

Interpreting Interpretability: Understanding Data Scientists’ Use of Interpretability Tools for Machine Learning

Harmanpreet Kaur¹, Harsha Nori², Samuel Jenkins²,
Rich Caruana², Hanna Wallach², Jennifer Wortman Vaughan²

¹University of Michigan, ²Microsoft Research
harmank@umich.edu, {hanori,sajenkin,rcaruana,wallach,jenn}@microsoft.com

ABSTRACT

Machine learning (ML) models are now routinely deployed in domains ranging from criminal justice to healthcare. With this newfound ubiquity, ML has moved beyond academia and grown into an engineering discipline. To that end, interpretability tools have been designed to help data scientists and machine learning practitioners better understand how ML models work. However, there has been little evaluation of the extent to which these tools achieve this goal. We study data scientists’ use of two existing interpretability tools, the InterpretML implementation of GAMs and the SHAP Python package. We conduct a contextual inquiry (N=11) and a survey (N=197) of data scientists to observe how they use interpretability tools to uncover common issues that arise when building and evaluating ML models. Our results indicate that data scientists over-trust and misuse interpretability tools. Furthermore, few of our participants were able to accurately describe the visualizations output by these tools. We highlight qualitative themes for data scientists’ mental models of interpretability tools. We conclude with implications for researchers and tool designers, and contextualize our findings in the social science literature.

Author Keywords

interpretability; machine learning; user-centric evaluation

CCS Concepts

•Computing methodologies → Machine learning;
•Human-centered computing → User studies;

INTRODUCTION

Machine learning (ML) has become ubiquitous in our everyday lives in domains ranging from criminal justice and public policy to healthcare and education. Modern ML models show promise on problems in computer vision and natural language processing for which solutions were only recently out of reach. Moreover, while ML was once confined to the academic community, it has now grown into a full engineering discipline.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

CHI '20, April 25–30, 2020, Honolulu, HI, USA.

Copyright is held by the owner/author(s). Publication rights licensed to ACM.

ACM ISBN 978-1-4503-6708-0/20/04 ...\$15.00.

<http://dx.doi.org/10.1145/3313831.3376219>

These developments create countless opportunities for impact, but with these opportunities come new challenges. ML models have been found to amplify societal biases in datasets and lead to unfair outcomes [4, 9, 29]. When ML models have the potential to affect people’s lives, it is critical that their developers are able to understand and justify their behavior. More generally, data scientists and machine learning practitioners cannot debug their models if they do not understand their behavior. Yet the behavior of complex ML models like deep neural networks and random forests is notoriously difficult to understand.

Faced with these challenges, the ML community has turned its attention to the design of techniques aimed at *interpretability*¹ [14, 39]. These techniques generally take one of two approaches. First, there are ML models that are designed to be inherently interpretable, often due to their simplicity, such as point systems [25, 68] or generalized additive models (GAMs) [10]. Second, there are techniques that provide post-hoc explanations for the predictions made by complex models, such as local interpretable model-agnostic explanations (LIME) [55] and Shapley additive explanations (SHAP) [44].

Although the number of proposed techniques continues to grow, there has been little evaluation of whether they help stakeholders achieve their desired goals. User studies of interpretability are notoriously challenging for a number of reasons. They require expertise in the mathematics underlying ML models and in human–computer interaction (HCI), as well as knowledge of both the academic literature and day-to-day engineering practices. To paint a full picture, studies must rely on qualitative methods to understand the nuances of how tools are used in context, and quantitative methods to scale up findings. They must also mimic realistic settings, yet not be too cumbersome (e.g., take over an hour to complete).

We study the effectiveness of interpretability tools for one key stakeholder group: data scientists and machine learning practitioners. (For simplicity, we refer to this group simply as “data scientists” throughout the paper.) We perform a human-centric evaluation of two existing interpretability tools, the InterpretML implementation of GAMs and the SHAP Python package, in the context of building and evaluating ML models. Our work consists of three components that build on each

¹There is not yet consensus within the research community on the distinction between the terms *interpretability*, *intelligibility*, and *explainability*, and they are often, though not always, used interchangeably. Throughout this paper, we stick with *interpretability*, which is more commonly used within the machine learning community.

other: 1) a series of pilot interviews ($N = 6$) to identify common issues faced by data scientists in their day-to-day work; 2) a contextual inquiry ($N = 11$) to observe data scientists' abilities to use interpretability tools to uncover these issues, and 3) a large-scale survey ($N = 197$) to scale up and quantify the main findings from our contextual inquiry and shed more light on data scientists' mental models of interpretability tools.

Our results indicate that the visualizations output by interpretability tools can sometimes—though not always—help data scientists to uncover the kinds of issues that they deal with on a day-to-day basis in their work, such as missing values in a dataset that have been filled in incorrectly. We found that the choice of interpretability tool matters, with participants performing better using one than the other. However, for both tools, the existence of visualizations and the fact that the tools were publicly available led to cases of over-trust and misuse. Furthermore, despite being provided with standard tutorials, few of participants were able to accurately describe what the visualizations were showing. Participants were also biased toward model deployment, despite recognizing suspicious aspects of the ML models. This was true even when we showed them manipulated, nonsensical explanations, though we observed this less with data scientists who were more experienced. Our results highlight challenges for designing interpretability tools for data scientists, and, in line with prior work [49, 62], we advocate for similar user-centric evaluations to be conducted for all stakeholders of interpretability tools and ML models.

RELATED WORK

Machine Learning Research on Interpretability

Within the ML community, there has been a surge of research on interpretability techniques. These techniques generally take one of two approaches: “glassbox” ML models that are designed to be inherently interpretable (e.g., simple point systems [25, 68], GAMs [10]) and post-hoc explanation techniques that are designed to make the predictions of “black-box” models more interpretable (e.g., LIME [55], SHAP [44]). Gilpin et al. [19] provide an overview of different interpretability techniques. However, despite this proliferation of techniques, there is still debate about what interpretability should entail [14, 39, 56]. Rudin [56] 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. Doshi-Velez et al. [15] propose that an explanation for a particular prediction should include not only a justification, but also a description of the decision-making process followed by the model. Lipton [39] surveys different criteria for assessing interpretability, such as simulatability, as well as different goals that interpretability may be used to achieve.

Only recently has the ML community begun to evaluate interpretability techniques via user studies. Tan et al. [61] use publicly available datasets to test the feasibility of a new GAM-based post-hoc explanation technique with ML experts. Kim et al. [31] do the same for a technique based on Bayesian model criticism, intended to add criticisms to example-based explanations. Poursabzi-Sangdeh et al. [54] test the impact of two factors often thought to affect interpretability—number of input

features and model transparency (i.e., glassbox vs. blackbox). 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 can cause people to incorrectly follow a model when it makes a mistake, due to information overload. Lage et al. [36] study two aspects of an explanation (length and complexity) via a wizard-of-oz approach in two domains, finding that longer explanations overload people's cognitive abilities.

Social Science Research on Explanations

Complementary to the ML community's research on interpretability, social science research has focused on how to craft and present explanations in such a way that people are able to understand and use them. Most notably, Miller [47] defines overarching principles from the social science literature on explanations: explanations are contrastive, social, and selected by people in a biased manner (according to cognitive heuristics); referring to probabilities or statistical generalizations is usually unhelpful. To that end, Miller [47, 48] and Lombrozo [41] propose 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 et al.'s maxims of quality, quantity, relation, and manner [20], which form the core of a good conversation, should therefore be followed when designing explanations [34, 46, 59]. Leake's goal-based approach to explanation evaluation further includes metrics such as the timeliness of an explanation in providing the opportunity to deal better with the prediction being explained, knowability and the features responsible for “knowing,” and the independence of individual explanations [38]. Explanations that follow this goal-based approach must include grounding in some common demonstrative reference between people and the explanation system [11, 45].

Although this line of work provides guidance for designing explanation systems that work best for humans—and some ML researchers have begun to incorporate this guidance into interpretability techniques [2]—it also criticizes the development of explanation systems by ML researchers, citing this as an example of “inmates running the asylum” [48] because of a lack of user-centric evaluation. Our work addresses this critique by conducting user studies of two existing interpretability tools.

HCI and Interpretability

HCI has a long-standing tradition of studying complex systems from a user-centric perspective. Bellotti and Edwards [6] were the first to consider intelligibility and accountability, providing guidelines for system designers. These guidelines include clarifying the system's capabilities, providing feedback, navigating privacy via personalized settings, and providing control and interactive guidance for edge cases. This line of work has been extended to ML-based systems, with a focus on interactive ML (iML). The term was coined by Fails and Olsen Jr. [18] to describe an approach where people are involved in an iterative process of training, using, and correcting an ML model, requiring interpretability for effective corrections [13]. Several examples of iML systems now exist, for applications including annotation of animal behavior [26], academic citation review [63], and on-demand personalized group formation [3].

