A Missing Piece in the Puzzle
Considering the Role of Task Complexity in Human-AI Decision Making

Salimzadeh, Sara; He, Gaole; Gadiraju, Ujwal

DOI
10.1145/3565472.3592959

Publication date
2023

Document Version
Final published version

Published in
UMAP 2023 - Proceedings of the 31st ACM Conference on User Modeling, Adaptation and Personalization

A Missing Piece in the Puzzle: Considering the Role of Task Complexity in Human-AI Decision Making

Sara Salimzadeh
Delft University of Technology
Delft, The Netherlands
s.salimzadeh@tudelft.nl

Gaole He
Delft University of Technology
Delft, The Netherlands
g.he@tudelft.nl

Ujwal Gadiraju
Delft University of Technology
Delft, The Netherlands
u.k.gadiraju@tudelft.nl

ABSTRACT

Recent advances in the performance of machine learning algorithms have led to the adoption of AI models in decision making contexts across various domains such as healthcare, finance, and education. Different research communities have attempted to optimize and evaluate human-AI team performance through empirical studies by increasing transparency of AI systems, or providing explanations to aid human understanding of such systems. However, the variety in decision making tasks considered and their operationalization in prior empirical work, has led to an opacity around how findings from one task or domain carry forward to another. The lack of a standardized means of considering task attributes prevents straightforward comparisons across decision tasks, thereby limiting the generalizability of findings. 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. To retrospectively explore how different HCI communities have considered the influence of task complexity in designing experiments in the realm of human-AI decision making, we survey literature and provide an overview of empirical studies on this topic. We found a serious dearth in the 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 with respect to three dimensions (`component`, `coordinative`, and `dynamic`) and quantified the complexity of decision tasks in existing work accordingly. We then summarized current trends and proposed research directions for the future. Our study highlights the need to account for task complexity as an important design choice. This is a first step to help the scientific community in drawing meaningful comparisons across empirical studies in human-AI decision making and to provide opportunities to generalize findings across diverse domains and experimental settings.

CCS CONCEPTS

• **Human-centered computing** → **User studies**; **Empirical studies in HCI**.

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

UMAP '23, June 26–29, 2023, Limassol, Cyprus

ACM ISBN 978-1-4503-9932-6/23/06.
https://doi.org/10.1145/3565472.3592959

ACM Reference Format:
Sara Salimzadeh, Gaole He, and Ujwal Gadiraju. 2023. A Missing Piece in the Puzzle: Considering the Role of Task Complexity in Human-AI Decision Making. In *UMAP '23: Proceedings of the 31st ACM Conference on User Modeling, Adaptation and Personalization (UMAP '23), June 26–29, 2023, Limassol, Cyprus.* ACM, New York, NY, USA, 13 pages. https://doi.org/10.1145/3565472.3592959

1 INTRODUCTION

Recent advances in the performance of machine learning algorithms have led to a rise in human-AI decision making in a wide variety of domains. For example, recidivism prediction algorithms have been used to help judges determine whether defendants are likely to re-offend [26, 31, 65, 69], medical diseases are being diagnosed with AI systems [19, 57, 62], and loan risk prediction algorithms are employed to approve or reject loan applications [11, 25, 38, 95].

To take advantage of AI systems and achieve an ideal complementary team performance, human decision makers need to recognize the strengths and weaknesses of AI systems and effectively use AI advice to optimize their decision making. To this end, a wide variety of mechanisms have been proposed to facilitate effective human-AI collaboration such as increasing transparency of AI systems, and their interpretability. For instance, many studies provide explanations along with AI decisions to help humans interpret AI systems' decisions [11, 12, 38, 40, 68]. It is also common to present information about the AI systems to create a better perception of their functionality among users [18, 45, 45, 55]. Prior work has also examined how human trust and reliance on AI systems is affected by different design choices through empirical studies [16, 52, 67, 76].

Apart from different features of AI systems and inherent human factors, the choice of decision tasks also affects the performance of human-AI teams [4, 8, 95]. Although several studies have examined the role of human factors in shaping interactions with AI systems, there is a limited understanding of task characteristics in the human-AI decision making context [2, 7, 83, 93]. Even though some studies incorporate tasks with different characteristics [53, 55, 61], task attributes haven't been identified systematically in the literature, so their impact on human-AI complementary performance has not been fully investigated. Consequently, there is no standard and coherent way to compare decision tasks, hindering research efforts and preventing generalizability across domains explored in empirical studies. For example, it is difficult to say how human trust shapes in the context of recidivism prediction task [97] compares to movie recommendation task [52]. Although this is not a straightforward endeavor, being able to make such comparisons will allow us to build a deeper understanding of when, why, and how humans rely on AI systems and how users can be best supported in their interactions. Doshi-Velez and Kim [30] have argued that to

create and advance a 'rigorous science' in the realm of human-AI decision making, there is room for empirical work that considers functionally-grounded explanations with proxy users and tasks, human-grounded evaluation with real users and simple tasks, and application-grounded evaluation with real users and real tasks. In practice, however, it is difficult to understand the transferability of findings across these levels of empirical work. Moreover, our exploratory analysis of rationales reported for the tasks considered in recent empirical work on human-AI decision making revealed a lack of depth. For instance, in a recent study, a specific task was selected due to the abundance of datasets in its domain [95], and in another because it has been used in previous studies [85]. We believe that this contributes to – and is indicative of – the opacity around tasks and the transferability of concomitant findings.

To facilitate comparison across distinct human-AI decision making tasks, we propose the lens of **task complexity** in this paper. Complexity of a task is influenced by task characteristics which increase information load, information diversity, or rate of information change. The complexity of tasks is an important dimension differentiating one task from the other, playing a significant role in determining the performance of a human-AI team [2, 43, 66]. It has been found to be an essential predictor of human performance and behaviour [2, 21, 66], affecting the success of team work [7]. Task complexity can also impact trust and reliance on AI systems. Intuitively, more complex tasks demand more effort from decision makers to complete and one can expect that human decision makers perform worse on highly complex tasks. On one hand, more complex tasks may imply a greater need for humans to rely on AI systems [27] as a result of increased information overload in such tasks [21]. On the other hand, human decision makers may struggle to identify errors created by AI systems on complex tasks, leading to over-reliance [7, 83]. Note that we consider the construct of task complexity independently from the users' standpoint or abilities, *i.e.*, independently from factors which influence the perceived task complexity [20].

In this paper, we first shed light on the extent to which task complexity has been considered in the design of recent empirical studies across research communities that have explored human-AI decision making. Next, we propose a means to operationalize task complexity to facilitate comparisons across empirical works and provide us with an instrument to gauge potential transferability of findings along this axis. We thereby address the following research questions:

- **RQ1:** How has recent research in human-AI decision making considered the influence of task complexity?
- **RQ2:** How can task complexity facilitate a comparative lens for empirical work on human-AI decision making?

To answer the RQs, we provide an overview of the current state of human-AI decision making research through a retrospective study. We focus on studies in which decision tasks were adopted within the human-AI team setting to evaluate or improve their performance, either as the team or individual component. We limited our scope to articles published in HCI conferences and journals in the last four years, considering most relevant articles have been published in the last four years based on our preliminary analysis on Google Scholar hits. We found little evidence of task complexity being considered or controlled as a factor within the study design. Inspired by Robert Wood's seminal construct of task complexity [100], we coded different aspects of existing decision tasks based on three dimensions of complexity — `component`, `coordinative`, and `dynamic` complexity. Next, we annotated the empirical study setups in different articles in our corpus ($N = 127$) corresponding to each dimension of task complexity, highlighted current trends, and proposed research directions for the future.

