Interpretability Gone Bad: The Role of Bounded Rationality in How Practitioners Understand Machine Learning

HARMANPREET KAUR*, University of Minnesota, USA
MATTHEW R. CONRAD† and DAVIS RULE†, University of Michigan, USA
CLIFF LAMPE, University of Michigan, USA
ERIC GILBERT, University of Michigan, USA

While interpretability tools are intended to help people better understand machine learning (ML), we find that they can, in fact, impair understanding. This paper presents a pre-registered, controlled experiment showing that ML practitioners (N = 119) spent *5x less time* on task, and were *17% less accurate* about the data and model, when given access to interpretability tools. We present bounded rationality as the theoretical reason behind these findings. Bounded rationality presumes human departures from perfect rationality, and it is often effectuated by satisficing, i.e., an inclination towards "good enough" understanding. Adding interactive elements—a strategy often employed to promote deliberative thinking and engagement, and tested in our experiment—also does not help. We discuss implications for interpretability designers and researchers related to how cognitive and contextual factors can affect the effectiveness of interpretability tool use.

CCS Concepts: • **Human-centered computing** → **Empirical studies in HCI**; • **Computing methodologies** → **Machine learning**.

Additional Key Words and Phrases: interpretability, explainability, bounded rationality, cognitive science

ACM Reference Format:
Harmanpreet Kaur, Matthew R. Conrad, Davis Rule, Cliff Lampe, and Eric Gilbert. 2024. Interpretability Gone Bad: The Role of Bounded Rationality in How Practitioners Understand Machine Learning. *Proc. ACM Hum.-Comput. Interact.* 8, CSCW1, Article 77 (April 2024), 34 pages. https://doi.org/10.1145/3637354

1 INTRODUCTION

Harmful use of ML can only be avoided if the people who build and use ML models understand the reasoning behind their predictions. The ML community has developed approaches like interpretability and explainability to help people understand ML outputs and reasoning. These include models that are inherently interpretable (e.g., decision trees [101], generalized additive models (GAMs) [14, 41]) and post-hoc explanations for the predictions made by blackbox models (e.g., LIME [104], SHAP [77]). Furthermore, interpretability and explainability approaches have been incorporated in tools that not only provide text or visual explanations, but also engage people with interactive features for data and model exploration (e.g., counterfactual analysis, responsive UI

*This work was completed while the author was a PhD student at the University of Michigan.
†Both authors contributed equally to this research.

Authors' addresses: Harmanpreet Kaur, harmank@umn.edu, University of Minnesota, Minneapolis, Minnesota, USA; Matthew R. Conrad, conradmr@umich.edu; Davis Rule, djrule@umich.edu, University of Michigan, Ann Arbor, Michigan, USA; Cliff Lampe, cacl@umich.edu, University of Michigan, Ann Arbor, Michigan, USA; Eric Gilbert, eegg@umich.edu, University of Michigan, Ann Arbor, Michigan, USA.

ACM 2573-0142/2024/4-ART77
https://doi.org/10.1145/3637354

elements for comparing individual datapoints, error cohorts, etc.). Toolkits like Google's What-if Tool, Microsoft's Responsible AI Toolbox, and IBM's AI Fairness 360, all offer a variety of these interactive features to promote model interpretability and explainability.

However, several studies with both ML practitioners and lay end-users have shown that interpretability and explainability approaches, and tools that implement these, do not work as intended. Evidence suggests that people misuse and over-trust interpretability tools,[1] and are unable to make accurate judgements about the data and model despite having access to the additional information provided by these tools [9, 13, 59, 62]. These user studies validate in practice the work of several scholars who have translated theories from the social sciences (e.g., [43, 118]), cognitive science (e.g., [68, 75]), philosophy (e.g., [42, 73, 97]), and organizational science (e.g., [81, 134]) for the ML context [30, 58, 84, 86]. These scholars assert that we need to learn from how people make sense of something—individually and collectively—and design solutions that support similar sensemaking for ML outputs and reasoning. This line of prior work proposes a shift in focus from generating different types of explanations to understanding what people need from an explanation. Our research is an extension of this human-centered shift in perspective. We consider the question of *why* interpretability tools are inadequately used.

We hypothesize *bounded rationality* as being the underlying reason for inadequate use of interpretability tools. Bounded rationality suggests a "kind of rational behavior that is compatible with the access to information and the computational capacities that are *actually* possessed by organisms, including man, in the kinds of *environments* in which such organisms exist" (emphasis our own) [114, p99]. Under this model of decision-making, people select "good enough" options rather than considering the utility of all alternatives. For example, a diner may default to the most commonly purchased or highly-rated items in a food delivery app, rather than evaluate all possible menu items. Bounded rationality is an innate feature of human decision-making; people can rarely consider the utility of all possible choices before coming to a decision—it would lead to information overload and decision paralysis. However, whether the outcomes of bounded rationality are good or bad is dependent on the heuristics that people apply to select a good enough option. When these heuristics are inaccurate, bounded rationality can lead to and propagate harmful judgements. Therefore, for this ML-based setting, we ask the following questions: *To what extent do people apply bounded rationality when using interpretability tools? Does the application of bounded rationality help or hurt in this context?*

For a concrete example of bounded rationality in the ML context, imagine you are a ML practitioner analyzing the Titanic survival dataset.[2] This dataset is used to predict which passengers survived the sinking of the Titanic based on demographic and socio-economic features. Say we built a blackbox model for this dataset and are using the SHAP [77] Python package as a post-hoc explainer. Figure 1 presents the three types of visuals available via SHAP (and most interpretability tools): (1) global explanation, showing the average impact of each feature on the model's predictions; (2) partial dependence plots, showing the relationship between one input feature and the output variable; and (3) local explanations, describing how an individual prediction was made. As a ML practitioner trying to understand the model outputs, one approach you might take is to explore counterfactuals: what is the smallest change in input features that would cause a prediction to flip? Consider this question for the datapoint presented in the bottom row of Figure 1. Here are some options and reasoning possibilities for switching the prediction of this datapoint from 1 (survived) to 0 (did not survive): (1) change "TotalFamily" from 3 to 5, because if you had more people in your

[1]We use the term "interpretability" to indicate both interpretability and explainability approaches and tools throughout the paper. For the purposes of our study and the tools that we employ, the two terms can be considered interchangeable.
[2]https://www.kaggle.com/c/titanic

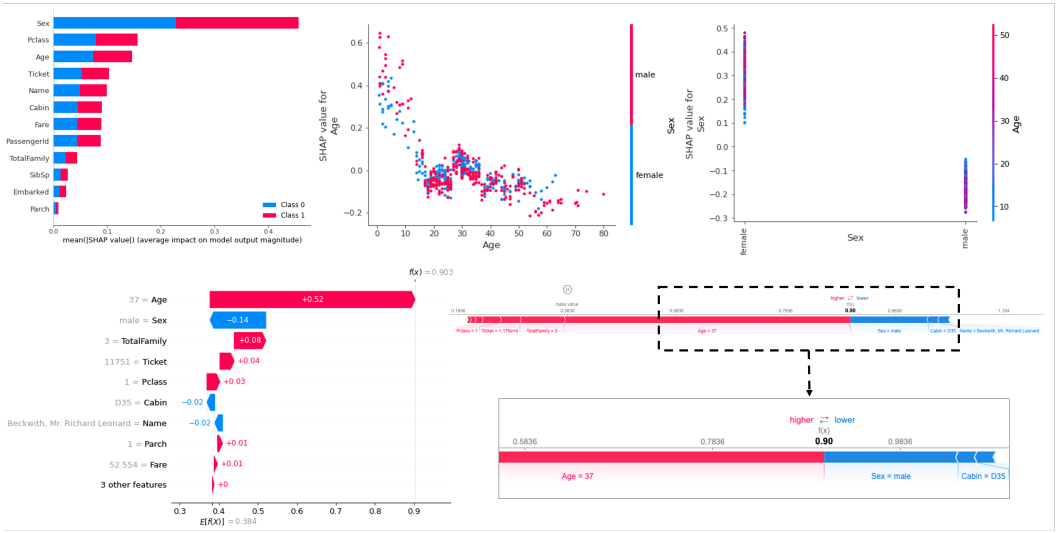

Fig. 1. Visualizations output by the SHAP Python package, a post-hoc explainer for blackbox models. These are generated for the Titanic survival dataset using a LightGBM model. Top (left to right): Global explanation; Partial dependence plot for a continuous input feature, age; Partial dependence plot for a categorical input feature, sex. Bottom (left to right): Waterfall plot and Force plot, both types of local explanations for an individual data point.

family, your attention would have been divided in trying to make sure they all reached the rafts; (2) change "Fare" from \$52.55 to \$10, because a lower fare would likely correspond to a lower passenger class, who were assigned cabins in the lower decks; or (3) change "Age" from 37 to 70, because the likelihood of survival is lower for older passengers. Which option would you pick? Under bounded rationality, people look for a plausible answer, which can be accurate or inaccurate. Of these options, (1) and (2) are accurate; (2) is also plausible, thus a case of good, accurate outcomes; and (3) is plausible but inaccurate (due to the mediating effect of "Pclass" being first), thus a case of bad outcomes from applying bounded rationality.

To observe the role of bounded rationality in the ML context, we conducted a between-subjects, pre-registered, controlled experiment with ML practitioners (N = 119),[3] asking them to perform exploratory data analysis and answer questions about the data and model that closely resembled the example above. Before we briefly describe our setup, it is worth highlighting the trade-offs between internal and ecological validity that make these kinds of evaluations hard to conduct. On the one hand, any experiment requires consistent setup and data collection mechanisms. On the other hand, ML practitioners typically have unique and personalized work setups, making it hard to simulate something realistic but also consistent for everyone in the experiment. This is only made more challenging by the diversity of tools available for ML and interpretability, all with their unique, sometimes incompatible, setup requirements. In sum, our domain (ML and interpretability), participant pool (ML practitioners), research goals (hypothesis testing), and construct (the abstract and individualized notion of bounded rationality), all have conflicting yet critical requirements.

[3]We use the broad category term ML practitioners to represent people with prior experience in ML. These include data scientists, practitioners, software designers/developers for ML-based systems, ML researchers, and ML engineers.

We worked around these issues by using a joint setup where participants were given access to a Google Colab notebook with the relevant code and asked questions (multiple-choice questions (MCQs), ratings, open-text) concurrently via a Qualtrics survey. In terms of interpretability approaches, our experiment design was split into five conditions: one representing the normal ML pipeline without built-in interpretable outputs (control); two conditions representing visual explanations output by implementations of interpretability approaches, one glassbox model (GAMs) and the other a post-hoc explainer for a blackbox model (SHAP); and two conditions representing the full scope of interactive features offered by interpretability toolkits, one built on a glassbox model (Microsoft's Explanation Dashboard) and the other relying on a post-hoc explainer for a blackbox model (Google's What-if Tool). Our glassbox conditions relied on the same underlying model, and the blackbox conditions both used the same post-hoc explainer and blackbox model. The questions about the data and model were also consistent across all conditions.

We found significant evidence showing that interpretability tools lead to bounded rationality with inaccurate outcomes. People who relied on visual explanations from static interpretability tools spent *5x less time* on the data science task and had *17% less accurate* answers compared to control. We had hoped and hypothesized that interactive interpretability tools would promote engagement and more deliberative thinking, leading to accurate outcomes from bounded rationality. However, in practice, not only did participants in our interactive tools conditions rely on bounded rationality with inaccurate outcomes, but they did so under higher cognitive effort and lower reported usability scores. Our exploratory analyses also indicate that prior experience with ML or interpretability and a more accurate mental model of the setup used do not disengage bounded rationality modes. Instead, seemingly negative user experience design attributes like lower confidence, lower usability, and higher skepticism appear to coincide more with accurate answers about the data and model. Given these results, we argue that interpretability needs to be reconsidered at a paradigmatic level. We pose the following question for interpretability tool designers and researchers: how do we design for interpretability and explainability knowing that people will never pay attention to all the information presented to them? We discuss design implications based on our exploratory analysis of cognitive (e.g., prior experience) and contextual (e.g., usability, confidence) factors, and how these affect bounded rationality in interpretability tool use.

2 RELATED WORK

2.1 Interpretability

Interpretability is defined from a model's perspective as the "ability to explain or to present in understandable terms to a human" [25, p2]. It is necessary because "if the system can explain its reasoning, we can verify whether that reasoning is sound with respect to [important, predetermined] auxiliary criteria" [25, p1]. These auxiliary criteria include things like "safety [6, 93, 126], nondiscrimination [12, 38, 107], avoiding technical debt [110], or providing the right to explanation [34]" [25, p1]. Interpretability also serves as a proxy for other desiderata for ML-based systems such as reliability, robustness, informativeness, etc. which, in turn, promote trustworthiness, accountability, and fair and ethical decision-making [74].

There are two approaches for achieving model interpretability. First, using "glassbox" ML models that are designed to be inherently interpretable due to their simplicity. These include simple point systems [53, 138], decision trees [101] and sets [67], and generalized additive models (GAMs) [14, 40, 89]. The second approach is to train post-hoc explainers that are designed to make the predictions of "blackbox" models more interpretable. These include local interpretable model-agnostic explanations (LIME) [104], Shapley additive explanations (SHAP) [77, 112], and other approaches for local explanations [4, 111, 117]. For an overview of interpretability techniques for shallow and deep learning

models, see the comprehensive reviews by Gilpin et al. [33] published in 2018, Arrieta et al. [7] published in 2020, Liao and Varshney [72] published in 2021, and Dwivedi et al. [29] published in 2023.

Despite this proliferation of techniques, there is still debate about what interpretability should entail. In particular, Rudin [106] argues against the use of post-hoc explanation techniques for ML models deployed in high-stakes domains because they may not faithfully represent the models' behavior. Jin et al. [52] call these type of explanations "plausible" rather than accurate, and similarly express their concern for presenting these explanations to help people reason about model outputs. Doshi-Velez et al. [26] suggest that interpretability should include not only a justification for the prediction, but also a description of the decision-making process followed by the model. Similarly, Lipton [74] surveys different criteria for assessing interpretability, such as simulatability and decomposability. Adbul et al. [2] highlight interactivity and learnability as cornerstones for designing visualizations that better support interpretability. Dourish [27] adds scrutability as a necessary component of interactivity for interpretability.

2.2 Explainability

Although sharing similar underlying goals as interpretability, explainability is more human-centered and is "associated with the notion of an explanation as an interface between humans and a decision maker that is, at the same time, both an accurate proxy of the decision maker and comprehensible to humans" [7, p85]. With its focus on designing explanations in ways that are understandable to people, explainability relies on insights from the social sciences, in how people explain things to each other, and cognitive science, in how people individually make sense of the world.

Miller [84] and Mueller et al. [87] provide comprehensive reviews of the properties of explanations from the philosophy (e.g., [35, 73, 96, 125]) and social science (e.g., [43, 68, 76, 79, 88, 118]) literatures. They note that explanations are contrastive, social, and selected by people in a biased manner (e.g., in accordance with cognitive or social heuristics), and that referring to probabilities or statistical generalizations in explanations is usually unhelpful. To that end, Miller [84], Miller et al. [86], and Lombrozo [75] suggest simplicity, generality, and coherence as the main evaluation criteria for explanations.