More recently, the HCI community has sought to improve the relationship between people and machines in iML systems. Stumpf et al. [60] study the viability of including rich human feedback in an ML-based email filtering system, finding the lack of common language shared by people and machines to be a key challenge. Kulesza et al. [35] suggest that accurate mental models can help people to better use a music recommender system, leading to greater satisfaction. Kocielnik et al. [33] find that people’s mental models and subsequent expectations of accuracy from an ML-based system can be controlled via different design strategies. Patel et al. [53] and Zhu et al. [69] provide domain-specific challenges to designing iML systems for software developers and game designers, respectively.

Finally, there has been recent research within the HCI community on defining new metrics for human–ML collaboration. Abdul et al. [1] highlight interactivity and learnability as cornerstones for designing visualizations that better support interpretability. Dourish [16] adds scrutability as an important component of interactivity. Most closely related to our work, Hohman et al. [23] follow these guidelines to design visualizations for GAMs, confirming that interactivity plays an important role in data scientists’ understanding and satisfaction.

PILOT INTERVIEWS

The goal of our work is to study data scientists’ use of two existing interpretability tools. To better understand the issues that data scientists face in their data-to-day work—i.e., the setting in which the interpretability tools will be used—we first conducted semi-structured interviews with six data scientists at a large technology company. The interview protocol (included in the supplementary material) was designed to surface common issues that arise when building and evaluating ML models. On average, each interview lasted about 40 minutes.

Based on an inductive thematic analysis of the interview transcripts, conducted via open coding followed by affinity diagramming [7], we identified six themes capturing common issues faced by data scientists. We present these themes in the first two columns of Table 1. Five correspond to issues with data itself: missing values, temporal changes in the data, duplicate data masked as unique, correlated features, and ad-hoc categorization. The sixth theme relates to the difficulty of trying to debug or identify potential improvements to an ML model based on only a small number of data points. With only six interviews, we cannot claim that this list is at all exhaustive, but it is consistent with previous research on ML pitfalls [42].

STUDY 1: CONTEXTUAL INQUIRY

With these common issues in mind, we designed a contextual inquiry, intended to put data scientists in a realistic setting: exploring a dataset and an ML model in a hands-on fashion. We recruited eleven participants, each of whom was given a Jupyter notebook that included a dataset, an ML model that we had trained using that dataset, an interpretability tool that we had set up, and several questions to answer. Our goal was to observe whether they were able to use the interpretability tool to uncover the issues identified via our pilot interviews. With participants’ consent, we recorded both audio and video,

and saved all responses provided in the Jupyter notebooks for analysis. The scenario was approved by our internal IRB.

Dataset

We derived the dataset from the Adult Income dataset,² a publicly available ML dataset based on 1994 US census data. Each data point corresponds to a person. The input features include age, education, marital status, native country, and occupation. Each label is a binary value indicating whether or not the person in question made >\$50k in 1994 (equivalent to ~\$86.5k when adjusted for inflation). We synthetically manipulated a subset of the features to incorporate the common issues identified via our pilot interviews. For example, to incorporate 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 >\$50k. We provide the details of all manipulations in Table 1.

ML Models and Interpretability Tools

We used two existing interpretability tools: one that implements GAMs, an inherently interpretable technique, and one that implements SHAP, a post-hoc explanation technique.

GAMs are a class of ML models, rooted in statistics, that decompose a learned predictor into additive components that are functions of one input feature each [22]. Each component can be complex and non-linear, but because it is a function of only a single input feature it can be easily visualized. GAMs can be as accurate as more complex ML models such as random forests or boosted decision trees. Because GAMs are glassbox models that are designed to be inherently interpretable, they do not require post-hoc explanations. We used the InterpretML³ implementation of GAMs. InterpretML provides built-in plotting functionality, allowing each individual component to be visualized (see Figure 1, top middle). InterpretML also provides global explanations (see Figure 1, top left) and local explanations (see Figure 1, top right) by ranking and sorting the contributions made by each input feature to the predictions [50].

SHAP is a post-hoc explanation technique for blackbox ML models. It assigns each input feature an importance score for each prediction [44]. These scores are based on the notion of Shapley values from cooperative game theory [58]; for each prediction, they capture a fair distribution of “credit” over the input features. We used the implementation of SHAP in the SHAP Python package.⁴ The importance scores computed by this package directly translate to local explanations for individual predictions (see Figure 1, bottom right). By aggregating the importance scores for many predictions, the SHAP Python package can also produce global explanations (see Figure 1, bottom left) and dependence plots for single input features (see Figure 1, bottom middle). Ideally, we would have used the same underlying ML model—i.e., a GAM—with both interpretability tools; however, it was not computationally feasible to generate explanations for GAMs using the SHAP Python package. As a result, we used LightGBM [30], an implementation of gradient boosted decision trees, to create the underlying model to be explained using the SHAP Python package. This

²<https://archive.ics.uci.edu/ml/datasets/Adult>

³<https://github.com/interpretml/interpret>

⁴<https://github.com/slundberg/shap>

Theme	Description	Incorporation into Contextual Inquiry	Num.
Missing values	Many methods for dealing with missing values (e.g., coding as a unique value or imputing with the mean) can cause biases or leakage in ML models.	Replaced the value for the “Age” feature with 38 (the mean) for 10% of the data points with an income of >\$50k, causing predictions to spike at 38. Asked about the relationship between “Age” and “Income.”	4 of 11
Changes in data	Data can change over time (e.g., new categories for an existing feature).	Asked whether the model (trained on 1994 data) would work well on current data after adjusting for inflation.	10 of 11
Duplicate data	Unclear or undefined naming conventions can lead to accidental duplication of data.	Modified the “WorkClass” feature to have duplicate values: “Federal Employee,” “Federal Worker,” “Federal Govt.” Asked about the relationship between “WorkClass” and “Income.”	1 of 11
Redundant features	Including the same feature in several ways can distribute importance across all of them, making each appear to be less important.	Included two features, “Education” and “EducationNum,” that represent the same information. Asked about the relationships between each of these and “Income.”	3 of 11
Ad-hoc categorization	Category bins can be chosen arbitrarily when converting a continuous feature to a categorical feature.	Converted “HoursPerWeek” into a categorical feature, binning arbitrarily at 0–30, 30–60, 60–90, and 90+ hours. Asked about the relationship between “HoursPerWeek” and “Income.”	3 of 11
Debugging difficulties	Identifying potential model improvements based on only a small number of data points is difficult.	Asked people to identify ways to improve accuracy based on local explanations for 20 misclassified data points.	8 of 11

Table 1: Six themes capturing common issues faced by data scientists. We incorporated each issue into the contextual inquiry as described in the third column. The fourth column contains the number of participants who identified the corresponding issue.

was for three reasons: the InterpretML implementation of GAMs is based on gradient boosted decision trees, the SHAP Python package has a highly optimized routine for computing explanations for the predictions made by a LightGBM model [43], and LightGBM is widely used. We emphasize that the comparisons that we make between GAMs and SHAP are comparisons between the InterpretML implementation of a GAM and the SHAP Python package used to explain a LightGBM model. The two trained models had similar test-set accuracies (.907 and .904, respectively) for our dataset.

We chose these two interpretability tools because they are publicly available, widely used, and provide both local and global explanations. In contrast, LIME, another popular post-hoc explanation technique, provides only local explanations. Each participant used only one interpretability tool, selected at random; 6 participants used GAMs, while 5 used SHAP.

We provided each participant with a print-out of a tutorial, written by us based on READMEs and examples included with the interpretability tools, containing a light overview of the math behind the interpretability technique implemented in the tool that they were to use and information on the tool’s visualizations. These tutorials are in the supplementary material.

Contextual Inquiry Protocol

First, we asked each participant to sign a consent form and answer some questions. We followed a semi-structured interview protocol with questions about (1) their background in ML; (2) their team and role; (3) their typical ML pipeline, including how they make decisions about data and models; (4) any checks they typically perform on data or models; (5) if they work in customer-facing scenarios, what makes them feel confident enough about a model to deploy it; and (6) their awareness of and prior experience with interpretability tools.