Original Contributions. We analyzed recent empirical studies of human-AI decision making from an under-explored but important perspective of *task complexity*. To the best of our knowledge, this is the first systematic analysis of task complexity across empirical human-AI decision making studies. We operationalized task complexity in decision tasks, measured and annotated task complexity of decision tasks considered in recent literature across research communities. We found that tasks in the literature are distributed across all levels and dimensions of complexity. Based on our analysis, most tasks that have been considered in empirical studies have a low or medium level of component complexity. We found that highly-complex tasks generally represent real-world problems by incorporating higher risk levels and requiring domain expertise that demands a greater level of trust and reliance by humans. Despite existing limitations in operationalizing task complexity such as difficulty in accounting for features like task stakes we argue that task complexity can provide us with an axis along which we can engage in comparisons across decision tasks in empirical human-AI studies. Our work offers a starting point on which we hope that future work can build upon, extend our framework, and model various aspects and attributes of decision-making tasks in greater depth. Our findings can assist researchers in making meaningful comparisons across studies, provide opportunities to generalize findings across diverse domains, and inspire future work to tackle issues pertaining to transferability of findings in empirical human-AI decision making research.

2 RELATED WORK

2.1 Human-AI Decision Making

Since AI systems have shown promising performance on various intelligent tasks like financial risk estimation [78] and medical diagnosis [6], a growing number of researchers and practitioners have begun to propose such AI systems in augmenting human decision making [54]. One main goal of such human-AI collaboration is to achieve complementary team performance [65]. For this purpose, human decision makers are expected to identify when they should rely on AI and when they should work on the tasks themselves, thereby exhibiting '*appropriate reliance*' on AI systems [58]. Only a few empirical studies have reported such appropriate reliance [54, 65]. However, there is substantial evidence that corroborates how challenging it is to foster appropriate reliance among users on AI systems [67, 101]. To promote appropriate reliance on AI systems, different interventions including explanations of AI advice [97], cognitive forcing functions [16], and user tutorials [23, 24] have been proposed in empirical studies of human-AI decision making with varying extent of success and befitting varying contexts.

Existing studies have also found that human-AI decision making is affected by a number of factors. The information shown to users along with AI advice can greatly impact their trust and reliance.

Explanations [65, 97], stated performance [7, 102, 106], risk perception [39], and uncertainty [92] have been studied extensively in this context. User factors like expertise [28], machine learning literacy [24], and task characteristics like task subjectivity and proximity [15], and task types [34, 49, 55, 61, 74] have also been broadly investigated.

Despite the significance of tasks in the human-AI decision making field, a limited number of studies have focused on explicitly considering task complexity and understanding the impact of varying task complexity. Bansal et al. [7] defined the number of features in each task instance as task dimensionality and conducted a user study controlling the number of human-visible features. They found that human-AI team performance decreases as task dimensionality increases. Similarly, Poursabzi-Sangdeh et al. [83] investigated how the number of features impacts the capability of participants to simulate AI predictions. Participants struggled detecting errors when faced with more task features due to information overload, which can be detrimental to the complementary human-AI team performance. In contrast, having considered two levels of complexity in their task design, Tolmeijer et al. [91] found that complexity does not impact participants' performance due to a learning effect. In terms of comparing human and AI systems, according to Lin et al. [64], AI systems outperform humans when they have access to extensive amount of information.

In this paper, we specifically focus on the task complexity, which is under-explored in human-AI decision making studies. We first reviewed studies published in recent four years to evaluate the extent to which they take task complexity into account while designing user studies. Towards this goal, we adapted a framework to conceptualize different dimensions of task complexity and annotated the tasks accordingly. We then evaluated how tasks with various levels of complexity have been distributed in the past based on the proposed framework. Based on this review, we present our findings on how various dimensions of task complexity could influence human-AI decision making.

2.2 Task Complexity

Task complexity became a point of interest for over 50 years. In the late-1980s, some frameworks were proposed to define and analyze task complexity; they were adapted in many domains such as psychology, management, information systems, and etc. [20, 48, 66, 100]. Among all works, the theories introduced by Campbell [20], Wood [100] gained popularity with more than 2000 citations and became the basis of other frameworks. According to Campbell [20], complexity of a task is influenced by task characteristics that increase information load, information diversity, or rate of information change. More importantly, task complexity is defined independently of any task doer's ability [20]. Aligned with this definition, Wood [100] recognized three factors contributing to task complexity which are (i) the number of distinct pieces of information required to complete the task, (ii) the number of steps, and (iii) any changes in either piece of information or steps over time. They named these factors as component complexity, coordinative complexity, and dynamic complexity.

Through adapting the framework proposed by Wood [100], we operationalize task complexity in empirical studies in a human-AI decision making context. Note that we study tasks given to humans, not AI systems in our paper. We then explore how decision tasks are distributed in existing work and discuss the limitations of current experimental studies and the implication for researchers to consider task complexity as their design choice.

3 METHOD

3.1 Scoping Our Literature Review

We followed a semi-systematic literature review, widely adopted in prior studies [77, 84], including the four stages summarized in Figure 1 below.

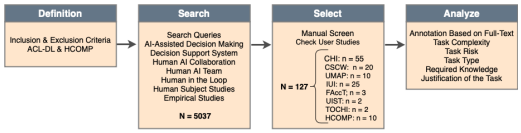


Figure 1: A workflow diagram of the semi-systematic literature review process that we followed.

3.1.1 Define Inclusion and Exclusion Criteria. The purpose of this study is to examine empirical human-subject studies pertaining to human-AI decision making, which evaluate or improve the performance of human-AI as a team or individual component. We applied the following inclusion and exclusion criteria to filter the articles.

Human-AI Decision Making: Selected articles need to include at least one empirical human-subjects study in which humans are asked to accomplish a decision task with the aid of an AI system. We thus exclude non-empirical articles or articles focusing on tasks such as debugging, creativity, and sketching.

Qualitative Human-Subjects Studies: Human-subjects studies must be evaluated quantitatively in the selected articles. Therefore, studies considering only interviews with humans to determine design decisions or asking about their preferences and understandings resulting in only filling questionnaires were excluded.

Proceedings: Selected articles are published in HCI conferences or journals, including CHI, CSCW, IUI, UMAP, FAccT, TOCHI, HCOMP, and UIST within recent four years, as of January 2019 up to August 2022.

Format: We included only full papers in our collection.

Most HCI conferences and journal articles are published through ACM Digital Library, so we identify it as our source. For the articles in Proceedings of AAAI Conference on Human Computation and Crowdsourcing, which do not exist in ACM Digital Library, we retrieve the articles from their proceedings.

3.1.2 Search. We conducted an exploratory search in the ACM Digital Library to determine search queries. We searched for articles that included user studies in which participants were tasked to complete decision tasks. We retrieved 50 articles from the proceeding mentioned in our inclusion criteria using "empirical studies" and "human-AI decision making" as keywords. We manually analyzed these articles and extracted seven common keywords. We then utilize the keywords as our final search query. The search query included the following terms, "AI-assisted decision making," "decision support systems," "human AI collaboration," "human AI team," "human in the loop," "human subject studies," and "empirical studies." An initial search yielded 5037 articles after limiting the proceedings specified in our criteria.

3.1.3 Select. We manually screened the articles according to our inclusion and exclusion criteria. We looked for articles containing empirical studies by searching through the full texts of all articles with the keywords "user study" and "empirical study". We then examined the detail and type of study to decide whether we could add this article to the final collection. For instance, we removed articles containing only interviews or surveys as user studies. After excluding out-of-scope studies, we reached a collection of 127 articles.

