery, New York, NY, USA, 295–305. https://doi.org/10.1145/3351095.3372852