Next, we let each participant explore the dataset, model, and interpretability tool on their own. For each tool, we provided examples of all three types of visualization—i.e., global explanations, components (GAMs) or dependence plots (SHAP), and local explanations, as depicted in Figure 1. After this, we

asked each participant to complete the trust questionnaire of Jian et al. [24] with respect to the interpretability tool. We then asked them to answer ten questions about the dataset and model. Four were general questions about the visualizations (e.g., “What are the most important features that affect the output Income, according to the explanation above?”), while the remaining six were designed to get at the issues identified via our pilot interviews; a full list is in the supplementary material. Answering these questions required participants to use all three types of visualization. For each question, we also asked each participant to rate their confidence in their understanding of the visualizations and their confidence that these explanations were reasonable, on a scale of 1 (not at all) to 7 (extremely). After answering the questions, we asked each participant to complete the trust questionnaire again, allowing us to observe whether their self-reported trust in the interpretability tool had changed. We ended with a short interview, asking each participant about their experience with the tool and whether it would be useful in their typical ML pipeline.

Participants and Data

We recruited participants via an internal mailing list at a large technology company. In order to filter out participants with no prior experience with ML, the recruitment email included a short survey asking people about their background in ML, the extent to which they had used interpretability tools before, their familiarity with GAMs or SHAP, and their familiarity with the Adult Income dataset. Out of 24 potential participants, all passed this initial filter, but we subsequently excluded several based on their location because we needed to conduct the contextual inquiry in person. This left 11 participants (4 women, 7 men; self-reported). Participants’ roles included ML researcher, applied data scientist, and intern in ML team. On average, participants had been in their current role for 2 years (min = 2 months, max = 6 years). Most participants were not familiar with the Adult Income dataset (average familiarity = 2 on a scale of 1–7) and moderately familiar with GAMs or SHAP (average = 4 on a scale of 1–7). All participants were compensated with a \$20 lunch coupon or gift certificate upon completion of the contextual inquiry.

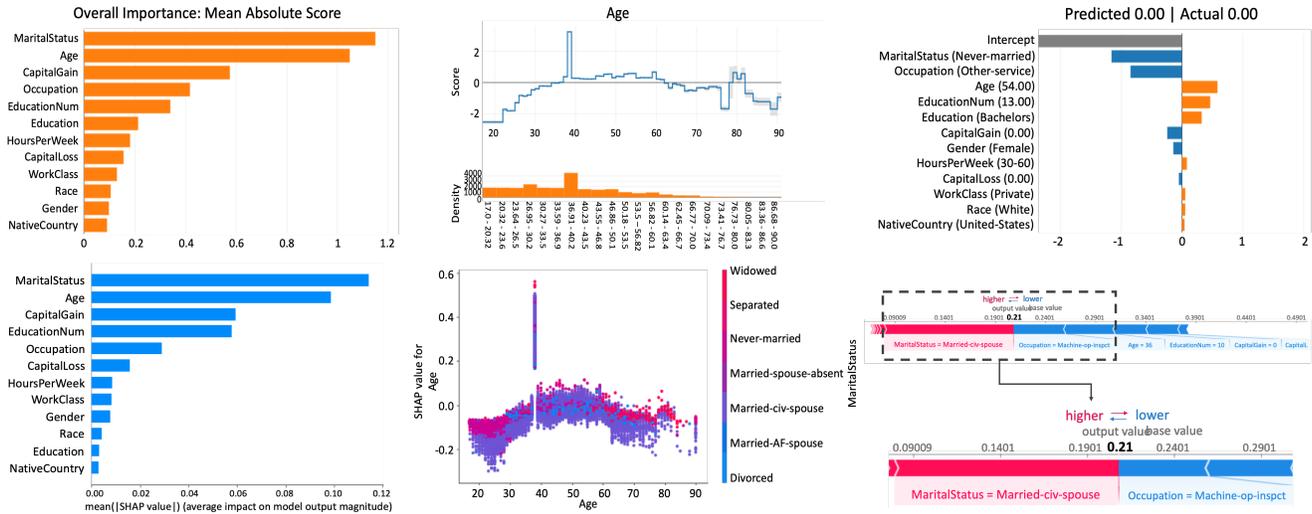


Figure 1: Visualizations output by the InterpretML implementation of GAMs (top) and the SHAP Python package (bottom). Left column: global explanations. Middle column: component (GAMs) or dependence plot (SHAP). Right column: local explanations.

We used speech recognition software to generate transcripts from the video files and hand-corrected any errors. We qualitatively coded these transcripts and participants’ open-ended responses to the questions using inductive thematic analysis [7]. We gave participants credit for uncovering an issue if there was any mention of confusion, suspicion, or a need for more testing in their response to the question about that issue. We also obtained descriptive statistics from the trust questionnaire and the questions about their background, etc.

Results

Our contextual inquiry reveals a misalignment between data scientists’ understanding of interpretability tools and these tools’ intended use. Participants either over- or under-used the tools. In some cases, they ended up over-trusting the dataset or the underlying ML model. Participants trusted the tools because of their visualizations and their public availability, though participants took the visualizations at face value instead of using them to uncover issues with the dataset or models.

The final column in Table 1 contains the number of participants who identified the corresponding issue. Each issue was identified by at least one participant. However, each participant only identified 2.5 issues on average (s.d.=1.4). Participants provided high ratings for their confidence in their understanding of the visualizations (mean=5.6, s.d.=0.8) and for their confidence that these explanations were reasonable (mean=5.0, s.d.=0.7). The only question for which participants’ average confidence rating was less than 5 (on a scale of 1–7) was one in which participants were asked to use local explanations for 20 misclassified data points to suggest ways to improve the model. Most participants (8 out of 11) recognized that this could not be done effectively. We did not observe a substantial difference in participants’ self-reported trust in the interpretability tools before and after using them, though the sample size is too small to make claims about significance; participants’ average trust (measured via Jian et al.’s trust questionnaire [24]) was 3.70 (s.d.=0.4) before using the tools and 3.90 (s.d.=0.6) after.

Misuse and Disuse

Most participants relied too heavily on the interpretability tools. Previous work categorizes such over-use as *misuse* [17, 52]. Here, the misuse resulted from over-trusting the tools because of their visualizations; participants were excited about the visualizations and took them at face value instead of using them to dig deeper into issues with the dataset or model:

“Age 38 seems to have the highest positive influence on income based on the plot. Not sure why, but the explanation clearly shows it... makes sense.” (P9, GAMs)

Although interpretability tools are meant to help data scientists understand how ML models work, some participants used the tools to rationalize suspicious observations instead. After conducting several exploratory tests on the dataset, P8 said “Test of means says the same thing as SHAP about Age. All’s good!” (P8, SHAP), and gave confidence ratings of 7 (extremely).

In contrast, two participants under-used the tools because they did not provide explanations with the content or clarity that they expected. P7 noted that “This is not an explanation system. It’s a visualization. There was no interpretation provided here” (P7, GAMs). Similarly, P4 became skeptical when they did not fully understand how SHAP’s importance scores values were being calculated, eventually leading to disuse [17, 52]:

“[The tool] assigns a value that is important to know, but it’s showing that in a way that makes you misinterpret that value. Now I want to go back and check all my answers”... [later] “Okay, so, it’s not showing me a whole lot more than what I can infer on my own. Now I’m thinking... is this an ‘interpretability tool’?” (P4, SHAP)

Social Context is Important

We found that social context was important to participants’ perception and use of the interpretability tools. Both InterpretML and the SHAP Python package are publicly available and widely used, which swayed several participants to trust the tools without fully understanding them. P8 said, “I guess this

is a publicly available tool... must be doing something right. I think it makes sense” (P8, SHAP). Meanwhile, P6 noted:

“I didn’t fully grasp what SHAP values were. This is a pretty popular tool and I get the log-odds concept in general. I figure they were showing SHAP values for a reason. Maybe it’s easier to judge relationships using log-odds instead of predicted value. Anyway, so it made sense I suppose.” (P6, SHAP)

Participants also relied too heavily on the interpretability tools because they had not encountered such visualizations before: “[The tool] shows visualizations of ML models, which is not something anything else I have worked with has done. It’s very transparent, and that makes me trust it more” (P9, GAMs).

Visualizations can be Misleading

The visualizations output by both interpretability tools lack details about importance scores and other values shown. These details were available in our tutorials, but it is not clear that participants internalized the tutorials enough to interpret the visualizations as intended. Most participants mentioned some confusion around the seemingly arbitrary values shown in the visualizations: “It shows all these values, and I’m not sure what they correspond to because they’re just written on the plot with no context for what they are” (P2, SHAP). However, participants continued to use the visualizations despite the missing details, which in turn led to incorrect assumptions about the dataset, models, and interpretability tools, as discussed above.