The social science literature proposes that we think of explanations as a conversation. Grice's maxims of quality, quantity, relation, and manner [35], which form the core of a good conversation, should therefore be followed when designing explanations [64, 79, 118]. Leake's goal-based approach to explanation evaluation adds metrics such as the timeliness of an explanation in providing the opportunity to deal with the prediction being explained, knowability and the features responsible for "knowing," and the independence of individual explanations [68]. Explanations that follow this goal-based approach must include grounding in some common demonstrative reference between people and the explanation system [20, 78]. It is due to these insights that, increasingly, interpretability and explainability tools include characteristics such as interactivity [5, 44], counterfactual "what-if" outputs [85, 130], and modular and sequential explanations [82].

2.3 Human Evaluations of Interpretability and Explainability

User studies of interpretability and explainability tools have shown limited efficacy towards their goal of helping people understand ML. Only recently has the research community begun to evaluate the tools via user studies. Yet, already, there is a consistent pattern of inadequate use of these tools. Sometimes, this is seemingly due to information overload. For example, Poursabzi-Sangdeh et al. [99] test the impact of two factors often thought to affect interpretability—number of input features and model transparency (i.e., glassbox vs. blackbox models). They find that it is easier to simulate models with a small number of features, but that neither factor impacts people's willingness to follow a model's predictions. Moreover, too much transparency causes people to incorrectly follow a

model even when it makes a mistake. Springer and Whittaker [119] find that explanations can create information overload and distract people from forming a useful mental model of how a system operates. Similarly, Lage et al. [65] study the length and complexity of an explanation, finding that longer explanations overload people's cognitive abilities. Other times, it seems that people are not critical enough of the explanation outputs and consequently over-trust them. Kaur et al. [59] and Bansal et al. [9] evaluate practitioners' and lay users' use of explanations, both finding similar patterns of over-use and people using the existence of explanations as a signal for high model accuracy.

Between 2018 and 2021 alone, over 100 peer-reviewed papers published results from human subjects evaluations of AI outputs and explanations in decision-making contexts [66], and the number has only grown since. This thread of work helps answer questions like: which explanation types/formats are easier to understand [37, 71, 140]; what works best in higher stakes settings [16, 95, 128, 137]; how to present explanations to lay users, novices, and experts [100, 133]; which types of uncertainty information are helpful [48, 103, 139]; what causes information overload [13, 119]; whether explanations should presented as questions, answers, or a dialogue [23, 69, 70]; etc.

At a high level, these evaluations reveal a gap between our theoretical understanding of what an explanation should include and whether that works in practice. Although we know the properties of good explanations that would support effective human–machine collaboration, the other element of that relationship is relatively unexplored: we do not fully understand how facets about humans affect this relationship. Some recent work relies on cognitive science theories to assert that cognitive heuristics can affect the effectiveness of explainability tools in various domain-specific settings [3, 11, 13, 32, 58, 90, 103, 103]. Consolidating these theories, Wang et al. [131] even present a framework for how interpretability and explainability should be shaped, from the get-go, with human cognitive needs in mind. However, it remains unclear why and in what settings people apply these cognitive heuristics, and whether the outcomes of this application are harmful for establishing a good understanding of ML predictions. We extend this work by delving further into the fundamentals of human cognition.

2.4 Bounded Rationality

Bounded rationality describes human beings as rational agents functioning within cognitive and informational constraints [116]. These constraints separate the boundedly rational person from homo econmicus, described by Mill [83] in 1836 as "a being who desires to possess wealth, and who is capable of judging of the comparative efficacy of means for obtaining that end." It is this latter half with which bounded rationality most directly disagrees. Bounded rationality thus seeks to explain and predict human behavior in a way that more closely matches reality than the wholly rational view of human decision making encapsulated by the concept of homo economicus [114].

Given the constrained cognitive abilities and limited information available to people, they often employ a component of bounded rationality called satisficing in lieu of maximizing. Under a maximizing framework, people process all relevant information about a set of options and choose the optimal option in view of available global information and boundless cognition [115]. In reality, humans have neither the cognitive capacity nor the requisite information about most choices to maximize in this way. As a consequence, boundedly rational decision makers satisfice: they choose the option that suffices to meet a (consciously or unconsciously) predetermined set of criteria to a satisfactory degree (thus, the portmanteau of satisfy and suffice) [116].

Simply put, satisficing is a process by which people choose an option that is "good enough" as defined by their own criteria. Satisficers do not randomly select from a list of potential choices, but instead employ rational inattention that allows them to reduce cognitive overhead in the decision-making process [80]. The explicit rationality of both bounded rationality and satisficing distinguishes a satisficer from a random chooser. Satisficing appears to be a useful descriptive model

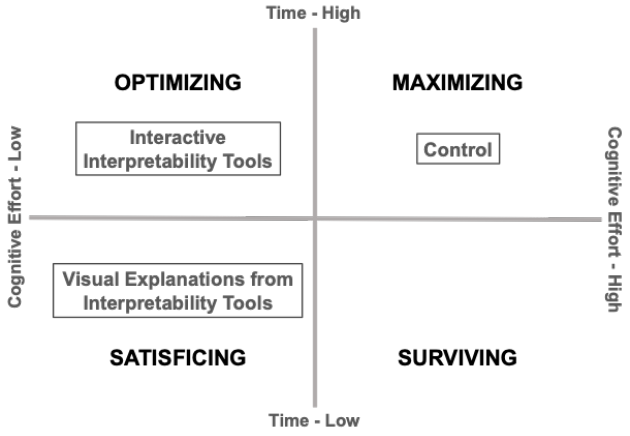

Fig. 2. We use two common proxies for studying bounded rationality, the time and cognitive effort spent on a task, to define quadrants for human cognition in Machine Learning settings. These proxies are borrowed from prior work in cognitive science and behavioral economics (e.g., [21, 47, 54]), and help situate our work in this existing literature on human-human decision-making settings.

for many cognitive processes including split-second decision making [92], patterned and random repeated choice scenarios [109], discrete choice scenarios [80], and explanations of theory of mind under uncertainty [98]. The salient commonalities across domains in which satisficing is observed or usefully descriptive are the presence of uncertainty or a prohibitive cost of information accrual.

In this work, we propose bounded rationality as being critical for how (much) people understand ML. Further, we evaluate the role of output explanations and features of interpretability tools in supporting or undermining human cognition as described by bounded rationality.

3 RESEARCH GOALS AND HYPOTHESES

We seek to examine the role that bounded rationality plays in how ML practitioners use interpretability tools. To do so, we compare ML practitioners' performance on an exploratory data science task between conditions where they had access to interpretability tools vs. their normal ML pipeline sans interpretability (i.e., a control condition). Table 1 presents a full list of our hypotheses corresponding to our five dependent variables: time, cognitive effort, hints, accuracy, and response type. Two of these—time and cognitive effort—are established proxies for bounded rationality in human-human settings. We introduce the additional three proxies as being specifically relevant for our setup, and ML and interpretability settings. We describe these hypotheses and dependent variables in more detail below. Before collecting any data, we pre-registered our intent to study these hypotheses on AsPredicted.[4]

In cognitive science and behavioral economics, the bounded nature of human rationality is often expressed as a function of time and effort [21, 47, 54]. These form the basis of our cognitive framework for the ML context (Figure 2). In an ideal world with infinite information processing capabilities, one would consider the utility of all information and alternatives before making decisions. This is referred to as *maximizing*, which requires significant time and cognitive effort [123]. However, prior work in cognitive science and behavioral economics shows that people *satisfice* to conserve time and effort in decision-making settings [55, 56, 114].

Translating and expanding the motivations behind these cognitive modes to the ML setting, we hypothesize that interpretability solutions engage the bounded cognitive modes in ML practitioners

[4] https://aspredicted.org/ry99g.pdf

more so than when they do not have access to interpretable outputs. These explanations essentially bypass the process that practitioners have to follow to understand the data and model on their own, which includes finding the right approach for generating interpretable outputs, writing code to make it work, and only then having access to the explanations. However, not all interpretability solutions are so simple in their working—some require deliberate engagement with interactive features and show the data and model in a multitude of visual formats. We hypothesize that this type of engagement might instead lead to *optimizing* behavior wherein, similar to satisficing, practitioners spend less cognitive effort on the task, but differently spend more time being engaged in understanding the content. The remaining quadrant in our cognitive framework represents low time–high effort situations. The domains of these situations require urgent decision-making with less time at hand (e.g., healthcare, aviation). Although ML practitioners are responsible for the models used in these settings, they are rarely actively involved in the day-to-day of this type of work. That is, the urgency in less acute for practitioners as stakeholders. Therefore, we do not test this cognitive mode in our experiment.

Overall, we test three types of ML setups: (1) a control condition sans interpretability, (2) visual explanations from interpretability tools that are static in nature, and (3) interactive interpretability tools. In line with the quadrants that each of these conditions belong to, we hypothesize that:

H1a People will spend *less time* on the data science task when using visual explanations from interpretability tools.

H2a People will expend *less cognitive effort* on the data science task when using visual explanations from interpretability tools.

H1b People will spend *more time* on the data science task when using interpretability tools with their full range of interactive features.

H2b People will expend *less cognitive effort* on the data science task when using interpretability tools with their full range of interactive features.

In addition to the traditional metrics used in cognitive science and behavioral economics, we also rely on metrics that seem relevant specifically for our task setup and the ML and interpretability settings, based on findings from prior work. One such metric is the number of hints used during the study. To avoid making our task setup too cumbersome for participants, we also provide hints for how to use and interpret the ML models and explanation outputs (setup details in Section 4.4). Compared to a setup with no built-in interpretability options (control), we anticipate that explanations and interactive elements would make the setup more straightforward, and the data and model easy to explore, thus lessening the need for hints. Therefore, we hypothesize that:

H3a People will rely on *fewer hints* to complete the data science task when using visual explanations from interpretability tools.

H3b People will rely on *fewer hints* to complete the data science task when using interpretability tools with their full range of interactive features.

Time, cognitive effort, and hints reflect the *process* behind bounded rationality, but it also affects the *outcomes* of a task. Under bounded rationality, people often apply heuristics, i.e., automated processes that circumvent the need for conscious deliberation of information. While this has its benefits (e.g., avoiding information overload), it can have negative consequences when people apply inaccurate heuristics or overly rely on the automaticity afforded by them. In the ML setting, practitioners are responsible for making accurate decisions about the data and model. As such, we measure the impact of bounded rationality on the decisions made by practitioners by defining two *outcome*-based proxies: (1) the accuracy of their answers about the data and model; and (2) the type of responses they select, where the types can be accurate, plausible, or randomly inaccurate responses.

	Dependent Variable	Hypothesis	Cognition Mode
1	Time	**a)** *Lower* when using visual explanations from interpretability tools.	Satisficing
		b) *Higher* when using interpretability tools with interactive features.	Optimizing
2	Cognitive Effort	**a)** *Lower* when using visual explanations from interpretability tools.	Satisficing
		b) *Lower* when using interpretability tools with interactive features.	Optimizing
3	Hints	**a)** *Fewer* when using visual explanations from interpretability tools.	Satisficing
		b) *Fewer* when using interpretability tools with interactive features.	Optimizing
4	Accuracy	**a)** *Lower* when using visual explanations from interpretability tools.	Satisficing
		b) *Higher* when using interpretability tools with interactive features.	Optimizing
5	Response Type	**a)** *Plausible* when using visual explanations from interpretability tools.	Satisficing
		b) *Accurate* when using interpretability tools with interactive features.	Optimizing

Table 1. An overview of the ten hypotheses for our five dependent variables. Each dependent variable corresponds to two hypotheses, one for visual explanations from interpretability tools and the other for interpretability tools with interactive features. We hypothesize that the former represents a satisficing cognition mode and the latter, optimizing.

Prior work claims that ML practitioners overly trust and rely on static explanations from interpretability tools [9, 59]. This suggests that people might be applying incorrect heuristics that lead to satisficing in these cases. Therefore, we hypothesize that:

H4a People's responses to questions about the data and model will be *less accurate* when using visual explanations from interpretability tools.

H5a People will select responses to questions about the data and model that are *plausible* (rather than accurate or inaccurate) when using visual explanations from interpretability tools.

On the other hand, interactive interpretability tools, in leading to the hypothesized optimizing behavior and more deliberate engagement, might resolve the potentially negative outcomes of bounded rationality. In his dual-process theory of cognition, Kahneman cites heuristics-based automated reasoning as the use of System 1 (of the brain), compared to System 2 which is a more deliberative reasoning unit [54]. It follows that one way to combat the application of potentially inaccurate heuristics for bounded rationality is to engage people in deliberative reasoning modes. Prior work in HCI shows that we can promote this deliberative thinking and engagement by making interpretability tools more interactive [5, 44, 102]. Therefore, we hypothesize that:

H4b People's responses to questions about the data and model will be *more accurate* when using interpretability tools with their full range of interactive features.

H5b People will select responses to questions about the data and model that are *accurate* (rather than plausible or inaccurate) when using interpretability tools with their full range of interactive features.

The five metrics described above form the core of our study of bounded rationality in the ML and interpretability contexts. We included other variables of interest for exploratory analysis in our pre-registration, such as usability of the task setup, validity of our dataset and model in the wild, etc. (Table 2 provides an overview of these).

4 METHODS

We conducted a pre-registered controlled experiment with ML practitioners to study our hypotheses. The experiment was between-subjects, split across five conditions, each representing a different ML + interpretability pipeline: (1) normal ML pipeline without any interpretability tools (control); (2) visual explanations from a glassbox model; (3) visual explanations from a post-hoc explainer

for a blackbox model; (4) interactive interpretability tool which used a glassbox model; and (5) interactive interpretability tool which used a post-hoc explainer for a blackbox model.

4.1 Experimental Setup

Our experiment was conducted using a Qualtrics survey with links to Google Colab notebooks for access to the data, model, and interpretability tools. The setup was designed after carefully considering the trade-offs between internal and ecological validity, both of which are individually hard to achieve in the ML setting with this participant pool [24, 59, 99]. On the one hand, studying an abstract construct via an experiment requires proxies that can be consistently captured through direct data collection or logging, for hypothesis testing. There is no way to log information like individual answers for accuracy or the time spent on each question in an open-ended Colab notebook. On the other hand, a consistent, purely quantitative experimental setup would take away both: (1) the context in which bounded rationality would normally occur, and (2) the ability to study the role of certain relevant features of interpretability tools (e.g., interactivity). For example, Kaur et al. [59] copied the visual outputs from interpretability tools within a Qualtrics survey. But, as they note, this did not allow for direct participant interaction with the data science setup.

