

# **Where are the Humans in Human-AI Interaction: The Missing Human-Centered Perspective on Interpretability Tools for Machine Learning**

by

Harmanpreet Kaur

A dissertation submitted in partial fulfillment  
of the requirements for the degree of  
Doctor of Philosophy  
(Information and Computer Science and Engineering)  
in The University of Michigan  
2023

## Doctoral Committee:

Associate Professor Eric Gilbert, Co-Chair  
Professor Cliff Lampe, Co-Chair  
Professor Mark Ackerman  
Associate Professor Eytan Adar  
Dr. Shamsi Iqbal, Microsoft Research  
Dr. Jennifer Wortman Vaughan, Microsoft Research

*In a world of diminishing mystery, the unknown persists.*

—Jhumpa Lahiri, *The Lowland*

Harmanpreet Kaur  
harmank@umich.edu  
ORCID iD: 0009-0009-8239-937X

© Harmanpreet Kaur 2023

To Mom, Dad, Avleen, & my grandparents

## ACKNOWLEDGEMENTS

First and foremost, I would like to thank my advisors Eric Gilbert and Cliff Lampe. There are no words for how grateful I am for your support and encouragement. Whatever I know about being a researcher, mentor, and an academic, I have learnt from the two of you. Cliff, thank you for completely changing my worldview in the best possible way. I would not have discovered my passion for (obscure) academic reading and making cross-disciplinary connections had you not handed me “Sensemaking in Organizations” during my first semester, and encouraged me to take classes in other departments. Eric, thank you for your kindness and unwavering belief that I could do this—there were many times when your belief was much stronger than mine and kept me going. You have been the best example of how to do an academic job in a balanced, happy way. I hope I can emulate the two of you as advisors, and come even slightly close to having the kind of impact that you have had on my research and future. Our weekly meetings were my favorite thing on my schedule and will be sorely missed.

I would also like to thank my dissertation committee, Mark Ackerman, Eytan Adar, Shamsi Iqbal, and Jenn Wortman Vaughan. Mark, your “social-technical gap” paper had such a significant impact on how I thought about my research area. I never imagined that I would get to talk to you about the paper or about my own work; it was truly a privilege to have your guidance and perspective. Eytan, thank you for sharing all your wisdom on finding the “right” projects, and for pushing me to think about the implications of my theoretical work. I so appreciate the “what would Eytan ask about this theory?” question in my head as I attempt more of this translation work. Shamsi, I first became a researcher under your guidance, dur-

ing the internships early on in my PhD. Thank you for taking a chance on a first year student and teaching me everything I needed to know to be successful at the job. Thank you, also, for being there for every single moment of my journey since: celebrating the good, cheering me on, and supporting me through the rough patches. Jenn, again, talk about being lucky with the internship mentors I have had. I owe my choice of dissertation topic to you. You not only shaped my thinking on interpretability, but also helped me see the kind of contribution I could make in this field. Thank you for the incredible professional and personal guidance.

I have been extremely fortunate to have several amazing mentors throughout my research journey. Jaime Teevan, who, along with Shamsi, was there for all the ups and downs since that first summer internship. Jaime pushed me to think outside the box and has been an exemplary role model of how to be a badass woman in tech. Nicole Ellison and Sarita Schoenebeck, who were both there with their support and advise, and offered me safe havens when I needed them most. I also cherish the time spent with Nicole's pets (Coco, Sasha, Mr. Friendly, and Theo)—being around them was the best way to forget all my problems. I am also grateful for having Hanna Wallach as a mentor, not only for all the things I learnt from her given the amazing researcher that she is, but also for the fun NYC times, cute pet pictures, and our shared interest in finding the best buffalo chicken and babka in NYC. I am thankful for the collaboration opportunities with Dan Weld, Doug Downey, Jonathan Bragg, Harsha Nori, Rich Caruana, Samuel Jenkins, David Alvarez-Melis, Hal Daumé III, Daniel McDuff, and Mary Czerwinski, which influenced my research in significant ways. Of course, I cannot forget where this all started: being a part of the GroupLens Research Lab at the University of Minnesota as an undergrad. Loren Terveen, Joe Konstan, Brent Hecht, and Lana Yarosh shaped my understanding of HCI research, and have continued to be mentors throughout this journey. I am beyond excited to be going back to GroupLens for my next academic adventure.

I could not have done this without the support system I had in my friends, both near and far. There are no words that can adequately express how thankful I am for my friendships with Megan Shearer, Preeti Ramaraj, and Jaylin Herskovitz—my absolute essentials in Ann

Ann Arbor. Megan is a force to be reckoned with, and one that has always been firmly on my side. I can also count on her for adventures, fun, and carefree times. Our long walks, cooking sessions, trips and future planning, and shared passions about all things in life, kept me sane. Preeti is one of the kindest souls I know, an optimist at heart, trying to make the world a better place. She is fiercely protective of her friends (I would know). She was also my go-to person for passionate discussions on our shared research interests, reminiscing about all things India, and dosas. Jaylin has seen me through the many ups and downs of grad school. I met her very early into my time in Ann Arbor. Lucky, because you will be hard-pressed to find a better cheerleader, listener, or friend. Our adventures in Seattle, Montreal, and Scotland are some of my happiest grad school memories. I was also lucky to have fantastic labmates like Laura Kurek, Qiwei Li, Nasanbayar Ulzii-Orshikh, Josh Ashkinaze, Elizabeth Whittaker, Song Mi Lee, Han Na Shin, and Rebecca Krosnick, who made Ann Arbor and grad school fun. Thank you, all of you, for helping me see this through.

Despite being in different cities, Stevie Chancellor was always there for me when I needed a friend, a research buddy, a brainstorming partner, or a mentor. Visiting her felt like having a home away from home, as did being around Hannah Miller, Sarah Lewis, and their families. Thank you for the friendship, love, and support, always. I am also beyond grateful for my friendships with Mitchell Gordon, Michael Madaio, Lindsay Blackwell, Alex Williams, Lisa Elkin, and Joseph Seering. You all represent some of the best parts of grad school. My friendships with Neera Munjal, Vivek Nair, Gurpreet Singh, Zahaib Mateen, VS Karthick, and Lakshita Mehrotra began back home in India, but have been close to my heart all this time, all these miles away. Their love and support follows me wherever I go, for which I am immensely thankful.

Finally, I want to thank my family for their love and support. My mother, Vimmi Hora, who first encouraged me to apply to universities in the US for undergrad, and then to follow my PhD dreams. Thank you for your constant reassurances that I could do this, for taking my calls day or night (despite the time difference), and for doing all that you do to take care

of me. My father, Balbir Singh, who has always been the driving force behind my passion for learning. Thank you for encouraging me to explore and ask questions, in everyday life and in my work. My sister, Avleen Kaur, who has this innate ability for knowing exactly when I need her. With all our fights as kids, I could never have imagined the closeness we share now. Thank you for being there for me, every single step of the way, as my biggest cheerleader and best friend. My grandfather, Manohar Singh Hora, whom I deeply wish I could have shared this with. He was the first to encourage me to chase my dreams, and the happiest to see me do it even if it meant I was far away from him. My aunt, Priyanka Arora, who always goes above and beyond in making sure I am happy and cared for. Thank you, maasi, for always wishing the best for me and being one of my biggest supporters. I am also immensely grateful to my grandparents and the rest of my family for their love and blessings. Last, but certainly not the least, I am so incredibly thankful for our family's crew of crazy pets, Saaya, Rhys, Biscuit, and Bruno, who made everything better, every day. Thank you for all the joy you have brought in my life. I am who I am because of this wonderful family.



# TABLE OF CONTENTS

<b>DEDICATION</b> . . . . .	ii
<b>ACKNOWLEDGEMENTS</b> . . . . .	iii
<b>LIST OF FIGURES</b> . . . . .	xi
<b>LIST OF TABLES</b> . . . . .	xiv
<b>LIST OF APPENDICES</b> . . . . .	xvi
<b>ABSTRACT</b> . . . . .	xvii
<b>CHAPTER</b>	
<b>I. Introduction</b> . . . . .	1
1.1 The Human in Human-Machine Collaboration . . . . .	2
1.1.1 The “Why”: Why should we care about helping people understand ML? . . . . .	3
1.1.2 The “What”: What can we do to help people understand ML? . . . . .	4
1.1.3 The “How”: How should we design ways to help people understand ML? . . . . .	5
1.1.4 The Vital “Who”: Do these solutions work for who they are intended for? . . . . .	6
1.2 Interpreting Interpretability: Understanding Practitioners’ Use of Interpretability Tools for Machine Learning . . . . .	7
1.3 Interpretability Gone Bad: The Role of Bounded Rationality in How Practitioners Understand Machine Learning . . . . .	8
1.4 Sensible AI: Re-Imagining Interpretability and Explainability using Sensemaking Theory . . . . .	9
1.5 Contributions of this Dissertation . . . . .	10
1.5.1 Summary of Contributions . . . . .	12
<b>II. Background</b> . . . . .	13

2.1	The “What”, “Why”, and “How” of Interpretability and Explainability	14
2.1.1	Machine Learning Approaches	14
2.1.2	Cognitive Science, Philosophy, and Social Science Influences	16
2.1.3	HCI Background on Human–Machine Collaboration	18
2.2	Understanding the “Who” in Interpretability and Explainability	20
2.2.1	Cognitive Factors	20
2.2.2	Prior Experience	21
2.2.3	Job- and Task-based Information Needs	22
2.2.4	Social, Organizational, and Socio-Organizational Context	23
2.2.5	Summary	23

**III. Interpreting Interpretability: Understanding Practitioners’ Use of Interpretability Tools for Machine Learning** . . . . . 24

3.1	Pilot Interviews	25
3.2	Contextual Inquiry	26
3.2.1	Dataset	27
3.2.2	ML Models and Interpretability Tools	27
3.2.3	Contextual Inquiry Protocol	29
3.2.4	Participants and Data	30
3.2.5	Results	31
3.3	Large-Scale Survey	34
3.3.1	Experimental Conditions	34
3.3.2	Components of the Survey	34
3.3.3	Participants	36
3.3.4	Preregistration	37
3.3.5	Methods	37
3.3.6	Results	38
3.4	Discussion and Future Work	45
3.4.1	Bridging the Gap Between the ML and HCI Communities	45
3.4.2	Designing Interactive Interpretability Tools	45
3.4.3	Designing Tools for Deliberative Reasoning	46
3.5	Limitations	46
3.6	Conclusion	47

**IV. Interpretability Gone Bad: The Role of Bounded Rationality in How Practitioners Understand Machine Learning** . . . . . 48

4.1	Bounded Rationality	49
4.1.1	Example Setting	50
4.2	Research Goals and Hypotheses	52
4.3	Methods	56

4.3.1	Experimental Setup . . . . .	56
4.3.2	Choice of Dataset . . . . .	57
4.3.3	Choice of Model and Interpretability Tools . . . . .	58
4.3.4	Components of the Survey . . . . .	63
4.3.5	Dependent and Independent Variables . . . . .	68
4.3.6	Analysis Methods . . . . .	68
4.3.7	Participants and Data . . . . .	69
4.4	Results . . . . .	69
4.4.1	Hypothesis Testing . . . . .	69
4.4.2	Perceptions of the Setup . . . . .	73
4.4.3	Exploratory Analyses . . . . .	75
4.4.4	Summary of Results . . . . .	78
4.5	Discussion . . . . .	78
4.5.1	Implications for Design . . . . .	80
4.6	Limitations . . . . .	82
4.7	Conclusion . . . . .	84
<b>V.</b>	<b>Sensible AI: Re-Imagining Interpretability and Explainability using Sensemaking Theory . . . . .</b>	<b>85</b>
5.1	Sensemaking . . . . .	88
5.1.1	Grounded in Identity Construction . . . . .	90
5.1.2	Social . . . . .	93
5.1.3	Retrospective . . . . .	95
5.1.4	Enactive of Sensible Environments . . . . .	98
5.1.5	Ongoing . . . . .	100
5.1.6	Focused on and by Extracted Cues . . . . .	103
5.1.7	Driven by Plausibility rather than Accuracy . . . . .	106
5.1.8	Summary . . . . .	109
5.2	Discussion . . . . .	110
5.2.1	Seamful Design . . . . .	111
5.2.2	Inducing Skepticism . . . . .	112
5.2.3	Adversarial Design . . . . .	113
5.2.4	Continuous Monitoring and Feedback . . . . .	113
5.3	Conclusion . . . . .	114
<b>VI.</b>	<b>Conclusion . . . . .</b>	<b>115</b>
6.1	Designing Inefficiencies for Effective Human-Machine Collaboration . . . . .	117
6.2	Generative AI as a New Interaction Modality . . . . .	119
6.3	Rethinking Evaluation Metrics for Human-AI Partnerships . . . . .	120
6.4	Translating and Applying Relevant Theories . . . . .	121
<b>APPENDICES</b>	<b>. . . . .</b>	<b>122</b>

**BIBLIOGRAPHY** . . . . . 150

## LIST OF FIGURES

### Figure

1.1	Common types of explanations output by interpretability tools. Left: Summary feature importances (global explanation). Middle: Component or partial dependence analysis (global explanation). Right: Feature attributions (local explanation). These particular visualizations are output by the interpretML implementation of a popular type of glassbox models, Generalized Additive Models (GAMs) [167]. The underlying dataset used for these visuals is the Adult Income dataset compiled over 1994 census data. It predicts income based on demographic features. . . . .	4
1.2	Explainable AI, as conceptualized in a DARPA program report [72]. . . .	5
3.1	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. . . . .	27
3.2	Percentage of participants that selected each option when asked about the interpretability tools' capabilities. . . . .	36
3.3	Percentage of participants with low, neutral, and high deployment scores per condition, and a total percentage per score type across all participants. . . . .	40
4.1	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): Overall feature importances; Partial dependence plot for a continuous input feature, age; Partial dependence plot for a categorical input feature, sex (all global explanations). Bottom (left to right): Waterfall plot and Force plot, both types of local explanations for an individual data point. . . . .	51
4.2	Using the framework of bounded rationality to create quadrants for human cognition in Machine Learning settings, defined by the amount of time and cognitive effort spent on the task. The experimental conditions I tested map to three of these quadrants. . . . .	52

4.3 The landing interface of Microsoft’s Explanation Dashboard with minor edits to reduce whitespace. The dashboard includes several exploration sections. Here I 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. . . . . 61

4.4 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 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 this case, I set up the tool to also provide 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). . . . . 62

4.5 Visualizations output by interpretML’s implementation of GAMs, i.e., Explainable Boosting Machines. These ones are generated for the Adult Income dataset, same as the study task. Top (left to right): Overall feature importances (global explanation); Feature attributions for an individual datapoint (local explanation). Bottom: Partial dependence plot for a continuous input feature, age (global explanation). . . . . 67

4.6 Results of my 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. . . . . 71

4.7 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 4.7. . . . . 72

4.8 Descriptive statistics for independent variables with noteworthy differences. 74

4.9 Updated overview of the study conditions using my proposed framework for cognition in Machine Learning settings. The unexpected results are highlighted in color. . . . . 79

5.1 Left: DARPA’s conceptualization of Explainable AI, adapted from [72]. Right: Weick’s sensemaking properties (1–7) categorized using the high-level Enactment-Selection-Retention organizational model, adapted from [103]. Enactment includes properties about perceiving and acting on environmental changes; Selection, properties related to interpreting what the changes mean; and Retention, properties that describe storing and using prior experiences [122]. My proposed Sensible AI framework extends the existing definition of interpretability and explainability to include Weick’s sensemaking properties. . . . . 88

5.2 Saliency maps for chest radiographs, adapted from [48]. . . . . 96

6.1 An initial prototype for FrictionBot, an application that supports more deliberate engagement with ML and interpretability outputs using design guidelines from my Sensible AI framework, powered by LLMs. . . . . 119

F.1 Local explanations included with the multiple-choice question about predicting the outcome given input feature values of a datapoint. These were included by default for the interpretability conditions. These explanations are output by: (a) interpretML’s version of GAMs, called Explainable Boosting Machines (EBMs); (b) SHAP applied to a lightGBM model; (c) Explanation Dashboard, using an EBM; and (d) What-If Tool, using SHAP on a lightGBM model. (a) and (b) are examples of static interpretability tools whereas (c) and (d) are from interactive tools. (a) and (c) are glassbox approaches, and the other two are post-hoc explainers for blackbox models. . . . . 146

F.2 Local explanation for the multiple-choice question about predicting the outcome given input feature values of a datapoint. This chart is output by the XGBoost Explainer and included as a hint for the control condition participants. . . . . 147

G.1 Descriptive statistics for all the independent variables in the study. . . . . 149

## LIST OF TABLES

### Table

3.1	Six themes capturing common issues faced by data scientists. Each issue was synthesized in a dataset for the contextual inquiry, as described in the third column. The fourth column contains the number of participants in the contextual inquiry who identified the corresponding issue. . . . .	26
3.2	The results of the 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 ( $\mu_1$ and $\mu_2$ are the means of conditions 1 and 2 in the header, with standard deviations). Significant differences are highlighted in gray along with 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.	38
4.1	An overview of the ten hypotheses corresponding to my dependent variables. Each dependent variable is split into two hypotheses, one for static visual explanations from interpretability tools and the other for interpretability tools with interactive features. The former represents a satisficing cognition mode and the latter, optimizing. . . . .	55
4.2	Questions included under the <b>first</b> survey component on <b>setup familiarity</b> , and the corresponding independent variables. . . . .	63
4.3	Multiple choice questions included under the <b>second</b> survey component on <b>the data and model</b> , and the corresponding dependent variables. . . .	64
4.4	Questions included under the <b>third</b> survey component on <b>high-level task evaluation</b> , and the corresponding dependent and independent variables. .	64
4.5	Questions included under the <b>fourth and fifth</b> survey components on <b>mental models and setup engagement</b> , and the corresponding independent variables. . . . .	65



4.6 The results of my 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  $\eta^2$  value for effect size; suggested norms: small = 0.01, medium = 0.06, large = 0.14. All of the dependent variables show a large effect size. . . . . 70

4.7 Contingency table with counts for response types—accurate, plausible, and randomly inaccurate—for all five conditions. . . . . 71

5.1 Weick’s basic descriptions of the properties of sensemaking for the human-human context [247, pp.61-62], and my proposed claims from translating these properties to the human-machine context. . . . . 89

5.2 Principles for high-reliability organizations (columns) that inspired my design ideas (rows) under the Sensible AI framework. . . . . 111

## LIST OF APPENDICES

### Appendix

A.	Semi-Structured Interview Protocol for the Pilot Interviews . . . . .	123
B.	Introduction to Generalized Additive Models (GAMs) . . . . .	126
C.	Introduction to SHAP . . . . .	129
D.	Contextual Inquiry Questions about the Dataset and the Model . . . . .	132
E.	Screenshots from a Google Colab Notebook used for the Experiment . . . . .	134
F.	Multiple-Choice Questions about the Dataset and the Model . . . . .	138
G.	Descriptive Statistics for All Independent Variables . . . . .	148

## **ABSTRACT**

This dissertation aims to provide a richer understanding of the extent to which people understand complex AI and ML outputs and reasoning, what influences their understanding, and how we can continue to enhance it going forward. AI- and ML-based systems are now routinely deployed in real-world settings, including sensitive domains like criminal justice, healthcare, finance, and public policy. Given their rapidly growing ubiquity, understanding how AI and ML work is a prerequisite for responsibly designing, deploying, and using these systems. With interpretability and explainability approaches, these systems can offer explanations for their outputs to aid human understanding. Though these approaches rely on guidelines for how humans explain things to each other, they ultimately solve for improving an artifact (e.g., an explanation or explanation system). My dissertation makes the argument that helping people understand AI and ML is as much a human problem as a technical one. Detailing this vital human-centered piece, I present work that shows that current interpretability and explainability tools do not meet their intended goals for a key stakeholder, ML practitioners, who end up either over- or under-utilizing these tools. Investigating the reasons behind this behavior, I apply the cognitive model of bounded rationality to the human-machine setting. Under this model of decision-making, people select plausible options based on their prior heuristics rather than internalizing all relevant information. I find significant evidence showing that interpretability tools exacerbate the application of bounded rationality. As a solution, I present a new framework for re-imagining interpretability and explainability based on sensemaking theory from organizational studies. This Sensible AI framework prescribes design guidelines grounded in nuances of human cognition—facets of the individual and their environmental, social, and organizational context.

# CHAPTER I

## Introduction

Human-machine partnerships are increasingly commonplace. What began as an automation of routine tasks (e.g., software testing, data analytics, enterprise resource planning) has grown into algorithms that manage the content we see (e.g., news feeds on social media, information retrieval via search engines, and recommender systems like Netflix and Amazon) and AI that is capable of generating new data—text, images, videos—given sufficient context and prompts (e.g., chatbots grounded in large language models). In effect, human-machine collaboration has changed the way we think, communicate, and collaborate in personal and professional settings, at a societal level.

Yet, systems that rely on human-machine partnerships are unable to consistently capture the dynamic needs of people, or explain complex machine reasoning and outputs. In practice, a human-machine partnership is bound by both: the extent to which the machine can infer people’s needs and people’s ability to articulate and process information. When not taken into consideration, this *socio-technical gap*—a mismatch between the dynamic social needs of people and the brittleness of a machine’s representation of them [5]—has led to harmful outcomes such as propagation of biases against marginalized populations and missed edge cases in sensitive domains.

How do we fix this? Taking inspiration from prior work in Human-Computer Interaction (HCI) and the social sciences, my research follows the belief that technical development must

come in concert with an understanding of human-centric cognitive, social, and organizational phenomena. Put another way, systems that rely on human-machine partnerships cannot only be technically sound. They must also adapt to the nuances of human behaviors, as individuals and collectively. Investigating these nuances and designing for them are the core themes of this dissertation.

## **1.1 The Human in Human-Machine Collaboration**

The ubiquity of human-machine collaboration has snowballed with the recent developments in AI and ML. This has allowed the machine counterpart to, for example, infer patterns in past data to forecast outcomes for future datapoints, simulate seemingly-realistic content based on prompts, and communicate with end users in natural language—to the point where it has become challenging to distinguish when the AI is completely fabricating information that has no basis in reality [23, 104]. Moreover, while AI- and ML-based systems were once confined to the academic community, they have now grown into an industry that is far more accessible to everyone. Even people with very little expertise on AI and ML can now interact with these technologies, especially since the boom in generative AI. These systems are routinely deployed in real-world settings, including sensitive domains ranging from criminal justice and public policy to healthcare, education, and finance.

However, there has not been enough parallel growth in helping people keep up with AI- and ML-based human-machine systems. We are no longer certain of what a system will do next, explain why it reached a certain conclusion, or what it understands about the data it was fed. This lack of human understanding has enlarged the delta between innovating a new human-machine system and its secure deployment in the real world. Over the next few sections, I will further describe why this lack of human understanding of their complex machine counterparts is problematic, what solutions currently exist, and the critical piece that this dissertation contributes.

To concretize the problem and solutions, I study human understanding of complex

machines in the context of ML-based decision-support systems (DSSs). An ML-based DSS is an information system that gathers and analyzes data to help people determine future courses of action in rapidly changing settings. DSSs are widely used in everyday settings (e.g., which restaurant to try in a new city or which research paper to read next) to critical domains (e.g., what is the best course of treatment for a patient’s symptoms or should a loan application be approved). I focus on DSSs developed for a specialized setting, i.e., where the ML model is tailored for the given task as opposed to utilizing a pre-trained, generative model. This type of DSS requires at least three human roles with varying ML expertise: the practitioner who manages the ML-based system backend, the system designer who designs the interface for user interaction, and the end-user. Of these stakeholders, I focus on ML practitioners who make decisions such as the data to be used, the ML algorithm being applied, and when the model is ready for deployment to end-users—all of these decisions could benefit from a better understanding of the underlying ML model, its inputs and outputs.