Some of the visualizations do not follow usability guidelines. P4, observing different axis ranges in different visualizations, remarked “cardinal sin of visualization when scales are not compatible.” However, many participants did not notice this and therefore made erroneous judgments about the contribution of each input feature to individual predictions. In fact, P4’s frustration is evident from their attempt to extract concrete information from SHAP’s local explanations: “Am I supposed to have some sort of calipers? How can anyone infer the weight (magnitude) [of each feature] from this (force plot)?”

STUDY 2: SURVEY

Following the contextual inquiry, we designed a survey to scale up and quantify our main findings and shed light on data scientists’ mental models of interpretability tools. Similar to our contextual inquiry, the survey placed data scientists in a realistic setting. The dataset, models, and interpretability tools used were identical to those used in the contextual inquiry. We ran the survey through Qualtrics. All participants were compensated with a \$20 gift card. Additionally, three participants were selected at random from those with high-quality open-ended responses to win a pair of Surface headphones.

Experimental Conditions

As in our contextual inquiry, each participant used only one interpretability tool (either the InterpretML implementation of GAMs or the SHAP Python package used to explain a LightGBM model, as described above), selected at random. We also showed participants either “normal” or “manipulated” visualizations, again selected at random. In the normal-visualization condition, we showed participants the visualizations output

by the interpretability tools. However, in the manipulated-visualization condition, we instead showed participants global and local explanations where the input feature names had been rearranged, resulting in the input features with smallest contributions to the predictions being displayed as the most important, and vice versa. We designed this manipulation to test the extent to which participants’ perception and use of the interpretability tools depend on how reasonable their explanations are (as opposed to the mere existence of visualizations).

Components of the Survey

First, we asked each participant to sign a consent form and gave them a brief introduction to the survey. We then asked them to answer some questions about their demographics and background, including (1) their current role and how long they had been in this role; (2) the extent to which ML was a part of their day-to-day work; (3) how long they had been using ML; (4) their familiarity with interpretability, and with GAMs or SHAP; (5) the approximate number of hours that they had spent using interpretability tools, and using GAMs or SHAP; and (6) their familiarity with the Adult Income dataset.

Because it was not possible to provide Jupyter notebooks, participants were not able to explore the dataset, models, and interpretability tools on their own. Instead, we showed them the results of common exploration commands that had been run previously by participants in the contextual inquiry. We also gave each participant access to a description of the dataset and a tutorial on the interpretability tool that they were to use.

Next, we asked each participant to answer four blocks of questions about the dataset and the model, covering global feature importance, the relationship between the age and the output variable (i.e., whether or not the person in question made >\$50k), the local explanation for a correctly classified data point, and the local explanation for a misclassified data point, respectively. Each of these blocks contained seven questions: (1) a multiple-choice question with a ground-truth correct answer, which was designed to quantify the participants’ accuracy at reading the visualizations (e.g., “Which is the 3rd most important feature for the underlying model, according to the explanation system?”); (2) an open-ended question designed to test how well participants understood the visualizations and whether any suspicions arose; (3) a question about which visualizations they had used to answer the previous questions; (4) their stated confidence in their understanding of the visualizations (on a scale of 1–7); (5) their stated confidence that these explanations were reasonable (on a scale of 1–7); (6) their stated confidence that the underlying model was reasonable (on a scale of 1–7); and (7) an optional open-ended text field for comments or concerns.

After answering the questions and familiarizing themselves with the visualizations, we asked each participant to select the capabilities of the interpretability tool from a list of options (shown in Figure 2). To better understand participants’ mental models of the tools, we asked them to describe what the x- and y-axes represented in each of the visualizations they had seen. We also asked them whether the visualizations were useful and, if so, how they would use them in their typical ML pipeline.

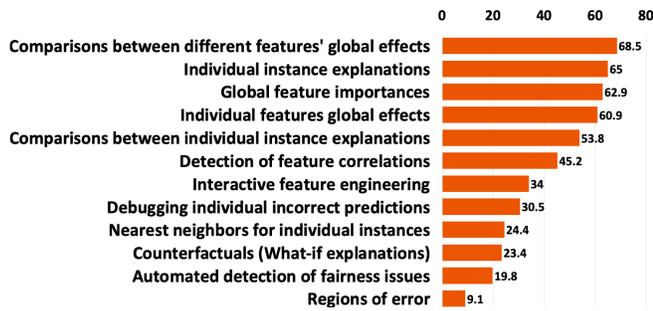


Figure 2: Percentage of participants that selected each option when asked about the interpretability tools' capabilities.

Finally, to encourage each participant to think critically about issues with the dataset or model, we asked them to rate the extent (on a scale of 1–7) to which they thought the model was ready for deployment and to explain this rating. We also asked them to describe how they would convince a customer that this was the right model to deploy (or not, as appropriate). We concluded with a NASA-TLX cognitive load index [21].

Participants

We advertised the survey via internal mailing lists at a large technology company and publicly via social media. To filter out participants with no prior experience with ML, we asked people about their ML experience, and only surveyed people who rated their experience as at least 3 on a scale of 1–7. We collected survey responses from 253 participants. After filtering out responses with exactly the same content for every open-ended text field or other gibberish text, we were left with responses from 197 participants. Of these, 49 participants were assigned to GAM-Normal, 51 to GAM-Manipulated, 51 to SHAP-Normal, and 46 to SHAP-Manipulated. On average, participants took about 44 minutes to complete the survey (s.d.=28), excluding clear outliers. All participants were compensated with a \$20 gift card upon completion of the survey.

Preregistration

Before conducting any analyses, we preregistered our intent to analyze eight outcome variables: (1) participants' accuracy on the four multiple-choice questions with ground-truth correct answers; (2) their stated confidence in their understanding of the visualizations; (3) their stated confidence that the explanations were reasonable; (4) their stated confidence that the underlying models were reasonable; (5) their NASA-TLX cognitive load index; (6) the extent to which they thought the models were ready for deployment; (7) whether they expressed any suspicions about the dataset or models; and (8) whether they expressed any suspicions about the interpretability tools.

For each outcome variable, the comparisons that we preregistered were (1) the main effect of normal vs. manipulated visualizations; (2) GAM-Normal vs. GAM-Manipulated; (3) GAM-Normal vs. SHAP-Normal; (4) SHAP-Normal vs. SHAP-Manipulated; and (5) GAM-Manipulated vs. SHAP-Manipulated. These comparisons allow us to understand differences that arise because of the tools, as well as differences that arise based on how reasonable the explanations are. We intentionally omitted comparing the main effect of GAMs

vs. SHAP to avoid lumping together data from the normal-visualization and manipulated-visualization conditions. We noted our intent to conduct additional exploratory analyses. The preregistration document is available on AsPredicted.⁵

Methods

We used two-way ANOVAs to compare the main effect of normal vs. manipulated visualizations and unpaired t-tests for the more specific comparisons. Following convention, we did not apply Bonferroni correction because only four comparisons were performed for each outcome variable. We coded the content from the open-ended responses via open and axial coding [12]. We used inductive, data-driven coding to code participants' open-ended responses (in sum) for any mention of suspicions about the dataset or models, or about the interpretability tools. Two authors coded these responses with an inter-rater reliability of 1, measured using Cohen's kappa on 12% of the data. Once coded, these suspicion variables were compared via Fisher's exact test; we used this over the preregistered chi-squared test due to a class imbalance for these variables. We conducted exploratory analyses using the participants' ML experience by fitting multiple linear regression models and calculating Pearson correlation coefficients. We also noted descriptive means, standard deviations, and counts.

Results

Table 2 summarizes the results of our preregistered analyses. For three outcome variables—participants' accuracy on the four multiple-choice questions with ground-truth correct answers, their stated confidence in their understanding of the visualizations, and their NASA-TLX cognitive load index—there are significant differences between GAM-Normal and SHAP-Normal and between GAM-Manipulated and SHAP-Manipulated. Specifically, participants who used GAMs had higher accuracy, higher stated confidence in their understanding of the visualizations, and lower cognitive load than participants who used SHAP. This suggests that explanations based on GAMs are easier to understand than explanations based on SHAP. There are no differences between normal and manipulated visualizations for these outcome variables.

Although there are no differences between conditions for participants' stated confidence that the explanations were reasonable, we do see differences in their stated confidence that the underlying models were reasonable, both as a main effect of normal vs. manipulated visualizations in an ANOVA ($F(1, 101) = 25.05, p \ll 0.001$) and when comparing GAM-Normal and GAM-Manipulated, SHAP-Normal and SHAP-Manipulated, and GAM-Manipulated and SHAP-Manipulated. These results indicate that it is not the mere existence of visualizations that matters. Reassuringly, participants were less confident that the underlying models were reasonable when shown manipulated visualizations. The difference between GAM-Manipulated and SHAP-Manipulated suggests that participants who used GAMs were more likely to be skeptical of the models when shown manipulated visualizations than participants who used SHAP. This is an argument in favor of GAMs.