For our research goals, both access to a realistic setup and consistency of data logging are (conflicting) critical requirements. To account for these, our setup employs both mechanisms: a Qualtrics survey to consistently capture quantitative metrics for hypothesis testing, and a Colab notebook with the relevant ML components. This works with all our constraints because questions about the data and model were asked in the Qualtrics survey, which allows easy metric logging. The Colab notebooks allow for exploration while answering these questions, which enables relevant context. Each Colab notebook presented an overview of the various ML elements included in it, followed by a dataset description, model overview and train/test accuracy numbers, and an overview of the interpretability option in the condition.[5] It is worth noting that, although this setup resolved issues with internal and ecological validity, it only works for the specific task, setting, and stakeholders we employed in our study. As such, the external validity of our study and participant pool remains an open question that we hope to examine in future work (see Section 7 for more details).

4.2 Choice of Dataset

We used the Adult Income dataset[6] with some modifications for our data science task. The Adult Income dataset is based on 1994 census data, publicly available via the UCI Machine Learning Repository. Each of its 45,000 instances represent a person, with 14 attributes that relate to demographic and socio-economic features such as their age, education, marital status, and occupation. The binary output variable records whether or not the person made <=$50,000 (converted to 0) or >$50,000 (converted to 1). This threshold would be equivalent to ~$100,850 in 2022 when adjusted for inflation.

We selected this dataset because it: (1) did not make the data science task overly cumbersome due to esoteric data, and (2) was on a topic that people would have encountered before and formed heuristics about. Similar to prior work [59], we synthetically introduced errors in this dataset to quantitatively capture faulty reasoning about the data and model. We included two errors that commonly occur in people's day-to-day ML work: missing values and redundant features [59]. To synthesize missing values, we replaced the age value with 38, the mean for all data points, for 10% of the data points with an income of >$50,000. For redundant features, we relied on the pre-existing redundancy in two features of the dataset, *Education* (a categorical variable) and *Education-Num* (a nominal representation of the categories for Education).

[5] Colab notebooks available via this link. Additionally, Jupyter notebook versions of the Colab notebooks are included in the supplementary material.

[6] https://archive.ics.uci.edu/ml/datasets/adult

4.3 Choice of Model and Interpretability Tools

We selected ML models and interpretability tools based on their consistency in outputs and features, and compatibility with the Colab environment. Our two glassbox conditions were built on the same underlying model (GAMs). The blackbox conditions both used the same post-hoc explainer (SHAP) and blackbox model (LightGBM). The LightGBM model outputs also matched the XGBoost model used in our control condition. Additionally, the glassbox and blackbox conditions were consistent in their explanation types and features. More details on these selections are included below.

4.3.1 Control Condition. Our control condition simulated a normal ML pipeline with no interpretability options. We relied on a boosted tree model using the XGBoost library.[7] Control was intended to be the hardest of the five conditions—participants had to find and add code for any interpretable outputs they wanted for the model. XGBoost was selected primarily for the availability of interpretable outputs for these models, with additional code. We included hints leading to these outputs to avoid high drop off rates due to the condition being too challenging. For example, code for global feature importances (a simple built-in function of the XGBoost model) and for partial dependence values (a basic function from the scikit-learn library) was included. However, the latter was a set of complex raw values in array form—participants would have had to search for the equivalent plotting function to convert these into a visual output. The local explanations were the most challenging missing piece in this condition. Ideally, the participants could have searched and found an XGBoost explainer developed for local explanations on their own. To make this slightly easier (and avoid high drop-off), we provided a link to the explainer,[8] but no code was included. Participants were not required to use this feature and could rely on other Python libraries or built-in functions. Indeed, while some participants used the linked explainer, others relied on their prior knowledge of relevant statistical tests and descriptive plots.

4.3.2 Visuals-only Conditions. Static implementations of interpretability tools present visual outputs for global and local explanations and partial dependence plots per feature. These are available for both types of interpretability approaches: glassbox models and post-hoc explainers for blackbox models. We considered several options for both types, eventually selecting the following implementations due to their underlying consistency (as noted in prior work [59]).

Generalized Additive Models (GAMs). GAMs are a class of glassbox models which are inherently interpretable. They explain predictions based on additive components, where each component is a function that models an input feature. We used the interpretML implementation of GAMs called Explainable Boosting Machines (EBMs) [89].[9] EBMs have built-in visualizations for global feature importances, partial dependence plots, and local explanations (see Figure 5).

SHapley Additive exPlanations (SHAP). SHAP is a post-hoc explanation approach for blackbox models [77]. It explains each prediction by assigning optimal credit to each input feature using Shapley values from cooperative game theory [112, 120]. We used the SHAP Python package[10] which provides local explanations for each data point. These are then aggregated to also present global feature importances and partial dependence plots per feature (see Figure 1). LightGBM[11] served as our underlying blackbox model explained by SHAP. It follows a highly optimized tree-based gradient boosting approach which makes training the model extremely fast [60]. Additionally, the SHAP implementation offers a separate, high-speed algorithm for tree ensemble methods like LightGBM.

[7] https://xgboost.ai/

[8] XGBoost explainer was originally developed for R (the package and the corresponding blog post). The Python community replicated this functionality in their module.

[9] https://github.com/interpretml/interpret/

[10] https://github.com/slundberg/shap

[11] https://lightgbm.readthedocs.io/en/v3.3.2/

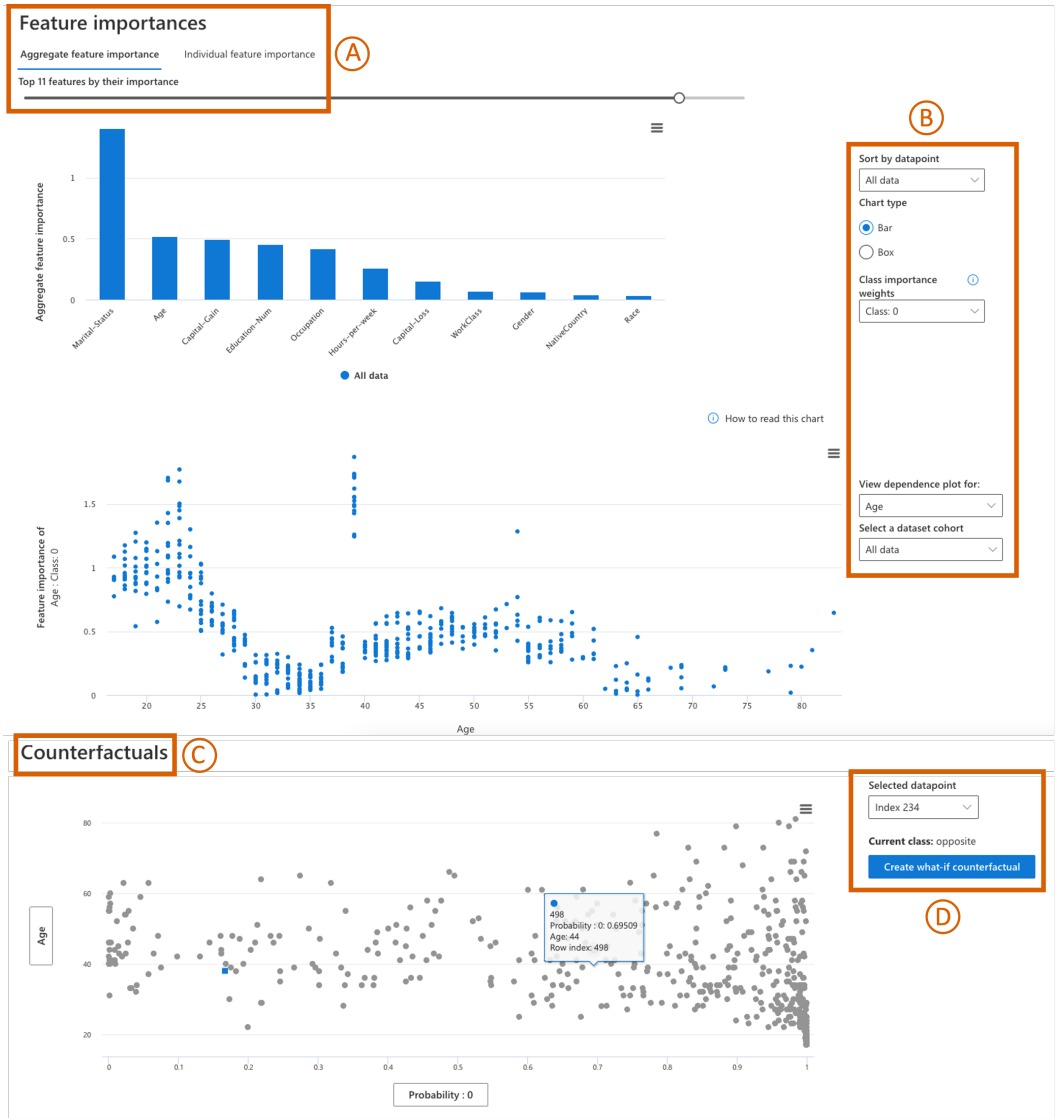

Fig. 3. The landing interface of Microsoft's Explanation Dashboard with minor edits to reduce whitespace. The dashboard includes several exploration sections. Here we show two of them, for Feature Importances (A) and Counterfactuals (C). Each section has its own set of interactive visuals and controls. Users can view aggregate or individual feature importances, and clicking on a feature bar shows the partial dependence plot (the middle one here). There are several sorting and binning options for displaying this data (B). The counterfactuals section allows users to select datapoints, view their local explanations (not shown here), and create what-if counterfactuals (D) which opens a new pane with various editing and sorting options for the suggested counterfactuals.

Fig. 4. The landing interface of Google's What-if Tool showing the Datapoint Editor. Users can access different tabs related to the data, model performance, and features (A). Each tab has its own interactive elements with a similar density of features as the Datapoint Editor. The scatter plot representing the data can be updated based on several options, such as binning x and y axes and scattering data using different labels (B). Users can also compute additional information about the datapoint, such as counterfactuals, local partial dependence plots, comparison with the same datapoint's prediction from a different model (C). Individual input feature values are editable for interactive what-if testing, for the selected datapoint on the scatter plot (D). In our case, the tool also provides SHAP attribution values as local explanations (D). The inference section for the selected datapoint (E) allows users to get real-time predictions for any edited input features from (D).

4.3.3 Interactive Tools Conditions. Interactive interpretability tools present many of the same visual outputs that static options do, but embed them in interactive features. These tools offer more fine-grained exploration of the data and model, responsive UIs for comparing several datapoints and features, additional metrics (e.g., fairness performance), etc. We picked the following two tools over other options because they were the most similar in features, and they ensured maximum consistency with the visuals-only conditions by supporting the use of the same underlying models.

Explanation Dashboard (ED). We used Microsoft's Explanation Dashboard[12] with EBMs for consistency with our other glassbox condition. The Explanation Dashboard includes interactive features for four avenues of exploration: (1) model overview, i.e., performance metrics and probability distributions for the data as a whole and for any user-defined data cohorts of interest; (2) data analysis, i.e., data-centric statistics and plots (e.g., scatter, density) based on user-selected filters for x and y axes; (3) feature importances, i.e., global explanations and partial dependence plots, and local explanations and individual conditional expectation plots; and (4) counterfactuals, i.e., answers to what-if questions about individual datapoints and perturbation-based analysis for how changes in input features would affect the model's prediction. Figure 3 presents a subset of these features.

[12] https://github.com/microsoft/responsible-ai-toolbox/blob/main/docs/explanation-dashboard-README.md

What-If Tool (WIT). We used Google's What-If Tool[13] with the same underlying elements as our post-hoc visuals-only condition: a LightGBM model and a SHAP explainer. The tool is divided into three main sections: (1) a datapoint editor, for visualizing datapoints using scatter plots with several options for binning the x and y axes, editing individual input feature values to test changes in prediction, finding the nearest counterfactual, analyzing local feature importances, etc.; (2) a performance and fairness tab for global metrics; and (3) a features tab, for data distributions and descriptive statistics for each feature. Figure 4 presents the datapoint editor tab for WIT.

4.4 Components of the Survey

Table 2 provides an overview of the survey components in the order in which they were presented to the participants, along with the corresponding dependent and independent variables. First, we provided a brief overview of task and setup and obtained consent from the participants. This was followed by questions about demographics (e.g., age, self-reported gender) and educational background (e.g., level of education, occupation, time in their current job role, the extent to which ML is part of their job, and familiarity with ML interpretability tools). Participants were next introduced to the data science task setup, provided access to the Colab notebooks, and asked about their familiarity with the model and interpretability options being used in their condition's setup.

Next, in the main part of the survey, we asked them five multiple choice questions (MCQs) about the data and model. Since this was the main data science element of the task, the questions immediately after it covered self-reports on cognitive load using the six-item NASA-TLX questionnaire [39] and the usability of the setup using the eight-item SUPR-Q scale [108]. We also asked them high-level evaluation questions about using the data and model in real-world applications, and their confidence in whether the data and outputs from their setup were error-free.

By now, participants had some experience with the task setup. Therefore, we next asked an exploratory multiple-answer question (borrowed from [59]) to understand their mental model of the setup in their condition. We also included an open-ended free response question asking participants to reflect on how they answered the questions in the study and their experience with the overall setup. The survey concluded with three questions to contextualize the extent of participants' engagement with the setup: a rating question about the extent to which they relied on the setup; and two yes-no questions on whether they read any documentation included for the various models and tools used in the study, and whether they wrote any code of their own. We include additional relevant details on the main elements below.

4.4.1 Multiple Choice Questions. The main portion of the survey consisted of a set of five MCQs about the data and model.[14] These questions were related to the common ways in which ML practitioners use interpretability tools [44, 59] (Table 2). Each MCQ had five choices that matched the categories of our nominal dependent variable (response type), presented in a randomized order. The response type can be accurate, plausible, or randomly inaccurate. Of these, the plausible category—which represents a boundedly rational state—can occur due to the application of various heuristics. We focused on two common heuristics relevant in this setting: the anchoring heuristic (wherein people make decisions based on the piece of information they notice first [19]) and the availability heuristic (wherein people make decisions based on incomplete information, using whatever comes to mind immediately [124]). Naturally, the MCQ choices we generated for these plausible answers were somewhat dependent on our own heuristics about income data. We tried to pick the heuristics-based choices that were also mentioned most frequently in our pilot studies with ML practitioners.

[13] https://pair-code.github.io/what-if-tool/

[14] All MCQs with their corresponding answer options and hints are included in the supplementary material.

Table 2. Questions included under each survey component and their corresponding dependent and independent variables. Dependent variables are highlighted in gray.