### **1.1.1 The “Why”: Why should we care about helping people understand ML?**

To answer this question, consider what happens when people do not have a sufficient understanding of a ML-based decision-support system, but it is deployed in the real world regardless. One glaring example of harmful outcomes was the use of scores from COMPAS, a risk assessment instrument used in the criminal justice system to establish the likelihood of committing future crimes. The model amplified societal biases in the dataset, wrongly labeling black defendants as more likely to commit crimes in the future [14]. Another known example is the social biases that are transmitted from existing data into loan approvals at banks, causing future predictions to overfit on the gender or ethnicity of loan seekers [77]. This potential for harm is why it is imperative for stakeholders to have some understanding of how ML models work.

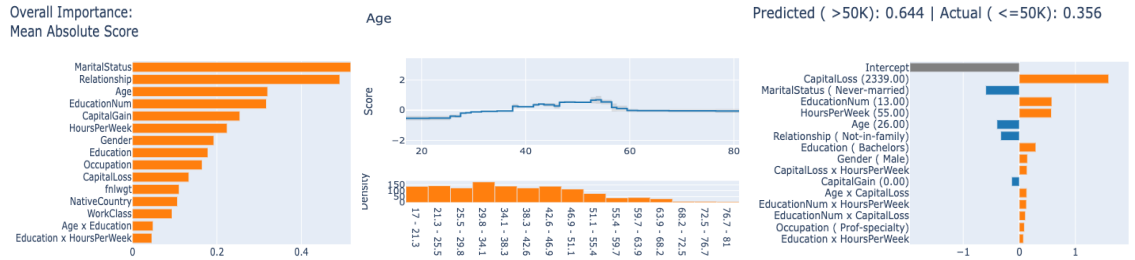


Figure 1.1: Common types of explanations output by interpretability tools. Left: Summary feature importances (global explanation). Middle: Component or partial dependence analysis (global explanation). Right: Feature attributions (local explanation). These particular visualizations are output by the interpretML implementation of a popular type of glassbox models, Generalized Additive Models (GAMs) [167]. The underlying dataset used for these visuals is the Adult Income dataset compiled over 1994 census data. It predicts income based on demographic features.

### 1.1.2 The “What”: What can we do to help people understand ML?

Approaches like interpretability have been proposed as a way helping people better understand ML. Interpretability is defined from a ML model’s perspective as the “ability to explain or to present in understandable terms to a human” [54, p2]. It serves as a proxy for other important desiderata for ML-based systems such as reliability, robustness, transferability, informativeness, etc. These properties in turn promote trustworthiness, accountability, and fair and ethical decision-making based on ML outputs [54, 142]. There are two common ways to support model interpretability: (1) designing glassbox models that are inherently interpretable (e.g., simple point systems [106, 256] or generalized additive models GAMs [39]), and (2) providing post-hoc explanations for the predictions made by complex, blackbox models (e.g., local interpretable model-agnostic explanations (LIME) [190], Shapley additive values (SHAP) [149]). There are two types of explanations that are commonly output by these approaches: global explanations that describe the overall logic of the model (e.g., a summary of feature importances, partial dependence plots, component analysis) and local explanations pertaining to individual datapoints (e.g., feature attributions that illustrate the importance of each feature in making a specific prediction). Figure 1.1 shows some examples of these explanations in a visual format.

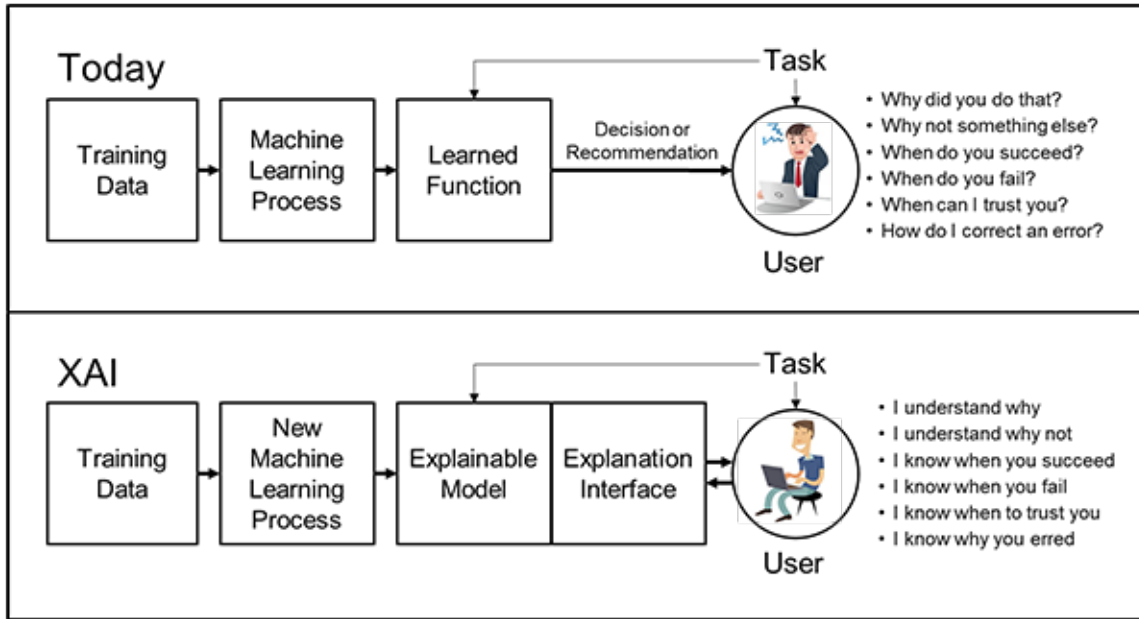


Figure 1.2: Explorable AI, as conceptualized in a DARPA program report [72].

### 1.1.3 The “How”: How should we design ways to help people understand ML?

Instantiating the aforementioned interpretability approaches and types of explanations into user-facing tools can be accomplished in several ways. The question is: *how* should we design them? What started as static explanations output by mathematical representations of interpretability now includes interactive visual explanations output by explainable AI. Although similarly defined, this idea of explainability is more human-centered. It 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 human” [16, p85]. Figure 1.2 shows DARPA’s conceptualization of explainable AI and the needs of a user that the model should address when explaining itself [72]. This conceptualization is further supported by research. Scholars have incorporated prior work from philosophy and the social sciences with the motivation that by translating ideas from how people explain things to each other, we can design better solutions for how ML models can be explained to people [161]. As a result, increasingly, interpretability and explainability tools include characteristics such as interactivity [10], counterfactual “what-if” outputs [162, 241],



and modular and sequential explanations [158].

#### **1.1.4 The Vital “Who”: Do these solutions work for who they are intended for?**

So far, using arguments from prior work, I have discussed why we should care about helping people understand ML, what we can do about this problem, and how researchers have designed solutions for this problem. My work follows a simple observation: where is the *who* in all this? When designing to help people understand ML, we cannot assume that people are entirely rational agents. Much of the early scholarship on interpretability focused on developing approaches that explain model behavior and outputs. However, the more recent thrust of research on explainable AI has argued for a human-centered perspective on this problem, noting that helping people understand ML is as much a human problem as a technical one. This emergent research focuses on investigating whether interpretability and explainability approaches help people understand the model behavior and outputs that they intend to explain, *in practice*. My dissertation work is a part of this human-centered research thrust. It challenges the assumption that people (can) perfectly internalize the information from interpretability and explainability tools and, as a result, better interact with ML-based systems. Instead, I empirically show that this problem is bound as much by people’s information processing capabilities and the nuances of sensemaking, as it is by the quality of information ML models and explanations can provide. My work is grounded in cognitive science and organizational research, which has abundantly posited that human behavior includes descriptive, normative, and prescriptive considerations that are not accounted for in technical solutions.

While existing interpretability and explainability approaches explain how a ML model works, they assume that people adequately use this information to make decisions. In my work, I test the validity of this assumption and propose solutions that take into account the complexities human behavior in the human-machine context. This takes the form of the following research questions:

- **Observing Existing Tool Use:** How do ML practitioners perceive and use interpretability tools? What are the key challenges to their use of these tools?
- **Understanding the Bounds of Human Cognition:** What are the underlying causes for the challenges that ML practitioners face when using interpretability tools?
- **Alternate Pathways to Support Human Understanding:** What are the characteristics of solutions that account for human-centered cognitive principles when designing interpretable ML outputs?

I explore these themes via a combination of methodologies, incorporating qualitative and quantitative methods, and translating theory from other, relevant fields.

## **1.2 Interpreting Interpretability: Understanding Practitioners' Use of Interpretability Tools for Machine Learning**

In Chapter III, I present studies on observing the use of interpretability tools by a key stakeholder—ML practitioners—in context. Prior to this study, there had been little to no evaluation of interpretability tools in a realistic context. One reason for this is the challenge of the context itself: designing and running user studies that reflect a realistic data science context is notoriously difficult. These studies require expertise in the mathematics underlying ML models and in HCI, as well as knowledge of both the academic literature and day-to-day engineering practices. To paint a full picture, these 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).

Working with an interdisciplinary team that met these criteria, I designed a three-phase setup to study the effectiveness of interpretability tool use by ML practitioners. This included pilot interviews with practitioners to identify common issues faced by them in

their day-to-day ML pipeline, followed by a contextual inquiry designed to observe whether interpretability tools could help practitioners overcome the common issues surfaced during the pilots, and finally a large-scale survey to generalize the findings from the contextual inquiry. These studies highlighted a trend of over-trust and misuse of interpretability tools. Often, the novelty and publicly available nature of the tool, as well as the ability to visualize complex ML models, served as the reasons people trusted tool outputs rather than using them to further reason about the underlying data and the accuracy of the model itself. This study set the stage for the needs of the stakeholders—despite presenting helpful information, interpretability tools were evaluated by this key stakeholder based on other characteristics that had little to do with the information being presented.

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

A natural follow-up to the results of the previous studies is: why do people over-trust and misuse interpretability tools? I hypothesize *bounded rationality* as being the underlying reason for inadequate use of interpretability tools. Bounded rationality is a model of decision-making under which 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.

In Chapter IV, I describe a between-subjects, pre-registered controlled experiment

with ML practitioners, investigating the role of bounded rationality in people’s use of interpretability tools and whether it leads to positive or negative outcomes. The experiment covers a variety of interpretability approaches and tools, including both glassbox and blackbox models, and static and interactive tools. The results provide significant evidence showing that interpretability tools lead to bounded rationality with bad outcomes. A control condition sans interpretability outperforms all interpretability conditions in terms of task accuracy by  $\sim 17\%$ . Given these results, I argue that interpretability needs to be reconsidered at a paradigmatic level. I 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?

#### **1.4 Sensible AI: Re-Imagining Interpretability and Explainability using Sensemaking Theory**

Through various user studies, we now know that factors such as stakeholders’ prior experience [61], attitude towards AI (e.g., algorithmic aversion [35, 51]), socio-organizational context [60], etc., all affect how (much) people understand ML, even with interpretability and explainability tools at hand. In the previous work, I studied the role that bounded rationality—an innate characteristic of people in decision-making settings—plays in all this. While these types of studies will continue to help us develop a better understanding of the importance of stakeholders and learn about their characteristics that are relevant in this setting, there is also an opportunity to learn from prior work in other fields and translate theory to develop a research agenda that is truly focused on who interpretability tools are intended for.

To that end, I present a new theoretical framework called Sensible AI, which takes into consideration principles of human cognition in proposing guidelines for effective human-machine collaboration in decision-support settings. This framework is grounded in Karl Weick’s sensemaking theory from organizational studies [247], which describes

several individual, environmental, social, and organizational properties that are relevant when people are trying to make sense of an object or event. In Chapter V, I first explain these properties and their relevance to the human-machine context, and then use an application of sensemaking in organizations as a template for discussing design guidelines for Sensible AI. This new framework complements the recent evaluations with stakeholders by unifying them based on an explanatory theory from organizational studies: indeed, several properties of sensemaking (e.g, social and organizational context) have already been shown to be important by the recent human evaluations of interpretability and explainability tools, and ML-based systems more broadly [60, 151].

## 1.5 Contributions of this Dissertation

**First, it presents empirical evidence that existing interpretability tools do not work as intended.** While existing tools do meet the criteria of an artifact which is designed to support exploration of how ML models work, why models make certain predictions, and an overall understanding of the inputs and outputs, these tools do not cater to what we know about how people internalize information. This is particularly true for the human-machine context where the information is often complex and can sometimes present conflicted outcomes. For example, imagine a local explanation for which the order of feature importance is quite different from the global feature importance order, making it harder to gauge whether the model’s prediction is dependent on accurate feature understanding or randomness. In this regard, my work offers insights into tool use via human evaluations that, although more common than before, continue to be a rarer thread of research in the community. Not only that, these empirical studies are a methodological contribution in that they are designed to simulate a realistic context for ML practitioners, taking into account ML theory, HCI experimental design guidelines, and considerations from both the engineering and academic practices surrounding data science.

**Second, it provides a richer understanding of the cognitive frameworks under-**

**pinning human behavior in the human-machine context.** As we move towards human-centered design and evaluation of interpretability tools, my work is grounded in prior work in cognitive and organizational sciences. For example, bounded rationality is a well-known cognitive framework which I am using to explain how ML practitioners currently use interpretability tools. These cognitive frameworks have been extensively used to explain human behavior in human-only decision-making settings. My work extends them to the AI and ML settings. I do so by conducting both large-scale quantitative studies to test the validity of these frameworks and smaller-scale qualitative studies to capture ways in which these frameworks must be updated for the human-machine context. With this theory-grounded evaluative work, I answer questions about: 1) how our knowledge of human decision-making must change to accommodate the role that AI and ML now play in it, and 2) which characteristics of human-only decision-making continue to be relevant in designing for effective human-machine collaboration.

**Third, it presents a new theoretical framework that prescribes design guidelines for re-imagining interpretability and explainability with human cognition and sense-making at its core.** Only recently have we, as a community, acknowledged the critical role that stakeholders play in determining the effectiveness of interpretability and explainability approaches. My work continues this new thread of descriptive and evaluative work. In addition to that, I also take a prescriptive stance with a research agenda based on properties of human sensemaking from organizational science. This includes a template for designing AI- and ML-based systems that account for individual, social, and organizational properties of sensemaking. Prior work in organizational science has suggested ways in which we can account for sensemaking in settings that require high reliability (e.g., nuclear power plants, aircraft carriers [248]). I translate these to design guidelines for the human-machine context.

### **1.5.1 Summary of Contributions**

Overall, my work presents empirical evidence that current interpretability tools do not meet their intended goal of helping people understand ML, and highlights the need for additional in-context human evaluations of these tools. I empirically show that the cognitive framework of bounded rationality shapes current over-reliance on interpretability tools. Additionally, I present sensemaking as a new, unifying theoretical framework for designing solutions that account for important individual, environmental, social, and organizational properties of human cognition, translated for ML-based decision-support settings.

## CHAPTER II

### Background

With ML-based systems routinely being deployed in the wild—including in sensitive domains such as criminal justice, education, and healthcare—it is increasingly important for stakeholders of these systems to have some understanding of how they work. Depending on stakeholder needs and prior experiences, this includes understanding why a certain prediction was made based on the input features provided, understanding the overall picture that the model has about the data, understanding specific relationships between the input and output variables, being able to debug the model, being able to communicate model behavior with customers and end-users, etc. [92, 143, 224]. Important factors such as trust, reliability, accountability, transparency, etc., have all been observed as a consequence of applying interpretability and explainability in AI and ML settings [55, 142]. In the next few sections, I discuss in detail the questions of *what* these interpretability and explainability approaches are, *why* they are important, and *how* they are instantiated in existing systems. This is followed by a discussion of *who* these approaches are intended for, an area that is the primary focus and contribution of my work.



## 2.1 The “What”, “Why”, and “How” of Interpretability and Explainability

There is not yet consensus within the research community on the distinction between the terms interpretability and explainability, and they are often, though not always, used interchangeably. When reviewing these terms in the context of the research conducted under each, *interpretability* is often defined from a model’s perspective as the “ability to explain or to present in understandable terms to a human” [54, p2]. Approaches for model interpretability have been the focus of ML research for quite some time now. Although similarly defined, *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 human” [16, p85]. Given its human-centered focus, research on explainability has made great strides by learning from prior work in cognitive science, philosophy, and the social sciences. Both interpretability and explainability have also been evaluated from a user-centered perspective in HCI research. Below, I present details on each of these research thrusts.

### 2.1.1 Machine Learning Approaches

The basic premise of interpretability and why we should care about it has been succinctly described by Doshi-Velez and Kim as “if the system can explain its reasoning, we can verify whether than reasoning is sound with respect to [important, pre-determined] auxiliary criteria”. These criteria include things like “safety [13, 172, 235], nondiscrimination [27, 75, 196], avoiding technical debt [203], or providing the right to explanation [69]” [54, p1]. Interpretability also serves as a proxy for other desiderata for ML-based systems such as reliability, robustness, transferability, informativeness, etc. These properties in turn promote trustworthiness, accountability, and fair and ethical decision-making based on ML outputs [142].

At a high-level, there are two approaches for achieving interpretability of ML models.

First, there are “glassbox” ML models that are designed to be inherently interpretable, often due to their simplicity. These include simple point systems [106, 256], decision trees [185] and sets [127], and generalized additive models (GAMs) [39, 78, 167]. The second approach is to train post-hoc explanation techniques that are designed to make the predictions of “blackbox” models more interpretable. These include local interpretable model-agnostic explanations (LIME) [190], Shapley additive explanations (SHAP) [150, 206], and other approaches for local explanations [9, 204, 212]. For an overview of interpretability techniques for shallow and deep learning models, see the comprehensive reviews by Gilpin et al. [68] published in 2018, Arrieta et al. [16] published in 2020, Liao and Varshney [138] published in 2021, and Dwivedi et al. [58] published in 2023.

Despite this proliferation of techniques, there is still debate about what interpretability should entail [54, 142, 195]. In particular, Rudin [195] 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. [55] propose that an explanation for a prediction should include not only a justification, but also a description of the decision-making process followed by the model. Lipton [142] surveys different criteria for assessing interpretability, such as simulatability and decomposability, 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. For example, Tan et al. [229] use publicly available datasets to test the feasibility of a new GAMs-based post-hoc explanation technique with ML experts. Kim et al. [115] do the same for a technique based on Bayesian model criticism, intended to add criticisms to example-based explanations. Poursabzi-Sangdeh et al. [183] 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 can cause people to incorrectly

follow a model when it makes a mistake, due to information overload. Lage et al. [125] 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. I discuss more of these factors that shape how interpretability techniques are used by people, as well as human evaluations of interpretability, in the subsections that follow. Cau et al. [40] note the difference on task performance when using inductive, abductive, and deductive reasoning based explanations, finding that the latter two improve performance for the high-uncertainty domain of stock trading. Meta-reviews like [58, 126] highlight several of these studies.

### **2.1.2 Cognitive Science, Philosophy, and Social Science Influences**

What started as static explanations output by mathematical representations of interpretability now includes interactive text and visual explanations. Several scholars have sought to incorporate prior work from cognitive science, philosophy, and the social sciences in designing new approaches and tools, the motivation for doing so being that by translating ideas from how people explain things to each other, we can design better solutions for how ML models can be explained to people. Let us consider this work in more detail.

Hempel and Oppenheim [81] and van Fraassen [233] define an explanation as consisting of two main pieces: the *explanandum*, a description of the phenomenon to be explained, and the *explanans*, the facts or propositions that explain the phenomenon, which may rely on relevant aspects of context. It is the explanans that we refer to as the *explanation* in the explainable AI literature. Different ways of formalizing the explanation have given rise to various theories, ranging from logical deterministic propositions [81] to probabilistic ones [198, 233]. A historical overview of these terms is included in the surveys by Pitt [180] and Miller [161].

Miller [161] further reviews the properties of explanations from the philosophy (e.g., [70, 141, 178, 233]) and social science (e.g., [83, 130, 146, 154, 166, 214]) literature. He notes 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 [161, 163] and Lombrozo [145] 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 [70], which form the core of a good conversation, should therefore be followed when designing explanations [123, 154, 214]. Leake’s goal-based approach to explanation evaluation adds 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 [130]. Explanations that follow this goal-based approach must include grounding in some common demonstrative reference between people and the explanation system [44, 152].

The guidelines from this work can be distilled into five overarching human-centered principles for designing explanations: (1) explanations should be contrastive, i.e., explain why the model predicted a certain outcome over others; (2) explanations should be exhaustive, i.e., provide a justification for why other alternatives were not predicted; (3) explanations should be modular and compositional, i.e., break up predictions into their basic, simple components; (4) explanations should rely on easily-understandable quantities, i.e., ensure that each component is understandable; and (5) explanations should be parsimonious, i.e., they should include only the most relevant facts about the prediction [158]. It is as a result of these principles that, increasingly, interpretability and explainability tools include characteristics such as interactivity [10, 85], counterfactual “what-if” outputs [162, 241], and modular and sequential explanations [158].

Although this line of work provides guidance for designing explanation systems that work best for humans—and ML researchers have begun to incorporate this guidance for interpretability [8, 58]—it also criticizes the development of explanation systems by researchers, citing this as an example of “inmates running the asylum” [163] because of a lack of user-centric evaluation. My dissertation addresses this critique by: (1) conducting

stakeholder evaluations of interpretability and explainability tools; and (2) proposing new frameworks for interpretability and explainability grounded in theories about the individual, environmental, social, and organizational factors that affect human understanding.

### **2.1.3 HCI Background on Human–Machine Collaboration**

HCI has a long-standing tradition of studying complex systems from a user-centric perspective. Bellotti and Edwards [21] were the first to define 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. Weld and Bansal extend this notion of intelligibility for ML-based systems [251]. In addition to the ones described in [21], they note the following benefits of intelligibility for ML-based systems: ensuring that the ML model is using adequate features for predictions, helping people accept the model’s predictions, improving human insight into situations for human-in-the-loop ML systems, and for legal auditing.

The HCI community has sought to improve the relationship between people and machines by providing richer human feedback to the models being used by these systems [223]. This line of work focuses on building interactive ML (iML) systems. The term iML was coined by Fails and Olsen Jr. [63] 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 [49]. Several examples of iML systems now exist, for applications including annotation of animal behavior [107], academic citation review [242], and on-demand personalized group formation [11]. Although helpful, this approach does have several challenges. For example, the need for complexity, introspection, and deliberative thinking, which is not supported by reactive elements since these are observation- rather than action-centric. Patel et al. [177] and Zhu et al. [261] note these challenges for the software development and game designing domains, respectively.

There has also been an effort within the HCI community on defining new metrics

for human–ML collaboration. Abdul et al. [4] highlight interactivity and learnability as cornerstones for designing visualizations that better support interpretability. Dourish [56] adds scrutability as an important component of interactivity. Hoffman et al. provide an overview of potentially relevant metrics for evaluating explainable AI systems: goodness of explanations, whether users are satisfied by them, how well they understand the AI systems, how curiosity influences the search for explanations, appropriate trust and reliance, and how the human-AI system performs overall [84]. Recent tool designs that have followed these metrics have shown positive outcomes in terms of stakeholders’ understanding and satisfaction when using these new tools [28, 33, 85, 117, 158, 213, 243, 257].