⁵<https://aspredicted.org/ek2bm.pdf>

	GAM-Normal vs. GAM-Manipulated		GAM-Normal vs. SHAP-Normal		SHAP-Normal vs. SHAP-Manipulated		GAM-Manipulated vs. SHAP-Manipulated	
	μ_1	μ_2	μ_1	μ_2	μ_1	μ_2	μ_1	μ_2
Accuracy of answers	$\mu_1 = 78.1 \pm 25.1$	$\mu_2 = 74.0 \pm 28.4$	$\mu_1 = 78.1 \pm 25.1$ $t(98) = 3.83, p < 0.001, \text{Cohen's } d = 0.8$	$\mu_2 = 58.8 \pm 24.7$	$\mu_1 = 58.8 \pm 24.7$	$\mu_2 = 54.4 \pm 28.2$	$\mu_1 = 74.0 \pm 28.4$ $t(94) = 3.38, p < 0.001, \text{Cohen's } d = 0.7$	$\mu_2 = 54.4 \pm 28.2$
Confidence: understand explanation(s)	$\mu_1 = 5.7 \pm 0.6$	$\mu_2 = 5.3 \pm 1.1$	$\mu_1 = 5.7 \pm 0.6$ $t(93) = 8.77, p < 0.001, \text{Cohen's } d = 1.7$	$\mu_2 = 4.4 \pm 0.8$	$\mu_1 = 4.4 \pm 0.8$	$\mu_2 = 4.1 \pm 0.9$	$\mu_1 = 5.3 \pm 1.1$ $t(95) = 6.04, p < 0.001, \text{Cohen's } d = 1.2$	$\mu_2 = 4.1 \pm 0.9$
Confidence: explanation(s) are reasonable	$\mu_1 = 5.3 \pm 0.6$	$\mu_2 = 5.3 \pm 1.1$	$\mu_1 = 5.3 \pm 0.6$	$\mu_2 = 5.1 \pm 0.8$	$\mu_1 = 5.1 \pm 0.8$	$\mu_2 = 5.1 \pm 1.0$	$\mu_1 = 5.3 \pm 1.1$	$\mu_2 = 5.1 \pm 1.0$
Confidence: underlying model is reasonable	$\mu_1 = 4.8 \pm 0.9$ $t(98) = 5.14, p < 0.001, \text{Cohen's } d = 1.0$	$\mu_2 = 3.8 \pm 0.9$	$\mu_1 = 4.8 \pm 0.9$	$\mu_2 = 5.0 \pm 1.1$	$\mu_1 = 5.0 \pm 1.1$ $t(95) = 2.48, p < 0.05, \text{Cohen's } d = 0.5$	$\mu_2 = 4.4 \pm 1.1$	$\mu_1 = 3.8 \pm 0.9$ $t(89) = -3.03, p < 0.05, \text{Cohen's } d = 0.6$	$\mu_2 = 4.4 \pm 1.1$
Cognitive load	$\mu_1 = 3.8 \pm 0.9$	$\mu_2 = 4.2 \pm 1.0$	$\mu_1 = 3.8 \pm 0.9$ $t(96) = -6.40, p < 0.001, \text{Cohen's } d = 1.3$	$\mu_2 = 4.9 \pm 0.8$	$\mu_1 = 4.9 \pm 0.8$	$\mu_2 = 5.0 \pm 0.9$	$\mu_1 = 4.2 \pm 1.0$ $t(95) = -4.57, p < 0.001, \text{Cohen's } d = 0.9$	$\mu_2 = 5.0 \pm 0.9$
Deployment score	$\mu_1 = 4.9 \pm 1.6$	$\mu_2 = 4.8 \pm 1.6$	$\mu_1 = 4.9 \pm 1.6$	$\mu_2 = 5.3 \pm 1.6$	$\mu_1 = 5.3 \pm 1.6$	$\mu_2 = 5.0 \pm 1.5$	$\mu_1 = 4.8 \pm 1.5$	$\mu_2 = 5.0 \pm 1.5$
Suspicious data or model	9 out of 49	5 out of 51	9 out of 49	7 out of 51	7 out of 51	3 out of 46	5 out of 51	3 out of 46
Suspicious tool	1 out of 49	0 out of 51	1 out of 49	2 out of 51	2 out of 51	1 out of 46	0 out of 51	1 out of 46

Table 2: The results of our preregistered analyses. Each column is a pair of conditions, while each row is an outcome variable. Each cell contains the mean of the outcome variable in that row for one of the conditions being compared in that column (μ_1 and μ_2 are the means of conditions 1 and 2 in the header, with standard deviations). We highlight significant differences in gray and list details of the t-test. Cohen's d values: 0.2–0.5 = small effect size, 0.5–0.8 = medium effect size, > 0.8 = large effect size.

Even though participants were, on average, not very confident that the underlying models were reasonable, few explicitly mentioned suspicions about the dataset or models, or about the interpretability tools. Furthermore, they generally thought that the models were ready for deployment. There are no differences between conditions for these outcome variables.

Factors that Affect Willingness to Deploy

To explain why participants thought, on average, that the underlying models were ready for deployment, we present a selection of themes that emerged from their open-ended responses.

Intuition. Most participants gave the models high deployment ratings based on intuition, driven by their prior experience with ML, rather than careful consideration of the explanations:

“I think it'll be good to test this model in practice. The numbers [for performance metrics] seem good, and based on my experience with such numbers, I would deploy it and see if it works.” (P102, SHAP-Normal)

A subset of these participants also said that they would attempt to convince a customer that this was the right model to deploy by simply asking the customer to trust their judgment.

Superficial Evaluation of Explanations. Instead of critically evaluating the explanations, some participants took the visualizations at face value, using their existence to convince themselves that the underlying models were ready for deployment.

Participants in all conditions mentioned things like, “the results of the comprehensive chart display make it easy to make effective judgments” (P28, GAM-Normal), and crafted narratives to convince themselves about the reasonableness of the models: “The charts in combination help you infer reasonable things about the model. Person has college level education, working in private sector full-time, having married to civilian spouse and white race indicates high income which makes actual sense” (P148, SHAP-Manipulated). They relied on these narratives, along with the visualizations from the tools, when explaining how they would attempt to convince a customer that the model was ready for deployment. There was an element of only superficial evaluation to these responses:

“Considering plot A, the top features are reasonable, and the model does not seem to be very much impacted by ethnicity or gender bias. Plot B looks reasonable, since it provides a bonus to individuals in working age range and penalizes others. It also correctly considers widowed people to be more likely to earn more, since they are more likely to inherit assets.” (P137, SHAP-Normal)

Perceived Suspicions. Not all participants gave the underlying models high deployment ratings: 14.2% were neutral and 18.3% gave low ratings (see the figure in the supplementary material). One reason for these ratings was that participants were suspicious about the models, and felt that they could be biased in several ways: “what the heck is happening with the 37/38 year olds?” (P50, GAM-Manipulated); “Marital status as the top-most predictor of income? Should we approve loans to married people and not single people (or vice-versa?)” (P98, SHAP-Normal). The other reason was that participants were unsure about deployment without running more comprehensive tests: “No R-square value, no confidence interval, no overall test score. Far away from deployment” (P19, GAM-Normal). These participants were the ones who used the interpretability tools in their intended ways: to investigate the datasets and underlying models, uncovering issues that required deeper investigation. When asked how they would convince a customer to deploy the model (or not, as appropriate), these participants tended to argue against deployment.

Mental Models of Interpretability Tools

The themes above make it clear that participants, for the most part, did not use the interpretability tools as intended. The HCI literature refers to this kind of behavior as a mismatch between the participants' mental models of the tools and the conceptual models of the tools. A mental model is based on someone's perceptions of a tool, while the conceptual model is the intended use that the tool's designer had in mind [51].

Qualitative analysis of participants' descriptions of the visualizations indicates that most participants did not have an accurate understanding of the visualizations. For all three types of visualizations, one of the axes represents a score (e.g., SHAP's importance score), and is titled as such (“Score” for GAMs and

“Shap value” for SHAP). Three authors iteratively coded participants’ open-ended responses until complete agreement was achieved. A response was considered to be “accurate” when the values represented were explained (e.g., a clear description of what “Score” represents), “partially accurate” when the description was accurate but incomplete, and “superficially accurate” when the axis title was used as-is as the description.