3.1.4 Analyze. In order to evaluate how each article considered the influence of task complexity, we first reviewed the full text of the articles. We then started annotating the decision tasks by extracting relevant information such as: what kind of decision tasks, the risk of the decision tasks, how much knowledge is required to perform tasks, explicit justification of choosing decision tasks, and whether tasks are proxy tasks or actual decision making task [15]. Furthermore, we coded the component, coordinative, and dynamic complexity of decision tasks based on how many information cues are required to accomplish the tasks and the number of steps required to complete the tasks according to our rubrics explained in section 3.3. In case of any changes in a number of information cues or steps, we reported dynamic complexity. We created our rubrics for operationalizing task complexity in a decision making context while annotating the articles and observing new scenarios. As identification of information cues and steps could be subjective, we discussed the rubrics iteratively along the way to ensure the integrity of the process. Authors of this paper iterated over 30 articles, before finalizing and converging on the rubrics. The **full list of articles and our annotations** are publicly accessible for the benefit of the research community and in the spirit of open science.[1]

3.2 Operationalizing Task Complexity

In this section, we, introduce the general definitions represented by Wood [100] and clarify the concepts using an example in the context of human-AI decision making. In the next sub-section 3.3, we leverage these terminologies to explain how our framework to models task complexity in the context of human-AI decision making. Note that all of the terminologies and definitions in this section are adopted from the Wood [100] work.

All tasks contain three essential components: products, required acts, and information cues. We follow the example in Figure 2 to elaborate on each component and introduce terminologies. The example presents a two-stage decision making process for recidivism prediction task. According to the defendant's profile, a human decision maker has to decide whether the defendant re-offends the crime in two years. We identified 3 constructs to calculate task complexity:

Product: Products are entities created by behaviours that can be identified separately from behaviours that produce them. They are identified as a set of assembled attributes such as an object or event and contain some defining attributes like quantity, quality, and cost. The final decision of a human is the Product in Figure 2.

Acts: Acts serve as a specific activity or process carried out with some identifiable purpose. Acts are defined as the component of the task which is independent of an individual who performs them. In figure 2, making an initial decision, reviewing model prediction, and making a final decision are classified as acts.

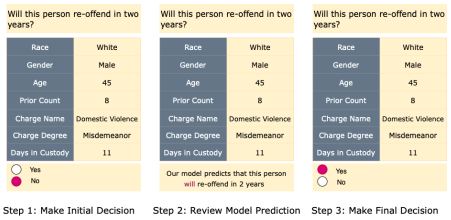


Figure 2: A decision task study. The features of the defendant profile are recognized as information cues, each step is an act, and the final decision is considered the product.

Information Cues: Information cues are pieces of information upon which an individual can make judgments during the performance of the tasks. Each variable in the defendant's profile, such as race, gender, etc., is considered an information cue. The model prediction is also a distinct information cue in Figure 2.

Acts and information cues are referred to as task inputs that determine the complexity of tasks. In other words, **task complexity** describes the relationship between task inputs and will be a significant determinant of individual performance. Task complexity is defined with three dimensions:

Component Complexity: It refers to the total number of distinct information cues that need to be processed to perform the task. In our example, race, gender, age, prior count, charge name, charge degree, days in custody, and model predictions form component complexity as 8.

Coordinative Complexity: It is defined by a number of sequences of acts that are required in the task performance. The number of steps to accomplish the task is three in figure 2.

Dynamic Complexity: Changes in either value of information cues or number of acts lead to dynamic complexity. We count the number of information cues with variable quantities or additional steps required for accomplishing the task as dynamic complexity. In our example, both component and coordinative complexities are static during the process of decision making, indicating that the dynamic complexity is 0.

Note that a task can be a combination of multiple sub-tasks. So, these definitions can be assessed at both task and sub-task levels. As a result, the overall complexity of the task in each dimension is the aggregation of the complexity across all sub-tasks. According to Wood [100], the overall task complexity is expressed as the linear combination of component, coordinative, and dynamic complexities: $TC_{overall} = \alpha \cdot TC_{component} + \beta \cdot TC_{coordinative} + \gamma \cdot TC_{dynamic}$.

However, it is not evident how the three dimensions of relate to one another Wood [100]. Therefore, we consider each dimension separately in our study.

3.3 Measuring Task Complexity

We operationalize the theoretical model of task complexity proposed in seminal work by Wood [100] in the realm of human-AI decision making. We model the information cues and required acts defined in previous Section 3.2 to decision tasks in our article collection and calculate three dimensions of complexity (i.e., component,

[1] https://osf.io/9bg8c/?view_only=7c0fedff68514fca892b16afa385a0e8

coordinate, and dynamic) for each task. Such a framework can assist us in comparing complexity across various tasks and domains.

As a first step, we need to identify the information cues and required acts in the human-AI decision making context so we can determine each dimension of complexity defined in the previous section. To identify information cues, we created a set of rubrics. Additionally, the mandatory steps that are required to complete the task are the required acts.

We categorize our rubrics into two groups: general rules applied to all cases and specific rules depending on each dimension of task complexity. In order to illustrate how these rubrics can be applied in practice, Figure 3 shows seven conditions of decision tasks along with their complexity.

3.3.1 General Rules. **Rule 1**: We excluded all task-independent components when calculating the task complexity: such as pilot studies, questionnaires for user factor assessment, tutorials before the actual decision task. As all of these factors do not contribute to the task complexity, they are discarded in our measurement.

Rule 2: When dynamic complexity is not zero, (due to changes in component or coordinative complexities), we report the minimum static complexity for component complexity/coordinative complexity. The dynamic dimension is indicated as the differential between maximum and minimum component/coordinative complexity.

Rule 3: Different experimental conditions of a decision task can vary in each dimension of complexity. We only consider the condition in which the authors investigate the effectiveness of their proposed approach or evaluate their primary hypothesis. Such a condition is typically the condition with the maximum complexity, among others.

Rule 4: We consider explanation methods as information cues. Although they are supposed to assist humans to interpret AI decisions, they augment task complexity as humans should digest them along with the AI decisions. However, such methods affect the complexity differently; one can directly increase component complexity,and the other may dynamically change leading to dynamic complexity.

Rule 5: Information cues can be presented in various ways, such as plots, paragraphs, tables, and images, each requiring a different amount of steps to interpret. Using a table as an information cue might be easier to digest than using a sophisticated plot. Since we do not have any references to determine the number of steps each require, we assume all of them have a similar coordinative complexity.

Rule 6: Tasks with different stakes (risks) may intuitively have different complexity levels. However, there is no way to map risk levels to either information cues or required actions. So, we consider it as the limitation of this framework as it can not capture them.

Rule 7: For each task, a set of features is required to make an informed decision. Missing any of these features will cause the task to be complex and error-prone. However, with this framework, we can not account for this type of complexity; since we cannot figure out this set of features for each specific task.

3.3.2 Component Complexity. **Rule 1**: We count the total number of distinct human-visible features, considering each as an information cue. The number of information cues indicates component complexity. Note that redundant information cues are not counted according to the definition.

Rule 2: In addition to features, the correlation between each combination of them is also considered a distinct information cue. If we have n features and their correlations, then the number of distinct information cues equals $2^n - 1$. For e.g., for $n = 2$, the component complexity would equal three as we have three distinct information cues: feature_1, feature_2, and correlation between them.

Rule 3: Each of the following factors is examined as one information cue: 1) model prediction, 2) model uncertainty score, 3) model performance, and 4) overview of the model or algorithm distribution.

Rule 4: Each explanation method is counted as one or more information cues. 1) feature importance highlighting key features is considered as one information cue. 2) feature contribution showing top key features and their coefficients are counted as two information cues. 3) counterfactual explanation focusing on what changes in feature values result in an opposite AI prediction are recognized as two distinct information cues - as they provide both what features and which new values, 4) demographic-based information is one information cue, 5) example-based explanation such as nearest neighbour methods is considered as one information cue if only examples with similar predictions are presented; in case of providing examples with different predictions, based on the number of various predictions, they can be counted 2 to n distinct information cues.

Rule 5: The feedback regarding the performance of the humans or AI is also considered as one information cue.

3.3.3 Coordinative Complexity. **Rule 1**: We count the total number of steps to accomplish the decision task as coordinative complexity.

Rule 2: Each task instance is recognized as one separate step.

3.3.4 Dynamic Complexity. **Rule 1**: When the quantity of any feature changes during the process of decision making (any changes in component complexity), the dynamic complexity should be greater than zero. Otherwise, the dynamic complexity is zero.

Rule 2: There is dynamic complexity if any feature affects the sequence of performing the task (changes in coordinative complexity).