Survey Component: Setup Familiarity	
- Use of ML in their daily job (scale 0–7) - Total time estimate for how long they have been practicing ML (in months)	Prior Experience with ML
- Familiarity with interpretability tools in general (scale 0–7) - Estimated hours spent using interpretability tools (categories based on an upward-facing parabola, i.e., never, less than 10 hrs, 10–20hrs, 20–50hrs, 50–100hrs, and more than 100hrs; borrowed from [59])	Prior Experience with Interpretability Tools
- Familiarity with the specific model and interpretability option (scale 0–7) - Estimated hours spent using this model and interpretability option (categories based on an upward-facing parabola, same as above)	Prior Experience with Task Setup
Survey Component: Multiple Choice Questions	
- Global Feature Importance: If you were forced to remove a feature from this model, which of the following features would you remove? - Partial Dependence for a Feature: Which of the following ranges for Age values has the most likelihood of making a high income? - Predict the Outcome: Given the following input feature values and importances, what do you think the model predicted for this individual and why? - Explain Misclassification: The model misclassified this datapoint with the given input feature values. Why do you think that happened? - What-if Question: A person with the following input features makes <=50k income. A change to which of the following feature values would cause the prediction to become 1 (i.e., >50k income)?	Time, Number of Hints Used, Accuracy, Response Type
Survey Component: High-Level Task Evaluation	
- Cognitive Load using NASA-TLX questionnaire [39]	Cognitive Effort
- Usability of the setup using the SUPR-Q scale [108]	Usability Score
- Use of this dataset for a loan approval prediction tool for a bank now, in 2022 (scale 0–7) - Possibility of this model's deployment in the wild (scale 0–7) - Use of accuracy as a key performance indicator (scale 0–7)	Hypothetical Use Rating
- Confidence in their answers (scale 0–7) - Confidence in the setup (scale 0–7)	Task Confidence
- Presence of errors related to missing values (scale 0–7) - Presence of errors related to redundant features in the dataset (scale 0–7)	Error Recognition
Survey Component: Mental Models	
- Multiple-answer question with options representing established and intended capabilities of interpretability solutions (e.g., "individual instance explanations," "counterfactuals," "regions of error", etc.); borrowed from [59]	Mental Model Accuracy
Survey Component: Setup Engagement	
- Extent to which they relied on the setup (scale 0–7)	Setup Reliance
- Read additional documentation about the dataset, model, or interpretability options included in the setup (binary)	Read Documentation
- Wrote their own code in the Colab notebook (binary)	Wrote Code

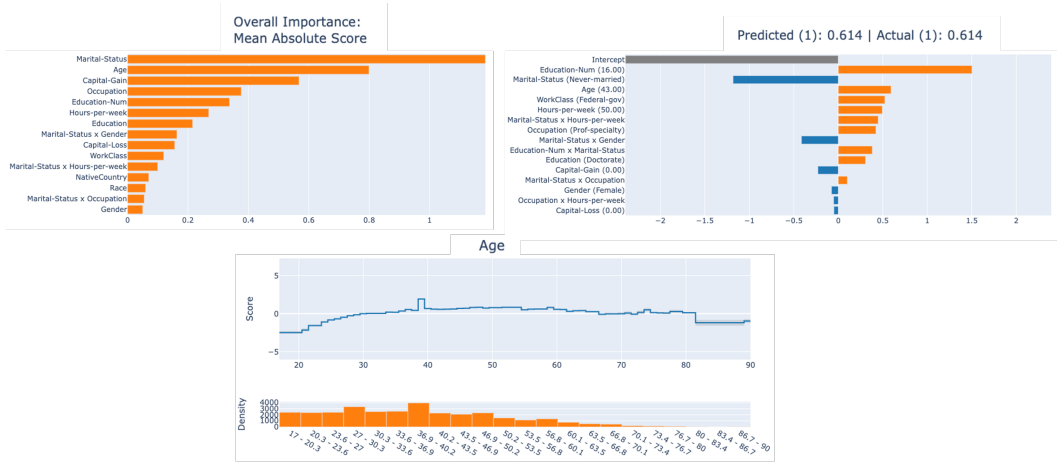

Fig. 5. Visualizations output by interpretML's implementation of GAMs. These ones are generated for the Adult Income dataset, same as our study task. Top (left to right): Global explanation; Local explanation for an individual datapoint. Bottom: Partial dependence plot for a continuous input feature, age.

Therefore, our MCQs included choices that were: (1) *accurate*; (2) *inaccurate*; (3) *plausible and accurate*, i.e., a response that is accurate and is also easier to reach based on a heuristic that people commonly apply about the relationship between income and demographics; (4) *visually plausible but inaccurate*, i.e., a response that is inaccurate but, when looking at the visual explanation charts, easy to anchor to as the most obvious choice; and (5) *heuristically plausible but inaccurate*, i.e., a response that is inaccurate but easy to reach based on a heuristic that people commonly apply about the relationship between income and demographics. Additionally, each question included an optional hint, which participants could access by clicking a "Hint" button. We recorded this button click to calculate the total number of hints used by a participant.

For a concrete example of the answer choices, consider one of our MCQs: "Given the following input feature values, what do you think the model predicted for this individual and why?" For the interpretability conditions, this question included a visual of the local explanation for the datapoint being considered, with the model's prediction cropped out (e.g., Figure 5-top right). For the control condition, we only provided the visual in the hint since local explanations are not featured in a normal ML pipeline. The answer options for this question were: (1) *accurate*, ">50K income because most of the features have a positive influence and it adds up to greater than negative influence;" (2) *plausible and accurate*, ">50K because the values of input features for this person correspond to those that have positive influence on income in the partial dependence plots;" (3) *visually plausible but inaccurate*, "<=50k because the intercept has a significant negative influence;" (4) *heuristically plausible but inaccurate*, "<=50k because the Marital-Status is 'Never-married';" and (5) *inaccurate*, "<=50K income because the sum of all negative importances is greater than positive importances."

4.4.2 High-level Task Evaluation. We included hypothetical questions asking participants to rate: (1) the use of the dataset in the wild ("Consider the case of a loan approval prediction tool for use by a bank now (in 2022). How would you rate the likelihood of this dataset being applicable for predicting income in that setting?"); (2) the model's readiness for deployment in the wild; and (3) the use of accuracy as a key performance indicator. We also used self-reported ratings to establish participants' confidence in their answers and their data science setup. Recall that our

modified dataset also included two errors, missing values and redundant features. We asked about participants' confidence that the dataset was error-free, specifically from these two errors (rated individually), and any other errors that they might have noticed (open-text). All rating questions were on an eight-point Likert scale to avoid neutral responses.

4.5 Dependent and Independent Variables

The five **dependent variables** were calculated based on people's answers for the MCQs about the data and model: (1) *time*, i.e., how long participants took to answer all five MCQs on average;[15] (2) *cognitive effort*, i.e., total effort applied in answering questions, measured using the NASA-TLX questionnaire [39]; (3) *hints*, i.e., the number of hints used (out of a total of five, one for each MCQ); (4) *accuracy*, i.e., average correctness of responses across all five MCQs, where both accurate and plausible and accurate responses were considered correct; and (5) *response type*, i.e., counts for whether the MCQ option selected was accurate, plausible, or randomly inaccurate, across all MCQs. The **independent variables** were calculated based on the remaining survey questions. Table 2 includes a full list of all variables of interest.

4.6 Analysis Methods

We used our pre-registered analysis methods for our dependent variables: one-way ANOVAs followed by TukeyHSD post-hoc tests for continuous variables and Chi-square test of independence for the nominal variable. We also include descriptive statistics and results from exploratory multiple linear regressions and Pearson's correlations for our independent variables. Several of these independent variables were calculated based on responses to a set of similar questions. We established internal consistency for these metrics using Cronbach's alpha and Pearson's correlation values, and used averages depending on consistency outcomes. Participants answered one open-ended question reflecting on how they accomplished the study task. We coded these responses using an inductive thematic analysis [22]: open coding followed by axial coding, with themes generated by affinity diagramming the axial codes.

4.7 Participants and Data

We advertised the survey on social media (e.g., Reddit, Twitter), messaging platforms (e.g., Discord, Slack), and via mailing lists. People could only participate if they were over 18 years old and rated their ML experience as at least 2 on a scale of 0–7; 21 people were excluded based on this criteria. After filtering out 7 responses with spam-like text for our open-ended question, we were left with 119 total participants. These were split into conditions as: 25 to control, 24 to GAMs, 25 to SHAP, 22 to Explanation Dashboard, and 23 to What-if Tool. Participants rated their ML experience as 4.5 on average (σ=1.7; scale 0–7) and had practiced ML for 40.2 months on average (σ=22.65). The most commonly listed job roles for them were data scientist (35 out of 119 participants), ML practitioner (27), software designer/developer for a ML-based system (23), ML researcher (22), and ML engineer (12). Participants were compensated with a $25 Amazon gift card upon completion of the study. The study protocol was approved by the University of Michigan IRB.

5 RESULTS

5.1 Hypothesis Testing

Our results demonstrate a significant difference across conditions (control, visuals-only from GAMs and SHAP, and interactive tools ED and WIT) for all dependent variables. Table 3 presents the

[15]We did not compare the overall study time across conditions to maintain procedural consistency with other metrics. We were focused on the interplay between bounded rationality and the decision-making element of the task for all metrics.

means, standard deviations, one-way ANOVA and TukeyHSD results for significance testing for our continuous dependent variables (time, cognitive effort, hints, and accuracy). Further, a Chi-square test indicates that our nominal dependent variable—a response type of accurate, plausible, or randomly inaccurate, selected for the MCQs—is significantly different by condition with a large effect size ($\chi^2(8, N=119) = 51.33$, p«0.001, Cramer's v=0.47). Figure 6 shows two plots: (1) residuals for the three response types per condition, and (2) the relative contribution of each response type per condition to the total Chi-square score (calculated as a percentage $\frac{residual^2}{\chi^2 statistic}$ for each cell). Additionally, Table 4 provides exact counts for each response type per condition in a contingency table.

Dependent Variable	Control	GAMs	SHAP	Explanation Dashboard (ED)	What-if Tool (WIT)	Pairwise Significance
Time (in minutes) F(4,114) = 4.15, p<0.01 partial η^2 = 0.13	19.65 ± 17.34	3.90 ± 2.08	4.63 ± 2.87	3.25 ± 2.84	4.64 ± 2.33	Control > GAMs* Control > SHAP* Control > ED* Control > WIT*
Cognitive Effort (1-7) F(4,114) = 14.62, p<0.001 partial η^2 = 0.34	4.16 ± 0.77	3.29 ± 0.96	3.49 ± 0.62	4.47 ± 0.86	4.72 ± 0.67	Control > GAMs** Control > SHAP* ED > GAMs*** ED > SHAP*** WIT > GAMs*** WIT > SHAP***
Hints (0-5) F(4,114) = 14.45, p<0.001 partial η^2 = 0.34	3.08 ± 1.12	1.42 ± 1.06	1.48 ± 1.66	3.18 ± 1.30	3.52 ± 1.24	ED > GAMs*** ED > SHAP*** Control > GAMs*** Control > SHAP*** WIT > GAMs*** WIT > SHAP***
Accuracy (0-100) F(4,114) = 3.92, p<0.01 partial η^2 = 0.12	59.20 ± 15.79	44.17 ± 19.54	42.40 ± 15.62	40.91 ± 22.66	41.74 ± 21.67	Control > GAMs* Control > SHAP* Control > ED** Control > WIT*

Table 3. The results of our pre-registered analysis (one-way ANOVAs and TukeyHSD post-hoc tests) for the continuous dependent variables. Each row represents a dependent variable with numbers for the ANOVA results, means and standard deviations for each condition, and the conditions with a significant pairwise difference based on the TukeyHSD tests. Significance levels are indicated as: *=p<.05, **=p<.01, ***=p<.001. ANOVA results include a partial η^2 value for effect size; suggested norms: small = 0.01, medium = 0.06, large = 0.14. All of our dependent variables show a large effect size.

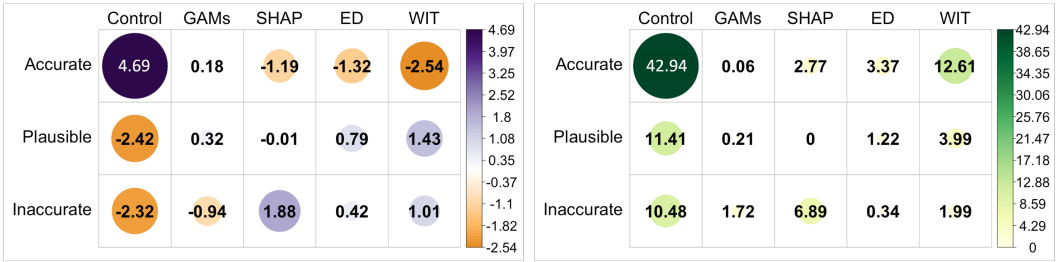

Fig. 6. Results of our pre-registered Chi-square test for the nominal dependent variable, response type. Left: Residuals from the test, indicating directionality and effect of each response type-condition combination for the magnitude of the resulting chi-square statistic. Right: contribution of each cell to the chi-square statistic calculated as a percentage.

	Control	GAMs	SHAP	Explanation Dashboard (ED)	What-if Tool (WIT)	Total
Accurate	69	39	32	27	21	188
Plausible	49	69	69	67	75	329
Inaccurate	7	12	24	16	19	78
Total	125	120	125	110	115	595

Table 4. Contingency table with counts for response types—accurate, plausible, and randomly inaccurate—for all five conditions.

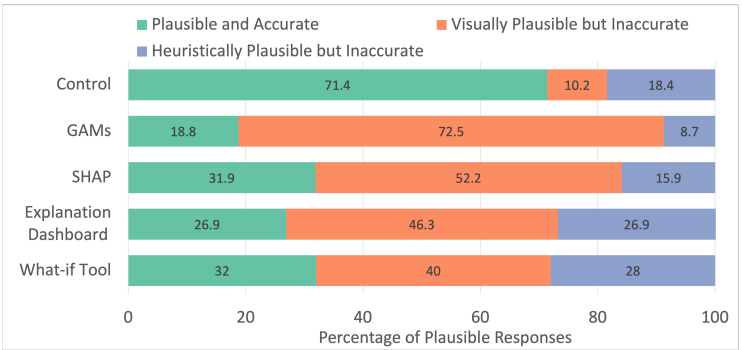

Fig. 7. Fine-grained breakdown of the type of responses selected under the plausible response type: plausible and accurate, visually plausible but inaccurate, and heuristically plausible but inaccurate. The percentages are calculated using the raw plausible response numbers in Table 4.

As hypothesized, participants spend significantly less time and cognitive effort when the setup provides visuals from GAMs and SHAP, i.e., our visuals-only conditions (**H1-2a**). They also use significantly fewer hints when they have access to these visuals (**H3a**). This supports our theory that the cognitive framework of bounded rationality via satisficing is applicable in this ML and interpretability context. With satisficing, it comes as no surprise that participants select plausible response types when answering questions about the data (**H4a**). Table 4 presents the numbers for each high-level response type: accurate, plausible, or inaccurate. As we noted in our setup, a plausible response can also be accurate, leading to good outcomes from bounded rationality and suggesting an optimizing cognitive mode. Unfortunately, the breakdown of the plausible response type category in Figure 7 signifies that a large percentage of the plausible responses selected are also inaccurate. This is primarily because participants selected the answer option that was visually the most obvious choice, and sometimes because they selected the option that was inaccurate but easy to reach based on a common heuristic. Further supporting this, the accuracy numbers—calculated as the percentage of accurate and plausible+accurate responses to MCQs—are also significantly lower for the visuals-only conditions (**H5a**).