HCI studies have also contributed insights into explanation formats and modalities that suit various stakeholders. For example, Liao et al. [136] formalize this in the form of a question bank to determine the right explanation format for different information needs. Similarly, Wang et al. [244] outline how desiderata related to human understanding and reasoning can inform AI explanations. Considering AI systems and explanations more broadly, Amershi et al. [12] consolidate interaction design guidelines for AI system designers based on stakeholder needs. This thread of HCI work helps answer questions like: which explanation types/formats are easier to understand [74, 137, 259]; what works best in higher stakes settings [40, 175, 238, 255]; how to present explanations to lay users, novices, and experts [184, 246]; which types of uncertainty information are helpful [92, 187, 258]; what causes information overload [33, 219]; whether explanations should be presented as questions, answers, or a dialogue [47, 131, 132]; etc.

HCI research has always argued for a human-centered approach to designing and evaluating intelligent systems. Now, with these intelligent systems becoming increasingly powerful and the subsequent need for interpretability and explainability, these design principles are even more essential. Investigating stakeholders’ use of these systems is an avenue where we can learn from the rich history of prior work in HCI. Next, I discuss these human-centered factors in the context of research on explainable AI systems. My

dissertation directly builds on this burgeoning research thrust.

## **2.2 Understanding the “Who” in Interpretability and Explainability**

The last few years have seen exponential growth in research on explainable AI (XAI), an emerging area with the goal of “making AI understandable by people” [138, p2]. Paez calls this switch from explainability as a technical or definitional question to a stakeholder-driven functionalist approach the “pragmatic turn in explainable AI” [174]. Scholars in all communities—ML, HCI, and social science—have advocated for the importance of understanding *who* the explanations are intended for, based on evaluations with ML experts and practitioners, non-ML domain experts, end-users, and various other stakeholders. Between 2018 and 2021, over 100 peer-reviewed papers published results from human subjects evaluations of AI outputs in decision-making contexts [126], and the number has only grown since. Here, I discuss important principles about stakeholders that have been identified by this thrust of research on the “who” in interpretability and explainability.

### **2.2.1 Cognitive Factors**

On a cognitive front, one critical element to consider is the role of heuristics and biases. People are prone to applying these heuristics and biases as shortcuts to avoid conscious deliberation of information. For example, when applying the anchoring heuristic, people assign more value to the first piece of information—the anchor—provided to them in a decision-making setting [110]. Similarly, with the imitate-the-majority heuristic, when given a choice, people pick one that is popular in their socially proximal reference group [79]. In his dual-process theory of cognition, Kahneman [108] 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—in other words, “thinking, fast and slow” [109]. Recent studies with stakeholders have confirmed that these factors are also at play when people use AI/ML and explanations. Springer and Whittaker [219] find that explanations can create information

overload and distract people from forming a useful mental model of how a system operates. Nourani et al. [169] focus on the anchoring bias, and show that the order of positive and negative information about the system can shape over- or under-reliance. Bertrand et al. [25] provide an overview of all relevant cognitive heuristics and biases that can apply to how people use interpretability and explainability tools.

Appropriately handling these cognitive factors (e.g., via accurate mental models, improving the type of reasoning applied in making decisions) has been shown to be critical for how (much) people understand the outputs of interpretability and explainability tools. For example, accurate mental model formation and deliberative reasoning can help avoid ML practitioners' misuse of, and over-reliance on, interpretability outputs [111]. This applies to end-users without ML expertise as well, otherwise explanations increase the likelihood that an end-user will accept an AI's output, regardless of its correctness [18, 36]. For end-users, completeness (rather than soundness) is a key property of explanations that help them form accurate mental models [124]. Accuracy and example-based explanations can similarly shape people's mental models and expectations, albeit in different ways [120]. Although system-focused information helps in mental model formation, recent work has also shown that people prefer social information (i.e., how other people are using a system) when trying to form an idea of how a system works [29]. In decision-making settings, it is important to de-anchor people from only considering certain pieces of information which were selected due to the application of biases. Rastogi et al. [187] find that increasing interaction time with an AI system reduces system 1 biases like anchoring. More generally, when cognitive factors are taken into account and explanations are designed to support deliberative reasoning, over-reliance on explanations is reduced for end-users [33].

### **2.2.2 Prior Experience**

Prior experience with ML is important in determining which explanations are easier to understand for a stakeholder. For example, evidence from recent studies has shown



that, compared to ML practitioners and non-ML domain experts, lay users prefer counterfactuals and example-based explanations over high-dimensionality visuals and graphical interfaces [73, 217, 218]. Prior experience or background in ML can result in preset expectations from explanations [61]. Whether these expectations lead to positive or negative outcomes remains unclear. Some evaluations with ML practitioners have shown that prior experience can lead to over- or under-reliance on these explanations [61, 111]. With non-ML domain experts, there are some positive signs of accurate decision-making as long as the person reviewing the predictions from the ML-based decision support system has plenty of prior experience with the domain [102]. When presented with explanations in a visual format (which is the case for most interpretability tool outputs), novices perform worse than experts with prior experience, and are more likely to have “illusory satisfaction” [228].

### **2.2.3 Job- and Task-based Information Needs**

Different job roles and task needs require different kinds of information from explanations [140, 255]. For daily use applications (e.g., social media, recommender systems), end users prefer to not have detailed explanations. In fact, only a very small percentage of people look at explanations for recommendations or timeline content at all [34]. For applications that support task productivity (e.g., email classification systems, task or break recommenders), people prefer detailed explanations and, more importantly, the opportunity to collaborate with the underlying model being used in these systems via an interactive explanation interface [112, 120, 223]. Further, explanations from glassbox models with fewer number of features are easier for end-users to understand [183]. For ML practitioners, specific types of visuals of explanations (e.g., local vs. global, sequential vs. collective) differ in how much they help them understand and debug models, and explain them to customers [85, 158]. Lim et al. [140] review various intelligibility needs (e.g., to filter a cause, generalize and learn, predict and control) with ML-based systems and suggest adequate explanation types corresponding to each of these needs. Similarly, Suresh et al. [224] break

down explainability needs based on two axes: stakeholders' knowledge and their objectives. Knowledge options include formal, instrumental, and personal gains; objectives can be long-term, immediate, and specific tasks to perform with explanations.

#### **2.2.4 Social, Organizational, and Socio-Organizational Context**

For stakeholders that work with AI- or ML-based systems within an organization, their social, organizational, and socio-organizational context has been found to be an important external factor that the field had not considered thus far [60, 144]. This is particularly true for technology companies with ML practitioners that are either tasked with deploying a ML-based system for the company's products (e.g., content moderation teams for social media platforms) or responsible for building models for decision-support systems being used by external organizations (e.g., technology companies building healthcare support systems) [86, 261]. In these settings, prior work has found there to be a conflict in the goals of various stakeholders within and outside the organization. This conflict makes the adoption of responsible and ethical AI practices challenging for practitioners working within organizations [151, 237]. Adequate use of interpretability and explainability tools requires cooperation and mental model comparison between people operating at different job roles, both within and outside an organization [92].

#### **2.2.5 Summary**

Studies from ML, HCI, and the social sciences have all highlighted relevant factors about the "who" in interpretability and explainability. Simply updating explanations is not enough to account for these external factors that impact people. My dissertation builds on this idea. I present empirical studies and new frameworks that explain how individual, social, and organizational factors can affect the human-machine context, and provide a path forward that accounts for these *who*-centered factors.

## CHAPTER III

# Interpreting Interpretability: Understanding Practitioners’ Use of Interpretability Tools for Machine Learning

Machine learning (ML) has become ubiquitous in our everyday lives in domains ranging from criminal justice and public policy to healthcare and education. The recent developments in ML 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 [14, 38, 113]. 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 ML 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 [195].

Faced with these challenges, the ML community has turned its attention to the design of techniques aimed at *interpretability* [54, 142]. 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 [106, 256] or generalized additive models (GAMs) [39]. Second, there are techniques that provide post-hoc explanations for the predictions made by complex models, such as local interpretable model-agnostic explanations (LIME) [190] and Shapley additive explanations (SHAP) [149]. Although the number of proposed techniques continues to grow, there has been little evaluation of whether

they help stakeholders achieve their desired goals.

In this chapter, I present studies that I conducted to observe the effectiveness of interpretability tools for one key stakeholder group: data scientists and ML practitioners<sup>1</sup>. The studies involved 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. These studies took the form of three components that build on each 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.

### 3.1 Pilot Interviews

To better understand the issues that data scientists face in their data-to-day work—i.e., the setting in which the interpretability tools would ideally be used—I first conducted semi-structured interviews with six data scientists at a large technology company. The interview protocol was designed to surface common issues that arise when building and evaluating ML models (see Appendix A). On average, each interview lasted  $\sim 40$  minutes.

Based on an inductive thematic analysis of the interview transcripts, conducted via open coding followed by affinity diagramming [30], I identified six themes capturing common issues faced by data scientists (see Table 3.1). Five of these themes correspond to issues with the 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, this list is not exhaustive, but it is consistent with previous research on ML pitfalls [147].

---

<sup>1</sup>For simplicity, I refer to this group as “data scientists” throughout the chapter.

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 “Work-Class” 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 “Education-Num,” 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 3.1: Six themes capturing common issues faced by data scientists. Each issue was synthesized in a dataset for the contextual inquiry, as described in the third column. The fourth column contains the number of participants in the contextual inquiry who identified the corresponding issue.

### 3.2 Contextual Inquiry

With these common issues in mind, I 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. Eleven participants took part in the study, each of whom was given a Jupyter notebook that included a dataset, an ML model which had been trained using that dataset, an interpretability tool, and several questions to answer. The goal was to observe whether they were able to use the interpretability tool to uncover the issues identified during the pilot interviews. With participants’ consent, I recorded both audio and video, and saved all responses provided in the Jupyter notebooks for analysis. The scenario was approved by the internal IRB at the technology company at which these studies were performed<sup>2</sup>.

<sup>2</sup>I conducted these studies over the course of an internship at the same technology company.

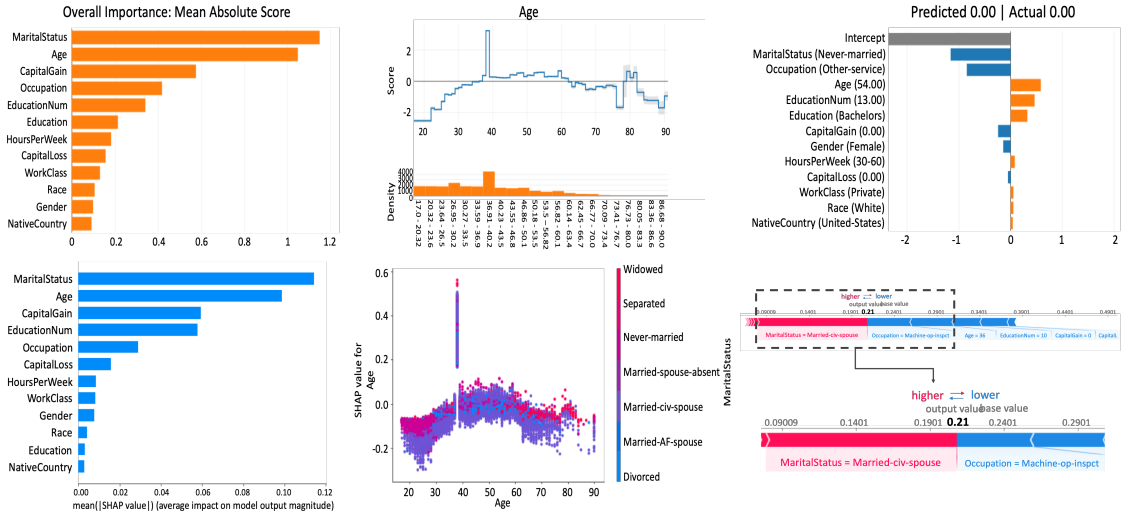


Figure 3.1: Visualizations output by the InterpretML implementation of GAMs (top) and the SHAP Python package (bottom). Left column: global explanations. Middle column: component (GAMs) or dependence plot (SHAP). Right column: local explanations.

### 3.2.1 Dataset

I derived the dataset from the Adult Income dataset,<sup>3</sup> 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  $\sim \$86.5k$  when adjusted for inflation). I synthetically manipulated a subset of the features to incorporate the common issues identified via our pilot interviews. For example, to incorporate missing values, I replaced the age value with 38, the mean for all data points, for 10% of the data points with an income of  $> \$50k$ . Table 3.1 includes the details of all manipulations performed on the dataset.

### 3.2.2 ML Models and Interpretability Tools

I 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

<sup>3</sup><https://archive.ics.uci.edu/ml/datasets/Adult>

into additive components that are functions of one input feature each [78]. 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. I used the InterpretML<sup>4</sup> implementation of GAMs. InterpretML provides built-in plotting functionality, allowing each individual component to be visualized (see Figure 3.1, top middle). InterpretML also provides global explanations (see Figure 3.1, top left) and local explanations (see Figure 3.1, top right) by ranking and sorting the contributions made by each input feature to the predictions [167].

SHAP is a post-hoc explanation technique for blackbox ML models. It assigns each input feature an importance score for each prediction [149]. These scores are based on the notion of Shapley values from cooperative game theory [206]; for each prediction, they capture a fair distribution of “credit” over the input features. I used the implementation of SHAP in the SHAP Python package.<sup>5</sup> The importance scores computed by this package directly translate to local explanations for individual predictions (see Figure 3.1, bottom right). By aggregating the importance scores for many predictions, the SHAP Python package can also produce global explanations (see Figure 3.1, bottom left) and dependence plots for single input features (see Figure 3.1, bottom middle). Ideally, I 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, I used LightGBM [114], an implementation of gradient boosted decision trees, to create the underlying model to be explained using the SHAP Python package. This 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 [148], and

---

<sup>4</sup><https://github.com/interpretml/interpret>

<sup>5</sup><https://github.com/slundberg/shap>

LightGBM is widely used. Please note that the comparisons that I 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 the modified dataset.

I chose these specific 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, only has local explanations. Each participant used one interpretability tool, selected at random; 6 participants used GAMs, while 5 used SHAP.

I also provided each participant with a print-out of a tutorial which was prepared 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 included as Appendices B and C.

### **3.2.3 Contextual Inquiry Protocol**

First, I asked each participant to sign a consent form and answer some questions. These questions followed a semi-structured interview protocol 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, I let each participant explore the dataset, model, and interpretability tool on their own. For each tool, the Jupyter notebook included examples of all three types of visualization—i.e., global explanations, components (GAMs) or dependence plots (SHAP), and local explanations, as depicted in Figure 3.1. After this, I asked each participant to complete the trust questionnaire of Jian et al. [105] with respect to the interpretability tool. This was followed by 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 the pilot interviews; a full list is included in Appendix D. Answering these questions required participants to use all three types of visualization. For each question, I 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, each participant completed the trust questionnaire again, allowing me to observe whether their self-reported trust in the interpretability tool had changed. The inquiry 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.

### **3.2.4 Participants and Data**

I 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 I subsequently excluded several based on their location because I 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.

I used speech recognition software to generate transcripts from the video files and hand-corrected any errors. These transcripts and participants’ open-ended responses were

qualitatively coded using inductive thematic analysis [30]. I 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. I also obtained descriptive statistics from the trust questionnaire and the questions about their background, etc.

### **3.2.5 Results**

The contextual inquiry revealed 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 3.1 presents 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. I 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 [105]) was 3.70 (s.d.=0.4) before using the tools and 3.90 (s.d.=0.6) after.

### 3.2.5.1 Theme 1: Misuse and Disuse

Most participants relied too heavily on the interpretability tools. Previous work categorizes such over-use as *misuse* [59, 176]. 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 confident).

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 [59, 176]:

“[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)

### 3.2.5.2 Theme 2: Social Context is Important

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).

### **3.2.5.3 Theme 3: 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 my 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)?”

### **3.3 Large-Scale Survey**

Following the contextual inquiry, I designed a survey to scale up and quantify our main findings and shed light on data scientists’ mental models of interpretability tools. Similar to the 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. I 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.

#### **3.3.1 Experimental Conditions**

As in the 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 in Section 3.2.2 above), selected at random. I also showed participants either “normal” or “manipulated” visualizations, again selected at random. In the normal-visualization condition, participants were shown the visualizations output by the interpretability tools. However, in the manipulated-visualization condition, they were shown the 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. I 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).

#### **3.3.2 Components of the Survey**

First, I asked each participant to sign a consent form and gave them a brief introduction to the survey. I 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, I showed them the results of common exploration commands that had been run previously by participants in the contextual inquiry. I also gave each participant access to a description of the dataset and a tutorial on the interpretability tool that they were to use. These tutorials were the same as the ones used in the contextual inquiry (see Appendices B and C).

Next, I 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, I asked each participant to select the capabilities of the interpretability tool from a list of options (shown in Figure 3.2). To better understand participants' mental models of the tools, I asked them to describe what the x- and y-axes represented in each of the visualizations

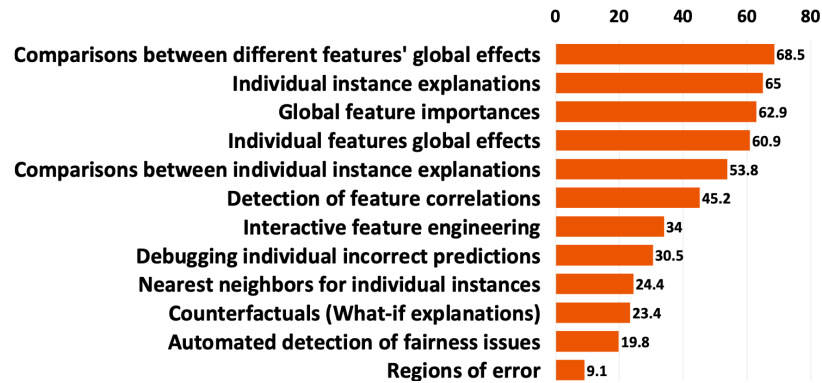


Figure 3.2: Percentage of participants that selected each option when asked about the interpretability tools' capabilities.

they had seen. I also asked them whether the visualizations were useful and, if so, how they would use them in their typical ML pipeline.

Finally, to encourage each participant to think critically about issues with the dataset or model, I 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. I also asked them to describe how they would convince a customer that this was the right model to deploy (or not, as appropriate). The survey concluded with the NASA-TLX cognitive load questionnaire [1].

### 3.3.3 Participants

I 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, I asked people about their ML experience, and only surveyed people who rated their experience as at least 3 on a scale of 1–7. The survey was completed by 253 participants. After filtering out responses with exactly the same content for every open-ended text field or other gibberish text, I was 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.

### 3.3.4 Preregistration

Before conducting any analyses, I 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 preregistered comparisons 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. I intentionally omitted comparing the main effect of GAMs vs. SHAP to avoid lumping together data from the normal-visualization and manipulated-visualization conditions. The preregistration also noted my intent to conduct additional exploratory analyses. The full document is available on AsPredicted.<sup>6</sup>

### 3.3.5 Methods

I 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, I did not apply Bonferroni correction because only four comparisons were performed for each outcome variable. The content from the open-ended responses was coded via open and axial coding [46]. I 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 interpretabil-

---

<sup>6</sup><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	
Accuracy of answers	$\mu 1 = 78.1 \pm 25.1$	$\mu 2 = 74.0 \pm 28.4$	$\mu 1 = 78.1 \pm 25.1$	$\mu 2 = 58.8 \pm 24.7$ t(98) = 3.83, p < 0.001, Cohen's d = 0.8	$\mu 1 = 58.8 \pm 24.7$	$\mu 2 = 54.4 \pm 28.2$	$\mu 1 = 74.0 \pm 28.4$	$\mu 2 = 54.4 \pm 28.2$ t(94) = 3.38, p < 0.001, Cohen's d = 0.7
Confidence: understand explanation(s)	$\mu 1 = 5.7 \pm 0.6$	$\mu 2 = 5.3 \pm 1.1$	$\mu 1 = 5.7 \pm 0.6$	$\mu 2 = 4.4 \pm 0.8$ t(93) = 8.77, p < 0.001, Cohen's d = 1.7	$\mu 1 = 4.4 \pm 0.8$	$\mu 2 = 4.1 \pm 0.9$	$\mu 1 = 5.3 \pm 1.1$	$\mu 2 = 4.1 \pm 0.9$ t(95) = 6.04, p < 0.001, Cohen's d = 1.2
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	$\mu 1 = 4.8 \pm 0.9$	$\mu 2 = 3.8 \pm 0.9$ t(98) = 5.14, p < 0.001, Cohen's d = 1.0	$\mu 1 = 4.8 \pm 0.9$	$\mu 2 = 5.0 \pm 1.1$	$\mu 1 = 5.0 \pm 1.1$	$\mu 2 = 4.4 \pm 1.1$ t(95) = 2.48, p < 0.05, Cohen's d = 0.5	$\mu 1 = 3.8 \pm 0.9$	$\mu 2 = 4.4 \pm 1.1$ t(89) = -3.03, p < 0.05, Cohen's d = 0.6
Cognitive load	$\mu 1 = 3.8 \pm 0.9$	$\mu 2 = 4.2 \pm 1.0$	$\mu 1 = 3.8 \pm 0.9$	$\mu 2 = 4.9 \pm 0.8$ t(96) = -6.40, p < 0.001, Cohen's d = 1.3	$\mu 1 = 4.9 \pm 0.8$	$\mu 2 = 5.0 \pm 0.9$	$\mu 1 = 4.2 \pm 1.0$	$\mu 2 = 5.0 \pm 0.9$ t(95) = -4.57, p < 0.001, Cohen's d = 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 3.2: The results of the 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 ( $\mu 1$  and  $\mu 2$  are the means of conditions 1 and 2 in the header, with standard deviations). Significant differences are highlighted in gray along with 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.

ity tools. My collaborator and I 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; I used this over the preregistered chi-squared test due to a class imbalance for these variables. I conducted exploratory analyses using the participants' ML experience by fitting multiple linear regression models and calculating Pearson correlation coefficients. I also noted descriptive means, standard deviations, and counts, whenever applicable.

### 3.3.6 Results

Table 3.2 summarizes the results of the 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, there are differences in their stated confidence that the underlying models were reasonable, both as a main effect of normal vs. manipulated visualizations in an ANOVA ( $F(1, 101) = 25.05, p << 0.001$ ) 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.

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

### 3.3.6.1 Factors that Affect Willingness to Deploy

To explain why participants thought, on average, that the underlying models were ready for deployment, next I present a selection of themes 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 try to convince a customer that

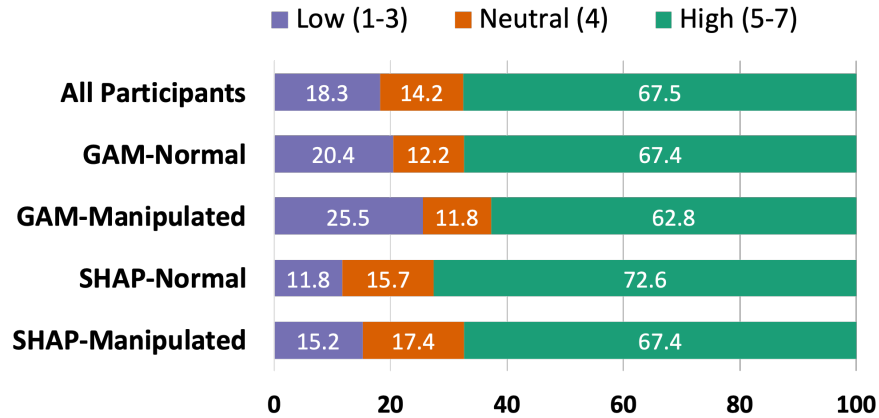


Figure 3.3: Percentage of participants with low, neutral, and high deployment scores per condition, and a total percentage per score type across all participants.