Rule 3: The additive feature attribution explanation method contributes to dynamic complexity. Providing additive feature attribution method, a human decision maker can modify the values of any features (Rule 1) and observe their correlation among other features and their impact on AI predictions. Based on Rule 2 in component complexity, the maximum component complexity with n features, given their correlation, is calculated as $2^n - 1$. As we report the differential of maximum ($2^n - 1$) and minimum (n) component complexity as dynamic complexity, then, the dynamic complexity would be $2^n - 1 - n$.

Rule 4: If we let humans choose whether and when to see the AI recommendation, then the steps required to accomplish the task (coordinative complexity) would be dynamic depending on whether the decision makers request AI recommendation or not.

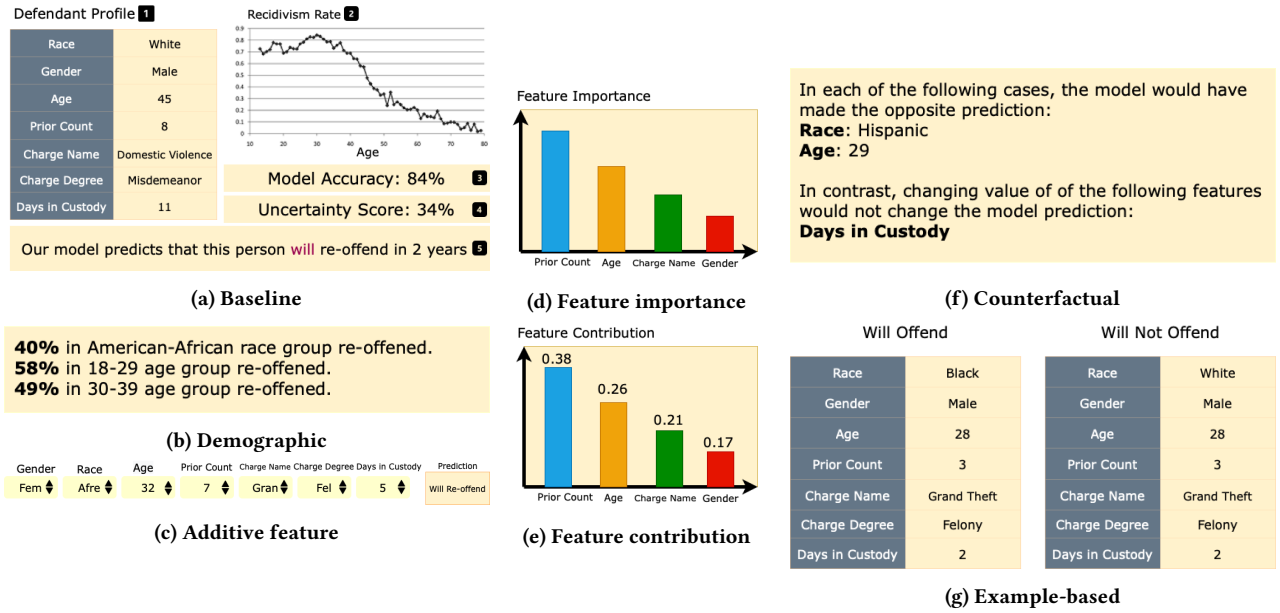


Figure 3: The complexity of different experimental conditions of a decision Task. Participants are asked to make a prediction on whether this defendant would re-offend within two years on 30 trials. The study contains seven experimental conditions providing different types of explanations. The component complexity of each conditions is: a)12, b) 13, c) 12, d) 13, e) 14, f) 14, and g) 14 . This task has coordinative and dynamic complexities of 30 and 0, respectively. Except for condition 3, the dynamic complexity is 120. More details on how to calculate each dimension will be found on the companion page.

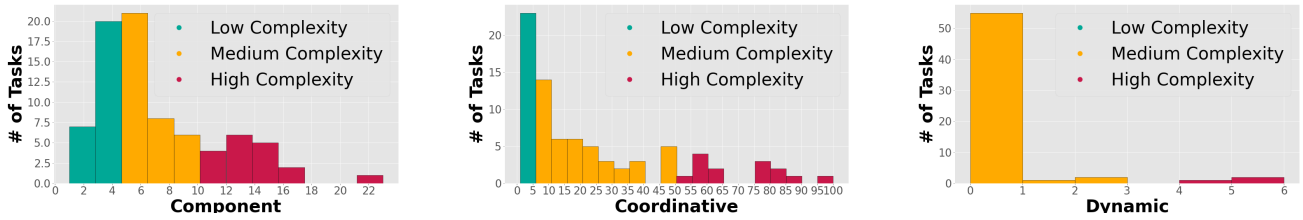
4 RESULTS

RQ1 asks to what extent recent HCI literature has considered the impact of task complexity in the design of decision making tasks. Among all the relevant articles we collected, a limited number of studies have considered task complexity in designing their decision tasks [7, 83, 91]. This finding corroborates that there is no standard framework to quantify the complexity of decision tasks. We analyzed existing tasks according to the framework we proposed in Section 3.3.2, operationalizing the measurement of task complexity. We first shed light on the descriptive statistics; distribution across the component, coordinative, and dynamic complexities, the extremities observed in our data, and the outliers.

Component Complexity: Component complexity was found to be within the range of 1 to 23, shown in Figure 4a. The task with the component complexity score of 1 is related to a mind wandering detection task in which crowd workers were asked to identify whether people's attention in the presented video clip drifted away [14]. On the other end of the spectrum, a task related to music recommendation was found to have a component complexity of 23 [70]. In this study, a wide range of features associated with user's preferences, attributes of artists, and explanations for suggested songs were incorporated. The average component complexity of tasks in our data was found to be 6.9 (±4.3).

Research in neuroscience led by Miller [71] revealed that the average human information processing capacity ranges between 5 and 9, which is the number of objects an average human can hold in their short-term memory. This indicates the range of component complexity for human decision makers. Based on this, we consider three bins of component complexity. First, tasks with the number of information cues (indicating component complexity) below 5 as those corresponding to low complexity. Next, tasks with 5-9 information cues are considered to have a medium component complexity, while those with more than 9 being highly complex. We note that the decision tasks considered in recent literature have a medium level of component complexity on average (6.9). In total, 33.7% of tasks have a low level, 40% have a medium level, and 26.2% have a high level of component complexity. Furthermore, 12% of tasks were found to be outliers with a high level of complexity between 24 to 132. The decision task with a component complexity of 132 relates to predicting the risk of not paying back a loan and a convict's chance of recidivism [95] within the same task. The study included 18 variables in which the three-dimensional relationship between some features was presented, including two decision tasks from different domains and many variables (shown as scatter plots) which increased the task complexity. All outliers leveraged specific datasets, included many features, and employed sophisticated plots.

Coordinative Complexity: We found that the coordinative complexity of tasks considered in our data lies between 1 and 100. There are four tasks with the lowest coordinative complexity, where participants were asked to react to the hypothetical scenario in which their Facebook account is suspended by an algorithmic content moderation system [94], movie recommendation [104, 104], and medical diagnosis [9]. The task with the highest coordinative complexity was found to be bail decision making for 50 cases. We found that the coordinative complexity was 25.1±25.9 on average, meaning that participants must follow 25.1 steps to accomplish



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

Figure 4: Distribution of component, coordinative, and dynamic complexity in the decision tasks corresponding to our corpus.

the task. Based on the Figure 4b, 75.6% of the complexity of tasks distributed between a range of 1 to 40.

We divided the level of coordinative complexity as low, medium, and high based on the quartiles; with the bottom quartile corresponding to low, top quartile corresponding to high, and the other two constituting the medium level. Tasks in the bin of low complexity corresponded to a coordinative complexity below 5; those with a medium level of complexity corresponded to between 5 to 50; highly complex tasks corresponded to a coordinative complexity of over 50. In total, 25.9% of tasks were found to correspond to a low level of coordinative complexity, 56.8% have a medium level of complexity, and 17.2% have a high level of complexity.