Only 5.6% of participants were able to accurately describe the score axis for local explanations, 1.9% for components (GAMs) or dependence plots (SHAP), and 1.9% for the global explanations. Another small percentage of participants (16.4% for local, 7.5% for components of dependence plots, and 1.9% for global) provided partially accurate descriptions of these axes, giving a general outline of what they represent, but no details. A large percentage of participants (23.4% for local, 43% for components or dependence plots, 48.2% for global) indicated only a superficial understanding of the axes. Furthermore, the largest percentage of participants did not understand the visualizations at all (54.6% for local, 47.6% for components or dependence plots, and 48% for global). These participants often suggested that the scores represented the data points’ labels or the underlying models’ predictions.

These results indicate that participants did not fully understand the visualizations output by the interpretability tools. However, despite this, they had high expectations for these visualizations, above and beyond the tools’ capabilities. When asked to explain how they would use these visualizations in their typical ML pipeline, participants listed uses that ranged from actual capabilities of these visualizations (e.g., understanding the underlying model and its most important features, understanding how a prediction was made) to uses that no interpretability tool could currently provide (e.g., automated checking for societal biases in the dataset or unfair outcomes). Figure 2 depicts the percentage of participants that selected each of the options provided when asked about the interpretability tools’ capabilities.

Tension between Cognitive and Social Factors

Our survey captured contextual information about both cognitive factors (e.g., prior experience with ML, familiarity with interpretability) and social factors (e.g., confidence ratings for participants’ understanding of the visualizations, the reasonableness of the explanations, the reasonableness of the underlying models). Below we discuss how these factors affected participants’ use of the tools and their deployment ratings for the underlying models. These are exploratory analyses; although we report p-values, we did not preregister these analyses.

Accuracy at Reading Visualizations. To explore the relationship between participants’ prior experience with ML and their accuracy at reading the interpretability tools’ visualizations, we fit a multiple linear regression (MLR) using participants’ accuracy on the four multiple-choice questions with ground-truth correct answers as the dependent variable and the following independent variables: (1) how long they had been in their current role; (2) the extent to which ML was a part of their day-to-day work; (3) how long they had been using ML; (4) their familiarity with interpretability, and with GAMs or SHAP; and (5) their familiarity with the Adult Income dataset. The second and third of these independent variables signifi-

cantly predicted participants’ accuracy at reading the visualizations ($b = 5.77$, $t(189) = 3.44$, $p \ll 0.001$ and $b = 0.39$, $t(189) = 5.04$, $p \ll 0.001$, respectively, where b is the corresponding coefficient). The MLR was effective at predicting participants’ accuracy (adjusted $R^2 = 0.27$, $F(7, 189) = 11.62$, $p \ll 0.001$; $R > 0.5$ represents a large effect size).

We used Pearson correlation coefficients to explore the relationship between social factors and participants’ accuracy at reading the interpretability tools’ visualizations because this relationship is more symmetric. Of the three questions about confidence ratings, only participants’ stated confidence in their understanding of the visualizations is strongly correlated with their accuracy (Pearson’s $r(195) = 0.49$, $p \ll 0.001$). This result confirms that participants’ confidence ratings were high when they were accurately reading the visualizations.

Deployment Ratings. We fit an MLR using participants’ deployment ratings as the dependent variable and the following independent variables: the cognitive factors, the social factors, and participants’ accuracy on the four multiple choice questions with ground-truth correct answers. There are several significant predictors for participants’ deployment ratings: how long they had been using ML ($b = -0.02$, $t(184) = -3.75$, $p < 0.001$), their stated confidence that the explanations were reasonable ($b = 0.37$, $t(184) = 3.02$, $p < 0.01$), their stated confidence that the underlying models were reasonable ($b = 0.23$, $t(184) = 2.43$, $p < 0.05$), and their accuracy at reading the visualizations ($b = -0.01$, $t(184) = -2.61$, $p < 0.01$). More ML experience and higher accuracy at reading the visualizations have a negative effect on deployment ratings, whereas higher confidence ratings for the reasonableness of the explanations and the underlying models have a positive effect. The MLR was accurate at predicting deployment ratings (adjusted $R^2 = 0.37$, $F(12, 184) = 10.8$, $p \ll 0.001$).

These results suggest an inverse relationship between cognitive and social factors: participants with more ML experience had higher accuracy at reading the tools’ visualizations, but lower confidence ratings for the reasonableness of the explanations and the underlying models, and thus, lower deployment ratings. We confirmed this relationship using Pearson correlation coefficients and found them to match our expectations. Participants’ ML experience and their stated confidence that the explanations were reasonable are strongly negatively correlated (Pearson’s $r(195) = -0.17$, $p < 0.01$), as are their ML experience and their stated confidence that the underlying models were reasonable (Pearson’s $r(195) = -0.27$, $p < 0.001$).

The Role of Mental Models. Mental models play a crucial role in this inverse relationship between cognitive and social factors. The tension between these types of factors reflects a complicated relationship between two outlooks on the use of interpretability tools. When participants were able to form (partially) accurate mental models of the tools, they evaluated them in a more principled way, and therefore made careful decisions. For example, P88, who had only two months of ML experience, noted, “the spike in age around 35–39 worries me because it seems more representative of a boom that describes a very specific group of people. The model doesn’t account for the fact that that group of people will age past 35–39.”

In contrast, without (partially) accurate mental models of the tools, even the most experienced participants “don’t see any red flags as confirmed by the explanations” (P57, GAM-Manipulated, 4 years of ML experience). Worse, in some cases, their prior experience with ML led them to rely on their intuition and only superficially evaluate the explanations. Without accurate mental models, social factors can rationalize suspicious observations, leading to higher deployment ratings.

DISCUSSION AND FUTURE WORK

Bridging the Gap Between the ML and HCI Communities

To date, research on interpretability in the ML and HCI communities has evolved somewhat independently [66]. Our results highlight the value of user studies for evaluating interpretability techniques from the ML community with stakeholders, marrying the goals and methods of both communities. Indeed, we previously argued that user studies of interpretability require qualitative methods to understand the nuances of how tools are used in context, coupled with quantitative methods to scale up findings. One of our findings is that data scientists with different amounts of ML experience are unable to fully understand the visualizations output by two existing interpretability tools, in turn hindering their ability to understand the dataset and underlying models. Overcoming this challenge will require expertise in the mathematics underlying ML models and in communicating information to users (e.g., the design of tutorials, visualizations, or interactive tools). Ideally, members of the HCI and ML communities should work together from the start, with HCI methodologies applied at all stages of interpretability tool development: supporting need-finding studies (e.g., [8, 64]), designing tools that can be understood by users with different background (e.g., [37]), and undertaking user studies at each stage of tool development (e.g., [67]).

Designing Interactive Interpretability Tools

Interpretability is typically viewed as being unidirectional, with tools providing information to user. However, it may be better to design interpretability tools that allow back-and-forth communication [5]. As one of our participants said, “These explorations are like goal-based communication. If I go in without a hypothesis, it’s hard to evaluate what the tool tells me. When I do make an evaluation [based on the tool], can the tool follow up?” (P4, Contextual Inquiry). In essence, this participant was looking for interactivity from the interpretability tool. Social science and HCI research consider this kind of back-and-forth to be a key factor in making explanations accessible to people with different levels of expertise [23, 48]. Weld and Bansal [65] propose interactive interpretability tools that allow users to dig deeper into explanations or to compare explanations from multiple different interpretability techniques. One might also imagine a tool that could update its mode of interactivity based on users’ perceptions [40]. More generally, interpretability tools should be designed to adapt to users’ expectations. We describe one such design implication below.

Designing Tools for Deliberative Reasoning

Interpretability tools are designed to help stakeholders better understand how ML models work. However, as we found, these tools’ visualizations can encourage people to make quick

decisions instead of digging deeper. As P4 from our contextual inquiry said, “There is this concept in UX called thinking fast and slow. While these visualizations are made to make me think fast, every detail about them requires that I think slow.” This sentiment echoes Kahneman’s [27, 28] cognitive processes for humans: system 1, which tries to make quick, automatic decisions based on heuristics, and system 2, which performs deliberative reasoning and engages more deeply before making decisions. People are prone to make decisions using system 1, unless system 2 is engaged, because of missing heuristics. Designing interpretability tools so that they activate system 2 is therefore an important avenue for future work.