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 Figure 3.3). 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.

### **3.3.6.2 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 [168].

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). Two of my collaborators and I 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 3.2 depicts the percentage of participants that selected each of the options provided when asked about the interpretability tools' capabilities.

### **3.3.6.3 Tension between Cognitive and Social Factors**

The 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, I discuss how these factors affected participants' use of the tools and their deployment ratings for the underlying models. These

are exploratory analyses; although I report p-values, I 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, I 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 significantly 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).

I 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, 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.** I 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. I confirmed this relationship using Pearson correlation coefficients and found them to match my 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.

## **3.4 Discussion and Future Work**

### **3.4.1 Bridging the Gap Between the ML and HCI Communities**

The results from this chapter highlight the value of user studies—normally conducted in HCI and adjacent fields—for evaluating interpretability techniques designed by the ML community, marrying the goals and methods of both communities. These 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 the findings here 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., [37, 244]), designing tools that can be understood by users with different background (e.g., [128]), and undertaking user studies at each stage of tool development (e.g., [254]). Indeed, since I conducted these studies in 2019, there has been exponential growth in studying interpretability from a stakeholder perspective, with user studies being conducted by scholars in both ML and HCI communities, sometimes jointly. As detailed in Chapter II, these studies continue to provide invaluable insights into how different explanations are suited for various stakeholders in a multitude of settings.

### **3.4.2 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 [19]. As one of my 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 [85, 163]. Weld and Bansal [251] 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 [144]. More generally, interpretability tools should be designed to adapt to users' expectations. I describe one such design implication below.

### **3.4.3 Designing Tools for Deliberative Reasoning**

Interpretability tools are designed to help stakeholders better understand how ML models work. However, as I found, these tools' visualizations can encourage people to make quick decisions instead of digging deeper. As P4 from the 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 [109, 110] 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.

## **3.5 Limitations**

These studies have several limitations. First, although I tried to put data scientists in a realistic setting via Jupyter notebooks in the contextual inquiry and via visualizations and the results of common exploration commands in the survey, I cannot be certain that this was

sufficient. Second, longitudinal studies might reveal different or more nuanced patterns of behavior than either the contextual inquiry or the survey. Third, I 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. I therefore relied on qualitative findings to support and add nuance to my quantitative results. Fourth, I used a tabular dataset. Although there is research on interpretability techniques for deep learning and richer types of data (e.g., images [116, 204]), this was not the focus of my work.

### **3.6 Conclusion**

In this chapter, I presented empirical evidence that existing interpretability tools do not work as intended. I conducted studies with data scientists' regarding their 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). These studies included pilot interviews ( $N = 6$ ) to identify common issues faced by data scientists in their day-to-day work, followed by a contextual inquiry ( $N = 11$ ) to observe how data scientists use interpretability tools to uncover these issues, and finally a survey ( $N = 197$ ) to scale up and quantify the main findings from our contextual inquiry. The results presented 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, I discussed the need for members of the HCI and ML communities to work together, and some avenues for future exploration when designing interpretability tools. Overall, this work sets the foundation for my dissertation by identifying challenges to a key stakeholders' use of interpretability tools.

## CHAPTER IV

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

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. However, several studies with both ML practitioners and lay end-users—including the ones presented in Chapter III—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,<sup>1</sup> and are unable to make accurate judgements about the data and model despite having access to the additional information provided by these tools [18? , 111, 120]. In this chapter, I consider the question of *why* interpretability tools are inadequately used.

I 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

---

<sup>1</sup>I use the term “interpretability” to indicate both interpretability and explainability approaches and tools throughout the chapter. Interpretability has model-centric connotations whereas explainability is used to signal a focus on human-centered values. However, for the purposes of this study and the tools I employ, the two terms can be considered interchangeable.

*actually* possessed by organisms, including man, in the kinds of *environments* in which such organisms exist” (emphasis my own) [209, p99]. Under this model of decision-making, people select “good enough” options rather than considering the utility of all alternatives. 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, I ask the following questions: *How do people apply bounded rationality when using interpretability tools? Does it help or hurt in this context?*

To observe the role of bounded rationality in the ML context, I conducted a between-subjects, pre-registered, controlled experiment with ML practitioners ( $N = 119$ ),<sup>2</sup> asking them to perform exploratory data analysis and answer questions about the data and model. The experiment compared four interpretability conditions (two glassbox models and two post-hoc explainers for blackbox models) and a control condition sans interpretability. Next, I provide a background on bounded rationality along with an example in the ML setting, followed by a list of hypotheses generated for this setting, the study methods and results, and a discussion of implications for interpretability tool designers and researchers.

## **4.1 Bounded Rationality**

Bounded rationality describes human beings as rational agents functioning within cognitive and informational constraints [211]. These constraints separate the boundedly rational person from homo economicus, described by Mill [160] 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

---

<sup>2</sup>I use the broad category term ML practitioners to represent people with prior experience in ML. These include data scientists, practitioners, ML software designers and developers, and researchers.

encapsulated by the concept of homo economicus [209].

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 [210]. 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) [211].

Simply put, bounded rationality is a process by which people choose an option that is “good enough” as defined by their own criteria. People 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 [156]. Bounded rationality and satisficing appear to be a useful descriptive model for many cognitive processes including split-second decision making [171], patterned and random repeated choice scenarios [200], discrete choice scenarios [156], and explanations of theory of mind under uncertainty [182]. The salient commonalities across domains in which these are observed or usefully descriptive are the presence of uncertainty or a prohibitive cost of information accrual—both potential characteristics of an exploratory ML task (example below).

#### **4.1.1 Example Setting**

Imagine you are a ML practitioner analyzing the Titanic survival dataset.<sup>3</sup> This dataset is used to predict which passengers survived the sinking of the Titanic based on demographic and socio-economic features. Say you built a blackbox model for this dataset and are using the SHAP [150] Python package as a post-hoc explainer. Figure 4.1 presents the

---

<sup>3</sup><https://www.kaggle.com/c/titanic>

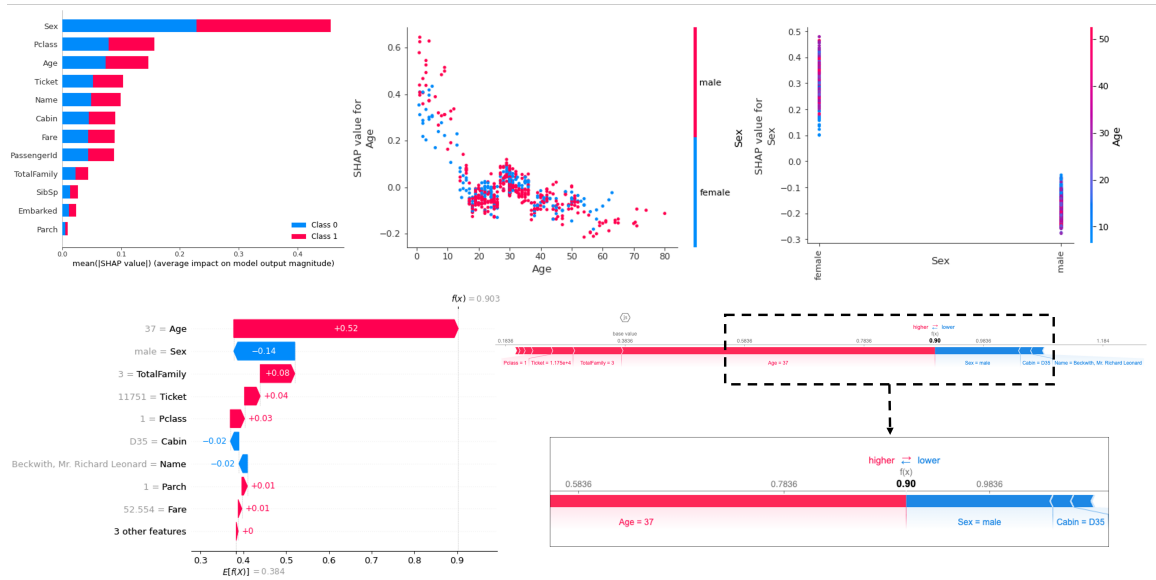


Figure 4.1: Visualizations output by the SHAP Python package, a post-hoc explainer for blackbox models. These are generated for the Titanic survival dataset using a LightGBM model. Top (left to right): Overall feature importances; Partial dependence plot for a continuous input feature, age; Partial dependence plot for a categorical input feature, sex (all global explanations). Bottom (left to right): Waterfall plot and Force plot, both types of local explanations for an individual data point.

types of visuals available via SHAP (and most interpretability tools): (1) the overall impact of each feature on the model’s predictions (global explanation); (2) partial dependence plots, showing the relationship between one input feature and the output variable (global explanation); and (3) feature attributions, describing how an individual prediction was made (local explanation). 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 4.1. 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 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

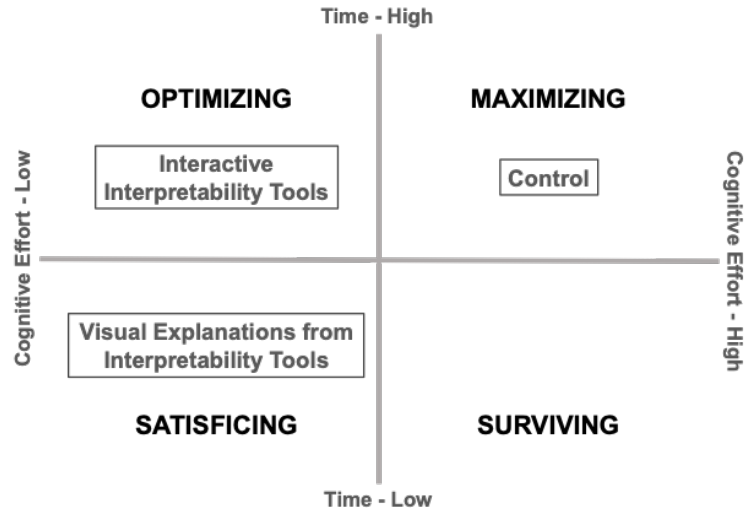


Figure 4.2: Using the framework of bounded rationality to create quadrants for human cognition in Machine Learning settings, defined by the amount of time and cognitive effort spent on the task. The experimental conditions I tested map to three of these quadrants.

to 70 because the likelihood of survival is lower for older passengers. Which option would you pick? With bounded rationality, people are looking 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.

## 4.2 Research Goals and Hypotheses

In cognitive science, the bounded nature of human rationality is often expressed as a function of time and effort. These form the basis of my cognitive framework for the ML context (Figure 4.2). 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. However, prior work in cognitive science and behavioral economics shows that people *satisfice* to conserve time and effort in decision-making settings [209, 110, 109].

Translating and expanding the motivations behind these cognitive modes to the ML

setting, I hypothesized that interpretability solutions engage the bounded cognitive modes in ML practitioners 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. I hypothesized 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 the 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 is less acute for practitioners as stakeholders. Therefore, I did not test this cognitive mode in the experiment.

Overall, I tested 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, I hypothesized 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.

To avoid making the task setup too cumbersome for participants, I also provided hints for how to use and interpret the ML models and explanation outputs (setup details in Section 4.3.4). Compared to a setup with no built-in interpretability options (control), I anticipated 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, I 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; 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, I measured the impact of bounded rationality on the decisions made by practitioners in the form of 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.

Prior work claims that ML practitioners overly trust and rely on static explanations from interpretability tools [18, 111]. This suggests that people might be applying incorrect heuristics that lead to satisficing in these cases. Therefore, I hypothesized that:

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

Table 4.1: An overview of the ten hypotheses corresponding to my dependent variables. Each dependent variable is split into two hypotheses, one for static visual explanations from interpretability tools and the other for interpretability tools with interactive features. The former represents a satisficing cognition mode and the latter, optimizing.

**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 [108]. It then 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 [10, 85, 186]. Therefore, I hypothesized 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 my study of bounded rationality in the ML and interpretability contexts. Table 4.1 presents a full list of my hypotheses. Before collecting any data, I pre-registered these hypotheses on AsPredicted.<sup>4</sup> I included other variables of interest for exploratory analysis in the pre-registration, such as usability of the task setup, validity of the dataset and model in the wild, etc. (Tables 4.2– 4.5 provide an overview of these).

## 4.3 Methods

I conducted a pre-registered controlled experiment with ML practitioners to study my hypotheses. The experiment was between-subjects, split across 5 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.

### 4.3.1 Experimental Setup

The experiment was conducted using a Qualtrics survey with links to Google Colab notebooks for access to the data, model, and interpretability tools. I designed the setup 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 [50, 183, 111]. On the one hand, studying an abstract construct via an experiment requires proxies that can

---

<sup>4</sup>An anonymized version of this pre-registration generated for peer review is available here: [https://aspredicted.org/462\\_XKP](https://aspredicted.org/462_XKP)

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, in the survey in Chapter III, I copied the visual outputs from interpretability tools within a Qualtrics survey. But, one limitation was that this did not allow for direct participant interaction with the data science setup.

For my research goals, access to a realistic setup and consistency of data logging were both critical (conflicting) requirements. To account for these, the setup employed both: a Qualtrics survey to consistently capture quantitative metrics for hypothesis testing, and a Colab notebook with the relevant ML components. This worked with all the 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. Screenshots of a Colab notebook are included in Appendix E.

### **4.3.2 Choice of Dataset**

Similar to the survey in Chapter III, I used the Adult Income dataset<sup>5</sup> with some modifications for the 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  $\leq \$50,000$  (converted to 0) or  $> \$50,000$  (converted

---

<sup>5</sup><https://archive.ics.uci.edu/ml/datasets/adult>

to 1). This threshold is equivalent to  $\sim$ \\$100,850 in 2022 when adjusted for inflation.

I 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 the previous study (Chapter III), I synthetically introduced errors in this dataset to quantitatively capture faulty reasoning about the data and model. I included two errors that commonly occur in people’s day-to-day ML work: missing values and redundant features [111]. To synthesize missing values, I 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, I 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).

### **4.3.3 Choice of Model and Interpretability Tools**

I selected ML models and interpretability tools based on their consistency in outputs and features. The 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. LightGBM model outputs also matched that of the XGBoost model used in the control condition. Additionally, the glassbox and blackbox conditions were consistent in their explanation types and features when compared to each other. More details on these selections are included below.

#### **4.3.3.1 Control Condition**

The control condition simulated a normal ML pipeline with no interpretability options. I relied on a boosted tree model using the XGBoost library.<sup>6</sup> 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. I included hints leading to

---

<sup>6</sup><https://xgboost.ai/>

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), I provided a link to the explainer,<sup>7</sup> 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.

#### 4.3.3.2 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. I considered several options for both types, eventually selecting the following implementations due to their underlying consistency (as noted in prior work [111]).

**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. I used the interpretML implementation of GAMs,<sup>8</sup> called Explainable Boosting Machines (EBMs) [167]. EBMs have built-in visualizations for global feature importances, partial dependence plots, and local explanations (see Figure 4.5).

---

<sup>7</sup>XGBoost explainer was originally developed for R (package and the corresponding blog post). The Python community replicated this functionality for their module.

<sup>8</sup><https://github.com/interpretml/interpret/>

**SHapley Additive exPlanations (SHAP).** SHAP is a post-hoc explanation approach for blackbox models [150]. It explains each prediction by assigning optimal credit to each input feature, using Shapley values from cooperative game theory [206, 222]. I used the SHAP Python package<sup>9</sup> 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 4.1). LightGBM<sup>10</sup> served as the underlying blackbox model explained by SHAP. It follows a highly optimized tree-based gradient boosting approach which makes training the model extremely fast [114]. Additionally, the SHAP implementation offers a separate, high-speed algorithm for tree ensemble methods like LightGBM.

#### 4.3.3.3 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. I 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).** I used Microsoft’s Explanation Dashboard<sup>11</sup> with EBMs for consistency with the 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

---

<sup>9</sup><https://github.com/slundberg/shap>

<sup>10</sup><https://lightgbm.readthedocs.io/en/v3.3.2/>

<sup>11</sup><https://github.com/microsoft/responsible-ai-toolbox/blob/main/docs/explanation-dashboard-README.md>



Figure 4.3: The landing interface of Microsoft’s Explanation Dashboard with minor edits to reduce whitespace. The dashboard includes several exploration sections. Here I 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.





Figure 4.4: 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 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 this case, I set up the tool to also provide 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).

Questions	Independent Variables
<ul style="list-style-type: none"> <li>- Use of ML in their daily job (scale 0–7)</li> <li>- Total time estimate for how long they have been practicing ML (in months)</li> </ul>	Prior Experience with ML
<ul style="list-style-type: none"> <li>- Familiarity with interpretability tools in general (scale 0–7)</li> <li>- 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 [111])</li> </ul>	Prior Experience with Interpretability Tools
<ul style="list-style-type: none"> <li>- Familiarity with model and interpretability option in the condition (scale 0–7)</li> <li>- Estimated hours spent using this model and interpretability option (categories based on an upward-facing parabola, same as above)</li> </ul>	Prior Experience with Task Setup

Table 4.2: Questions included under the **first** survey component on **setup familiarity**, and the corresponding independent variables.

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 4.3 presents a subset of these features.

**What-If Tool (WIT).** I used Google’s What-If Tool<sup>12</sup> with the same underlying elements as the 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 4.4 presents the datapoint editor tab for WIT.

#### 4.3.4 Components of the Survey

Tables 4.2– 4.5 provide an overview of the survey components in the order in which they were presented to the participants, along with the corresponding dependent and independent

<sup>12</sup><https://pair-code.github.io/what-if-tool/>

Multiple Choice Questions	Dependent Variables
<u>Global Feature Importance</u> : If you were forced to remove a feature from this model, which of the following features would you remove?	Time, Number of Hints Used, Accuracy, Response Type
<u>Partial Dependence for a Feature</u> : Which of the following ranges for Age values has the most likelihood of making a high income?	
<u>Predict the Outcome</u> : Given the following input feature values and importances, what do you think the model predicted for this individual and why?	
<u>Explain Misclassification</u> : The model misclassified this datapoint with the given input feature values. Why do you think that happened?	
<u>What-if Question</u> : A person with the following input features makes $\leq 50k$ income. A change to which of the following features would cause the prediction to become 1 (i.e., $>50k$ income)?	

Table 4.3: Multiple choice questions included under the **second** survey component on **the data and model**, and the corresponding dependent variables.

Questions	Dependent (in gray) and Independent Variables
- 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
Cognitive load using NASA-TLX questionnaire [1]	Cognitive Effort
Usability of setup using the SUPR-Q scale [199]	Usability Score

Table 4.4: Questions included under the **third** survey component on **high-level task evaluation**, and the corresponding dependent and independent variables.

Questions	Independent Variables
Multiple-answer questions with options representing established and intended capabilities of interpretability solutions (e.g., “individual instance explanations,” “counterfactuals,” “regions of error,” etc.); borrowed from [111]	Mental Model Accuracy
Extent to which they relied on the setup (scale 0–7)	Setup Reliance
Read additional documentation about the dataset, model, or the interpretability options included in the setup (binary)	Read Documentation
Wrote their own code in the Colab notebook (binary)	Wrote Code

Table 4.5: Questions included under the **fourth and fifth** survey components on **mental models and setup engagement**, and the corresponding independent variables.

variables. First, I provided a brief overview of task and setup, and obtained consent from participation. 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 Google 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, I asked them five multiple choice questions (MCQs) about the data and model; and high-level evaluation questions about using the data and model in real-world applications, and their confidence in the data being error-free and the outputs from their setup. Since this was the end of the data science element of the task, the questions immediately after it covered self-reports on cognitive load using the six-item NASA-TLX questionnaire [1] and the usability of the setup using the eight-item SUPR-Q scale [199].

By now, participants had used the task setup, so I next asked an exploratory multiple-answer question (same as Chapter III) to understand their mental model of the setup in their condition. I 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. I include additional relevant details on the main elements below.

#### 4.3.4.1 Multiple Choice Questions

The main portion of the survey consisted of a set of five MCQs about the data and model (Appendix F lists all MCQs along with their corresponding answer options and hints). These questions were related to the common ways in which ML practitioners use interpretability tools [85, 111] (Table 4.3). All MCQs had five choices based on the categories of the nominal dependent variable, response type (accurate, plausible, or randomly inaccurate), presented in a randomized order. The plausible choices represented bounded rationality. They were generated based on common types of heuristics such as the anchoring heuristic (wherein people make decisions based on the piece of information they notice first [43]) and the availability heuristic (wherein people make decisions based on incomplete information, using whatever comes to mind immediately [232]). Naturally, the MCQ choices I generated for these plausible answers were somewhat dependent on my own heuristics about income data. Of these, I picked the heuristics-based choices that were also mentioned most frequently in pilot studies with ML practitioners.

Therefore, the 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

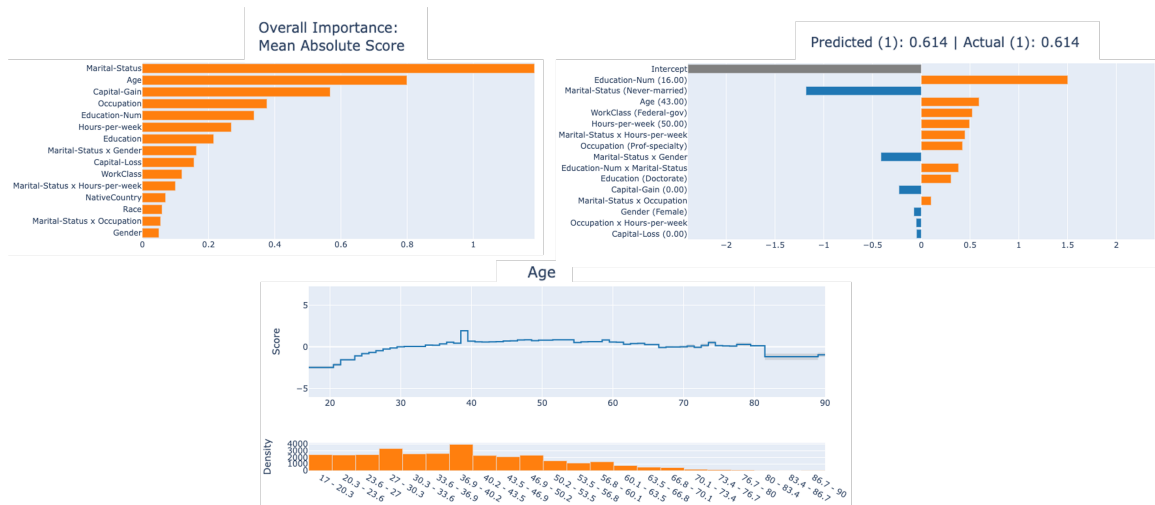


Figure 4.5: Visualizations output by interpretML’s implementation of GAMs, i.e., Explainable Boosting Machines. These ones are generated for the Adult Income dataset, same as the study task. Top (left to right): Overall feature importances (global explanation); Feature attributions for an individual datapoint (local explanation). Bottom: Partial dependence plot for a continuous input feature, age (global explanation).

participants could access by clicking a “Hint” button. I 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 the 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 4.5-top right). For the control condition, I 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.”