We also observed articles with coordinative complexity of 130 to 420 as outliers. In the task with a score of 420, participants were presented with 210 questions regarding quality control in a drinking glass-making factory scenario [103]. There is evidence to suggest that having more task instances, with a greater level of coordinative complexity, can cause mental fatigue. This is the result of prolonged periods of demanding cognitive activity [50] and has been shown to negatively affect performance [73, 99]. Therefore, it is important to set the number of task instances in empirical studies at a reasonable level to avoid the fatigue effect.

Dynamic Complexity: As explained, dynamic complexity depends on any changes in components or coordinative complexities. Our analysis revealed that dynamic complexity was distributed between 0 to 6 (cf. Figure 4c). We classified the level of dynamic complexity according to the bottom and top quartiles. As the result, tasks with a complexity of 0 corresponds to low dynamic complexity; task with dynamic complexity of 1 to 3 correspond to medium complexity; and tasks with dynamic complexity of 4 or more assigns to high dynamic complexity. We also observed that 95% of decision tasks have dynamic complexity between 0 to 2. This finding indicates that dynamic complexity is not common among decision tasks considered in empirical human-AI studies. The source of dynamic complexity was found to be the non-stationary nature of coordinative complexity. Some studies let the participants choose whether and when to see AI recommendations. This approach forced participants to be more cognitively involved in the decision making process by first probing the task inputs. This resulted in a variable number of steps depending on the participants, leading to dynamic complexity. We found outliers with a dynamic complexity ranging from 12 to 1890. The source of dynamicity for the study with the score of 1890 was changes in component complexity. This study includes a video activity searching tool to build specific queries and sort the number of videos about policies being followed by kitchen staff [76]. Dynamic complexity was a result of the fact that queries and responses were not constant.

Actual and Proxy Tasks: We also examined whether researchers conducted proxy tasks or actual tasks [15]. Participants in actual tasks are asked to make an informed decision with AI assistance, evaluating the performance of humans and AI as a team. In contrast, participants in proxy tasks have to simulate the model decision or decision boundaries. Buçinca et al. [15] showed how evaluations with proxy tasks do not predict the evaluation with actual tasks which can limit the generalizability of findings. In total, we observed that 86% of studies were conducted with actual tasks while the remaining 14% were proxy tasks in the set of articles in our corpus.

High-Stake and Low-Stake: We also analyzed the risk of tasks as this is identified as one of four dimensions that vary in decision tasks by Lai et al. [54]. Among all, 67.7% of studies did not specify how risky their task was. Although this aspect could be inferred from the context, this suggests a potential lack of explicit consideration of stakes. Of the remaining, 18.3% are classified as high-stake, 8.6% low-stake, and 5.4% set up their studies in both conditions, either by changing the decision task [5] or artificially modifying the scenario. For instance, in a study by Guttman et al. [42], participants were asked to check user requests for approval to run different software on company computers. In the high-stake domain, they were targeted by a malicious hacker, while in a low-stake setup, participants were told that they would be invited to a party if they performed well. As another approach for converting low-stake tasks to high-stake, participants were rewarded money/points in case of correct decisions and lost more amount of money/points serving as the punishment for incorrect decisions.

Note that we manually annotated tasks in which the stakes were not indicated. In total, 39.7% of tasks were found to be high-stakes, 54.8% were low-stake, followed by 5.4%, which contained both low-stake and high-stake scenarios. As there is limited understanding about the correlation between task stake and task complexity, future work could explore how these factors relate to each other and influence human-AI decision making. Example annotations of task complexity is shown in Tables 1a and 1b. The **full list of articles and our annotations** are publicly accessible for the benefit of the research community and in the spirit of open science.[2]

[2]https://osf.io/9bg8c/?view_only=7c0fedff68514fca892b16afa385a0e8

Decision Task	Complexity
Stroke Rehabilitation Assessment	(25,60,0)[59, 60]
Medical Image Retrieval	(14,6,1023)[19]
Medical Diagnosis	(7,4,2)[79],(6,1,0)[9], (6,240,0)[32]
Nutrition Prediction	(5,26,156)[16],(3,24,0)[15], (4,48,24)[33]
Recidivism Prediction	(10,64,32)[97],(11,200,50)[40], (8,12,6)[29],(132,10,0)[95], (11,50,0)[69]
Monitoring and Administration	(15,130,390)[81]
Job Application Approval	(5,24,0)[82]

(a) Task complexity in high-stake domains

Decision Task	Complexity
House Price Prediction	(9,15,0)[1],(11,24,12)[83], (12,40,10)[24]
Deceptive Review Prediction	(6,20,0)[55], (4,20,0)[56]
Sketch Recognition	(4,6,0)[18],(6,84,42)[105]
Movie Recommendation	(2,4,0)[52],(28,1,0)[104], (4,15,0)[63], (10,4,0)[10]
Place Recommendation	(12,2,0)[96], (51,12,0)[46]
Food Recommendation	(2,51,0)[37], (9,3,0)[72]
Image Classification	(3,12,0)[3],(5,216,0)[101], (4,90,0)[47], (3,40,0)[75]
Sentiment Analysis	(4,3,0)[89],(6,50,0)[8]

(b) Task complexity in low-stake domains

Table 1: Example annotations of task complexity in low-stake and high-stake domains. Task complexity is shown as a tuple (component, coordinative, dynamic).

RQ2 focuses on how task complexity can facilitate a comparative lens for empirical work on human-AI decision making. To address this research question, we examined tasks in each level of complexity, from low to high, and across all complexity dimensions. We found that there were some consistencies within each level of complexity and across dimension such as task stake, task expertise, and task type. However, there were also some differences across the levels in each dimension, which we discuss below.

Different Tasks Same Complexities: Our analysis has indicated that there are different decision tasks with the same complexity. For most score levels of component complexity, there are at least two studies with the same score. Among all cases, score 3 and 5 is dominant, with 11 and 12 decision tasks, respectively. Considering low-complex tasks, they are comparable in terms of their stake, domain expertise, and task types. More than 93% are low-stake tasks that can be accomplished without domain knowledge. A majority of these tasks involve recommendations or binary decisions. Such binary decision tasks are primarily artificial, with no explicit real-world applications. As the tasks are straightforward, they imply lower demand for humans to rely on [27].

Looking towards tasks with higher levels of complexity, we observe diversity among tasks in terms of their stake and expertise, which is not comparable. For instance, we found two scenarios of child clinical decision making [51] and nutrition prediction [17] tasks with the same component complexity; the first high-stakes task requires extensive domain knowledge, while the second task can be performed without any background and has a low risk. In addition to binary decision and recommendation tasks, tasks in the bin of medium complexity were found to include multi-class, regression, and retrieval tasks. Compared to low-complex tasks, the number of tasks resembling real-world problems was found to be higher in the bin of tasks with medium level of complexity. Our examination, established that around half of the tasks were still artificial [80, 105] or do not necessarily require human intervention. For instance, sentiment analysis [89] and text classification [86] tasks can be fully automatic; thereby, human intervention may not be needed.

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. The existing low-stake tasks in this bin are dedicated to recommendation systems tasks. Including a wide range of features to capture human preferences makes such recommendation tasks complex. It is important to point out that high-complex tasks are found to be explicit examples of real-life problems. In our study, we found that as tasks got more complex, they resembled real-world use cases more, demanded more domain knowledge, and had a bigger stake. To simulate real-world problems and human interaction with AI systems, it is pragmatic to adapt actual tasks in which humans may want to rely on AI support.