We had hypothesized that the interactive nature of some interpretability tools might promote more deliberative thinking and fewer instances of satisficing, i.e., encourage optimizing behavior instead **(H1-H5b)**. However, we do not find adequate support for this in our data. We do not see any significant differences for time, accuracy or response type between the visuals-only conditions and interactive interpretability tools (ED and WIT). All metrics continue to indicate satisficing with significantly inaccurate outcomes when compared to the control condition. While we had hoped that interactivity would make people spend more time with these tools, we instead find that participants spend significantly more cognitive effort in navigating these tools' features and need

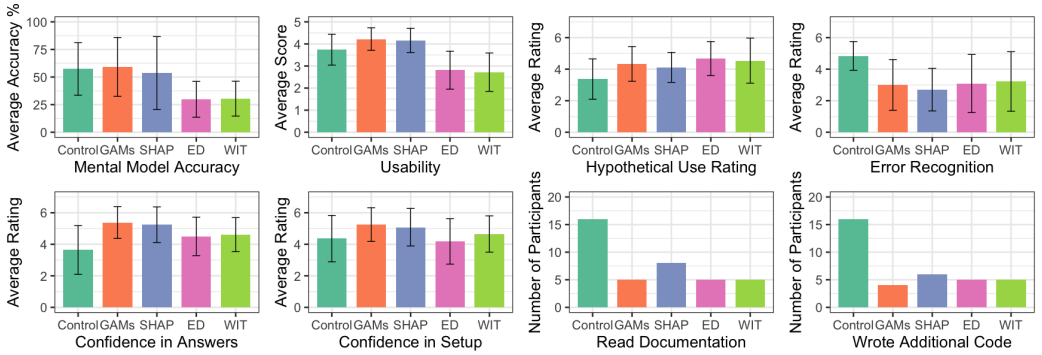

Fig. 8. Descriptive statistics for independent variables with noteworthy differences.

more hints to accomplish the task. The interactive tools are as challenging to use as the control condition with no interpretability options (i.e., no pairwise differences in cognitive effort and hints between the control condition and interactive tools). It is noteworthy that the averages for the control condition for cognitive effort and number of hints used are *lower* than the interactive tools.

Overall, we find evidence in support of all of our satisficing-related hypotheses for the visuals-only conditions and none for the optimizing-related hypotheses corresponding to our interactive tools conditions (Table 1). People satisfice—resulting in significantly low accuracy—when using *any* interpretability option.

5.2 Perceptions of the Setup

Figure 8 presents descriptive statistics for the independent variables for which we saw noteworthy differences.[16] These numbers reaffirm that, compared to the control condition, visual explanations from interpretability tools lead to satisficing and interactivity does not help in this setting. Participants in the control condition were more conservative with their ratings for hypothetical use, and their confidence in the task setup and their own answers. Further, these control condition participants read documentation and wrote their own code far more than those in the interpretability conditions. The statistics for the interactive tools were similar to that of the visuals-only conditions, with two exceptions: participants using interactive tools had lower mental model accuracy and usability scores. In fact, mental model accuracy and usability were lowest with interactive tools.

Participants felt that the interactive tools were far more challenging to use. One reason for this that came up frequently in answers to our open-ended question was that the these tools include a plethora of information and features, both presented in a way that was "difficult to understand", "stressful", and "burdensome".

> *I thought the data science setup [Colab notebook] was fairly intuitive (training, test, etc.) -> but the specific UI [interactive tool] to be extremely burdensome to navigate. There are a ton of colors + words + toggles going on, such that it almost felt like too much of a burden to actually derive the insights from the tool itself.* (P103, What-if Tool)

> *The UI of the tool felt a bit buggy and was kinda annoying to use, required too many clicks for one small change.* (P3, Explanation Dashboard)

While some participants appreciated all the information that they could glean from the interactive tools, it was often unclear how to interpret this information. Sometimes, it was hard to connect the

[16]Descriptive statistics for all independent variables are included as Appendix A.

different types of information presented across different features. Other times, the same information was presented in multiple ways without clarity on how to interpret it. The cognitive demands of using the interactive tools led some participants to not take advantage of their features and outputs, and instead use them in a limited way to confirm their own assumptions and narratives.

The reason behind satisficing seemed different for the visuals-only conditions. Participants reported higher usability scores for these and believed that they could understand everything very quickly. This fast-paced understanding of the model outputs "cross-verified their own understanding" and was in line with what "common sense would suggest." It seems that the intuitive type of information presented in visual explanations (e.g., global feature importance plots, partial dependence plots) closed participants off to further exploration of their own and they engaged in faster automatic processing under satisficing.

Surprisingly, the control condition seemed to be the best of both worlds: not too easy that people could quickly apply visual heuristics, and not too difficult that they would be frustrated by the cognitive effort needed and instead apply heuristics. The control condition was a normal ML setup sans interpretability, so there were no immediate explanations available or visuals-based judgements to be made. It required some cognitive effort, but seemingly not as high as the interactive tools. We provided some links to modules that might help with understanding the model outputs, but no actual code was included. The numbers from Figure 8-bottom row suggest that, in having to read documentation and write or debug code, participants remained invested enough to carefully look at any outputs.

Overall, our descriptive and qualitative analyses indicate that participants in all interpretability conditions applied bounded rationality, although with two distinct behaviors: the visuals-only interpretability tools resulted in quick, automated thinking and confirmation of people's own narratives (satisficing), whereas the interactive tools resulted in automated thinking to avoid the cognitively burdensome nature of the setup (surviving). The control condition served as a more optimally effortful intermediate, where participants used the setup more deliberately by reading documentation and writing code, and this did not feel too burdensome due to the hints included about the setup.

5.3 Exploratory Analyses

Now that we have established that people satisfice when using any interpretability options, we consider the question: what kind of internal or external factors can affect satisficing? We test two options: (1) cognitive factors (e.g., prior experience in ML), and (2) contextual factors (e.g., self-reported confidence, usability, etc.). Significance numbers reported in this section are only meant for generating concrete hypotheses for future work.

5.3.1 Relationship between cognitive factors and satisficing. Results from fitting multiple linear regression (MLR) models do not show any predictive relationship between our asymmetric cognitive factors (i.e., prior experience with ML, interpretability in general, and the specific task setup) and satisficing. That is, whether or not someone has this kind of prior experience does not change their likelihood of satisficing. We also do not find evidence in support of a symmetrical relationship between mental model accuracy and satisficing based on calculating Pearson's correlation coefficients. That is, whether or not someone has an accurate understanding (mental model) of the data science setup (i.e., the model and interpretability option) they are using has no bearing on if they satisfice while using the setup.

However, prior experience and mental model accuracy do prevent people from anchoring to the visually plausible but inaccurate response type. We know that an outcome of satisficing is that people select plausible (over accurate) response options to questions about the data and model. These are further categorized as: (1) plausible and accurate, (2) visually plausible but inaccurate, and

(3) heuristically plausible but inaccurate responses. Results from fitting three MLR models—one each for each of the three plausible response categories—show people with higher values for some prior experience variables are significantly less likely to select the visually plausible but inaccurate type of plausible answers ($F(6, 112)$=2.99, $p < 0.01$, Adj-$R^2 = 0.092$).[17] Note that the adjusted-R^2 value here is quite low—prior experience only explains a small amount of variance for visually plausible but inaccurate responses. We also find a similar relationship between mental model accuracy and visually plausible but inaccurate answers. People with higher mental model accuracy are also less likely to select visually plausible but inaccurate answers when satisficing (Pearson's $r(117)$=-0.27, $p < 0.01$).

Overall, cognitive factors do not prevent satisficing or support optimizing. But, more prior experience with the setup and higher mental model accuracy can both mitigate—to a small degree—the immediate type of satisficing that results from a quick skim of the visuals output by interpretability tools.

5.3.2 Relationship between contextual factors and satisficing. Results from fitting MLR models show that higher usability can lead to significantly more satisficing. People who rate their task setup as more usable expend significantly lower cognitive effort on the task and require fewer hints to complete it.[18] These cases represent satisficing since lower cognitive effort and the use of fewer hints are proxies for it. Conversely, people who rate their setup as less usable expend more cognitive effort and use more hints to complete the task. However, this is not necessarily a bad thing. As we see with our control condition, higher cognitive effort and use of more hints can be present in conjunction with higher accuracy. Additionally, when people with lower usability setups do satisfice by selecting plausible response types, these are significantly more likely to be plausible and accurate or heuristically plausible but inaccurate responses.[19] Therefore, with lower usability, we can perhaps avoid the visually plausible but inaccurate responses caused by a very quick perusal of the information being presented. The numbers in Table 4 and Figure 7 support this hypothesis.

For our symmetrical contextual factors, we find that higher accuracy significantly co-occurs with: (1) people having lower confidence in their own answers, (2) people having lower confidence in the study setup for their condition, (3) people thinking that the data and model have errors, and (4) people thinking that the data and model cannot be used in hypothetical real-world usage scenarios.[20] Even when people satisfice and select plausible responses, these factors co-occur with plausible and accurate response types rather than the plausible but inaccurate types.[21] People with these attributes either select accurate responses (maximizing) or plausible and accurate ones (optimizing).

[17]Coefficients for significant predictive relationships between the visually plausible but inaccurate response type and people's estimates for number of hours spent using: (1) interpretability tools in general (b=-0.13, $t(112)$=-3.03, $p < 0.01$); and (2) their condition's setup (b=-0.17, $t(112)$=-2.47, $p < 0.01$).

[18]Significant predictive relationships between usability and two dependent variables: (1) cognitive effort ($F(4, 112)$=5.54, $p < 0.001$, Adj-$R^2 = 0.14$; b=-0.36, $t(112)$=-4.13, $p < 0.001$), and (2) the number of hints used (($F(4, 112)$=6.02, $p << 0.001$, Adj-$R^2 = 0.15$); b=-0.58, $t(112)$=-4.03, $p < 0.001$).

[19]Significant predictive relationships between usability and two response types: (1) plausible and accurate responses (($F(1, 115)$=3.85, $p << 0.05$, Adj-$R^2 = 0.02$); b=-0.17, $t(115)$=-1.96, $p < 0.05$); and (2) heuristically plausible but inaccurate (($F(1, 115)$=7.79, $p << 0.01$, Adj-$R^2 = 0.06$); b=-0.19, $t(115)$=-2.79, $p < 0.001$).

[20]Significant correlations between accuracy and people's: (1) confidence in their answers to our study questions (Pearson's $r(117)$=-0.25, $p < 0.01$); (2) confidence in the study setup for their condition ($r(117)$=-0.28, $p < 0.01$); (3) error recognition ratings ($r(117)$=0.30, $p < 0.01$); and (4) hypothetical use ratings ($r(117)$=-0.28, $p < 0.001$).

[21]Significant correlations between plausible and accurate response type and people's: (1) confidence in their own answers ($r(117)$=-0.30, $p < 0.001$); (2) error recognition ratings ($r(117)$=0.21, $p < 0.05$); and (3) hypothetical use ratings ($r(117)$=-0.27, $p < 0.01$).

Overall, we find that contextual factors do affect satisficing and optimizing behaviors. Higher usability scores significantly predict satisficing. Lower usability scores are predictive of either a different kind of satisficing (heuristically plausible but inaccurate response selection) or, interestingly, optimizing (plausible and accurate response selection). Similarly, lower confidence, higher skepticism about errors, and lower hypothetical use ratings—all seemingly negative user experience design outcomes—are in fact related to selection of accurate or plausible and accurate responses.

5.4 Summary of Results

We find that people satisfice on an exploratory data science task when using visual explanations from interpretability tools. Interactivity—a strategy commonly employed to promote deliberation and engagement—does not help in this setting. Rather, interactive features make interpretability tools cognitively burdensome to use. Our exploratory analyses indicate that cognitive factors (e.g., prior experience in ML and interpretability) have no bearing on the likelihood of satisficing. Instead, certain seemingly negative contextual factors (e.g., lower confidence and usability, higher skepticism) co-occur with optimizing behavior (i.e., selection of accurate or plausible and accurate responses during the task).

6 DISCUSSION

The core tenet of interpretability is "to explain or to present in understandable terms to a human." [25, p2]. Interpretability approaches accomplish the "to explain" aspect of this—it is only due to these approaches that we can shed light on the reasoning behind model predictions. However, user studies with interpretability tools have all found them lacking on the "in understandable terms to a human" front [9, 24, 59, 99]. There has either been an over-reliance on outputs from interpretability tools or information overload based on their transparency, both findings also corroborated via different conditions in our study. Prior work has touted explainability as the human-centered counterpart of interpretability. Explainability incorporates insights from the social sciences and underscores the need for explanations to be contrastive, modular, parsimonious, etc.—all characteristics of how people explain things to each other [7, 82, 84, 87].

Regardless of a model-centric or human-centric approach for helping people understand ML, the focus of these approaches is to design a human-understandable explanation. What is not considered enough are fundamentals of how people understand. Bounded rationality is one such fundamental, a cognitive framework that describes human nature and sensemaking in a way that has little to do with the information that is presented to a human [114]. It asserts—and we have confirmed for our specific ML setting with practitioners—that people do not process or evaluate information in a perfectly rational way. This framework does not examine how people apply bounded rationality, only that they do, and that they do it irrespective of the setting. How much they apply it might vary depending on the criticality of the task or the type of information available, but people are always boundedly rational. Presenting new information about the model, presenting it in a way that is easier for people to parse, all of these facets of designing interpretability and explainability tools: they assume perfect rationality. They assume people's desire to obtain information. And yes, people do want information, but only insofar as they can use it to attain a "good enough" understanding of the model. Failing to recognize this is why interpretability is currently broken. These approaches do enable explanation and interpretation, but they do not ensure cognition.

We put forward the following question for the research community: what would it mean to design interpretability and explainability knowing what we now know about bounded rationality? Put another way, how do we present information to people knowing that they will satisfice and never pay attention to all of it?

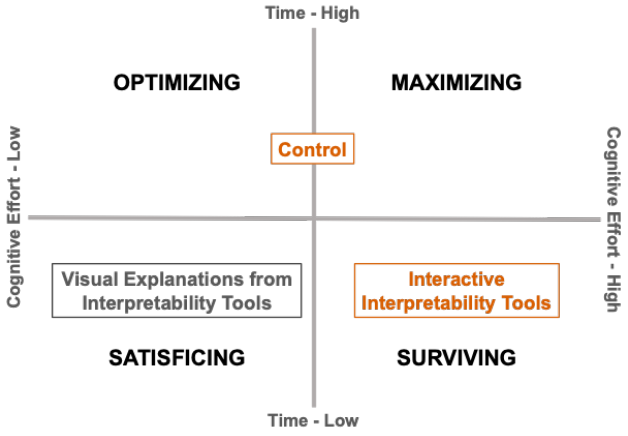

Fig. 9. Updated overview of our study conditions using our proposed framework for human cognition in Machine Learning settings. Compared to our hypothesized quadrants in Figure 2, the unexpected results are highlighted in color.