LIMITATIONS

Our work has several limitations. First, although we tried to put data scientists in a realistic setting via Jupyter notebooks in our contextual inquiry and via visualizations and the results of common exploration commands in our survey, we cannot be certain that this was sufficient. Second, longitudinal studies might reveal different or more nuanced patterns of behavior than either our contextual inquiry or our survey. Third, we found it challenging to distinguish between participants’ high-level understanding of the tools’ visualizations and participants’ deeper understanding of the importance scores shown. Indeed, research on mental models commonly faces this challenge. We therefore relied on our qualitative findings to support and add nuance to our quantitative results. Fourth, we used a tabular dataset. Although there is research on interpretability techniques for deep learning and richer types of data (e.g., images [32, 57]), this was not the focus of our work.

CONCLUSION

We study data scientists’ use of two existing interpretability tools: the InterpretML implementation of GAMs (glassbox models) and the SHAP Python package (a post-hoc explanation technique for blackbox models). We conduct pilot interviews ($N = 6$) to identify common issues faced by data scientists in their day-to-day work, a contextual inquiry ($N = 11$) to observe how data scientists use interpretability tools to uncover these issues, and a survey ($N = 197$) to scale up and quantify the main findings from our contextual inquiry. Our results indicate that the visualizations output by interpretability tools can sometimes help data scientists to uncover issues with datasets or models. However, for both tools, the existence of visualizations and the fact that the tools were publicly available led to cases of over-trust and misuse. Finally, we end by highlighting the need for members of the HCI and ML communities to work together, and discussing avenues for future exploration.

ACKNOWLEDGMENTS

We would like to thank the reviewers for their helpful comments, which improved the paper. We are also grateful to Scott Lundberg, Michael Madaio, Forough Poursabzi-Sangdeh, and especially Abigail Sellen for valuable feedback. This work was begun while the first author was an intern at Microsoft.

REFERENCES

- [1] Ashraf Abdul, Jo Vermeulen, Danding Wang, Brian Y. Lim, and Mohan Kankanhalli. 2018. Trends and

- Trajectories for Explainable, Accountable and Intelligent Systems: An HCI Research Agenda. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (CHI '18)*. ACM, New York, NY, USA, Article 582, 18 pages. DOI: <http://dx.doi.org/10.1145/3173574.3174156>
- [2] David Alvarez-Melis, Hal Daumé, III, Jennifer Wortman Vaughan, and Hanna Wallach. 2019. Weight of Evidence as a Basis for Human-Oriented Explanations. *arXiv preprint arXiv:1910.13503* (2019).
- [3] Saleema Amershi, James Fogarty, and Daniel Weld. 2012. Regroup: Interactive Machine Learning for On-demand Group Creation in Social Networks. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '12)*. ACM, New York, NY, USA, 21–30. DOI: <http://dx.doi.org/10.1145/2207676.2207680>
- [4] Julia Angwin, Jeff Larson, Surya Mattu, and Kirchner Lauren. 2016. Machine Bias: There's software used across the country to predict future criminals. And it's biased against blacks. *ProPublica*, May 23 (2016), 2016. <http://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing>
- [5] Dean C Barnlund. 2017. A transactional model of communication. In *Communication theory, Second edition*, C. David Mortensen (Ed.). Routledge, 47–57.
- [6] Victoria Bellotti and Keith Edwards. 2001. Intelligibility and Accountability: Human Considerations in Context-Aware Systems. *Human-Computer Interaction* 16, 2-4 (2001), 193–212. DOI: http://dx.doi.org/10.1207/S15327051HCI16234_05
- [7] Virginia Braun and Victoria Clarke. 2012. Thematic analysis. In *APA handbook of research methods in psychology, Vol 2: Research designs: Quantitative, qualitative, neuropsychological, and biological*. American Psychological Association, Washington, DC, US, 57–71. DOI: <http://dx.doi.org/10.1037/13620-004>
- [8] Carrie J. Cai, Samantha Winter, David Steiner, Lauren Wilcox, and Michael Terry. 2019. "Hello AI": Uncovering the Onboarding Needs of Medical Practitioners for Human-AI Collaborative Decision-Making. *Proc. ACM Hum.-Comput. Interact.* 3, CSCW, Article 104 (Nov. 2019), 24 pages. DOI: <http://dx.doi.org/10.1145/3359206>
- [9] Aylin Caliskan, Joanna J. Bryson, and Arvind Narayanan. 2017. Semantics derived automatically from language corpora contain human-like biases. *Science* 356, 6334 (2017), 183–186. DOI: <http://dx.doi.org/10.1126/science.aal4230>
- [10] 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 (KDD '15)*. ACM, New York, NY, USA, 1721–1730. DOI: <http://dx.doi.org/10.1145/2783258.2788613>
- [11] 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.
- [12] Juliet M. Corbin and Anselm Strauss. 1990. Grounded theory research: Procedures, canons, and evaluative criteria. *Qualitative sociology* 13, 1 (1990), 3–21. DOI: <http://dx.doi.org/10.1007/BF00988593>
- [13] Janez Demšar, Blaž Zupan, Gregor Leban, and Tomaz Curk. 2004. Orange: From Experimental Machine Learning to Interactive Data Mining. In *Knowledge Discovery in Databases: PKDD 2004*, Jean-François Boulicaut, Floriana Esposito, Fosca Giannotti, and Dino Pedreschi (Eds.). Springer Berlin Heidelberg, Berlin, Heidelberg, 537–539. DOI: http://dx.doi.org/10.1007/978-3-540-30116-5_58
- [14] Finale Doshi-Velez and Been Kim. 2017. Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608* (2017).
- [15] 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).
- [16] Paul Dourish. 2016. Algorithms and their others: Algorithmic culture in context. *Big Data & Society* 3, 2 (2016), 2053951716665128. DOI: <http://dx.doi.org/10.1177/2053951716665128>
- [17] Mary T Dzindolet, Hall P Beck, Linda G Pierce, and Lloyd A Dawe. 2001. *A framework of automation use*. Technical Report. Army Research Lab Aberdeen Proving Ground MD.
- [18] Jerry Alan Fails and Dan R. Olsen, Jr. 2003. Interactive Machine Learning. In *Proceedings of the 8th International Conference on Intelligent User Interfaces (IUI '03)*. ACM, New York, NY, USA, 39–45. DOI: <http://dx.doi.org/10.1145/604045.604056>
- [19] 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. DOI: <http://dx.doi.org/10.1109/dsaa.2018.00018>
- [20] Herbert P. Grice. 1975. Logic and Conversation. (1975), 41–58. DOI: http://dx.doi.org/10.1163/9789004368811_003
- [21] Sandra G. Hart and Lowell E. Staveland. 1988. Development of NASA-TLX (Task Load Index): Results of Empirical and Theoretical Research. In *Human Mental Workload*, Peter A. Hancock and Najmedin Meshkati (Eds.). Advances in Psychology, Vol. 52. North-Holland, 139–183. DOI: [http://dx.doi.org/10.1016/S0166-4115\(08\)62386-9](http://dx.doi.org/10.1016/S0166-4115(08)62386-9)