We also observed decision tasks with similar scores of coordinative complexity, representing the number of steps required to finish the tasks. Tasks with low levels of complexity consist of low-stake tasks without the expertise needed. On the other end, high-complex tasks were found to have higher risks and require expertise. Interestingly, these tasks also have a medium or high score of component complexity at the same time. That could be due to the fact that researchers may increase task instances for such tasks to examine human behaviors over time. Consequently, human decision-makers familiarize themselves with AI systems, form their mental models, and calibrate their trust. Nevertheless, having a higher level of component complexity and stake for these tasks, the cognitive load of performing tasks could grow simultaneously. Such cognitive load could lead to mental fatigue in participants earlier [50]. It seems that researchers might neglect the fatigue effect in their studies. Comparing these user studies with actual scenarios, it's also rare for a human decision maker to do 100 tasks concurrently in real-life cases. Instead of doing a hundred cases in one session, it's recommended to examine human behavior over time in different sessions to mitigate the fatigue effect and model the real world better.

Based on our analysis, low-complexity tasks have almost no expertise required, are low-stake, and are easy to do, so findings across studies can be generalizable. With the increasing complexity of tasks, we observe a wide variety of task types, task stakes, more features, and a variety of explanation methods adapted, which makes it hard to carry findings from one study to another.

Same Tasks Different Complexities: In contrast to studies with the same complexity score, we found some similar decision tasks with varied complexity scores. Recidivism prediction, loan approval, movie recommendation, and image classification tasks dominate our corpus. These tasks are presented in each level of component complexity, low to high. The underlying dataset is similar throughout the studies for the recidivism prediction [95, 98] and loan approval tasks [35, 38]. They incorporate different explanation methods, enriched with additional visualizations along with the AI decision to modify the component complexity of those tasks. For the movie recommendation [52, 104], the reason for having a spectrum of complexity is integrating different user preferences to improve the quality of AI recommendations. Lastly, we can see distinct types of images in image classification tasks [3, 13], from clinical, nutrition-related, animal, and animal images. Since the context of images differs, the type and number of component complexity vary among them. 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

Reflecting on tasks with a high complexity, we observed that interest in promoting the need to rely on AI or opportunities to propose explanation methods can typically inform such task design choices. Researchers have shown that explanations can effectively inform mental models of humans and improve their understanding, especially for laypeople [35, 36, 44, 90]. Additionally, to fill the knowledge gap between domain experts and laypeople or improve AI literacy, empirical studies engage with more explanations [22, 62, 87, 105]. A higher level of complexity can also result from adding more user preferences to improve the quality of recommendations [46, 72, 104].

Another reason to increase task complexity could be a need to study trust formation and reliance on AI systems in such contexts [24, 67, 83, 102, 104]. Incorporating various features can ensure that human decision-makers access salient features required to make better decisions, especially in high-stake domains [13, 19, 40, 60]. Missing a salient feature could be more harmful than presenting additional information.

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 [62, 62, 95]. Additionally, cognitive forcing interventions are applied in studies to engage human decision-makers more thoughtfully with AI systems. By increasing task complexity, such approaches affect human cognitive processes by: (I) asking humans to make decisions before seeing model predictions [97, 107], (II) varying AI systems response time [16, 80], and (III) providing feedback to humans [8, 40, 101, 103].

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 [15, 52, 89, 103]. Tasks with low complexity can help human decision makers have a better understanding of AI systems [18, 88].

5.2 Common Limitations and Challenges in Empirical Human-AI Studies

Our observations indicate an arbitrary selection of task design parameters like task type, number of features, etc is common in existing empirical studies. Depending on the factors that are investigated in experiments, task types can play a significant role. For instance, the pattern of reliance on AI systems in a decision task with high-stakes in healthcare domain could vary in comparison to a low-stakes task in the commercial domain. Features of a decision task are largely set from the dataset that the AI systems were trained on. However, not all of the features are relevant to a given task and some of them can even mislead users. Task parameters are also typically determined due to external factors such as associated costs, or available time. In many studies, the number of task instances, as well as the length of the study, are set according to the available budget. Furthermore, limited access to domain experts (where expertise is required) results in studies with fewer participants. Regarding whether tasks are actual or proxy tasks [15], we observed that 25% of studies in our corpus employed proxy tasks to evaluate their hypotheses. Evaluations of human-AI decision making using proxy tasks do not necessarily transfer to actual real-world tasks [15]. This pitfall can affect the generalizability and reliability of findings. A human-subject study often hinges on simulating real-world tasks accurately. While many parameters have to be simplified in isolated studies, the simulation still needs to be valid. There are sometimes tasks that are artificially created [18, 105], or tasks that do not fit into human-AI decision making are adapted [89, 101].

Explanation methods increase human understanding and transparency, but can inadvertently increase task complexity, which can be in conflict with what they are meant to achieve. It can be also detrimental to human-AI complementary performance if a lot of complicated and diverse visuals of task features are presented [90]. Although decision tasks are sometimes simplified for lay people to complete, some knowledge and familiarity, such as AI literacy [24], numeracy, and statistics background [41], may still be required to accomplish tasks, which may not be feasible to expect from all crowd workers — expert recruitment on-demand remains a challenge.

5.3 Caveats and Limitations of This Study

We limited our scope to articles published in HCI venues published in the last four years. Although our corpus is representative of literature, our sample frame may have resulted in not considering related articles published in other venues. We do not claim to provide exhaustive insights into why task complexity has not been widely considered in empirical human-AI studies. 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.

6 CONCLUSIONS AND FUTURE WORK

In our paper, we examined to what extent recent literature in human-AI decision making has considered task complexity in the design of empirical studies (**RQ1**) and how task complexity can facilitate comparisons across experimental settings (**RQ2**). To answer our research questions, 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 effect 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. Based on our analysis of empirical human-AI studies, we found that it is important to measure and report the different types of expertise or domain knowledge that the participants might have (numeracy, AI literacy, familiarity with statistics or information visualization), so that comparisons across studies can be made meaningfully. Future work in this realm can consider explicitly controlling the level of task complexity across the experimental conditions. In the imminent future, we aim to expand our operationalization of task complexity to account for other task features and build a tool that can automatically measure the complexity of tasks across Wood's three dimensions and inform researchers in their design of empirical human-AI studies.

REFERENCES

[1] Ashraf Abdul, Christian von der Weth, Mohan Kankanhalli, and Brian Y. Lim. 2020. COGAM: Measuring and Moderating Cognitive Load in Machine Learning Model Explanations. 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, 1–14. https://doi.org/10.1145/3313831.3376615

[2] 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

[3] Ahmed Alqaraawi, Martin Schuessler, Philipp Weiß, Enrico Costanza, and Nadia Berthouze. 2020. Evaluating Saliency Map Explanations for Convolutional Neural Networks: A User Study. In *Proceedings of the 25th International Conference on Intelligent User Interfaces* (Cagliari, Italy) *(IUI '20)*. Association for Computing Machinery, New York, NY, USA, 275–285. https://doi.org/10.1145/3377325.3377519

[4] Alessa Angerschmid, Jianlong Zhou, Kevin Theuermann, Fang Chen, and Andreas Holzinger. 2022. Fairness and Explanation in AI-Informed Decision Making. *Machine Learning and Knowledge Extraction* 4, 2 (2022), 556–579. https://doi.org/10.3390/make4020026

[5] Ariful Islam Anik and Andrea Bunt. 2021. Data-Centric Explanations: Explaining Training Data of Machine Learning Systems to Promote Transparency. 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 75, 13 pages. https://doi.org/10.1145/3411764.3445736

[6] Mihalj Bakator and Dragica Radosav. 2018. Deep learning and medical diagnosis: A review of literature. *Multimodal Technologies and Interaction* 2, 3 (2018), 47.

[7] 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

[8] 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

[9] Emma Beede, Elizabeth Baylor, Fred Hersch, Anna Iurchenko, Lauren Wilcox, Paisan Ruamviboonsuk, and Laura M. Vardoulakis. 2020. A Human-Centered Evaluation of a Deep Learning System Deployed in Clinics for the Detection of Diabetic Retinopathy. 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, 1–12. https://doi.org/10.1145/3313831.3376718

[10] Andrew Bell, Ian Solano-Kamaiko, Oded Nov, and Julia Stoyanovich. 2022. It's Just Not That Simple: An Empirical Study of the Accuracy-Explainability Trade-off in Machine Learning for Public Policy. In *2022 ACM Conference on Fairness, Accountability, and Transparency* (Seoul, Republic of Korea) *(FAccT '22)*. Association for Computing Machinery, New York, NY, USA, 248–266. https://doi.org/10.1145/3531146.3533090

[11] 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

[12] Or Biran and Kathleen McKeown. 2017. Human-Centric Justification of Machine Learning Predictions. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence* (Melbourne, Australia) *(IJCAI'17)*. AAAI Press, 1461–1467.