#### 4.3.4.2 High-level Task Evaluation

This 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. I also used self-reported ratings to establish participants’ confidence in their answers and their data science setup. Recall that the modified dataset included two errors, missing values and redundant features. I 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.

#### 4.3.5 Dependent and Independent Variables

The **dependent variables** were calculated based on people’s answers for the five MCQs about the data and model: (1) *time*, i.e., how long participants took to answer all five MCQs on average; (2) *cognitive effort*, i.e., total effort applied in answering questions, measured using the NASA-TLX questionnaire [1]; (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. Tables 4.2– 4.5 include a full list of all variables of interest.

#### 4.3.6 Analysis Methods

I used the pre-registered analysis methods for the dependent variables: one-way ANOVAs followed by TukeyHSD post-hoc tests for continuous variables and Chi-square test of inde-

pendence for the nominal variable. I also highlight descriptive statistics and results from exploratory multiple linear regressions and Pearson’s correlations for the independent variables. Several of these independent variables were calculated based on responses to a set of similar questions. I 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. I coded these responses using inductive thematic analysis [46]: open coding followed by axial coding, with themes generated via affinity diagramming the axial codes.

#### **4.3.7 Participants and Data**

I 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 the open-ended question, I was 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 ( $\sigma=1.7$ ; scale 0–7) and had practiced ML for 40.2 months on average ( $\sigma=22.65$ ). The most commonly listed job roles for them were data scientist, data analyst, ML researcher, and software developer. Participants were compensated with a \$25 Amazon gift card upon completion of the study. The study protocol was approved by IRB.

### **4.4 Results**

#### **4.4.1 Hypothesis Testing**

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



Dependent Variable	Control	GAM	SHAP	Explanation Dashboard (ED)	What-if Tool (WIT)	Pairwise Significance
<b>Time</b> (in minutes) F(4,114) = 4.15, p<0.01 partial $\eta^2$ =0.13	19.65 ± 35.68	3.90 ± 4.15	4.63 ± 5.74	3.25 ± 5.68	4.64 ± 4.66	Control > GAM* Control > SHAP* Control > ED* Control > WIT*
<b>Cognitive Effort</b> (1-7) F(4,114) = 14.62, p<0.001 partial $\eta^2$ =0.34	4.16 ± 0.77	3.29 ± 0.96	3.49 ± 0.62	4.47 ± 0.86	4.72 ± 0.67	Control > GAM** Control > SHAP* ED > GAM*** ED > SHAP*** WIT > GAM*** WIT > SHAP***
<b>Hints</b> (0-5) F(4,114) = 14.45, p<0.001 partial $\eta^2$ =0.34	3.08 ± 1.12	1.42 ± 1.06	1.48 ± 1.66	3.18 ± 1.30	3.52 ± 1.24	Control > GAM*** Control > SHAP*** ED > GAM*** ED > SHAP*** WIT > GAM*** WIT > SHAP***
<b>Accuracy</b> (0-100) F(4,114) = 3.92, p<0.01 partial $\eta^2$ =0.12	59.20 ± 15.79	44.17 ± 19.54	42.40 ± 15.62	40.91 ± 22.66	41.74 ± 21.67	Control > GAM* Control > SHAP* Control > ED** Control > WIT*

Table 4.6: The results of my 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  $\eta^2$  value for effect size; suggested norms: small = 0.01, medium = 0.06, large = 0.14. All of the dependent variables show a large effect size.

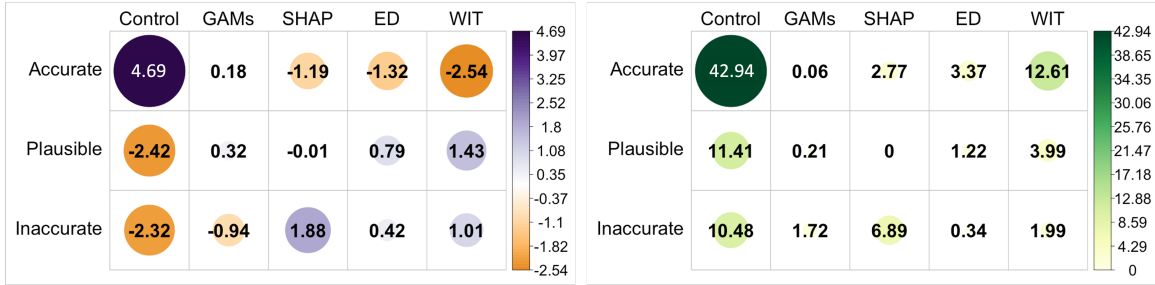


Figure 4.6: Results of my pre-registered Chi-square test for the nominal dependent variable, response type. Left: Residuals from the test, indicating directionality and effect of each response type-condition combination for the magnitude of the resulting chi-square statistic. Right: contribution of each 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
<b>Total</b>	125	120	125	110	115	595

Table 4.7: Contingency table with counts for response types—accurate, plausible, and randomly inaccurate—for all five conditions.

Table 4.6 presents the means, standard deviations, one-way ANOVA and TukeyHSD results for significance testing for the continuous dependent variables (time, cognitive effort, hints, and accuracy). Further, a Chi-square test indicates that the 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 4.6 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 4.7 provides exact counts for each response type per condition in a contingency table.

As hypothesized, participants spend significantly less time and cognitive effort when the setup provides visuals from GAMs and SHAP, i.e., the visuals-only conditions (**H1-2a**). They also use significantly fewer hints when they have access to these visuals (**H3a**). This supports my 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

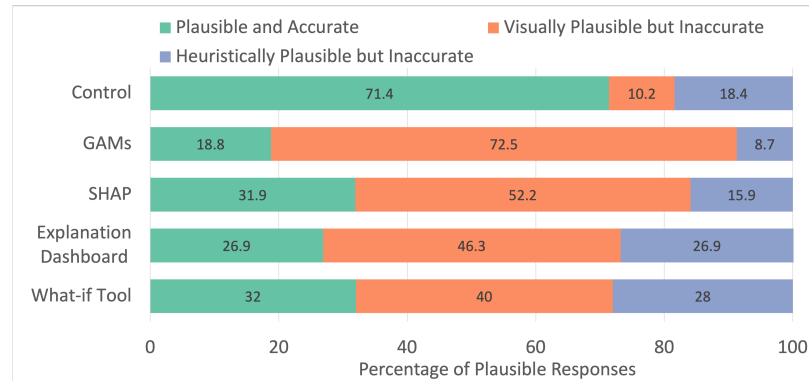


Figure 4.7: Fine-grained breakdown of the type of responses selected under the plausible response type: plausible and accurate, visually plausible but inaccurate, and heuristically plausible but inaccurate. The percentages are calculated using the raw plausible response numbers in Table 4.7.

select plausible response types when answering questions about the data (**H4a**). Table 4.7 presents the numbers for each high-level response type: accurate, plausible, or inaccurate.

As I noted in the setup details, 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 4.7 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 one 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**).

I 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 (H1-5b). However, I do not find adequate support for this in the data. There were no 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 bad outcomes when compared to the control condition. While I had hoped that interactivity would make these tools more engaging, I instead noted

that participants spend significantly more cognitive effort in navigating these tools' features and need 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, I find evidence in support of all of the satisficing-related hypotheses for the visuals-only conditions and none for the optimizing-related hypotheses corresponding to the interactive tools conditions (Table 4.1). People satisfice—resulting in significantly low accuracy—when using *any* interpretability option.

#### **4.4.2 Perceptions of the Setup**

Figure 4.8 presents descriptive statistics for the independent variables for which I observed noteworthy differences (descriptive statistics for **all** independent variables are included as Appendix G). These numbers reaffirm that, compared to the control condition, visual explanations from interpretability tools lead to satisficing and interactivity does not help. Participants in the control condition were more conservative with their ratings for hypothetical use, and 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 the open-ended question was that these tools include a plethora of information and features, both presented in a way that was “difficult to understand”, “stressful”, and “burdensome”.

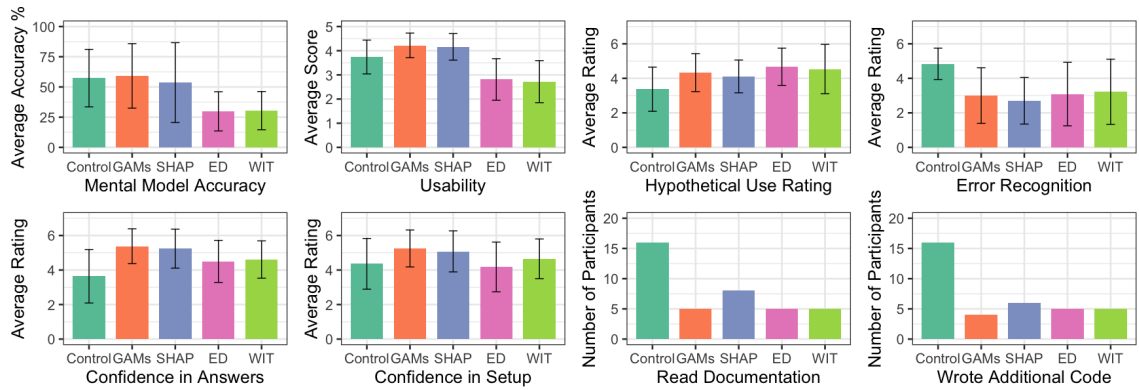


Figure 4.8: Descriptive statistics for independent variables with noteworthy differences.

*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. (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 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. They

were also more confident in their understanding of the setup and their answers for the task, despite not having read much documentation on the task setup or writing code in addition to what was provided. This fast and limited information processing under satisficing might also explain why prior work has found static interpretability tools (e.g., GAMs, SHAP, LIME) to often be misused and over-trusted.

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 visuals-based judgements to be made. It required some cognitive effort, but not as high as the interactive tools. I had provided some links to modules that might help with understanding the model outputs, but no actual code was included. The numbers from Figure 4.8-bottom row suggest that, in having to read documentation and write or debug code, participants remained engaged enough to carefully look at any outputs.

### **4.4.3 Exploratory Analyses**

Now that I have established that people satisfice when using any interpretability options, I consider the question: what kind of internal or external factors can affect satisficing? I 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.

#### **4.4.3.1 Relationship between cognitive factors and satisficing**

Results from fitting multiple linear regression (MLR) models do not show any predictive relationship between the 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.

I 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 as significantly less likely to select the visually plausible but inaccurate type of plausible answers ( $F(6, 112)=2.99, p < 0.01, \text{Adj-}R^2 = 0.092$ ).<sup>13</sup> 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. I 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, 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.

---

<sup>13</sup>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 ( $b=-0.17, t(112)=-2.47, p < 0.01$ ).

#### 4.4.3.2 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.<sup>14</sup> 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 one see with the 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.<sup>15</sup> 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 4.7 and Figure 4.7 support this hypothesis.

For the symmetrical contextual factors, I 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.<sup>16</sup> Even when people satisfice and select plausible responses, these factors co-occur with plausible and accurate response types rather than the plausible

---

<sup>14</sup>Significant predictive relationships between usability and two dependent variables: (1) cognitive effort ( $F(4, 112)=5.54, p < 0.001, \text{Adj-}R^2 = 0.14; b=-0.36, t(112)=-4.13, p < 0.001$ ), and (2) the number of hints used ( $F(4, 112)=6.02, p < 0.001, \text{Adj-}R^2 = 0.15; b=-0.58, t(112)=-4.03, p < 0.001$ ).

<sup>15</sup>Significant predictive relationships between usability and two response types: (1) plausible and accurate responses ( $F(1, 115)=3.85, p < 0.05, \text{Adj-}R^2 = 0.02; b=-0.17, t(115)=-1.96, p < 0.05$ ); and (2) heuristically plausible but inaccurate ( $F(1, 115)=7.79, p < 0.01, \text{Adj-}R^2 = 0.06; b=-0.19, t(115)=-2.79, p < 0.001$ ).

<sup>16</sup>Significant correlations between accuracy and people's: (1) confidence in their answers to the 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 ( $r(117)=-0.28, p < 0.001$ ).



but inaccurate types.<sup>17</sup> That is, people with these attributes seem to either select accurate responses (maximizing) or plausible and accurate ones (optimizing).

Overall, I 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.

#### **4.4.4 Summary of Results**

Results from a controlled experiment show that people satisfice on a 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 these tools cognitively burdensome to use. Exploratory analyses indicate that cognitive factors (e.g., prior experience in ML and interpretability) have no bearing on the likelihood of satisficing. Instead, 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).

## **4.5 Discussion**

The core tenet of interpretability is “to explain or to present in understandable terms to a human.” [54, 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

---

<sup>17</sup>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 ( $r(117)=-0.27, p < 0.01$ ).

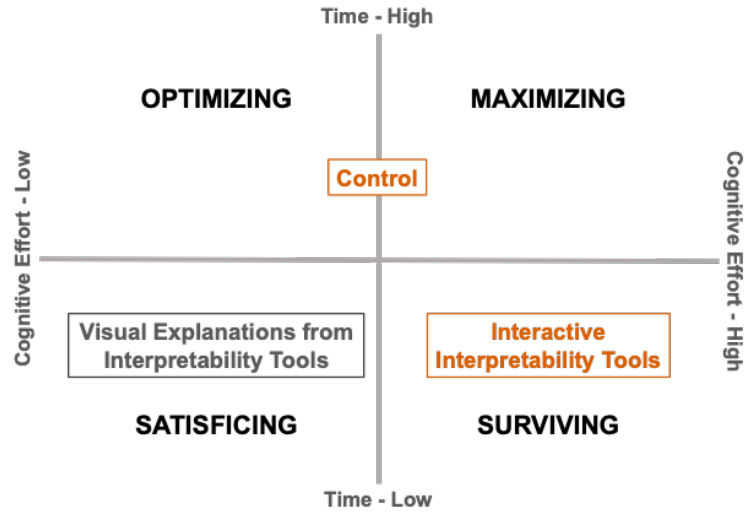


Figure 4.9: Updated overview of the study conditions using my proposed framework for cognition in Machine Learning settings. The unexpected results are highlighted in color.

on the “in understandable terms to a human” front [18, 50, 111, 183]. 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 my study presented here. 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 [16, 158, 161].

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 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 [209]. It asserts—and this study has confirmed it for the ML setting—that people do not process or evaluate information in a perfectly rational way. It does not examine how people satisfice, only that they do, and that they do it irrespective of the setting. How much they satisfice might vary depending on the criticality of the task or the type of information available,

but people always satisfice. 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.

I 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?

#### **4.5.1 Implications for Design**

**Interactivity does more harm than good in this setting.** In theory, interactivity should help improve engagement with a system [226]. Keeping people engaged should, in turn, have a higher likelihood of preventing heuristics-based, automated information processing [225, 173]. None of this is true for how interactivity is incorporated in existing interpretability tools [57]. Interactive features complicate the tool 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 [207]: the additional information and reactive interface elements do not add value corresponding to the effort needed to adequately understand them. Indeed, I believe the interactive tools conditions, as currently designed, end up falling under the *surviving* cognitive mode rather than my hypothesized *optimizing* one (Figure 4.9).

**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, has no bearing on their likelihood of satisficing. Mental model accuracy also has

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. 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, I believe, would result in the optimizing cognitive mode. My hypothesis here is grounded in the results for the control condition (Figure 4.9). Control was intended to be the most challenging condition; to avoid making it too cumbersome for a study, I included links to some helpful modules in the setup. It seems that this struck that balance between too helpful (satisficing) or too frustrating (surviving), and increased people’s likelihood of reading documentation, writing code, and staying engaged 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 [71, 93, 194, 245], or using surprise (e.g., Bayesian surprise that models information entropy [17, 22, 100]) and emotion [42, 121] 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 seamless, simplicity, consistency in features and interactions, and ease of use [41, 98, 239].

Balancing efficiency and ease-of-use with emphasizing negative aspects is known to be an effective strategy for sensemaking in both human-human and human-machine contexts. Indeed, organizations operating in critical domains employ red teaming and adversarial design to simulate failure situations for testing the limits of their human and machine operations [248, 249]. This class of solutions is virtually unexplored in the ML setting. I hope future work will incorporate these ideas towards designing more human (cognition)-centered interpretability and explainability.

## **4.6 Limitations**

Bounded rationality is an abstract construct. To study it, I made assumptions about proxies for bounded rationality and how I could operationalize it for the ML context. Although these assumptions were grounded in theory, I 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. I had to enforce consistency on an otherwise personalized construct; 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 I designed and explained the study and setup, rather than the specific outputs and interactivity of the interpretability tools. Although I believe this to not be the

case given the positive comments about the ease of the Colab + Qualtrics combination from the participants, I 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. 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, I had to find a balance between consistency and rigidity for the study (i.e., balance ecological and internal validity). Indeed, the generalizability of these results is dependent on how participants used my 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. The choice of task and questions also likely affected our results. While most ML practitioners have at some point conducted similar exploratory data analysis to what I asked for in the study, the majority of the results were calculated based on asking them very specific MCQs during and after their exploration. Although the study design decisions I made were necessary for a controlled experiment, the 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 I hoped with the optimizing quadrant of my bounded rationality framework. Adding monetary incentives based on performance might improve performance as well, although I believe these need to be quite significant if the participants involved are data scientists 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. [236] 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.

As researchers conduct more studies on interpretability and explainability using controlled experimental setups, consistency in the metrics used to evaluate these approaches is critical. I 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 [84, 88, 89, 260]. 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. Moreover, these results are specific to a particular stakeholder in one ML setting. The role of bounded rationality in human-machine interaction more generally is likely quite varied.

## 4.7 Conclusion

Interpretability and explainability are purported to be solutions for helping people understand ML models and their predictions. In this chapter, I presented 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. I had hypothesized that interactivity might prevent satisficing but did not find this to be true. Rather, interactivity increases cognitive burden to the extent that people satisfice—seemingly out of frustration—instead. I 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. Overall, this work investigates the role of human cognition and the bounded nature of human rationality in how a key stakeholder uses interpretability tools. The evidence from the experiment validates the significant role of cognitive frameworks in understanding the use of interpretability tools.

## CHAPTER V

# Sensible AI: Re-Imagining Interpretability and Explainability using Sensemaking Theory

With ML-based systems being deployed in the wild, it's imperative that all stakeholders of these systems have some understanding of how the underlying ML model works. From the experts who develop algorithms to practitioners who design and deploy ML-based systems, and end-users who ultimately interact with these systems—stakeholders require varying levels of understanding of ML to ensure that these systems are used responsibly. Approaches like interpretability and explainability have been proposed as a way to bridge the gap between ML models and human understanding [256, 150, 39]. Tools that implement these approaches have also been made available for public use. In light of this, recent work in HCI has evaluated the efficacy of these tools in helping people understand ML models. These findings suggest that ML practitioners [111] and end-users [18, 120] are not always able to make accurate judgments about the model, even with the help of explanations. In fact, having access to these tools often leads to over-trust in the ML models. Ultimately, noting that interpretability and explainability are meant for the stakeholders, recent work has proposed design guidelines for explanations based on research in the social sciences about how people explain things to each other [163? ]. Taking a human-centered or a model-centered approach, this prior work seeks to answer: *what are the characteristics of an explanation that can help people understand ML models?*



Consider a real-world setting. Imagine you are a doctor in a healthcare organization that has decided to use a ML-based decision-support software to help with medical diagnosis. The system takes as input information about patients' symptoms, demographics, family history, etc., and returns a predicted diagnosis. Naturally, you want to be able to overview why the software predicted a certain diagnosis before you suggest treatment based on its prediction. Further, you want to be able to explain to the patient why you (did not) trust and follow the predicted diagnosis. To aid with this, the software provider gives you access to an explanation system (e.g., LIME [190], SHAP [150]) which shows: (1) a local explanation (e.g., a bar chart) of the input features that were most important for the diagnosis made for a specific patient, (2) a global explanation for the features that are usually important to the model when making a prediction, and (3) an overview of each feature's relationship with the output classes. The explanation system also includes interactive elements so you can ask "what if" questions based on different combinations of input features.

Is this enough to ensure that the ML-based decision-support software can be reliably used by the doctor? I claim that the answer to this question is no. This chapter makes the argument that current interpretability and explainability solutions will always fall short of the task of helping people understand ML models due to their focus on designing better explanations—in other words, improving an artifact. For example, while the explanation might communicate the symptoms that were important to the model's prediction (i.e., a local explanation), it does not tell the doctor to be cautious that the patient's other symptoms are fluctuating, that the patient belongs to a sub-group for which the model has limited training data, or that the nurses have noticed other relevant symptoms in the visiting family. From the patient's perspective, the explanation does not convey why, for example, their own fear of having a particular disease (after an online symptom search or from rare family history) is unwarranted in this instance. These factors, that have little to do with the particular explanation or artifact, can alter the stakeholders' decision-making in significant ways. Here, *I propose a specific theoretical framework to shift from improving the artifact (e.g., an*

*explanation or explanation system) to understanding how humans make sense of complex, and sometimes conflicting, information.* Recent work supports this shift from *what* an explanation should look like to *who* it is intended for. Properties of the *who* such as, prior experience with AI and ML [61], attitude towards AI (e.g., algorithmic aversion [35, 51]), the socio-organizational context [60], have been observed as being critical to understanding AI and ML outputs. I extend this work by providing a framework for *how* to incorporate human-centered cognitive principles to interpretability and explainability.

I present Weick’s sensemaking as a framework for envisioning the needs of people in the human-machine context. Weick describes sensemaking as, quite literally, “the making of sense,” or “a developing set of ideas with explanatory possibilities” [247]. Although Weick’s definition is similar to that of prior work in HCI and information retrieval, the two deviate in their goals; the latter defines sensemaking for the purpose of finding representations that enable information foraging and question-answering [179, 197]. Weick’s sensemaking is more procedural: “placement of items into frameworks, comprehending, redressing surprise, constructing meaning, interacting in pursuit of mutual understanding, and patterning” [247, p6]. These processes are influenced by one’s identity, environment, social, and organizational context—Weick expands these into the seven properties of sensemaking (Figure 5.1-Right). For example, for the doctor trying to diagnose a patient with the help of an ML-based system (with explanations), their understanding of the predicted diagnosis can be influenced by questions such as, have they recently diagnosed another patient with similar symptoms; is the patient’s care team in agreement on a diagnosis; is the predicted diagnosis plausible; and, which symptoms are more visible and does the explanation present these as important to the prediction. The seven properties of sensemaking are a framework for identifying and understanding these contextual factors. I also discuss how this sensemaking perspective enables a new path forward in designing interpretability and explainability tools, grounded in prior work on how organizations have handled human sensemaking.

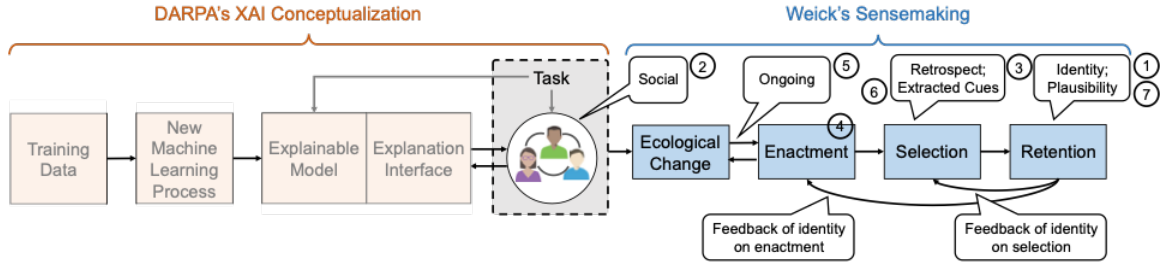


Figure 5.1: Left: DARPA’s conceptualization of Explainable AI, adapted from [72]. Right: Weick’s sensemaking properties (1–7) categorized using the high-level Enactment-Selection-Retention organizational model, adapted from [103]. Enactment includes properties about perceiving and acting on environmental changes; Selection, properties related to interpreting what the changes mean; and Retention, properties that describe storing and using prior experiences [122]. My proposed Sensible AI framework extends the existing definition of interpretability and explainability to include Weick’s sensemaking properties.