6.1 Implications for Design

We present the following implications for how existing interpretability tools are used by a key stakeholder—ML practitioners—for an exploratory data science task, and how we can improve their use of these tools going forward.

Interactivity can do more harm than good. In theory, interactivity should help improve engagement with any system [122]. Keeping people engaged should, in turn, have a higher likelihood of preventing heuristics-based, automated information processing [94, 121]. This is not quite true for how interactivity is incorporated in existing interpretability tools or how they are used by ML practitioners [28]. Interactive features complicate existing tools to the extent that people do not want to expend the high cognitive effort necessary to make sense of the features and the information they convey. Participants' responses to their experience with these tools suggest a case of the law of diminishing returns [113]: the additional information and reactive interface elements do not add value corresponding to the effort needed to adequately understand them. Indeed, we believe the interactive tools conditions, as currently designed, end up falling under the *surviving* cognitive mode rather than our hypothesized *optimizing* one (Figure 9). Based on our results, we propose thinking through the following initial questions for designing better interactivity for this setting: (1) how can we help people craft questions for further exploration using interactive features instead of simply seeing various interactive outputs?; (2) how can we design interactive features in conjunction with other forms of engagement (e.g., writing code)?; and (3) must interactivity be context-free or can there be a narrative to how interactive explanations are presented?

Experience-based cognitive factors do not help. Whether or not someone has prior experience in ML, interpretability in general, or the specific interpretability setup of their condition, had no bearing on their likelihood of satisficing in our study. Mental model accuracy similarly had no impact. Sure, experience might prevent the immediate satisficing that is a result of a quick perusal of information, but it does not guarantee the absence of other heuristics (e.g., "I've looked at similar datasets and models in the past, so I am able to guess which features may be correlated with each other and how they may be correlated with the response." (P34, GAMs); "I used most of my common sense to understand the questions. Income is going to increase with age, occupation

and hours of work per week. Even without looking at feature importance, I could make guesses of the prediction and answering the question." (P108, WIT)). The application of heuristics to answer questions seems related to cognitive effort. We find that people either apply heuristics because the cognitive effort needed to understand the visuals is so low that they make quick judgements based on the information, or so high that they are frustrated and do not want to use the tool for long.

Optimally effortful designs can help. The quadrant that remains unexplored with current interpretability solutions corresponds to low cognitive effort–high time which, we believe, would result in the optimizing cognitive mode. Our hypothesis here is grounded in the results for our control condition (Figure 9). Control was intended to be our most challenging condition; to avoid making it too cumbersome for a study, we included links to some helpful modules. It seems that this struck that balance between too helpful (satisficing) or too frustrating (surviving), and increased people's likelihood of reading documentation and writing code, i.e., staying invested enough to veer towards optimizing. The cognitive effort was not insignificant (4.16 on average on a scale of 1–7)—there is potential for interpretability tools to improve upon this control setup.

Seemingly negative contextual factors are surprisingly helpful. Features that are normally considered negative in the context of user experience design are correlated with higher accuracy in this context. Lower usability and confidence, and higher skepticism, were all related to accurate understanding of the data and model. Research and application areas that require optimal information processing have often relied on similar, seemingly negative, strategies to ensure user engagement. For example, the visualization community argues for highlighting the uncertainty of the underlying data in visuals [36, 49, 105, 132], or using surprise (e.g., Bayesian surprise that models information entropy [8, 10, 51]) and emotion [18, 63] to catch people's attention. Similarly, ubiquitous computing researchers and designers now emphasize seamfulness in their designs, adding uncertainty, ambiguity, and opportunities for user-appropriation instead of only prioritizing seamlessness, simplicity, consistency in features and interactions, and ease of use [17, 50, 129]. Even scholars in the explainable AI community have recently suggested that these negative contextual factors (seamfulness, criticism, forcing functions, etc.) might be a promising direction for appropriate use of AI explanations [13, 31, 58, 61].

Balancing efficiency and ease-of-use with emphasizing negative or imperfect aspects is known to be an effective strategy for sensemaking in both human-human and human-machine contexts. Indeed, organizations operating in critical domains are often failure-centric, and employ red teaming and adversarial design to simulate situations for testing the limits of their human and machine operations [135, 136]. Although it has shown merit as an approach for debugging intelligent systems and ML models [1, 15, 57, 69], this class of solutions is relatively unexplored for designing interpretability tools. We hope future work will incorporate these ideas towards designing more human (cognition)-centered interpretability and explainability.

7 LIMITATIONS

Bounded rationality is an abstract construct, and studying it quantitatively comes with the limitation of establishing construct validity. We made assumptions about proxies for bounded rationality and how we could operationalize these for the ML context. Although our assumptions were grounded in theory and prior work, we cannot be sure that these reflect bounded rationality in practice, especially for domains like ML and interpretability in which it has not been studied before. Specifically, recording time in a way that also represented the time spent on data exploration before answering each question was particularly hard. Additionally, the intricacies of bounded rationality are highly dependent on the varied heuristics that people develop (and update) over time and on the setting. We had to enforce consistency on an otherwise personalized construct, and it is unclear

what level of impact this had on our results. Similarly, it could be that the high values for cognitive effort and hints had more to do with how we designed and explained the study and setup, rather than the specific outputs and interactivity of the interpretability tools. Although we believe this to not be the case given the positive comments about the ease of the Colab + Qualtrics combination from the participants, we cannot be sure if there were any confounds introduced due to the setup.

Not only is human cognition via bounded rationality highly personalized, so is data science, making generalizability extremely challenging for studies like ours. Given our research goals of finding initial evidence of the role of bounded rationality in ML and interpretability settings, we focused on internal and ecological validity, but this affected the external validity of our setup and, consequently, its generalizability. ML Practitioners have a plethora of models and tools available to them and they use these in unique ways. Since it was infeasible to use people's varied local setups for a controlled experiment, we had to find a balance between consistency and rigidity for our study. Indeed, the generalizability of our results is dependent on how participants used our setup in comparison to their own ML pipelines. Perhaps future work can run similar studies within organizations that have uniform setups and benchmarks for their ML practitioners.

Our choice of task and questions also likely affected the generalizability of our results. While most ML practitioners have at some point conducted similar exploratory data analysis to what we asked for in our study, the majority of our results were calculated based on asking them very specific MCQs during and after their exploration. Although the study design decisions we made were necessary for a controlled experiment, our results could potentially change based on a different task, setup, or set of questions asked during the study. An easier task and setup might not require the same level of cognitive effort to begin with. This might in turn encourage people to spend more time interacting with the ML model and interpretability tool, as we hoped with the optimizing quadrant of my bounded rationality framework. Adding monetary incentives based on performance might improve performance as well, although we believe these need to be quite significant if the participants involved are ML practitioners with full-time jobs. Recent work does suggest that these variations can affect the metrics being studied here. For example, with a different task (solving a maze), changing monetary compensation based on performance, and varying explanation difficulty, Vasconcelos et al. [127] find there to be less over-reliance on AI outputs and explanations. However, they similarly note the need for something—adequate cognitive forcing functions, monetary rewards or other incentives, or task enjoyability—to ensure appropriate use of explanations.

Another type of variability that likely affected the generalizability of our findings is in the stakeholders themselves. We investigated the use of interpretability tools by ML practitioners, whom we defined very broadly as people with some prior ML experience. Changing this inclusion criteria and studying bounded rationality with other stakeholders will likely require different setups, and have slightly different outcomes. Given the evidence of bounded rationality in many human-human decision-making settings, we expect that this concept will be applicable to many human-machine settings like ours, but the exact mechanics of how (much) that will be are out of scope for this paper. The role of bounded rationality in human-machine interaction more generally is likely quite varied. We hope future work will dig deeper into this cognitive framework.

Finally, as we continue to study interpretability and explainability using controlled experimental setups, consistency in the metrics used to evaluate these approaches is also critical. We relied on one set of metrics and validated scales to do so. More recently, researchers have also proposed scales specific to human-AI interaction and explanations [45, 46, 141]. An interesting avenue of future work would be to compare and contrast interpretability tools using different scales. It would help us understand the kind of signals we can capture about people's use of interpretability and explainability tools in making decisions about their AI and ML decision-support counterparts.

8 CONCLUSION

Interpretability and explainability are purported to be solutions for helping people understand ML models and their predictions. We present results from a pre-registered controlled experiment with ML practitioners ($N = 119$) which provide significant empirical evidence that interpretability tools lead people to satisfice when trying to understand ML outputs. Compared to a control condition sans interpretability, people who rely on visual explanations from interpretability tools spend *5x less time* on the task, resulting in a 17% lower accuracy in answering questions about the data and model. We had hypothesized that interactivity might prevent satisficing but do not find this to be true. Rather, interactivity increases cognitive burden to the extent that people satisfice—seemingly out of frustration—instead. We argue for a paradigmatic shift in how interpretability solutions are currently designed, with the knowledge that people never pay attention to all the information available to them.

ACKNOWLEDGMENTS

We would like to thank the reviewers for their helpful comments. We are grateful to Eytan Adar, Jenn Wortman Vaughan, Preeti Ramaraj, Stevie Chancellor, Mitchell Gordon, and Elena Glassman for their feedback and support. We also want to thank the ML practitioners who participated in our study. Harmanpreet Kaur was supported by the Google PhD fellowship in HCI.

REFERENCES

[1] Hussein Abbass, Axel Bender, Svetoslav Gaidow, and Paul Whitbread. 2011. Computational red teaming: Past, present and future. *IEEE Computational Intelligence Magazine* 6, 1 (2011), 30–42. https://doi.org/10.1109/MCI.2010.939578

[2] Ashraf Abdul, Jo Vermeulen, Danding Wang, Brian Y. Lim, and Mohan Kankanhalli. 2018. Trends and Trajectories for Explainable, Accountable and Intelligible Systems: An HCI Research Agenda. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (Montreal QC, Canada) *(CHI '18)*. ACM, New York, NY, USA, Article 582, 18 pages. https://doi.org/10.1145/3173574.3174156

[3] 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*. 1–14. https://doi.org/10.1145/3313831.3376615

[4] David Alvarez-Melis and Tommi S Jaakkola. 2017. A causal framework for explaining the predictions of black-box sequence-to-sequence models. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. 412–421. https://doi.org/10.18653/v1/D17-1042

[5] Saleema Amershi, Maya Cakmak, William Bradley Knox, and Todd Kulesza. 2014. Power to the people: The role of humans in interactive machine learning. *AI Magazine* 35, 4 (2014), 105–120. https://doi.org/10.1609/aimag.v35i4.2513

[6] Dario Amodei, Chris Olah, Jacob Steinhardt, Paul Christiano, John Schulman, and Dan Mané. 2016. Concrete problems in AI safety. *arXiv preprint arXiv:1606.06565* (2016).

[7] Alejandro Barredo Arrieta, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador Garcia, Sergio Gil-Lopez, Daniel Molina, Richard Benjamins, Raja Chatila, and Francisco Herrera. 2020. Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion* 58 (6 2020), 82–115. https://doi.org/10.1016/J.INFFUS.2019.12.012

[8] Pierre Baldi and Laurent Itti. 2010. Of bits and wows: A Bayesian theory of surprise with applications to attention. *Neural Networks* 23, 5 (2010), 649–666. https://doi.org/10.1016/j.neunet.2009.12.007

[9] 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*. 1–16. https://doi.org/10.1145/3411764.3445717

[10] Nelly Bencomo and Amel Belaggoun. 2014. A world full of surprises: Bayesian theory of surprise to quantify degrees of uncertainty. In *Companion Proceedings of the 36th International Conference on Software Engineering*. 460–463. https://doi.org/10.1145/2591062.2591118

[11] Astrid Bertrand, Rafik Belloum, James R Eagan, and Winston Maxwell. 2022. How Cognitive Biases Affect XAI-assisted Decision-making: A Systematic Review. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*. 78–91. https://doi.org/10.1145/3514094.3534164

[12] Nick Bostrom and Eliezer Yudkowsky. 2014. The ethics of artificial intelligence. *The Cambridge handbook of artificial intelligence* 1 (2014), 316–334.

[13] 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. https://doi.org/10.1145/3449287

[14] Rich Caruana, Yin Lou, Johannes Gehrke, Paul Koch, Marc Sturm, and Noemie Elhadad. 2015. Intelligible Models for HealthCare: Predicting Pneumonia Risk and Hospital 30-day Readmission. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (Sydney, NSW, Australia) *(KDD '15)*. ACM, New York, NY, USA, 1721–1730. https://doi.org/10.1145/2783258.2788613

[15] Stephen Casper, Jason Lin, Joe Kwon, Gatlen Culp, and Dylan Hadfield-Menell. 2023. Explore, Establish, Exploit: Red Teaming Language Models from Scratch. *arXiv preprint arXiv:2306.09442* (2023).

[16] Federico Maria Cau, Hanna Hauptmann, Lucio Davide Spano, and Nava Tintarev. 2023. Supporting High-Uncertainty Decisions through AI and Logic-Style Explanations. In *Proceedings of the 28th International Conference on Intelligent User Interfaces*. 251–263. https://doi.org/10.1145/3581641.3584080

[17] Matthew Chalmers, Ian MacColl, and Marek Bell. 2003. Seamful design: Showing the seams in wearable computing. In *2003 IEE Eurowearable*. IET, 11–16. https://doi.org/10.1049/ic:20030140

[18] Tianlang Chen, Zhongping Zhang, Quanzeng You, Chen Fang, Zhaowen Wang, Hailin Jin, and Jiebo Luo. 2018. "Factual"or"Emotional": Stylized Image Captioning with Adaptive Learning and Attention. In *Proceedings of the European Conference on Computer Vision (ECCV)*. 519–535.

[19] Isaac Cho, Ryan Wesslen, Alireza Karduni, Sashank Santhanam, Samira Shaikh, and Wenwen Dou. 2017. The anchoring effect in decision-making with visual analytics. In *2017 IEEE Conference on Visual Analytics Science and Technology (VAST)*. IEEE, 116–126. https://doi.org/10.1109/VAST.2017.8585665

[20] Herbert H Clark, Robert Schreuder, and Samuel Buttrick. 1983. Common ground at the understanding of demonstrative reference. *Journal of verbal learning and verbal behavior* 22, 2 (1983), 245–258. https://doi.org/10.1016/S0022-5371(83)90189-5

[21] John Conlisk. 1996. Why Bounded Rationality? *Journal of Economic Literature* 34, 2 (1996), 669–700. http://www.jstor.org/stable/2729218

[22] Juliet M. Corbin and Anselm Strauss. 1990. Grounded theory research: Procedures, canons, and evaluative criteria. *Qualitative sociology* 13, 1 (1990), 3–21. https://doi.org/10.1007/BF00988593

[23] Valdemar Danry, Pat Pataranutaporn, Yaoli Mao, and Pattie Maes. 2023. Don't Just Tell Me, Ask Me: AI Systems that Intelligently Frame Explanations as Questions Improve Human Logical Discernment Accuracy over Causal AI explanations. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–13. https://doi.org/10.1145/3544548.3580672

[24] Wesley Hanwen Deng, Manish Nagireddy, Michelle Seng Ah Lee, Jatinder Singh, Zhiwei Steven Wu, Kenneth Holstein, and Haiyi Zhu. 2022. Exploring How Machine Learning Practitioners (Try To) Use Fairness Toolkits. In *2022 ACM Conference on Fairness, Accountability, and Transparency* (Seoul, Republic of Korea) *(FAccT '22)*. Association for Computing Machinery, New York, NY, USA, 473–484. https://doi.org/10.1145/3531146.3533113

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

[26] Finale Doshi-Velez, Mason Kortz, Ryan Budish, Chris Bavitz, Sam Gershman, David O'Brien, Stuart Schieber, James Waldo, David Weinberger, and Alexandra Wood. 2017. Accountability of AI under the law: The role of explanation. *arXiv preprint arXiv:1711.01134* (2017).