[13] 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

[14] Nigel Bosch and Sidney K. D'Mello. 2022. Can Computers Outperform Humans in Detecting User Zone-Outs? Implications for Intelligent Interfaces. *ACM Trans. Comput.-Hum. Interact.* 29, 2, Article 10 (jan 2022), 33 pages. https://doi.org/10.1145/3481889

[15] 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

[16] 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

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

[18] 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

[19] 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

[20] Donald J. Campbell. 1988. Task Complexity: A Review and Analysis. *The Academy of Management Review* 13, 1 (1988), 40–52. http://www.jstor.org/stable/258353

[21] 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

[22] 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

[23] Chun-Wei Chiang and Ming Yin. 2021. You'd Better Stop! Understanding Human Reliance on Machine Learning Models under Covariate Shift. In *13th ACM Web Science Conference 2021* (Virtual Event, United Kingdom) *(WebSci '21)*. Association for Computing Machinery, New York, NY, USA, 120–129. https://doi.org/10.1145/3447535.3462487

[24] 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

[25] Michael Chromik, Malin Eiband, Felicitas Buchner, Adrian Krüger, and Andreas Butz. 2021. I Think I Get Your Point, AI! The Illusion of Explanatory Depth in Explainable AI. In *26th International Conference on Intelligent User Interfaces* (College Station, TX, USA) *(IUI '21)*. Association for Computing Machinery, New York, NY, USA, 307–317. https://doi.org/10.1145/3397481.3450644

[26] 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

[27] 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

[28] Murat Dikmen and Catherine Burns. 2022. The effects of domain knowledge on trust in explainable AI and task performance: A case of peer-to-peer lending. *International Journal of Human-Computer Studies* 162 (2022), 102792.

[29] Jonathan Dodge, Q. Vera Liao, Yunfeng Zhang, Rachel K. E. Bellamy, and Casey Dugan. 2019. Explaining Models: An Empirical Study of How Explanations Impact Fairness Judgment. 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, 275–285. https://doi.org/10.1145/3301275.3302310

[30] Finale Doshi-Velez and Been Kim. 2017. Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608* (2017).

[31] Julia Dressel and Hany Farid. 2018. The accuracy, fairness, and limits of predicting recidivism. *Science Advances* 4, 1 (2018), eaao5580. https://doi.org/10.1126/sciadv.aao5580 arXiv:https://www.science.org/doi/pdf/10.1126/sciadv.aao5580

[32] Riccardo Fogliato, Shreya Chappidi, Matthew Lungren, Paul Fisher, Diane Wilson, Michael Fitzke, Mark Parkinson, Eric Horvitz, Kori Inkpen, and Besmira Nushi. 2022. Who Goes First? Influences of Human-AI Workflow on Decision Making in Clinical Imaging. In *2022 ACM Conference on Fairness, Accountability, and Transparency* (Seoul, Republic of Korea) *(FAccT '22)*. Association for Computing Machinery, New York, NY, USA, 1362–1374. https://doi.org/10.1145/3531146.3533193

[33] Krzysztof Z. Gajos and Lena Mamykina. 2022. Do People Engage Cognitively with AI? Impact of AI Assistance on Incidental Learning. In *27th International Conference on Intelligent User Interfaces* (Helsinki, Finland) *(IUI '22)*. Association for Computing Machinery, New York, NY, USA, 794–806. https://doi.org/10.1145/3490099.3511138

[34] Ella Glikson and Anita Woolley. 2020. Human trust in artificial intelligence: Review of empirical research. Academy of Management Annals (in press). *The Academy of Management Annals* (04 2020).

[35] 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

[36] 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

[37] Kazjon Grace, Elanor Finch, Natalia Gulbransen-Diaz, and Hamish Henderson. 2022. Q-Chef: The Impact of Surprise-Eliciting Systems on Food-Related Decision-Making. 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 11, 14 pages. https://doi.org/10.1145/3491102.3501862

[38] 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 (nov 2019), 24 pages. https://doi.org/10.1145/3359152

[39] Ben Green and Yiling Chen. 2020. Algorithmic risk assessments can alter human decision-making processes in high-stakes government contexts. *arXiv preprint arXiv:2012.05370* (2020).

[40] 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

[41] Shunan Guo, Fan Du, Sana Malik, Eunyee Koh, Sungchul Kim, Zhicheng Liu, Donghyun Kim, Hongyuan Zha, and Nan Cao. 2019. Visualizing Uncertainty and Alternatives in Event Sequence Predictions. 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/10.1145/3290605.3300803

[42] Rotem D. Guttman, Jessica Hammer, Erik Harpstead, and Carol J. Smith. 2021. Play for Real(Ism) - Using Games to Predict Human-AI Interactions in the Real World. *Proc. ACM Hum.-Comput. Interact.* 5, CHI PLAY, Article 228 (oct 2021), 17 pages. https://doi.org/10.1145/3474655

[43] 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

[44] 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

[45] 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

[46] 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

[47] Donald Honeycutt, Mahsan Nourani, and Eric Ragan. 2020. Soliciting Human-in-the-Loop User Feedback for Interactive Machine Learning Reduces User Trust and Impressions of Model Accuracy. *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing* 8, 1 (Oct. 2020), 63–72. https://doi.org/10.1609/hcomp.v8i1.7464

[48] THORVALD HÆREM, BRIAN T. PENTLAND, and KENT D. MILLER. 2015. TASK COMPLEXITY: EXTENDING A CORE CONCEPT. *The Academy of Management Review* 40, 3 (2015), 446–460. http://www.jstor.org/stable/43700530

[49] Mir Riyanul Islam, Mobyen Uddin Ahmed, Shaibal Barua, and Shahina Begum. 2022. A Systematic Review of Explainable Artificial Intelligence in Terms of Different Application Domains and Tasks. *Applied Sciences* 12, 3 (2022). https://doi.org/10.3390/app12031353

[50] 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

[51] Anna Kawakami, Venkatesh Sivaraman, Hao-Fei Cheng, Logan Stapleton, Yanghuidi Cheng, Diana Qing, Adam Perer, Zhiwei Steven Wu, Haiyi Zhu, and Kenneth Holstein. 2022. Improving Human-AI Partnerships in Child Welfare: Understanding Worker Practices, Challenges, and Desires for Algorithmic Decision Support. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (New Orleans, LA, USA) *(CHI '22)*. Association for Computing Machinery, New York, NY, USA, Article 52, 18 pages. https://doi.org/10.1145/3491102.3517439

[52] 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

[53] 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).

[54] Vivian Lai, Chacha Chen, Qingzi Vera Liao, Alison Smith-Renner, and Chenhao Tan. 2021. Towards a Science of Human-AI Decision Making: A Survey of Empirical Studies. *ArXiv* abs/2112.11471 (2021).

[55] 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

[56] Vivian Lai and Chenhao Tan. 2019. On Human Predictions with Explanations and Predictions of Machine Learning Models: A Case Study on Deception Detection. In *Proceedings of the Conference on Fairness, Accountability, and Transparency* (Atlanta, GA, USA) *(FAT* '19)*. Association for Computing Machinery, New York, NY, USA, 29–38. https://doi.org/10.1145/3287560.3287590

[57] Himabindu Lakkaraju, Stephen H. Bach, and Jure Leskovec. 2016. Interpretable Decision Sets: A Joint Framework for Description and Prediction. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (San Francisco, California, USA) *(KDD '16)*. Association for Computing Machinery, New York, NY, USA, 1675–1684. https://doi.org/10.1145/2939672.2939874

[58] John D Lee and Katrina A See. 2004. Trust in automation: Designing for appropriate reliance. *Human factors* 46, 1 (2004), 50–80.