## 5.1 Sensemaking

Sensemaking describes a framework for the factors that influence human understanding: “the sensemaking perspective is a frame of mind about frames of mind” [247, pxii]. It is most prominent in discrepant or surprising events. We try to put stimuli into frameworks particularly when predictions or expectations break down. That is, when people come across new or unexpected information, they like to add structure to this unknown. The process by which they do this, why they do it, and how it affects them and their understanding of the world are all central to sensemaking.

Sensemaking subsumes interpretability<sup>1</sup>. They share the same goals: both seek to understand an output (an outcome or experience, or an ML model) before using it to make decisions. If a ML-based system could explain its predictions, we can verify whether the reasoning is sound based on some auxiliary criteria (e.g., safety, nondiscrimination, etc.), and determine whether the system meets other desiderata such as fairness, reliability, causality, and trust [54, 142]. Sensemaking includes all of this and more. In describing how people understand something, sensemaking not only considers the information being

<sup>1</sup>As in previous chapters, I use the term “interpretability” to represent both interpretability and explainability approaches. I follow similar terminology choices with ML- (rather than AI-) based systems since interpretability is attributed to ML models.

<b>Property</b>	<b>Weick’s Description for the Human-Human Context</b>	<b>Proposed Claim for the Human-Machine Context</b>
Grounded in Identity Construction	“Sensemaking is a question about who I am as indicated by the discovery of how and what I think.”	Given multiple explanations, people will internalize the one(s) that support their identity in positive ways.
Social	“What I say and single out and conclude are determined by who socialized me and how I was socialized, as well as by the audience I anticipate will audit the conclusions I reach.”	Differences in micro- and macro-social settings affect the effectiveness of explanations.
Retrospective	“To learn what I think, I look back over what I said earlier.”	Providing explanations before people can reflect on the model and its predictions negatively affects sensemaking.
Enactive of Sensible Environments	“I create the object to be seen and inspected when I say or do something.”	The order in which explanations are seen affects how people understand a model and its predictions.
Ongoing	“My talking is spread across time, competes for attention with other ongoing projects, and is reflected on after it is finished, which means my interests may already have changed.”	The valence and magnitude of emotion caused by an interruption during the process of understanding explanations from interpretability tools change what is understood.
Focused on and by Extracted Cues	“The ‘what’ that I single out and embellish as the content of sensemaking is only a small portion of the utterance that becomes salient because of context and personal dispositions.”	Highlighting different parts of explanations will lead to varying understanding of the underlying data and model.
Driven by Plausibility rather than Accuracy	“I need to know enough about what I think to get on with my projects, but no more; sufficiency and plausibility take precedence over accuracy.”	Given plausible explanations for a prediction, people are not inclined to search for the accurate one amongst these.

Table 5.1: Weick’s basic descriptions of the properties of sensemaking for the human-human context [247, pp.61-62], and my proposed claims from translating these properties to the human-machine context.

presented to the person doing the meaning-making, but also additional contextual nuances that affect whether and how this information is internalized. This includes factors such as, the enacted environment, the individual’s identity, their social and organizational networks, prior experiences with similar information, etc.

In the subsections that follow, I describe the seven properties of sensemaking in the human-human context and translate them for the human-machine context. Table 5.1 presents an overview of these properties with Weick’s description for the human-human context and my proposed claims from translating these for the human-machine context. To concretize how these properties might affect stakeholders of ML-based systems, I present an example user vignette for each property. Prior work has proposed and applied similar methodology when translating theory [159, 6]. While the examples are based on popular press articles

and research papers about ML-based systems, they are not intended as being representative of these cases. I use them to highlight a sensemaking property, but do not claim that the property has a causal relationship with the example, i.e., there could be other reasons for why the ML-based systems functioned the way that they are described in these articles.

### **5.1.1 Grounded in Identity Construction**

Identity is critical for AI/ML sensemaking because people will only understand aspects of these systems that are congruent with their existing beliefs or those which update their beliefs in ways that shed a positive light on them.

#### **5.1.1.1 Identity Construction in the Human-Human Context**

Sensemaking begins with the sensemaker. In this way, sensemaking is innately human-centered: “how can *I* know what *I* think until *I* see what *I* say?” [247, p.18]. Sensemaking is grounded in the individual’s need to have a clear sense of identity. People make sense of something to either support their existing beliefs or update them in situations when applying their beliefs leads to a breakdown in their understanding. There are five things about identity construction that are relevant for sensemaking, according to Weick [247, pp.23-24]: (1) controlled, intentional sensemaking is triggered by a failure to confirm one’s self; (2) sensemaking is grounded in the desire to maintain a consistent, positive self-conception; (3) people learn about their identities by projecting them into an environment—which includes their social, organizational, and cultural contexts—and observing the consequences; (4) sensemaking by way of identity construction is a mix of proaction and reaction; and (5) despite the importance of the environment, sensemaking is self-referential in that the self is what ultimately needs interpreting—what a given situation means is defined by the identity that an individual relies on while dealing with it.

The relationship between identity and sensemaking is not limited to the individual sensemaker. The influence of social and environmental context becomes apparent as we consider

how identity is constructed. Weick describes this influence using three definitions of identity from the social sciences. First, Mead's claim that the mind and self are developed based on the communicative processes among people (i.e., social behaviorism). Each individual is comprised of "a parliament of selves" which reflect the individual's various social contexts [157]. Second, Knorr-Cetina's inclusion of social contexts based on the larger tapestry of social, organizational, and cultural norms, i.e., the macro-social [119]. Finally, Erez and Earley's three self-derived needs that shape identity, which include intrapersonal and interpersonal dynamics. These are: "(1) the need for *self-enhancement*, as reflected in seeking and maintaining a positive cognitive and affective state about the self; (2) the *self-efficacy* motive, which is the desire to perceive oneself and competent and efficacious; and (3) the need for *self-consistency*, which is the desire to sense and experience coherence and continuity" [62, p.28].

What ultimately makes sensemaking challenging with respect to identity is that the more identities that an individual has, the more ways in which they can assign meaning to something. Given the fluidity of identity construction, it is more commonly the case that people have to grapple with several, sometimes contradicting, ways of understanding. Sometimes, this flexibility and adaptability in one's identity can lead to better outcomes, for example, when a manager is brainstorming solutions for a personnel problem. However, in most cases, dealing with this identity-based equivocality can lead to confusion, cognitive burden, and, in turn, lead people towards heuristics-based understanding [188].

#### **5.1.1.2 Identity Construction in the Human-Machine Context**

Consider Platform X, a popular social media site which uses a ML model for content moderation, with two stakeholders in mind. First, Sharon, a 42 year old conservative in the U.S. who identifies as a Republican, and is against vaccination for COVID-19. Her recent posts include graphic descriptions and images of, what she claims, are the potential side-effects of getting vaccinated. Several of these posts have been flagged for removal by Platform X's content moderation ML model. Second, Avery, a 37 year old doctor

who believes it is their responsibility to share unfiltered information about the COVID-19 pandemic, including the efficacy of the vaccines. Several of their posts highlight the positives of getting vaccinated, and some of them present the rare potential side-effects that have been noted by medical professionals.

It is not uncommon for platforms to offer minimal explanation(s) for post removal. Sharon and Avery might simply be told that their post violated Platform X's content policy. With interpretability tools, we can support richer explanations. Based on the local feature importances output by an interpretability tool, Sharon is told that her post was removed due to its content type, the number of her previously flagged posts, her predicted political affiliation based on her posting history, and the topic being COVID-19. She might immediately latch on to the predicted political affiliation as *the* explanation, and not concern herself with understanding the removal any further (i.e., sensemaking is not triggered because her identity remains intact). For Avery, who simply wants to share all relevant information in line with their identity as a doctor, the post removal might attack their desire to maintain a positive self-conception, i.e., their needs for self-enhancement, self-efficacy, and self-consistency. When provided a more detailed explanation (let's assume it's the same one as Sharon's), they might assume that the content type being graphic is the main reason. This would support their positive self-conception, but effectively not require them to understand the model's reasoning any further.

Interpretability tools are designed to present information in a context-free, unbiased way. People, on the other hand, rarely internalize information in this static way. Every new piece of information has the potential to connect with or build upon different aspects of a person's identity. Weick argues that whether or how people internalize the contents of an explanation is dependent on their identity as an individual and as a part of their varying social contexts.

*Claim: Given multiple explanations, people will internalize the one(s) that support their identity in positive ways.*

## 5.1.2 Social

Social context is critical for AI/ML sensemaking because it represents the audience-oriented external factors that shape people’s understanding of the outputs of these systems.

### 5.1.2.1 Social Elements of Sensemaking in the Human-Human Context

Sensemaking describes human cognition. This might give it the appearance of being about the individual, but it is not. Weick notes the work on socially shared cognition (e.g., [189, 135]) which shows that human cognition and social functioning are essential to each other. Specifically, an individual’s conduct is dependent on their audience, whether this is an imagined, implied, or a physically present one [7, 32]. Regarding the lack of a need for a physically present audience, recall Weick’s reference to Mead’s work on the individual being “a parliament of selves” [157] (see Section 5.1.1 for details on socially-grounded identity construction).

While social sensemaking entails an audience—imagined, implied, or physically present—it does not require there to be a shared understanding between the individual and their audience(s) or even among the audience(s). Indeed, the social aspects of human cognition can be described as “socially shared cognition [189, 134], sociocognition [135]; situated cognition; shared reality [64]; naturalistic social cognition [96, 97]; group cognition; contextualized cognition; social cognition [95]; shared mental models [45]; team mental models [118]; distributed cognition [94]; the social science of cognition [82]; and collective identity [31]” [231, p3] (see [231] for a comprehensive review of these terms). Given this diversity in the form that social elements can take, Weick proposes that social sensemaking be equated with alignment instead of shared meaning since “alignment is no less social than is sharing [meaning], but it does suggest a more varied set of inputs and practices in sensemaking than does sharing” [247, p43].

A focus on social aspects of sensemaking naturally implies that modes of communication (e.g., speech, discourse) and tools that support these also get attention, since these



represent the ways in which social contact is mediated. Weick describes their importance on three levels, which exist beyond the individual: (1) inter-subjective, which represents conversations with others that can lead to alignment; (2) generic subjective, which represents socially-established norms (e.g., within organizations) when alignment has been achieved; and (3) extra-subjective, which represents culturally-established norms that do not necessarily require communication anymore. As we go from inter- to extra-subjective, the role of the implied or invisible audience becomes increasingly prominent. This, in turn, shapes the modes and tools of communication necessary for sensemaking.

### **5.1.2.2 Social Elements of Sensemaking in the Human-Machine Context**

As noted above, social elements of sensemaking are quite varied. Let's consider the macrosocial, i.e., generic- and extra-subjective outcomes. Most ML-based systems are grounded in specific datasets and practices of a particular organization or culture. When extended to a social context that is different, there is a high likelihood that these systems might break down, even if the model itself is generalizable.

Consider the model developed for predicting diabetic retinopathy (DR) based on healthcare data (predominantly eye fundus photos) collected in the US [20]. US healthcare system is consistent across organizations—low variability in input features makes it easy to develop and consistently use the model to screen for DR. When the same model was applied to a cultural context where the healthcare system is more varied—in Thailand, where healthcare is dependent on individual providers and patient needs in different regions—it failed in unanticipated ways.

First, there is the issue with the data itself. Several countries in Southeast Asia, including Thailand, do not have dedicated rooms for capturing fundus photos, making the photos inconsistent in opacity and leading to potentially inaccurate predictions. Second, there are established practices around the results of a DR screening test to consider. While it is fairly common to receive results immediately in healthcare systems like the US, it is less common

in countries with fewer technicians, doctors, and specialists. In Thailand, one might not expect to see the results for several weeks. As a result, a patient who is anticipating their DR result 4-5 weeks later might not have budgeted time to travel to a bigger hospital farther away, based on a referral on the same day as the DR screening visit.

While interpretability tools may offer an explanation for a prediction, these explanations are limited to the model and the training dataset. Weick's perspective suggests that these might not be enough to understand the prediction, due to the variability in people's social contexts which are also at play when using predictions in real-world settings.

*Claim: Differences in micro- and macro-social contexts affect the effectiveness of explanations.*

### **5.1.3 Retrospective**

Retrospection or reflective thinking is critical for AI/ML sensemaking because it engages people in deliberately thinking about diverse interpretations of outputs when trying to understand these systems, instead of following the more automated, heuristics-based reasoning pathways.

#### **5.1.3.1 Retrospective Sensemaking in the Human-Human Context**

Sensemaking is, by default, retrospective because the object of sensemaking is a *lived experience or outcome*. Weick describes the retrospective nature of sensemaking as the most important, but perhaps the least noticeable, property. The reason it so frequently goes unnoticed is because of how embedded retrospective thinking is in the sensemaking process. Retrospective sensemaking is derived from the work of Schutz, who believes that meaning is “merely an operation of intentionality, which...only becomes visible to the reflective glance” [202]. He claims that experiences are only meaningful, i.e., considered for sensemaking, when they are “lived experiences,” the key word being *lived* [201].

While experiences, events, or situations being considered for sensemaking are discrete

segments of time, they are best processed retrospectively, separate from their occurrence. This reflective process is seamless, to the point where we do not even notice it. Time not only “cover[s] appreciable intervals—a minute, a day, a year” but also the less obvious intervals such as a fraction of a second. These are “so much with us that we almost fail to notice [them] consciously...[for example] as I begin the latter portion of a long word, my utterance of the first part is already in the past” [76, p44].

Having established lived experiences as the objects of sensemaking, let us consider the process by which meaning is imparted to these experiences. The reflective process starts with an individual’s present circumstances, and those shape the past experiences selected for sensemaking. Reflection happens in the form of a cone of light that starts with the present and spreads backwards. In this way, the cues of the past lived experience that are paid attention to for sensemaking depend on how the present is shaped.

The challenge lies in *which* present to consider. People typically have several things on their mind at the same time, be it multiple projects at work or personal goals. As a result, they have a multitude of lenses that they could apply for the reflective sensemaking process—the object of their sensemaking thus becomes equivocal. When dealing with equivocality, people are already overwhelmed with information and providing more details is often not helpful. “Instead, they need values, priorities, and clarity about preferences to help them be clear about which projects matter” [247, pp27-28]. In looking for clarity on which meaning to select, people are prone to a hindsight bias [221]. They select the meaning with the most plausible story of causality for the outcome that they are trying to explain (I also discuss this in Section 5.1.7 which describes another property of sensemaking: being driven by plausibility over accuracy).

### **5.1.3.2 Retrospective Sensemaking in the Human-Machine Context**

In the situation where someone is trying to understand a ML model or a specific prediction, the model or the prediction are the “lived experiences.” Interpretability tools are

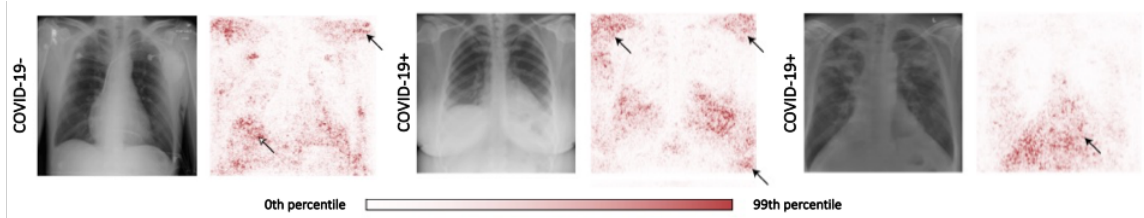


Figure 5.2: Saliency maps for chest radiographs, adapted from [48].

intended to aid with this task. For example, consider a radiologist who is tasked with reading chest radiographs to determine if a patient has COVID-19. Their hospital has purchased an ML-based image classification software to help with this task. The software also provides saliency maps (an interpretability approach) to help the radiologist determine if the ML model’s prediction for the MRI scan makes sense.

By immediately providing an explanation, the interpretability tool effectively disengages the retrospective process that helps with sensemaking. Figure 5.2 shows example explanations provided to the radiologist. As described in the accompanying research paper [48], these explanations show that the ML model sometimes relies on laterality markers to make the prediction. For example, in Figure 5.2, the saliency maps highlight not only the relevant regions in the lungs as being predictive, but also some areas (see pointers) that differ based on how the radiograph was taken. These, coincidentally, are also predictive of COVID-19 positive vs. negative results, leading to a spurious correlation.

Ideally, the radiologist evaluating the saliency map would be able to reach the same conclusion regarding these spurious correlations. However, the retrospective property would suggest that by providing this explanation without asking the radiologist to first think about what the explanation could be, the interpretability tool disengages their retrospective sensemaking process. This makes it easier for the radiologist to craft a plausible narrative that agrees with the model’s prediction instead of analyzing the radiograph in detail and accurately understanding the model.

Weick cautions against immediately providing additional information in sensemaking situations. When trying to understand something, people engage in a reflective process,

using their present to understand the outcome at hand. This reflective process helps them generate equivocal meanings (i.e., diverse interpretations). When they immediately have the explanation, there is no cognitive need for the radiologist to understand the intricacies of the model, which increases the likelihood of them missing the issues with the model. Prior work on stakeholders' use of interpretability tool corroborates this perspective: people expect far more from interpretability tools than their actual capabilities and, in doing so, often end up over-trusting and misusing them [18, 111].

*Claim: Providing explanations before people can reflect on the model and its predictions negatively affects sensemaking.*

#### **5.1.4 Enactive of Sensible Environments**

Enactment is critical for AI/ML sensemaking because it represents how (much) people understand these systems via an iterative process—it frames the parts of these systems that people understand and build on, over time. Enactment is the sense in which people create the environment they are trying to understand.

##### **5.1.4.1 Enactment in the Human-Human Context**

When we are tasked with making sense of something (an outcome, event, or situation), it might appear to belong to an external environment that we must observe and understand. Weick argues that this is not the case, that sensemaking works such that “people often produce part of the environment they face” [247, p30]. It is not just the person, rather, the person and their enacted environment that is the unit of analysis for sensemaking [181].

This environment that provides the necessary context for sensemaking is not a monolithic, fixed environment that exists external to people. Rather, people act, and their actions shape the environmental context needed for sensemaking: “they act, and in doing so create the materials that become the constraints and opportunities they face.” [247, p31]. Weick's understanding of enacted environments is influenced by Follett, who claims that there is no

subject or object when it comes to the activity of meaning-making. There is no meaning that one understands as the “result of the process;” there is just a “moment in process” [66, p60]. As such, this meaning is inherently contextual in that it is shaped by the cycle of action-enaction between the human and their environment.

Weick cautions against two things when it comes to the enactive nature of sensemaking. First, to not restrict our definition of action in shaping our environment. Action here could mean creating, reflecting, or interpreting: “the idea that action can be inhibited, abandoned, checked, or redirected, as well as expressed, suggests that there are many ways in which action can affect meaning other than by producing visible consequences in the world” ([26] as described by Weick [247, p37]). Second, the enacted environments do not need to embody existing ones. People want to believe that the world is defined using pre-given features, i.e., knowledge and meaning exist, we just need to find them. This is called Cartesian anxiety: “a dilemma: either we have a fixed and stable foundation for knowledge, a point where knowledge starts, is grounded, and rests, or we cannot escape some sort of darkness, chaos, and confusion” [234, p140]. In practice, when people are faced with equivocal meanings, they are inclined to select ones that reduce this Cartesian anxiety. But, in doing so, they also enable existing, socially constructed meanings to shape their sensemaking. This can be helpful in providing the clarity of values needed when faced with equivocality, or it can privilege some meanings over others, depending on agency and power [192].

#### **5.1.4.2 Enactment in the Human-Machine Context**

The most prominent examples of enactment in the human-machine context are those in which ML-based systems are used in urgent, reactive situations, such as predictive policing. Consider the use of PredPol, which uses location-based ML models that rely on connections between places and their historical crime rates to identify hot spots for police patrol [2]. Say a police officer is monitoring PredPol to allocate patrol units to various neighbourhoods in the city. As these patrol units are sent to specific locations based on the model’s predictions,

both the officer monitoring the software as well as those patrolling the neighbourhoods have updated their “environment” to be focused on certain neighborhoods. That is, they are primed to look for criminal activity in these neighborhoods. Additionally, when arrests are made using model predictions, they provide further evidence to the model that the patterns it has identified are accurate. The choices made by the monitoring and patrol officers shape the relevant context (i.e., enacted environment), which in turn shapes how the model is updated. In this way, the feedback loop causes the model to become increasingly biased [80]. Now imagine if the police officers were also provided an explanation for the model’s predictions. Each explanation and the order in which the explanations are accessed (e.g., global vs. local explanation first) changes the enacted environment for the officers. The sensemaking perspective presents several properties that describe how factors external to the information being presented (e.g., their identity, social network, etc.) affect people’s understanding of a situation.

Tools that implement interpretability approaches offer different types of information (e.g., global feature importances, local explanations, partial dependency plots, data distributions), but do not impose any order on how this information is explored. As a result, they do not account for the enactive nature of sensemaking. End users can take different paths to reaching conclusions about the model. Because sensemaking is sensitive to these enacted environments, it is important to remember that people treat explanations (or any information about the model) not as static and isolated; rather, as dynamic and use it to shape the environment (e.g., the prediction) that they must understand.

*Claim: The order in which explanations are seen affects how people understand a model and its predictions.*

### **5.1.5 Ongoing**

The ongoing nature of AI/ML sensemaking is critical because it highlights the ways in which interruptions and emotions can influence what is understood about these systems.

### **5.1.5.1 Sensemaking as an Ongoing Activity in the Human-Human Context**

Sensemaking never starts or stops. People are always in the middle of something. To think otherwise would suggest that people are able to chop meaningful moments from the flow of time, but that would be counter-intuitive because to determine whether something is “meaningful” would require sensemaking in the first place [191, 52].

Sensemaking is akin to being in situations of thrownness. Winograd and Flores describe these situations in terms of six different properties: (1) you cannot avoid acting; (2) you cannot step back and reflect on your actions; i.e., you have to rely on your intuitions; (3) the effects of action cannot be predicted; (4) you do not have a stable representation of the situation; (5) every representation is an interpretation, i.e., no objective analysis can be performed in the moment; and (6) language is action, i.e., people enact the situation via their descriptions of their environment, making it impossible to stay detached from it [253]. The feeling of thrownness comes from interruption of flow. Interruptions are common in our every day lives and people are rarely indifferent to them even when they are engaged in work. Interruptions to an ongoing process lead to an emotional response—in this way, emotion is able to influence sensemaking.

Emotion is embedded in sensemaking via the following process. Interruptions trigger arousal, i.e., a discharge in the autonomic nervous system, which convinces the individual that something in the environment has changed and that they must understand it and take appropriate action to get back to a state of flow [24, 155]. The higher the arousal post-interruption, the stronger the emotional response and, in turn, the stronger the affect of emotion on sensemaking. According to Weick, “arousal leads people to search for an answer to the question, ‘what’s up?’”—the longer it takes to answer that question, the longer people are in a state of interruption, and the stronger the emotional response.