[27] Paul Dourish. 2016. Algorithms and their others: Algorithmic culture in context. *Big Data & Society* 3, 2 (2016), 2053951716665128. https://doi.org/10.1177/2053951716665128

[28] John J Dudley and Per Ola Kristensson. 2018. A review of user interface design for interactive machine learning. *ACM Transactions on Interactive Intelligent Systems (TiiS)* 8, 2 (2018), 1–37. https://doi.org/10.1145/3185517

[29] Rudresh Dwivedi, Devam Dave, Het Naik, Smiti Singhal, Rana Omer, Pankesh Patel, Bin Qian, Zhenyu Wen, Tejal Shah, Graham Morgan, et al. 2023. Explainable AI (XAI): Core ideas, techniques, and solutions. *Comput. Surveys* 55, 9 (2023), 1–33. https://doi.org/10.1145/3561048

[30] Upol Ehsan, Q. Vera Liao, Michael Muller, Mark O. Riedl, and Justin D. Weisz. 2021. Expanding explainability: Towards social transparency in ai systems. *Conference on Human Factors in Computing Systems - Proceedings* 19 (5 2021). https://doi.org/10.1145/3411764.3445188

[31] Upol Ehsan, Q Vera Liao, Samir Passi, Mark O Riedl, and Hal Daume III. 2022. Seamful XAI: Operationalizing Seamful Design in Explainable AI. *arXiv preprint arXiv:2211.06753* (2022).

[32] Theodore Evans, Carl Orge Retzlaff, Christian Geißler, Michaela Kargl, Markus Plass, Heimo Müller, Tim-Rasmus Kiehl, Norman Zerbe, and Andreas Holzinger. 2022. The explainability paradox: Challenges for xAI in digital pathology. *Future Generation Computer Systems* 133 (2022), 281–296. https://doi.org/10.1016/j.future.2022.03.009

[33] Leilani H Gilpin, David Bau, Ben Z Yuan, Ayesha Bajwa, Michael Specter, and Lalana Kagal. 2018. Explaining explanations: An overview of interpretability of machine learning. In *2018 IEEE 5th International Conference on Data Science and Advanced Analytics (DSAA)*. IEEE, 80–89. https://doi.org/10.1109/dsaa.2018.00018

[34] Bryce Goodman and Seth Flaxman. 2017. European Union regulations on algorithmic decision-making and a "right to explanation". *AI magazine* 38, 3 (2017), 50–57. https://doi.org/10.1609/aimag.v38i3.2741

[35] Herbert P. Grice. 1975. Logic and Conversation. (1975), 41–58. https://doi.org/10.1163/9789004368811_003

[36] Henning Griethe and Heidrun Schumann. 2005. Visualizing uncertainty for improved decision making. In *Proceedings of 4th International Conference on Perspectives in Business Informatics Research (BIR 2005)*, Vol. 20.

[37] 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*. 1–9. https://doi.org/10.1145/3491102.3517650

[38] Moritz Hardt, Eric Price, and Nati Srebro. 2016. Equality of opportunity in supervised learning. *Advances in neural information processing systems* 29 (2016), 3315–3323.

[39] Sandra G Hart and Lowell E Staveland. 1988. Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. In *Advances in psychology*. Vol. 52. Elsevier, 139–183.

[40] Trevor Hastie and Robert Tibshirani. 1987. Generalized Additive Models: Some Applications. *J. Amer. Statist. Assoc.* 82, 398 (1987), 371–386. https://doi.org/10.1080/01621459.1987.10478440

[41] T J Hastie and R J Tibshirani. 1990. *Generalized Additive Models*. CRC Press.

[42] Carl G Hempel and Paul Oppenheim. 1948. Studies in the Logic of Explanation. *Philos. Sci.* 15, 2 (1948), 135–175. https://doi.org/10.1086/286983

[43] Denis J Hilton. 1996. Mental models and causal explanation: Judgements of probable cause and explanatory relevance. *Thinking & Reasoning* 2, 4 (1996), 273–308. https://doi.org/10.1080/135467896394447

[44] Fred Hohman, Andrew Head, Rich Caruana, Robert DeLine, and Steven M Drucker. 2019. Gamut: A design probe to understand how data scientists understand machine learning models. In *Proceedings of the 2019 CHI conference on human factors in computing systems*. Association for Computing Machinery, New York, NY, USA, 1–13. https://doi.org/10.1145/3290605.3300809

[45] Andreas Holzinger, André Carrington, and Heimo Müller. 2020. Measuring the quality of explanations: the system causability scale (SCS). *KI-Künstliche Intelligenz* 34, 2 (2020), 193–198. https://doi.org/10.1007/s13218-020-00636-z

[46] Andreas Holzinger, Georg Langs, Helmut Denk, Kurt Zatloukal, and Heimo Müller. 2019. Causability and explainability of artificial intelligence in medicine. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 9, 4 (2019), e1312. https://doi.org/10.1002/widm.1312

[47] Cars H Hommes and F Wagener. 2009. Bounded rationality and learning in complex markets. *Handbook of economic complexity* 87 (2009), 123. https://doi.org/10.4337/9781781952665.00011

[48] Sungsoo Ray Hong, Jessica Hullman, and Enrico Bertini. 2020. Human factors in model interpretability: Industry practices, challenges, and needs. *Proceedings of the ACM on Human-Computer Interaction* 4, CSCW1 (2020), 1–26. https://doi.org/10.1145/3392878

[49] Jessica Hullman, Xiaoli Qiao, Michael Correll, Alex Kale, and Matthew Kay. 2018. In pursuit of error: A survey of uncertainty visualization evaluation. *IEEE transactions on visualization and computer graphics* 25, 1 (2018), 903–913. https://doi.org/10.1109/TVCG.2018.2864889

[50] Sarah Inman and David Ribes. 2019. "Beautiful Seams" Strategic Revelations and Concealments. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. 1–14. https://doi.org/10.1145/3290605.3300508

[51] Laurent Itti and Pierre Baldi. 2009. Bayesian surprise attracts human attention. *Vision research* 49, 10 (2009), 1295–1306. https://doi.org/10.1016/j.visres.2008.09.007

[52] Weina Jin, Xiaoxiao Li, and Ghassan Hamarneh. 2023. Rethinking AI Explainability and Plausibility. *arXiv preprint arXiv:2303.17707* (2023).

[53] Jongbin Jung, Connor Concannon, Ravi Shroff, Sharad Goel, and Daniel G. Goldstein. 2017. Simple Rules for Complex Decisions. *SSRN Electronic Journal* (2 2017). https://doi.org/10.2139/SSRN.2919024

[54] Daniel Kahneman. 2003. A perspective on judgment and choice: mapping bounded rationality. *American psychologist* 58, 9 (2003), 697.

[55] Daniel Kahneman. 2011. *Thinking, fast and slow*. Macmillan.

[56] Daniel Kahneman, Stewart Paul Slovic, Paul Slovic, and Amos Tversky. 1982. *Judgment under uncertainty: Heuristics and biases*. Cambridge university press.

[57] Monique Kardos and Patricia Dexter. 2017. *A simple handbook for non-traditional red teaming*. Technical Report. Defence Science and Technology Group Edinburgh, SA.

[58] Harmanpreet Kaur, Eytan Adar, Eric Gilbert, and Cliff Lampe. 2022. Sensible AI: Re-Imagining Interpretability and Explainability Using Sensemaking Theory. In *2022 ACM Conference on Fairness, Accountability, and Transparency* (Seoul, Republic of Korea) *(FAccT '22)*. Association for Computing Machinery, New York, NY, USA, 702–714. https:

//doi.org/10.1145/3531146.3533135
[59] Harmanpreet Kaur, Harsha Nori, Samuel Jenkins, Rich Caruana, Hanna Wallach, and Jennifer Wortman Vaughan. 2020. Interpreting Interpretability: Understanding Data Scientists' Use of Interpretability Tools for Machine Learning. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–14. https://doi.org/10.1145/3313831.3376219
[60] Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. 2017. Lightgbm: A highly efficient gradient boosting decision tree. *Advances in neural information processing systems* 30 (2017).
[61] Been Kim, Rajiv Khanna, and Oluwasanmi O Koyejo. 2016. Examples are not enough, learn to criticize! Criticism for Interpretability. In *Advances in Neural Information Processing Systems 29*, D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett (Eds.). Curran Associates, Inc., 2280–2288. http://papers.nips.cc/paper/6300-examples-are-not-enough-learn-to-criticize-criticism-for-interpretability.pdf
[62] Rafal Kocielnik, Saleema Amershi, and Paul N Bennett. 2019. Will you accept an imperfect AI? exploring designs for adjusting end-user expectations of ai systems. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. 1–14. https://doi.org/10.1145/3290605.3300641
[63] Charles Kostelnick. 2016. The re-emergence of emotional appeals in interactive data visualization. *Technical Communication* 63, 2 (2016), 116–135.
[64] Todd Kulesza, Margaret Burnett, Weng-Keen Wong, and Simone Stumpf. 2015. Principles of Explanatory Debugging to Personalize Interactive Machine Learning. In *Proceedings of the 20th International Conference on Intelligent User Interfaces* (Atlanta, Georgia, USA) *(IUI '15)*. ACM, New York, NY, USA, 126–137. https://doi.org/10.1145/2678025.2701399
[65] Isaac Lage, Emily Chen, Jeffrey He, Menaka Narayanan, Sam Gershman, Been Kim, and Finale Doshi-Velez. 2019. Human Evaluation of Models Built for Interpretability. In *AAAI Conference on Human Computation and Crowdsourcing (HCOMP)*. https://doi.org/10.1609/hcomp.v7i1.5280
[66] 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).
[67] Himabindu Lakkaraju, Stephen H Bach, and L Jure. 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*. Association for Computing Machinery. https://doi.org/10.1145/2939672.2939874
[68] David B. Leake. 1991. Goal-based explanation evaluation. *Cognitive Science* 15, 4 (1991), 509–545. https://doi.org/10.1016/0364-0213(91)80017-Y
[69] Jaewook Lee, Jaylin Herskovitz, Yi-Hao Peng, and Anhong Guo. 2022. ImageExplorer: Multi-Layered Touch Exploration to Encourage Skepticism Towards Imperfect AI-Generated Image Captions. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. 1–15. https://doi.org/10.1145/3491102.3501966
[70] Yoonjoo Lee, Tae Soo Kim, Sungdong Kim, Yohan Yun, and Juho Kim. 2023. DAPIE: Interactive Step-by-Step Explanatory Dialogues to Answer Children's Why and How Questions. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–22. https://doi.org/10.1145/3544548.3581369
[71] Q Vera Liao, Hariharan Subramonyam, Jennifer Wang, and Jennifer Wortman Vaughan. 2023. Designerly understanding: Information needs for model transparency to support design ideation for AI-powered user experience. In *Proceedings of the 2023 CHI conference on human factors in computing systems*. 1–21. https://doi.org/10.1145/3544548.3580652
[72] Q Vera Liao and Kush R Varshney. 2021. Human-centered explainable ai (xai): From algorithms to user experiences. *arXiv preprint arXiv:2110.10790* (2021).
[73] Peter Lipton. 1990. Contrastive explanation. *Royal Institute of Philosophy Supplements* 27 (1990), 247–266.
[74] Zachary C. Lipton. 2018. The mythos of model interpretability. *Commun. ACM* 61 (9 2018), 36–43. Issue 10. https://doi.org/10.1145/3233231
[75] Tania Lombrozo. 2006. The structure and function of explanations. *Trends in Cognitive Sciences* 10, 10 (2006), 464–470. https://doi.org/10.1016/j.tics.2006.08.004
[76] Tania Lombrozo. 2012. Explanation and abductive inference. (2012).
[77] Scott M Lundberg and Su-In Lee. 2017. A Unified Approach to Interpreting Model Predictions. In *Advances in Neural Information Processing Systems 30*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (Eds.). Curran Associates, Inc., 4765–4774. http://papers.nips.cc/paper/7062-a-unified-approach-to-interpreting-model-predictions.pdf
[78] Prashan Madumal, Tim Miller, Frank Vetere, and Liz Sonenberg. 2018. Towards a Grounded Dialog Model for Explainable Artificial Intelligence. In *First international workshop on socio-cognitive systems at IJCAI 2018*. https://arxiv.org/abs/1806.08055
[79] Bertram F Malle. 2006. *How the mind explains behavior: Folk explanations, meaning, and social interaction*. Mit Press.
[80] Filip Matějka and Alisdair McKay. 2015. Rational Inattention to Discrete Choices: A New Foundation for the Multinomial Logit Model. *American Economic Review* 105, 1 (January 2015), 272–98. https://doi.org/10.1257/aer.