[59] Min Hun Lee, Daniel P. Siewiorek, Asim Smailagic, Alexandre Bernardino, and Sergi Bermúdez i Badia. 2020. Co-Design and Evaluation of an Intelligent Decision Support System for Stroke Rehabilitation Assessment. *Proc. ACM Hum.-Comput. Interact.* 4, CSCW2, Article 156 (oct 2020), 27 pages. https://doi.org/10.1145/3415227

[60] 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

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

[62] 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

[63] Mengqi Liao, S. Shyam Sundar, and Joseph B. Walther. 2022. User Trust in Recommendation Systems: A Comparison of Content-Based, Collaborative and Demographic Filtering. 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 486, 14 pages. https://doi.org/10.1145/3491102.3501936

[64] Zhiyuan "Jerry" Lin, Jongbin Jung, Sharad Goel, and Jennifer Skeem. 2020. The limits of human predictions of recidivism. *Science Advances* 6, 7 (2020), eaaz0652. https://doi.org/10.1126/sciadv.aaz0652 arXiv:https://www.science.org/doi/pdf/10.1126/sciadv.aaz0652

[65] Han Liu, Vivian Lai, and Chenhao Tan. 2021. Understanding the Effect of Out-of-Distribution Examples and Interactive Explanations on Human-AI Decision Making. *Proc. ACM Hum.-Comput. Interact.* 5, CSCW2, Article 408 (oct 2021), 45 pages. https://doi.org/10.1145/3479552

[66] 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

[67] 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

[68] 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

[69] Keri Mallari, Kori Inkpen, Paul Johns, Sarah Tan, Divya Ramesh, and Ece Kamar. 2020. Do I Look Like a Criminal? Examining How Race Presentation Impacts Human Judgement of Recidivism. 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/10.1145/3313831.3376257

[70] Martijn Millecamp, Nyi Nyi Htun, Cristina Conati, and Katrien Verbert. 2020. What's in a User? Towards Personalising Transparency for Music Recommender Interfaces. In *Proceedings of the 28th ACM Conference on User Modeling, Adaptation and Personalization* (Genoa, Italy) *(UMAP '20)*. Association for Computing Machinery, New York, NY, USA, 173–182. https://doi.org/10.1145/3340631.3394844

[71] George A. Miller. 1956. The Magical Number Seven, Plus or Minus Two: Some Limits on Our Capacity for Processing Information. *The Psychological Review* 63, 2 (March 1956), 81–97. http://www.musanim.com/miller1956/

[72] 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

[73] Tina Möckel, Christian Beste, and Edmund Wascher. 2015. The Effects of Time on Task in Response Selection - An ERP Study of Mental Fatigue. *Scientific Reports* 5 (03 2015). https://doi.org/10.1038/srep10113

[74] Meike Nauta, Jan Trienes, Shreyasi Pathak, Elisa Nguyen, Michelle Peters, Yasmin Schmitt, Jörg Schlötterer, Maurice van Keulen, and Christin Seifert. 2022. From Anecdotal Evidence to Quantitative Evaluation Methods: A Systematic Review on Evaluating Explainable AI. https://doi.org/10.48550/ARXIV.2201.08164

[75] Mahsan Nourani, Joanie King, and Eric Ragan. 2020. The Role of Domain Expertise in User Trust and the Impact of First Impressions with Intelligent Systems. *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing* 8, 1 (Oct. 2020), 112–121. https://doi.org/10.1609/hcomp.v8i1.7469

[76] 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* (College Station, TX, USA) *(IUI '21)*. Association for Computing Machinery, New York, NY, USA, 340–350. https://doi.org/10.1145/3397481.3450639

[77] 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

[78] Ahmet Murat Ozbayoglu, Mehmet Ugur Gudelek, and Omer Berat Sezer. 2020. Deep learning for financial applications: A survey. *Applied Soft Computing* 93 (2020), 106384.

[79] Cecilia Panigutti, Andrea Beretta, Fosca Giannotti, and Dino Pedreschi. 2022. Understanding the Impact of Explanations on Advice-Taking: A User Study for AI-Based Clinical Decision Support Systems. 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 568, 9 pages. https://doi.org/10.1145/3491102.3502104

[80] 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

[81] Andisheh Partovi, Ingrid Zukerman, Kai Zhan, Nora Hamacher, and Jakob Hohwy. 2019. Relationship between Device Performance, Trust and User Behaviour in a Care-Taking Scenario. 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, 61–69. https://doi.org/10.1145/3320435.3320440

[82] Andi Peng, Besmira Nushi, Emre Kıcıman, Kori Inkpen, Siddharth Suri, and Ece Kamar. 2019. What You See Is What You Get? The Impact of Representation Criteria on Human Bias in Hiring. *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing* 7, 1, 125–134. https://doi.org/10.1609/hcomp.v7i1.5281

[83] 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

[84] 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.

[85] Amy Rechkemmer and Ming Yin. 2022. When Confidence Meets Accuracy: Exploring the Effects of Multiple Performance Indicators on Trust in Machine Learning Models. In *CHI '22: CHI Conference on Human Factors in Computing Systems, New Orleans, LA, USA, 29 April 2022 - 5 May 2022*, Simone D. J. Barbosa, Cliff Lampe, Caroline Appert, David A. Shamma, Steven Mark Drucker, Julie R. Williamson, and Koji Yatani (Eds.). ACM, 535:1–535:14.

[86] 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

[87] 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

[88] 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

[89] 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

[90] 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

[91] Suzanne Tolmeijer, Ujwal Gadiraju, Ramya Ghantasala, Akshit Gupta, and Abraham Bernstein. 2021. Second Chance for a First Impression? Trust Development in Intelligent System Interaction. 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, 77–87. https://doi.org/10.1145/3450613.3456817

[92] Richard Tomsett, Alun Preece, Dave Braines, Federico Cerutti, Supriyo Chakraborty, Mani Srivastava, Gavin Pearson, and Lance Kaplan. 2020. Rapid trust calibration through interpretable and uncertainty-aware AI. *Patterns* 1, 4 (2020), 100049.

[93] Aybike Ulusan, Uttkarsh Narayan, Sam Snodgrass, Ozlem Ergun, and Casper Harteveld. 2022. "Rather Solve the Problem from Scratch": Gamesploring Human-Machine Collaboration for Optimizing the Debris Collection Problem. In *27th International Conference on Intelligent User Interfaces* (Helsinki, Finland) *(IUI '22)*. Association for Computing Machinery, New York, NY, USA, 604–619. https://doi.org/10.1145/3490099.3511163

[94] Kristen Vaccaro, Christian Sandvig, and Karrie Karahalios. 2020. "At the End of the Day Facebook Does What ItWants": How Users Experience Contesting Algorithmic Content Moderation. *Proc. ACM Hum.-Comput. Interact.* 4, CSCW2, Article 167 (oct 2020), 22 pages. https://doi.org/10.1145/3415238

[95] 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

[96] Sruthi Viswanathan, Behrooz Omidvar-Tehrani, and Jean-Michel Renders. 2022. What is Your Current Mindset?. 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 116, 17 pages. https://doi.org/10.1145/3491102.3501912

[97] 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

[98] 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

[99] Edmund Wascher, Björn Rasch, Jessica Sänger, Sven Hoffmann, Daniel Schneider, Gerhard Rinkenauer, Herbert Heuer, and Marie Gutberlet. 2013. Frontal theta activity reflects distinct aspects of mental fatigue. *Biological psychology* 96 (12 2013). https://doi.org/10.1016/j.biopsycho.2013.11.010

[100] 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

[101] 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

[102] 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

[103] 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

[104] 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

[105] 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

[106] 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

[107] 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