Even when interrupted, there are situations when arousal does not build and emotional response is not engaged. If there exist several plausible ways in which the interruption can be resolved, then it becomes easier to answer the “what’s up?” and “what next?”



questions. This would suggest that people who are generalists (in terms of skillsets) or able to improvise show less emotional behavior and less extreme emotions. On the other hand, when the interrupted action is tightly organized or interdependent, or when the interruption is pervasive, the arousal is quick and strong, as is the emotional response. As for the valence of the emotion, negative emotions are generally associated with interruptions that are unexpected or those that have a detrimental impact on whatever the person was doing in their state of flow. Positive emotions result from unexpected removal of a potentially interrupting stimulus, or when the original action being performed is accomplished at an unexpectedly accelerated pace, without interruptions.

Why does it matter if there is an emotional response during an ongoing sensemaking process? Emotions affect sensemaking in that recall and retrospect are dependent on one's mood [216]. Specifically, people recall events that are congruent with their current emotional valence. Of all the past events that might be relevant to sensemaking in a current situation, the ones we recall are not those that look the same, but those that feel the same.

#### **5.1.5.2 Sensemaking as an Ongoing Activity in the Human-Machine Context**

The ongoing property of sensemaking tells us that, while sensemaking is constantly happening, the times when we realize this explicitly are when we are interrupted. Interruptions create brackets of time that require sensemaking.

Consider the predictive policing example again. Let us assume the record of arrests shows that the likelihood of a legitimate arrest in an area predicted as a hot spot by the model is 40%. The officer monitoring the model outputs is made aware of this number every time they log into the system. Imagine this is what happens one day: the patrol officers allocated to one of the hot spots predicted by the model make a legitimate arrest. The monitoring officer is made aware of this and commended for their role in anticipating the situation. This happens several times during the day. As a result, the monitoring officer associates positive feedback for arrests based on the model's predictions. In writing their report about the

incidents, they use the explanations provided by the software to further justify their choices.

Next day, the patrol officers make another arrest in the same predicted hot spot. The monitoring officer is once again asked to record an explanation for selecting that area for patrol. Before they do so, they happen to look at social media and notice several posts showing outrage with regards to that arrest. This is an interruption, as described by the ongoing property of sensemaking. This time, when the monitoring officer is writing up their explanation, it could be that they mention that the model's predictions are not always right and highlight some other failure cases.

As I have noted before, information presented in explanations is rarely used in context-free settings. Despite being shown the same explanation, there is a possibility that the monitoring officer would notice different aspects of the explanation depending on whether they were interrupted or not, whether the interruption led to positive or negative emotional states, and the magnitude of those emotions. Sensemaking does not fix this affect of emotion on how people understand an outcome, but it accounts for it when observing the understanding that has been established.

*Claim: The valence and magnitude of the emotion caused by an interruption during the process of understanding explanations from interpretability tools change what is understood.*

### **5.1.6 Focused on and by Extracted Cues**

Recognizing extracted cues is critical for AI/ML sensemaking because they represent the (incomplete) bits of information that people use in trying to understand these systems.

#### **5.1.6.1 Extracting Cues in the Human-Human Context**

Weick describes extracted cues as “simple, familiar structures that are seeds from which people develop a larger sense of what may be occurring” [247, p50]. These extracted cues are important for sensemaking because they are taken as “equivalent to the entire datum

from which they come” and in being taken as such, they “suggest a certain consequence more obviously than it was suggested by the total datum as it originally came” [101, p340]. Sensemaking uses extracted cues like a partially completed sentence. The completed first half of the sentence constrains what the incomplete second half could be [208].

Extracting cues involves two processes—noticing and bracketing—which are both affected by the context in which sensemaking is occurring. First, context affects which cues are extracted based on what is noticed by the sensemaker. The term *noticing* is intended to imply an informal, even involuntary, observation of the environment that begins the process of sensemaking [220]. As a result, cues that are noticed are either novel, unusual, or unexpected in some way, or those that we are situationally or personally primed to focus on (e.g., recently or frequently encountered cues) [230]. Second, context affects how the extracted (noticed) cues are interpreted. Without context, any cues that are extracted lead to equivocal meanings [133]. As I noted with the retrospective property of sensemaking (Section 5.1.3), these situations of equivocality need a clarity of values instead of more information for sensemaking. Context can provide this clarity in the form of, for example, the setting or domain from which the cues are extracted, or the social and cultural norms that are relevant for that setting.

During the process of extracting cues, people are essentially forming a cognitive reference map that presumes that there is a connection between the situation/outcome and the cue. When they act based on this presumed connection, they enact an environment in which the connection is relevant, i.e., a self-fulfilling prophecy. To highlight this characteristic, Weick shares an incident that happened during military maneuvers in Switzerland, as described in a poem by Holub: “The young lieutenant of a small Hungarian detachment in the Alps sent a reconnaissance unit into the icy wilderness. It began to snow immediately, snowed for 2 days, and the unit did not return. The lieutenant suffered, fearing that he had dispatched his own people to death. But on the third day the unit came back. Where had they been? How had they made their way? Yes, they said, we considered ourselves lost and waited for the

end. And then one of us found a map in his pocket. That calmed us down. We pitched camp, lasted out the snowstorm, and then borrowed this remarkable map and bearings. And here we are. The lieutenant borrowed this remarkable map and had a good look at it. He discovered to his astonishment that it was not a map of the Alps, but a map of the Pyrenees.” [87].

While the example above sheds a positive light on this self-fulfilling aspect of extracted cues, this is not always the case. It can be helpful when seeded with the right cues, e.g., when people have encountered a situation before and have a plan of action in mind, they are able to extract relevant cues from the get go. However, there is potential for important cues to be missed when people do not have any prior experience with the situation.

#### **5.1.6.2 Extracting Cues in the Human-Machine Context**

Consider the example of a company which provides ML-based software to organizations to help with hiring decisions. A marketing company uses this software to shortlist candidates by sending some questions in advance—the candidates answer these questions in a video, the ML-based software analyzes these videos and provides a hiring score along with an explanation for its score. The kind of input data used by the model includes demographic information; prior experience from the candidate’s resume; and tone of voice, perceived enthusiasm, and other emotion data coded by the software after analyzing the recorded video.

Let us assume that the marketing company is using this software to shortlist candidates for the position of a sales representative. The software shows that A is a better candidate than B. They provide explanations for both A’s and B’s ratings (based on local explanations from interpretability tools). The HR folks responsible for the final decision see that A’s rating is based on their facial expressions during the interview (they were smiling, not visibly nervous, and seemed enthusiastic). They consider these to be good attributes for a sales representative and hire A even though B is more qualified. All additional information about both A’s and B’s qualifications is also noted in the local explanations but, as noted by Weick, might not be the cue that is extracted or focused on in this instance.

Sensemaking and other cognitive science theories tell us that people cannot pay attention to all the information provided to them. As such, how people make sense of something is dependent on what they extract as relevant cues for that event or experience, i.e., context is important. As I have noted for other sensemaking properties above, current interpretability tools present all types of information and let the user decide how to explore. Weick cautions against this unstructured exploration because it leads to equivocal alternatives for understanding a situation. Which one of these alternatives is ultimately selected can be a reasonable, reflective process or entirely arbitrary.

*Claim: Highlighting different parts of explanations can lead to varying understanding of the underlying data and model.*

### **5.1.7 Driven by Plausibility rather than Accuracy**

Recognizing that people are driven by plausibility rather than accuracy is critical for AI/ML sensemaking because we need to account for people's inclination to only search for good enough (i.e., plausible) explanations of these systems.

#### **5.1.7.1 Plausibility over Accuracy in the Human-Human Context**

Weick argues that accuracy is nice but not necessary for sensemaking. Even when accuracy is necessary, people rarely achieve it. Instead, people rely on plausible reasoning which is: (1) not necessarily correct but fits the facts, and (2) based on incomplete information [99]. Sensemaking is about “plausibility, pragmatics, coherence, reasonableness, creation, invention, and instrumentality” [247, p57]. When trying to understand something, people not only consider the sensory experience, but are also influenced by what is “interesting, attractive, emotionally appealing, and goal relevant” [65]. Weick notes eight reasons for why accuracy is secondary to sensemaking and what we can learn from these:

1. It is impossible to internalize the overwhelming amount of information available for sensemaking. To cope with this, people apply relevance filters to the information. It is

more productive to observe which filters are invoked and what is included or excluded as a result [67, 215].

2. There is no concrete object which needs to be sensed or understood accurately. Objects have multiple, often equivocal, meanings depending on the individual doing the sensemaking and their various contexts (Sections 5.1.3 and 5.1.6).
3. Sensemaking optimizes for continuation of flow, which does not always require or allow for accuracy considerations. The situations of thrownness that are associated with sensemaking are characteristically time-sensitive. There is a speed / accuracy trade-off, and sensemaking requires speed [65].
4. If a sensemaking situation needs accuracy, it is only for short periods of time or for specific questions. Swann distinguishes this type of accuracy as *circumscribed accuracy*, which is less sweeping than *global accuracy* [227]. In sensemaking situations, circumscribed accuracy is the best we can hope for, which, in practice, resembles plausible reasoning.
5. By virtue of being dependent on the individual (Section 5.1.1), the goal of sensemaking is the interpersonal perception of a situation rather than object perception. Only the latter requires accuracy.
6. Sensemaking perspective considers any meaning that helps counteract interruptions as accurate. This type of accuracy is action-oriented and, as a result, not absolute.
7. Trying to make sense of a situation by internalizing all the information is debilitating. When people filter what they notice, this biased noticing can be good for action, if not for deliberation. Deliberation is not the intended outcome: it is “futile in a changing world where perceptions, by definition, can never be accurate” [247, p60].
8. Accuracy is nice but not necessary because, at the time of sensemaking, it is impossible to tell if the sensemaker’s perceptions will be accurate. It is only in retrospect—after

the sensemaker has taken some action based on their understanding—that they evaluate their perceptions for accuracy.

With accuracy not being necessary for sensemaking, it is only natural to ask: what is? Weick claims that what is necessary for sensemaking is a good story, “something that preserves plausibility and coherence, something that is reasonable and memorable, something that embodies past experiences and expectations, something that resonates with other people, something that can be constructed retrospectively but also can be used prospectively, something that captures both feeling and thought, something that allows for embellishment to fit current oddities, something that is fun to construct” [247, pp60-61]. Stories help with sensemaking because they are templates generated from previous attempts at making sense of similar situations. Overall, this property of sensemaking is often amplified by the others in that the plausible narratives that might appeal to someone could depend on their identity, implied or actual audience, extracted cues, environmental context, emotional state, etc.

#### **5.1.7.2 Plausibility over Accuracy in the Human-Machine Context**

Interpretability outputs, such as text or visual explanations, inherently present a story. As long as this explanation / story is plausible, there is no reason for an individual to evaluate it for accuracy. Consider the example with the radiologist again. The radiologist is tasked with deciding whether a chest radiograph indicates that the patient has COVID-19. Their decision-making is supported by a ML-based software that is trained on publicly available chest radiograph datasets. To help them understand the model’s reasoning for a prediction, the radiologist has access to saliency maps as interpretable outputs (Figure 5.2).

According to Weick, when using the saliency map to determine whether the model’s prediction makes sense, the radiologist is essentially searching for a plausible story that explains the prediction. Figure 5.2 shows example explanations for a COVID-19 prediction model from [48]. These explanations show that the areas inside the lungs are relevant for the model’s prediction, a plausible reason for predicting COVID-19 positive or negative

outcomes. The radiologist could believe this plausible explanation and choose to follow it. Human evaluations of interpretability tools show that this confirmatory use of explanations is often the case, even when explanations reveal issues with the underlying model. People are prone to over-trusting interpretability outputs [111]—these explanations increase the likelihood that people will accept the prediction, without evaluating it for correctness [18].

Say that the radiologist was not immediately convinced that the prediction made sense after seeing the saliency map. Maybe they looked at the explanation (e.g., Figure 5.2-Middle) and noticed that the radiograph’s edges (by the person’s shoulders and diaphragm) were also salient for the prediction. Even with this observation, the radiologist is looking for a plausible story that explains the model’s prediction. Perhaps the patient was coughing and could not stay still when the radiograph was being captured, which could explain why the radiograph has different lateral markers for a COVID-19 positive patient. In reality, the model is relying on spurious correlations. With the role of plausibility in sensemaking, the radiologist might not try to accurately interpret the saliency map.

*Claim: Given plausible explanations for a prediction, people are not inclined to search for the accurate one amongst these.*

### **5.1.8 Summary**

When designing solutions for promoting human understanding of ML models, it is imperative that we consider the nuances of human cognition in addition to the technical solutions which explain ML models. Sensemaking provides a set of properties that describe these nuances, and each of these can be seen as a self-contained set of research questions and hypotheses that relates to the other six. As I have shown with examples from the human-machine context above, these sensemaking properties could explain some of the external factors that shape the information that is ultimately internalized by people when they use interpretability solutions to understand ML models.



## 5.2 Discussion

I propose a framework for Sensible AI to account for the properties of human cognition described by sensemaking. This has the potential to not only refine the explanations provided by interpretability tools for human consumption, but also to better support the human-centered criteria of ML-based systems. How do we do this? Once again, Weick (along with his colleagues) proposes a solution: to explicitly promote or amplify sensemaking, we can follow the model of *mindful organizing* [248]. Sensemaking and *organizing* are inextricably intertwined. While sensemaking describes the meaning-making process of understanding, organizing describes the final outcome (e.g., a map or frame of reference) that represents the understanding. They are a part of the same mutually interdependent, cyclical, recursive process—sensemaking is the process by which organizing is achieved [15, 250]. When one is observing an event for sensemaking, *mindfulness* is expressed by actively differentiating and refining existing categories that we use to assign meaning, and creating new categories as needed for events that have not been seen before [129, 248, 240].

Mindful organizing was proposed based on observations from high-reliability organizations (HROs). HROs are organizations that have successfully avoided catastrophes despite operating in high-risk environments [193, 249]. Examples of these include healthcare organizations, air traffic control systems, naval aircraft carriers, and nuclear power plants. Mindful organizing embodies five principles that were consistently observed in HROs: (1) **preoccupation with failure**, a focus on anticipating potential risks by always being on the lookout for failures, being sensitive to even the smallest ones; (2) **reluctance to simplify**, wherein each failure is treated as unique because routines, labels, and cliches can stop us from looking into further details of an event; (3) **sensitivity to operations**, having a heightened awareness of the state of relevant systems and processes because systems are not static or linear, and expecting uncertainty in anticipating how different systems within the organization will interact with each other in the event of a crisis; (4) **commitment to resilience**, prioritizing training for emergency situations by incorporating diverse testing

	Preoccupation with Failure	Reluctance to Simplify	Sensitivity to Operations	Commitment to Resilience	Deference to Expertise
Seamful Design		X	X	X	
Inducing Skepticism	X	X		X	
Adversarial Design		X		X	X
Continuous Monitoring and Feedback	X		X		X

Table 5.2: Principles for high-reliability organizations (columns) that inspired my design ideas (rows) under the Sensible AI framework.

pathways and team structures, and when a problematic event occurs, trying to absorb strain and preserve function, and learning from previous instances; and (5) **deference to expertise**, assuming that people who are in the weeds—often lower-ranking individuals—have more knowledge about the situation, and value their opinions. My proposal for Sensible AI encompasses designing, deploying, and maintaining systems that are reliable by learning from properties of HROs. Table 5.2 presents the corresponding principles of HROs that serve as inspiration for each idea.

### 5.2.1 Seamful Design

We can help people understand ML by giving them the agency to do so. Often, AI- and ML-based systems and interpretability tools are designed with seamless interaction and effortless usability in mind. However, this can engage people in automatic reasoning modes, leading them to using outputs from these systems without adequate deliberation [33, 111, 18]. Highlighting complex details of ML outputs and processes —seamful design [98]—can promote a reluctance to simplify that has helped HROs. It can also add a sensitivity to operations when changes to inputs for ML models can be clearly seen in the outputs. Enhancing reconfigurability of ML models and training people to understand their complexity can reduce the automatic, superficial evaluations on the part of the end-users. Increasing user control in the form of seamful design has the additional benefit of introducing opportunities for informational interruptions, which are necessary for the commitment to resilience seen in HROs. While current interpretability tools have features that allow for users to get additional

information if they want it, contextualizing this additional information by using narratives can help people maintain overall situational awareness and avoid dysfunctional momentum when using ML-based systems. For example, when a doctor is looking at a predicted patient diagnosis, a Sensible AI system could prompt them to view cases with similar inputs but different diagnoses. Next, I discuss some ways in which these systems can be designed without overloading the end-user with system features, interactivity, and information.

### **5.2.2 Inducing Skepticism**

One way to reduce over-reliance on known generalizations and information—both common outcomes of sensemaking—is to create situations in which people would ask questions. I call this inducing skepticism, an idea that has also been proposed in prior work as a strategy for promoting reflective design [205]. Inducing skepticism can be seen as a way to foster a preoccupation with failure, an HRO principle that encourages cultivating a doubt mindset in employees. HRO employees are always on the lookout for anomalies, they interpret any new cues from their systems in failure-centric ways, and collectively promote wariness. This can be incorporated in ML-based systems, for example, by suggesting that end-users ask about how a particular prediction is unique or similar to other data points, questioning outputs of interpretability tools every now and then (e.g., “does this feature importance value make sense?”), presenting bottom-n feature importances in an explanation instead of top-n, highlighting cases for which the model is unsure of its’ predictions, etc. Inducing skepticism can also be accomplished in social ways, by promoting diversity in teams, both in terms of skillsets and experience. For example, novices can prompt experts to view an ML output in more detail when they ask questions about it. This diversity is a common way in which HROs maintain their commitment to resilience. Ultimately, these technical and social forms of inducing skepticism have a common goal in mind, a reluctance to simplify by introducing complexity and diversity into a situation.

### 5.2.3 Adversarial Design

No one person can successfully anticipate all failures, even when the system supports inducing skepticism. Adversarial design suggests relying on social and organizational networks for this task. Adversarial design is a form of political design rooted in the theory of agonism, a condition of productive contestation and dissensus [53, 164, 252]. By designing Sensible AI systems with dissensus-centric features, we can increase the likelihood that *someone* raises a red flag given early signals of a failure situation or has a better understanding of why a particular model prediction is not accurate. Prior work has implemented adversarial design in the form of red teaming in technical and social ways (e.g., adversarial attacks for testing and promoting cybersecurity [3], and forming teams with collective diversity and supporting deliberation [91, 90], respectively). Here, HRO principles of reluctance to simplify, commitment to resilience, and deference to expertise can be observed in practice. I propose technical redundancies and social diversity to help capture unanticipated failures in model predictions and understanding, as one way of operationalizing adversarial design. Technical redundancies can be implemented in the form of system features wherein multiple people view the same output in different contexts, which gives the team a better chance of finding potential issues. Social or organizational diversity can be expanded by including people with different roles, skillsets, and opinions. The more diversity we have in people viewing outputs or explanations, the higher the likelihood that they collectively discover an issue, as long as deliberation is made easy [90].

### 5.2.4 Continuous Monitoring and Feedback

When ML-based systems are deployed in real-world settings, changes in data collection and distributional drifts are a given. However, a model that accurately finds patterns in data collected in the past is not necessarily generalizable to temporal changes in data. To manage these, researchers and industry practitioners have proposed MLOps—an extension of DevOps practices from software to ML-based settings—for incorporating continuous testing,

integration, monitoring, and feedback in maintaining the operation of ML-based systems in the wild [153]. I propose incorporating social features in this pipeline by designing for HRO principles such as preoccupation with failure, sensitivity to operations, and deference to expertise. This could be achieved by designing features for, for example, (1) continuous failure monitoring, effectively serving as distributed fire alarms that can be engaged by people at varying levels in an organization, or (2) model maintenance, by relying on people on the ground for detailed understanding of failure cases, as is done by several organizations in discussing and publishing the results of failure panels, audits, etc.

### **5.3 Conclusion**

Interpretability and explainability approaches are designed to help stakeholders adequately understand the predictions and reasoning of a ML-based system. Although these approaches represent complex models in simpler formats, they do not account for the contextual factors that affect whether and how people internalize information. In this chapter, I have presented an alternate framework for helping people understand ML models grounded in Weick’s sensemaking theory from organizational studies. Via its seven properties, sensemaking describes the individual, environmental, social, and organizational context that affects human understanding. I translated these for the human-machine context and presented a research agenda based on each property. I also proposed a new framework—Sensible AI—that accounts for these nuances of human cognition and presented initial design ideas as a concrete path forward. Overall, this theoretical framework prescribes design guidelines for re-imagining interpretability and explainability with human cognition at its core. It lays out my research agenda for future work in this domain.

## **CHAPTER VI**

### **Conclusion**

I began this dissertation with the goal of investigating how best to support the human in human-machine collaboration. With advances on the technical front, systems nowadays routinely employ complex AI and ML models in collaborative tasks with people (e.g., for decision-support, content moderation, information search, and now, even generating new content based on end-user prompts). However, the missing critical piece is a parallel growth in helping people understand these systems. As a result, despite this being an era of AI advancement like no other, we frequently see these system deployments resulting in massive failures including missed edge cases in critical domains, and propagation of fake news and societal biases. Through the work presented in this dissertation, I sought to examine how (much) people understood ML-based systems, their reasoning, and inputs and outputs; what (types of) factors affected people’s understanding; and what would be the characteristics of a solution that helped people better understand ML.

Working in the context of ML-based decision-support systems with a key stakeholder, ML practitioners, I first examined how practitioners perceive and use interpretability tools (Chapter III). My mixed-methods study setup helped identify the key challenges in practitioners’ use of these tools. I noted significant misuse and over-trust in interpretability tool outputs for the setting of a data science task, and a large-scale survey showed that these findings were generalizable. In Chapter IV, I followed up on these studies to investigate

the underlying cause of misuse and over-trust of interpretability tools. I found significant evidence in support of my hypothesis that the cognitive framework of bounded rationality was at play. Interpretability tools were significantly more likely to engage people in faster and less effortful thinking modes. This automated thinking is often a result of application of heuristics, which are grounded in prior experiences and knowledge. Delving deeper into these external factors that seemed relevant to human cognition, in Chapter V, I translated a theory about human sensemaking from its original socio-organizational context for the human-machine setting. This noted the various facets of human cognition that have nothing to do with the information being presented. Things like one’s identity, social influences, ongoing personal and professional tasks, human inclination towards plausible stories over searching for accuracy—these individual, environmental, social, and organizational characteristics shape how (much) people process information.

In concluding this document, there is one open question I continue to think about: **what next?** Or, what do the claims from this dissertation mean for future work on interpretability and explainability? How do we present information to people knowing they will never pay attention to all of it? How do we design for the incompatibilities in how people and machines process information? What other human-centered factors do we need to account for? What can we do to ensure that human-machine systems are not only effective at the task at hand, but also do not cause harm when deployed in the real world? These “what next” questions are even more essential to ask now, with AI and ML taking on a more significant role not only in our academic disciplines, but also in everyday applications. Pre-trained models have completely changed the landscape of what is possible with AI and ML technologies, and who has access to them. These questions apply for all stakeholders, and at a societal level.