- [22] Trevor Hastie and Robert Tibshirani. 1987. Generalized Additive Models: Some Applications. *J. Amer. Statist. Assoc.* 82, 398 (1987), 371–386. DOI: <http://dx.doi.org/10.1080/01621459.1987.10478440>
- [23] 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 (CHI '19)*. ACM, New York, NY, USA, Article 579, 13 pages. DOI: <http://dx.doi.org/10.1145/3290605.3300809>
- [24] Jiun-Yin Jian, Ann M. Bisantz, and Colin G. Drury. 2000. Foundations for an Empirically Determined Scale of Trust in Automated Systems. *International Journal of Cognitive Ergonomics* 4, 1 (2000), 53–71. DOI: http://dx.doi.org/10.1207/S15327566IJCE0401_04
- [25] Jongbin Jung, Connor Concannon, Ravi Shroff, Sharad Goel, and Daniel G Goldstein. 2017. Simple rules for complex decisions. Available at SSRN 2919024 (2017). <http://dx.doi.org/10.2139/ssrn.2919024>
- [26] Mayank Kabra, Alice A Robie, Marta Rivera-Alba, Steven Branson, and Kristin Branson. 2013. JAABA: interactive machine learning for automatic annotation of animal behavior. *Nature methods* 10, 1 (2013), 64–67. DOI: <http://dx.doi.org/10.1038/nmeth.2281>
- [27] Daniel Kahneman. 2011. *Thinking, fast and slow*. Macmillan.
- [28] Daniel Kahneman, Stewart Paul Slovic, Paul Slovic, and Amos Tversky. 1982. *Judgment under uncertainty: Heuristics and biases*. Cambridge university press.
- [29] Matthew Kay, Cynthia Matuszek, and Sean A. Munson. 2015. Unequal Representation and Gender Stereotypes in Image Search Results for Occupations. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems (CHI '15)*. ACM, New York, NY, USA, 3819–3828. DOI: <http://dx.doi.org/10.1145/2702123.2702520>
- [30] 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. 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., 3146–3154. <http://papers.nips.cc/paper/6907-lightgbm-a-highly-efficient-gradient-boosting-decision-tree.pdf>
- [31] 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>
- [32] Been Kim, Martin Wattenberg, Justin Gilmer, Carrie Cai, James Wexler, Fernanda Viegas, and Rory Sayres. 2018. Interpretability Beyond Feature Attribution: Quantitative Testing with Concept Activation Vectors (TCAV). In *Proceedings of the 35th International Conference on Machine Learning (Proceedings of Machine Learning Research)*, Jennifer Dy and Andreas Krause (Eds.), Vol. 80. PMLR, Stockholmsmässan, Stockholm Sweden, 2668–2677. <http://proceedings.mlr.press/v80/kim18d.html>
- [33] 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 (CHI '19)*. ACM, New York, NY, USA, Article 411, 14 pages. DOI: <http://dx.doi.org/10.1145/3290605.3300641>
- [34] 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 (IUI '15)*. ACM, New York, NY, USA, 126–137. DOI: <http://dx.doi.org/10.1145/2678025.2701399>
- [35] Todd Kulesza, Simone Stumpf, Margaret Burnett, and Irwin Kwan. 2012. Tell Me More?: The Effects of Mental Model Soundness on Personalizing an Intelligent Agent. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '12)*. ACM, New York, NY, USA, 1–10. DOI: <http://dx.doi.org/10.1145/2207676.2207678>
- [36] 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)*.
- [37] Himabindu Lakkaraju and Osbert Bastani. 2020. "How do I fool you?": Manipulating User Trust via Misleading Black Box Explanations. In *Proceedings of the 2020 AAAI/ACM Conference on AI, Ethics, and Society (AIES '20)*. ACM, New York, NY, USA.
- [38] David B. Leake. 1991. Goal-based explanation evaluation. *Cognitive Science* 15, 4 (1991), 509–545. DOI: [http://dx.doi.org/10.1016/0364-0213\(91\)80017-Y](http://dx.doi.org/10.1016/0364-0213(91)80017-Y)
- [39] Zachary C Lipton. 2016. The mythos of model interpretability. *arXiv preprint arXiv:1606.03490* (2016).
- [40] Stine Lomborg and Patrick Heiberg Kapsch. 2019. Decoding algorithms. *Media, Culture & Society* (2019). DOI: <http://dx.doi.org/10.1177/0163443719855301>
- [41] Tania Lombrozo. 2006. The structure and function of explanations. *Trends in Cognitive Sciences* 10, 10 (2006), 464–470. DOI: <http://dx.doi.org/10.1016/j.tics.2006.08.004>

- [42] Alexandra L'heureux, Katarina Grolinger, Hany F Elyamany, and Miriam AM Capretz. 2017. Machine learning with big data: Challenges and approaches. *IEEE Access* 5 (2017), 7776–7797. DOI: <http://dx.doi.org/10.1109/ACCESS.2017.2696365>
- [43] Scott M Lundberg, Gabriel G Erion, and Su-In Lee. 2018. Consistent individualized feature attribution for tree ensembles. *arXiv preprint arXiv:1802.03888* (2018).
- [44] 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>
- [45] 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>
- [46] Bertram F Malle. 2006. *How the mind explains behavior: Folk explanations, meaning, and social interaction*. Mit Press.
- [47] Tim Miller. 2018. Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence* (2018).
- [48] 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)*.
- [49] Michael Muller, Ingrid Lange, Dakuo Wang, David Piorkowski, Jason Tsay, Q. Vera Liao, Casey Dugan, and Thomas Erickson. 2019. How Data Science Workers Work with Data: Discovery, Capture, Curation, Design, Creation. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (CHI '19)*. ACM, New York, NY, USA, Article 126, 15 pages. DOI: <http://dx.doi.org/10.1145/3290605.3300356>
- [50] Harsha Nori, Samuel Jenkins, Paul Koch, and Rich Caruana. 2019. InterpretML: A Unified Framework for Machine Learning Interpretability. *arXiv preprint arXiv:1909.09223* (2019).
- [51] Donald A Norman. 2014. Some observations on mental models. In *Mental models*. Psychology Press, 15–22.
- [52] Raja Parasuraman, Robert Molloy, and Indramani L. Singh. 1993. Performance Consequences of Automation-Induced 'Complacency'. *The International Journal of Aviation Psychology* 3, 1 (1993), 1–23. DOI: http://dx.doi.org/10.1207/s15327108i_jap0301_1
- [53] Kayur Patel, James Fogarty, James A. Landay, and Beverly Harrison. 2008. Investigating Statistical Machine Learning As a Tool for Software Development. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '08)*. ACM, New York, NY, USA, 667–676. DOI: <http://dx.doi.org/10.1145/1357054.1357160>
- [54] Forough Poursabzi-Sangdeh, Daniel G Goldstein, Jake M Hofman, Jennifer Wortman Vaughan, and Hanna Wallach. 2018. Manipulating and measuring model interpretability. *arXiv preprint arXiv:1802.07810* (2018).
- [55] 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 (KDD '16)*. ACM, New York, NY, USA, 1135–1144. DOI: <http://dx.doi.org/10.1145/2939672.2939778>
- [56] 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. DOI: <http://dx.doi.org/10.1038/s42256-019-0048-x>
- [57] 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.
- [58] Lloyd S Shapley. 1997. A value for n-person games. *Classics in game theory* (1997), 69.
- [59] 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. DOI: <http://dx.doi.org/10.1002/ejsp.2420230302>
- [60] Simone Stumpf, Vidya Rajaram, Lida Li, Weng-Keen Wong, Margaret Burnett, Thomas Dietterich, Erin Sullivan, and Jonathan Herlocker. 2009. Interacting meaningfully with machine learning systems: Three experiments. *International Journal of Human-Computer Studies* 67, 8 (2009), 639–662. DOI: <http://dx.doi.org/10.1016/j.ijhcs.2009.03.004>
- [61] Sarah Tan, Rich Caruana, Giles Hooker, Paul Koch, and Albert Gordo. 2018. Learning global additive explanations for neural nets using model distillation. *arXiv preprint arXiv:1801.08640* (2018).
- [62] Richard Tomsett, Dave Braines, Dan Harborne, Alun Preece, and Supriyo Chakraborty. 2018. Interpretable to whom? A role-based model for analyzing interpretable machine learning systems. *arXiv preprint arXiv:1806.07552* (2018).

- [63] Byron C. Wallace, Kevin Small, Carla E. Brodley, Joseph Lau, and Thomas A. Trikalinos. 2012. Deploying an Interactive Machine Learning System in an Evidence-based Practice Center: Abstrackr. In *Proceedings of the 2Nd ACM SIGHIT International Health Informatics Symposium (IHI '12)*. ACM, New York, NY, USA, 819–824. DOI : <http://dx.doi.org/10.1145/2110363.2110464>
- [64] 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 (CHI '19)*. ACM, New York, NY, USA, Article 601, 15 pages. DOI : <http://dx.doi.org/10.1145/3290605.3300831>
- [65] Daniel S. Weld and Gagan Bansal. 2019. The Challenge of Crafting Intelligible Intelligence. *Commun. ACM* 62, 6 (May 2019), 70–79. DOI : <http://dx.doi.org/10.1145/3282486>
- [66] Qian Yang, Nikola Banovic, and John Zimmerman. 2018a. Mapping Machine Learning Advances from HCI Research to Reveal Starting Places for Design Innovation. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (CHI '18)*. ACM, New York, NY, USA, Article 130, 11 pages. DOI : <http://dx.doi.org/10.1145/3173574.3173704>
- [67] Qian Yang, Alex Scuito, John Zimmerman, Jodi Forlizzi, and Aaron Steinfeld. 2018b. Investigating How Experienced UX Designers Effectively Work with Machine Learning. In *Proceedings of the 2018 Designing Interactive Systems Conference (DIS '18)*. ACM, New York, NY, USA, 585–596. DOI : <http://dx.doi.org/10.1145/3196709.3196730>
- [68] 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. DOI : <http://dx.doi.org/10.1111/rssa.12227>
- [69] Jichen Zhu, Antonios Liapis, Sebastian Risi, Rafael Bidarra, and G. Michael Youngblood. 2018. Explainable AI for designers: A human-centered perspective on mixed-initiative co-creation. In *2018 IEEE Conference on Computational Intelligence and Games (CIG)*. IEEE, 1–8. DOI : <http://dx.doi.org/10.1109/CIG.2018.8490433>