20130047
[81] George Herbert Mead. 1934. *Mind, self and society*. Vol. 111. Chicago University of Chicago Press.
[82] David Alvarez Melis, Harmanpreet Kaur, Hal Daumé III, Hanna Wallach, and Jennifer Wortman Vaughan. 2021. From Human Explanation to Model Interpretability: A Framework Based on Weight of Evidence. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, Vol. 9. 35–47. https://doi.org/10.1609/hcomp.v9i1.18938
[83] John Stuart Mill. 1836. On the definition of political economy; and on the method of investigation proper to it. *London and Westminster Review* 4, October (1836), 120–164.
[84] Tim Miller. 2019. Explanation in Artificial Intelligence: Insights from the Social Sciences. *Artificial Intelligence* 267 (2019), 1–38. https://doi.org/10.1016/j.artint.2018.07.007
[85] Tim Miller. 2021. Contrastive explanation: A structural-model approach. *The Knowledge Engineering Review* 36 (2021). https://doi.org/10.1017/S0269888921000102
[86] Tim Miller, Piers Howe, and Liz Sonenberg. 2017. Explainable AI: Beware of inmates running the asylum or: How I learnt to stop worrying and love the social and behavioural sciences. In *IJCAI 2017 Workshop on Explainable Artificial Intelligence (XAI)*.
[87] Shane T Mueller, Robert R Hoffman, William Clancey, Abigail Emrey, and Gary Klein. 2019. Explanation in human-AI systems: A literature meta-review, synopsis of key ideas and publications, and bibliography for explainable AI. *arXiv preprint arXiv:1902.01876* (2019).
[88] Richard E Nisbett and Timothy D Wilson. 1977. Telling more than we can know: Verbal reports on mental processes. *Psychological review* 84, 3 (1977), 231.
[89] Harsha Nori, Samuel Jenkins, Paul Koch, and Rich Caruana. 2019. InterpretML: A Unified Framework for Machine Learning Interpretability. *arXiv preprint arXiv:1909.09223* (2019).
[90] 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*. 340–350. https://doi.org/10.1145/3397481.3450639
[91] Jum C Nunnally. 1978. An overview of psychological measurement. *Clinical diagnosis of mental disorders* (1978), 97–146.
[92] Jeffrey M; Zhu Pingping; Sommer Marc A; Ferrari Silvia; Egner Tobias Oh, Hanna; Beck. 2016. Satisficing in Split-Second Decision Making Is Characterized by Strategic Cue Discounting. , 1937–1956 pages. https://doi.org/10.1037/xlm0000284
[93] Clemens Otte. 2013. Safe and interpretable machine learning: a methodological review. *Computational intelligence in intelligent data analysis* (2013), 111–122.
[94] Heather O'Brien. 2016. Theoretical perspectives on user engagement. In *Why engagement matters*. Springer, 1–26.
[95] 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*. 1–9. https://doi.org/10.1145/3491102.3502104
[96] Charles Sanders Peirce. 1878. Illustrations of the Logic of Science: IV The Probability of Induction. *Popular Science Monthly* 12 (April 1878), 705–718.
[97] Joseph C Pitt. 1988. *Theories of explanation*. Oxford University Press.
[98] Jan Pöppel and Stefan Kopp. 2018. Satisficing Models of Bayesian Theory of Mind for Explaining Behavior of Differently Uncertain Agents: Socially Interactive Agents Track. In *Proceedings of the 17th International Conference on Autonomous Agents and MultiAgent Systems* (Stockholm, Sweden) *(AAMAS '18)*. International Foundation for Autonomous Agents and Multiagent Systems, Richland, SC, 470–478.
[99] 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*. 1–52. https://doi.org/10.1145/3411764.3445315
[100] Snehal Prabhudesai, Leyao Yang, Sumit Asthana, Xun Huan, Q Vera Liao, and Nikola Banovic. 2023. Understanding Uncertainty: How Lay Decision-makers Perceive and Interpret Uncertainty in Human-AI Decision Making. In *Proceedings of the 28th International Conference on Intelligent User Interfaces*. 379–396. https://doi.org/10.1145/3581641.3584033
[101] J R Quinlan. 1986. Induction of Decision Trees. *Mach. Learn.* (1986). https://doi.org/10.1023/A:1022643204877
[102] Gonzalo Ramos, Christopher Meek, Patrice Simard, Jina Suh, and Soroush Ghorashi. 2020. Interactive machine teaching: a human-centered approach to building machine-learned models. *Human–Computer Interaction* 35, 5-6 (2020), 413–451. https://doi.org/10.1080/07370024.2020.1734931
[103] Charvi Rastogi, Yunfeng Zhang, Dennis Wei, Kush R Varshney, Amit Dhurandhar, and Richard Tomsett. 2022. Deciding fast and slow: The role of cognitive biases in ai-assisted decision-making. *Proceedings of the ACM on Human-Computer Interaction* 6, CSCW1 (2022), 1–22. https://doi.org/10.1145/3512930

[104] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. " Why should I trust you?" Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*. 1135–1144. https://doi.org/10.1145/2939672.2939778

[105] Anna M Rose, Jacob M Rose, Kristian Rotaru, Kerri-Ann Sanderson, and Jay C Thibodeau. 2022. Effects of uncertainty visualization on attention, arousal, and judgment. *Behavioral Research in Accounting* 34, 1 (2022), 113–139. https://doi.org/10.2308/BRIA-2021-011

[106] Cynthia Rudin. 2019. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence* 1, 5 (2019), 206–215. https://doi.org/10.1038/s42256-019-0048-x

[107] Salvatore Ruggieri, Dino Pedreschi, and Franco Turini. 2010. Data mining for discrimination discovery. *ACM Transactions on Knowledge Discovery from Data (TKDD)* 4, 2 (2010), 1–40. https://doi.org/10.1145/1754428.1754432

[108] Jeff Sauro. 2015. SUPR-Q: A comprehensive measure of the quality of the website user experience. *Journal of usability studies* 10, 2 (2015).

[109] Wolfgang; Newell Ben R. Schulze, Christin; Gaissmaier. 2020. Maximizing as satisficing: On pattern matching and probability maximizing in groups and individuals. , 104382–104382 pages. https://doi.org/10.1016/j.cognition.2020.104382

[110] David Sculley, Gary Holt, Daniel Golovin, Eugene Davydov, Todd Phillips, Dietmar Ebner, Vinay Chaudhary, Michael Young, Jean-Francois Crespo, and Dan Dennison. 2015. Hidden technical debt in machine learning systems. *Advances in neural information processing systems* 28 (2015), 2503–2511.

[111] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. 2017. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*. 618–626.

[112] Lloyd S Shapley. 1997. A value for n-person games. *Classics in game theory* 69 (1997).

[113] Ronald W Shephard and Rolf Färe. 1974. The law of diminishing returns. In *Production theory*. Springer, 287–318.

[114] Herbert A Simon. 1955. A behavioral model of rational choice. *The quarterly journal of economics* 69, 1 (1955), 99–118.

[115] Herbert A. Simon. 1978. Rationality as Process and as Product of Thought. *The American Economic Review* 68, 2 (1978), 1–16. http://www.jstor.org/stable/1816653

[116] Herbert A. Simon. 1997. *Models of Bounded Rationality: Empirically Grounded Economic Reason*. The MIT Press. https://doi.org/10.7551/mitpress/4711.001.0001

[117] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. 2013. Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312. 6034* (2013).

[118] Ben R Slugoski, Mansur Lalljee, Roger Lamb, and Gerald P Ginsburg. 1993. Attribution in conversational context: Effect of mutual knowledge on explanation-giving. *European Journal of Social Psychology* 23, 3 (1993), 219–238. https://doi.org/10.1002/ejsp.2420230302

[119] 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*. 107–120. https://doi.org/10.1145/3301275.3302322

[120] Erik Štrumbelj and Igor Kononenko. 2014. Explaining prediction models and individual predictions with feature contributions. *Knowledge and information systems* 41, 3 (2014), 647–665. https://doi.org/10.1007/s10115-013-0679-x

[121] Alistair Sutcliffe. 2009. Designing for user engagement: Aesthetic and attractive user interfaces. *Synthesis lectures on human-centered informatics* 2, 1 (2009), 1–55.

[122] Alistair Sutcliffe. 2016. Designing for user experience and engagement. In *Why engagement matters*. Springer, 105–126.

[123] Ronald N Taylor. 1975. Psychological determinants of bounded rationality: Implications for decision-making strategies. *Decision Sciences* 6, 3 (1975), 409–429. https://doi.org/10.1111/j.1540-5915.1975.tb01031.x

[124] Amos Tversky and Daniel Kahneman. 1973. Availability: A heuristic for judging frequency and probability. *Cognitive psychology* 5, 2 (1973), 207–232.

[125] B van Fraassen. 1988. The Pragmatic Theory of Explanation. In *Theories of Explanation*, Joseph C Pitt (Ed.). Oxford University Press.

[126] Kush R Varshney and Homa Alemzadeh. 2017. On the safety of machine learning: Cyber-physical systems, decision sciences, and data products. *Big data* 5, 3 (2017), 246–255. https://doi.org/10.1089/big.2016.0051

[127] Helena Vasconcelos, Matthew Jörke, Madeleine Grunde-McLaughlin, Tobias Gerstenberg, Michael S Bernstein, and Ranjay Krishna. 2023. Explanations can reduce overreliance on ai systems during decision-making. *Proceedings of the ACM on Human-Computer Interaction* 7, CSCW1 (2023), 1–38. https://doi.org/10.1145/3579605

[128] Himanshu Verma, Jakub Mlynar, Roger Schaer, Julien Reichenbach, Mario Jreige, John Prior, Florian Evéquoz, and Adrien Depeursinge. 2023. Rethinking the role of AI with physicians in oncology: revealing perspectives from clinical and research workflows. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–19. https://doi.org/10.1145/3544548.3581506

[129] Janet Vertesi. 2014. Seamful spaces: Heterogeneous infrastructures in interaction. *Science, Technology, & Human Values* 39, 2 (2014), 264–284. https://doi.org/10.1177/0162243913516012
[130] Sandra Wachter, Brent Mittelstadt, and Chris Russell. 2017. Counterfactual explanations without opening the black box: Automated decisions and the GDPR. *Harv. JL & Tech.* 31 (2017), 841.
[131] Danding Wang, Qian Yang, Ashraf Abdul, and Brian Y. Lim. 2019. Designing Theory-Driven User-Centric Explainable AI. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) *(CHI '19)*. ACM, New York, NY, USA, Article 601, 15 pages. https://doi.org/10.1145/3290605.3300831
[132] Danding Wang, Wencan Zhang, and Brian Y Lim. 2021. Show or suppress? Managing input uncertainty in machine learning model explanations. *Artificial Intelligence* 294 (2021), 103456. https://doi.org/10.1016/j.artint.2021.103456
[133] 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*. 318–328. https://doi.org/10.1145/3397481.3450650
[134] Karl E Weick. 1995. *Sensemaking in organizations*. Vol. 3. Sage.
[135] Karl E Weick and Kathleen M Sutcliffe. 2015. *Managing the unexpected: Sustained performance in a complex world*. John Wiley & Sons.
[136] Karl E. Weick, Kathleen M. Sutcliffe, and David Obstfeld. 1999. Organizing for high reliability: Processes of Collective Mindfulness. *Research in Organizational Behaviour* 21 (1999), 81–123.
[137] Qian Yang, Aaron Steinfeld, and John Zimmerman. 2019. Unremarkable ai: Fitting intelligent decision support into critical, clinical decision-making processes. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. 1–11. https://doi.org/10.1145/3290605.3300468
[138] Jiaming Zeng, Berk Ustun, and Cynthia Rudin. 2017. Interpretable classification models for recidivism prediction. *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 180, 3 (2017), 689–722. https://doi.org/10.1111/rssa.12227
[139] Yunfeng Zhang, Q Vera Liao, and Rachel KE 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*. 295–305. https://doi.org/10.1145/3351095.3372852
[140] Yayan Zhao, Mingwei Li, and Matthew Berger. 2023. Graphical Perception of Saliency-based Model Explanations. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–15. https://doi.org/10.1145/3544548.3581320
[141] Jianlong Zhou, Amir H Gandomi, Fang Chen, and Andreas Holzinger. 2021. Evaluating the quality of machine learning explanations: A survey on methods and metrics. *Electronics* 10, 5 (2021), 593. https://doi.org/10.3390/electronics10050593

A DESCRIPTIVE STATISTICS

Figure 10 presents descriptive statistics for all the independent variables in Table 2. The values for two variables are averaged across the questions asked about them because of high internal consistency in their responses. First, hypothetical use rating, since Cronbach's Alpha for the three ratings questions (the use of the data and the model in the wild, and the use of accuracy as a key performance indicator) was α=0.73. This α value is acceptable for exploratory research [91]. Second, error recognition, since the ratings for the two error-related questions about missing values and redundant features were strongly correlated (Pearson's $r(115)$=0.76, $p < 0.001$).

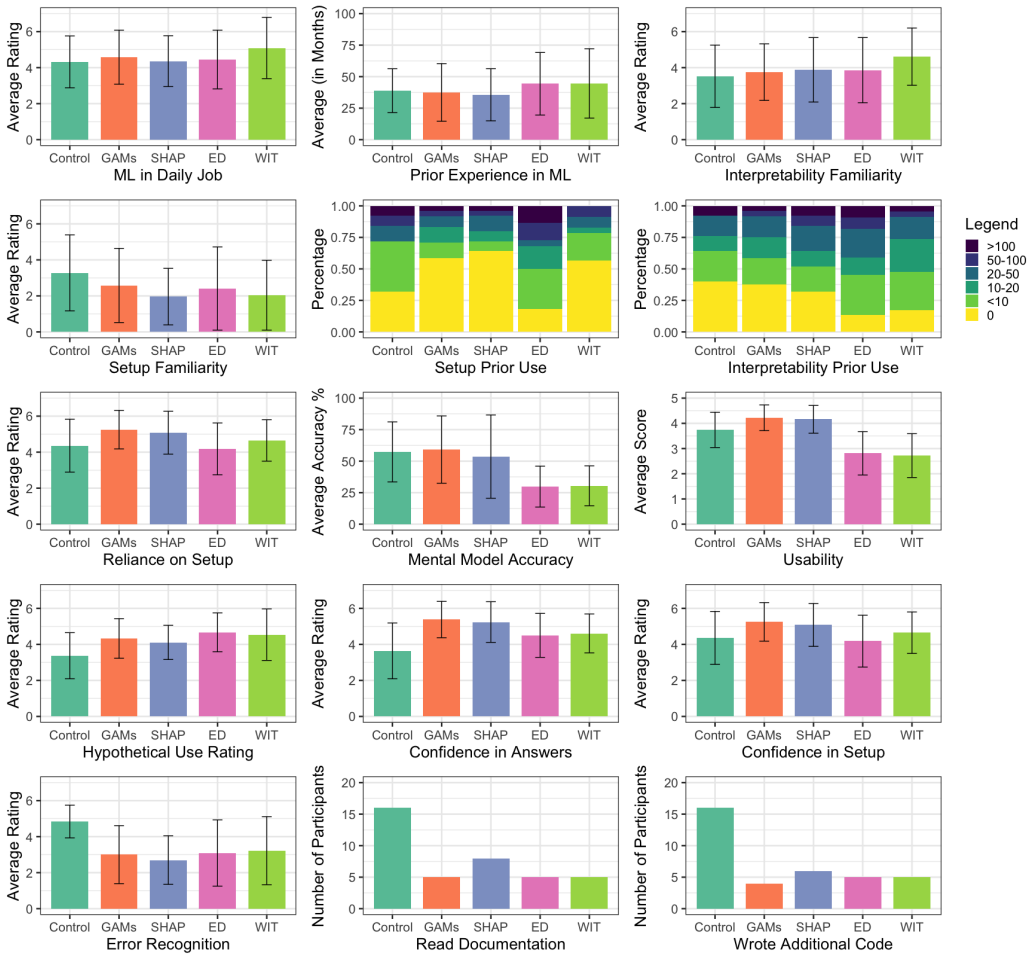

Fig. 10. Descriptive statistics for all the independent variables in our study.

Received January 2023; revised July 2023; accepted November 2023