Although my dissertation focused on a subset of ML models and DSSs that are more specialized and well-constrained, extrapolating from this work, I see the following findings becoming even more relevant in this new age of generative AI: (1) *growing over-reliance*, given the natural language interfaces for several new AI and ML technologies, I believe

systems that rely on these will appear to be far more capable than they are, which might further increase over-reliance; (2) *more interpretability issues*, given that most of the advanced new models are opaque—we cannot yet faithfully mimic how these models operate, making transparency and interpretability even harder than before; and (3) *the essentiality of understanding human cognition*, mental models, folk theories, and other sensemaking mechanisms will be critical for both evaluating and shaping future human-machine collaboration that does not cause harm. Liao and Vaughan’s recent article on transparency in the age of LLMs outlines a more exhaustive list of these anticipated challenges and some paths forward [139].

To make progress on these considerations, I propose an updated paradigm for human-machine collaboration: a paradigm that is human-centered, not only for integrating human capabilities in innovative human-machine systems, but also in accounting for the sometimes irrational nuances of how people think, communicate, and collaborate, and their affect on human-machine systems. Next, I discuss a future research agenda based on this paradigm shift: (1) using inefficiencies and friction in design to promote more deliberate human-machine engagement, (2) using generative AI as a new interaction modality, (3) rethinking ways of evaluating human-machine partnerships, and (4) continuing translation and application of relevant theories from other fields.

## **6.1 Designing Inefficiencies for Effective Human-Machine Collaboration**

Human-machine systems, especially when embedded with AI or ML, are designed with seamless interaction and effortless usability in mind. My work shows that efficiency and seamlessness in system design engage people’s automatic reasoning mode, leading them to use these systems in superficial ways (Chapters III– IV). There is also some initial evidence that the recent advances in AI are exacerbating this issue, their fabricated yet realistic-seeming outputs making it challenging to determine accuracy and ground truth [104]. I hypothesize that adding inefficiencies and friction in human-machine interaction might engage



people in deliberative reasoning modes. This would, in turn, increase the likelihood of an accurate assessment of AI/ML and their outputs. To that end, I have proposed a new class of SENSIBLE AI systems that take into account the nuances of human cognition and sensemaking (Chapter V). As a starting point, these systems could be designed based on principles of:

1. ***Seamful design***, which highlights some of the complexity in information and features of an AI system rather than simplifying everything. This can nudge people towards exploring more of the details available to them, especially when the nudges are surfaced in contextually relevant parts of the system. For example, in a conversational format with a large language model (LLM), instead of only seeing one utterance in response to a prompt, we could present an output showing how slightly changing the prompt would have resulted in different responses from the model.
2. ***Inducing skepticism***, which promotes a preoccupation with failure and anomalies in system outputs as a means of ensuring that people do not generalize the capabilities of their machine counterpart. Skepticism can minimize missed system edge cases. This can be incorporated by, for example, periodically questioning the AI explanations (e.g., presenting to the end-user the bottom-n features important for the model instead of the top-n), or highlighting cases for which a model is unsure of its outputs.
3. ***Adversarial design***, which suggests using diverse social and organizational networks for productive contestation and dissensus in understanding AI. For example, asking several people to explain their understanding of the same AI output, highlighting the diversity in responses, and designing system features that ensure engagement amongst those with differing opinions. Adversarial design is a distributed fail-safe—no one person is solely responsible for making decisions with the help of AI; technical and social redundancies can help.

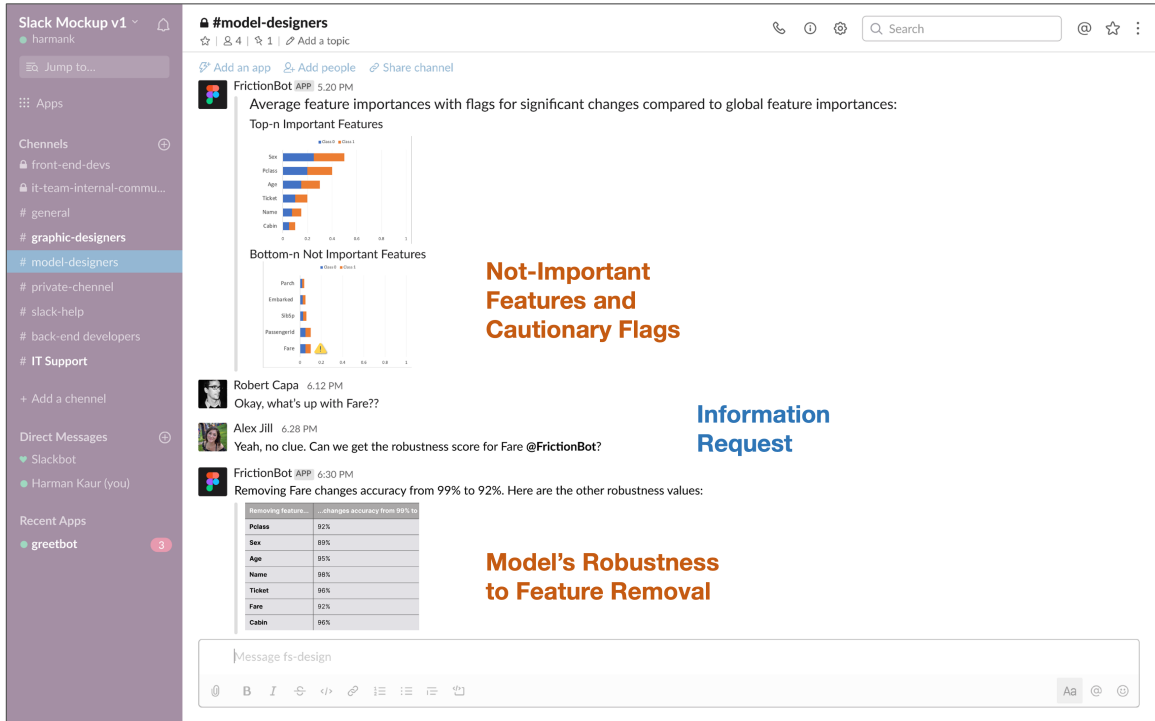


Figure 6.1: An initial prototype for FrictionBot, an application that supports more deliberate engagement with ML and interpretability outputs using design guidelines from my Sensible AI framework, powered by LLMs.

## 6.2 Generative AI as a New Interaction Modality

Although generative AI models pose enormous transparency risks, they can also be an opportunity for us to test new ways in which interpretability information about other models can be communicated. For example, with their affordance of natural language communication, LLMs could act as an intermediate layer, describing interpretability and explainability outputs to an end user in a conversational or visual narrative format. Recent work has shown that these different modalities—questions vs. answers vs. dialogue—can be effective ways of providing explanations in different settings [47, 132]. Prior work in the social sciences has similarly noted that a conversational format helps people form better mental models of each other [45, 163]. These LLM-infused applications could enable new interaction modalities for explanations that are more human-centered.

With the conversational interfaces enabled by LLMs, imagine the ease with which we

could design a system grounded in the ideas from Section 6.1. Figure 6.1 presents an initial prototype of a Slack application, FrictionBot, that I envision as an easy proof-of-concept and a test of these ideas for the setting of this dissertation: data science. FrictionBot brings together practitioners with varying expertise, working on different aspects of the same AI product for an organization (*adversarial design*). It initiates a conversation about the AI in question and, every time there is a request for information or conversation around it, it presents the information in both positive and negative framings (inducing skepticism), and such that there is always more information than was initially requested (seamful design). FrictionBot’s goal is to engage a diverse group of practitioners in a longer conversation about different aspects of the AI, with the intention of helping uncover any potential issues via this more deliberate engagement. One could also imagine extensions of this same design for other settings as well (e.g., a system feature that nudges the patient care team to communicate about a patient’s predicted diagnosis from an AI, and whether they should follow it).

### **6.3 Rethinking Evaluation Metrics for Human-AI Partnerships**

Although there are innumerable human-machine systems now deployed in the wild, we have yet to answer the question: what does a “good” human-machine partnership look like? What are the characteristics that define a successful human-machine team? My work argues for rethinking the primacy of the prevailing metrics for defining success (e.g., optimization, efficiency, seamlessness, and ease of use). But what should take their place?

One new *human-centric* metric could be the level of deliberative thinking and engagement, since cognitive and social theories suggest that to be a marker for adequate understanding of a technology. Several validated scales provide measures of user engagement with a system, cognitive effort in using it, broader acceptance of a technology, etc. However, there are no existing scales that measure *deliberative* thinking and engagement, especially for the human-machine context—an interesting avenue for future work. Another *system-centric* proxy for deliberative thinking might be the level of friction in a system or to what extent are

failures highlighted. For example, one could develop a friction scale by reversing the scale of the 10 heuristics by Nielsen [165], commonly used for usability evaluation. Similarly, a failure metric could be the number of GitHub issues that were raised before system deployment.

The innate conflict between these “positive” and “negative” metrics make it challenging to study and incorporate them. As an example, things that improve deliberative thinking (e.g., inefficiencies and friction in design) can negatively impact other metrics (e.g., trust, usability). We must further our understanding of these existing and potential metrics that represent a human-machine partnership, how these can be harmonious or dissonant, and how we might encapsulate these variabilities in one system.

## **6.4 Translating and Applying Relevant Theories**

My work follows the belief that, like AI, people are black boxes until we form functional mental models of those we collaborate with. As such, fields that study how people think and interact can provide initial guidance for designing effective human-AI partnerships. For example, I translated the seven properties of human sensemaking from organizational studies, and highlighted how these were relevant for the human-machine context using case studies from prior industry and academic work (Chapter V).

With theory translation, the somewhat harder question to answer is: how does this help future research and practice? When applied conscientiously, these theory translations can also offer a path forward. For example, with sensemaking, prior work in organizational studies had applied it as a framework to identify the ways in which teams and organizations must be made more reliable. These high-reliability organizations (HROs) have safeguards in place to account for the nuances of human cognition. I proposed new guidelines for designing interpretability and explainability using the principles of HROs, called Sensible AI.

We have barely scratched the surface in incorporating the decades of research on human sensemaking, communication, and collaboration into human-machine systems. Going forward, we must continue translating and applying relevant theories from these other fields.

## **APPENDICES**

## APPENDIX A

### Semi-Structured Interview Protocol for the Pilot Interviews

1. What is your background in Machine Learning and Data Science? How long have you been a data scientist?
  - What does your team do?
  - What type(s) of data do you usually use?
  - Do you build models for internal use or customer-facing prediction tasks?
2. At a high level, what is your usual ML / Data Science pipeline (starting from Data to Model to Usage / Prediction)
  - Are there any checks you perform at different stages? How do you generally evaluate the validity of the data and the model?
  - When are you satisfied with the results?
  - What makes you feel confident about them?
3. Do you know about any interpretability tools? What are your perceptions of interpretability tools or solutions (e.g., GAMs, black box models with SHAP, etc.)?
  - What do you think they are good for?

- Do you use interpretability tools at any point in your process, whether it's for internal checks or external pre-deployment checks?
  - If yes, continue to question 4 after asking: which tools do you use?
  - If no, ask them if there's any reason for not using interpretability tools. How do they evaluate if their model is ready for the prediction task at hand? Skip question 4.

4. For each of the interpretability tools someone mentions in question 3:

- How familiar are you with <x> tool?
- How frequently do you use it?
- According to you, what is the output or end-result of relying on <x> tool?
- How confident do you feel in using <x> tool? Do you feel like you understand what it does and the end-result / output from it?
- What is the overall value of using <x> tool in your day-to-day work?
- What kinds of questions do you try to answer using <x> tool?
  - If they don't mention any specific types of questions or outputs, prompt with: Do you look for global or local explanations, feature importance, or regions of error? Something else?
- Are there any challenges in using <x> tool or understanding its output?
- Is there other information that you wish <x> tool would tell you that it currently does not?

5. Retrospective think aloud: Think of an instance when you have had to debug a model recently. What were some challenges you faced in doing this?

- If there were no challenges: Are you generally comfortable with evaluating models and feel confident about it? Is there another instance you remember where the model-building and evaluation process was tricky?

- If there were challenges:
  - How did you overcome these challenges?
  - What was the outcome of the challenging situation? How did you “fix it”?
  - Did you use any interpretability tools or equivalent techniques in this process?

6. More generally, what are your thoughts on the accuracy-interpretability tradeoff? Do you cater to one or the other (or both) in your pipeline?



## APPENDIX B

### Introduction to Generalized Additive Models (GAMs)

A typical linear model assumes that the relationship between the target variable  $y$  and the input variables/features  $x_1, \dots, x_N$  for datapoint  $x$  is linear and can be captured with slope terms  $\beta_1, \beta_2, \dots, \beta_N$ :

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_N x_N$$

Generalized Additive Models (or GAMs) are a generalization of linear models in which the target variable depends linearly on smooth shape functions of the features,  $f_1, \dots, f_N$  that can be nonlinear and complex:

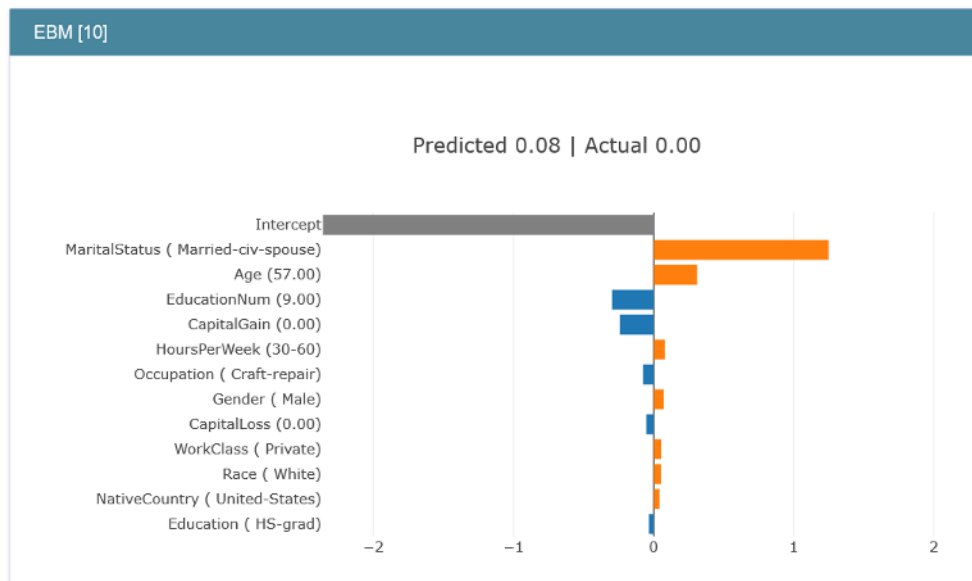
$$y = \beta_0 + f_1(x_1) + f_2(x_2) + \dots + f_N(x_N)$$

In both models  $\beta_0$  is the model intercept, and the relationship between the target variable and the features is still additive; however, the effect of each feature  $x_i$  in a GAM is captured by a shape function  $f_i$ . To use GAMs for classification, we replace  $y$  in the equation above with the log odds,  $\log\left(\frac{p}{1-p}\right)$ , where  $p$  is the predicted probability of a positive classification, analogous to logistic regression. GAMs allow us to visualize local explanations (how an individual prediction was made) as well as global explanations (what the model learned overall from training data). Given their ease of use for providing local and global explanations, GAMs are often considered to be inherently interpretable.

#### Local Explanations: How an Individual Prediction was made

Our GAM interface lets you select a datapoint and displays a local explanation for the datapoint that highlights how the prediction was made.

For example, let's consider this datapoint below with a predicted probability of 0.082. If we look at the local explanation, we can further understand how this prediction was made by observing how each feature either contributed positively or negatively to the prediction. The visualization shows a log odds score for each feature, which we refer to as the feature importance. In this instance, MaritalStatus of Married-civ-spouse (which means that the individual is married to a civilian spouse) had the most positive influence on the predicted value, whereas EducationNum of 9 had the most negative influence.

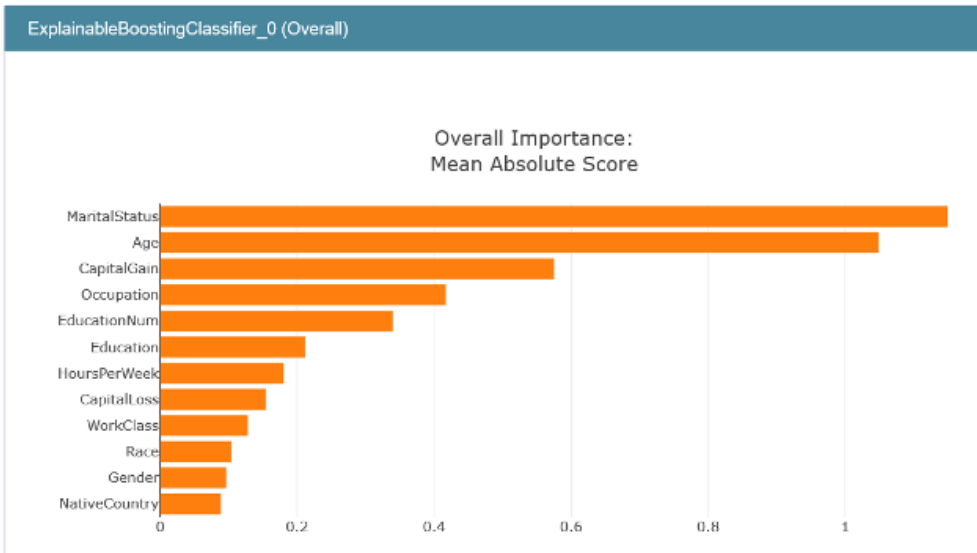


### Global Explanations: What the Model Learned Overall from Training Data

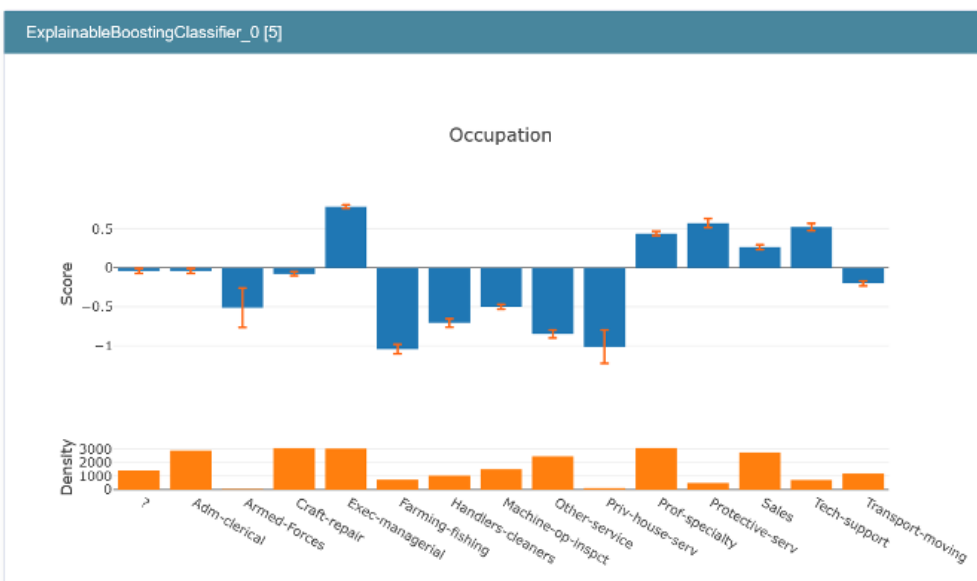
You can also use GAMs' global explanations to understand the overall feature importances, and how the different values that each feature can take affect the target variable.

For example, here (below) we can see a graph of all feature importances, ranked from highest to lowest. These are calculated by averaging the absolute value of local feature importances over all datapoints.

If we wanted to see the relationship between a specific feature and the output, we can select that feature to see the global trends for it. The plot also shows a density visualization



at the bottom to highlight the amount of data available for each value of a feature.



## APPENDIX C

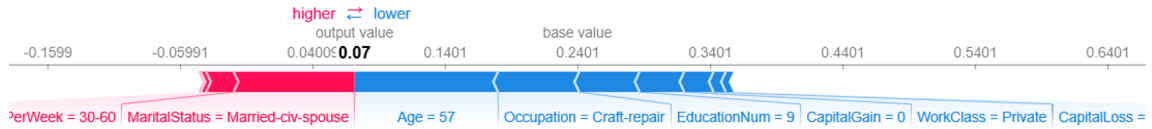
### Introduction to SHAP

SHAP (SHapley Additive exPlanations) is a tool that provides an explanation for the output of any ML model. SHAP is often applied on top of “black box” machine learning models to provide local explanations (how an individual prediction was made) as well as global explanations (what the underlying model learned overall from training data).

For every datapoint  $x$  with a predicted value  $f(x)$ , SHAP outputs a vector  $\phi(x)$  such that  $f(x) = \beta + \sum_{i=1}^N \phi_i(x)$ , where  $\beta$  is the base value—the output value (prediction) a datapoint would be assigned if all feature values were zero. Intuitively, the SHAP value  $\phi_i(x)$  represents the contribution that feature  $i$  for datapoint  $x$  makes to the prediction. The SHAP value for feature  $i$  will be different for each datapoint.

#### **Local Explanations: How an Individual Prediction was made**

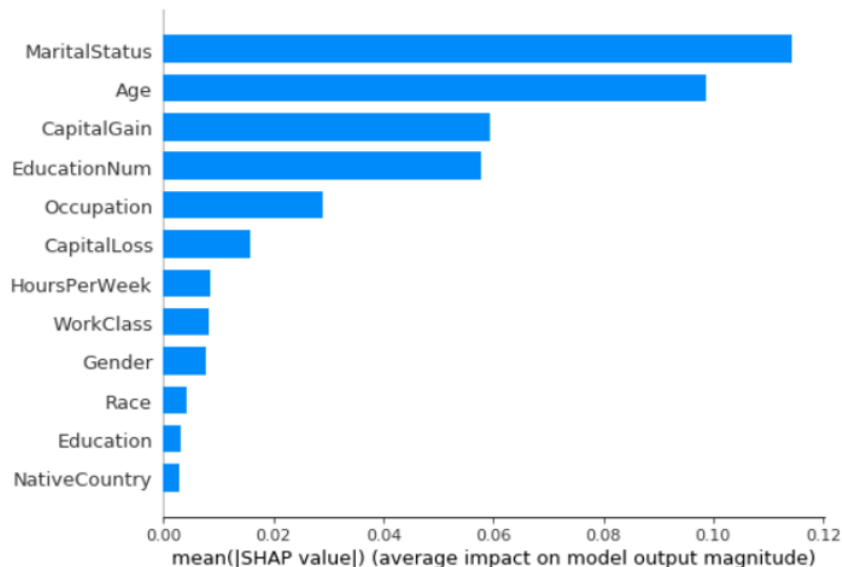
We can visualize how a prediction for an individual datapoint was made using SHAP’s force plots. Given a datapoint  $x$ , the force plots start at a base value  $\beta$ —the output value (prediction) a datapoint would be assigned if all feature values were zero. Then, as we add more specific information about this datapoint (that is, non-zero features  $x_i$ ), we add  $\phi_i(x)$  to the force plot, which either pushes the datapoint to have a higher value than the base value, or lower than it. Ultimately, the output value represents whichever force (higher or lower) wins, and that is equal to the prediction.



For example, if we look at the force plot above, we can see that the base value  $\beta$  is 0.2401. Input features such as Age (= 57), Occupation (= Craft-repair) and EducationNum (= 9) push the predicted output value to be lower, whereas input features such as MaritalStatus (= Married-civ-spouse, i.e., married to a civilian spouse) and HoursPerWeek (30-60) push the predicted output value to be higher. Ultimately, this results in the overall predicted output value of 0.07. The positive or negative scores here are SHAP values  $\phi_i(x)$  for each feature  $i$ .

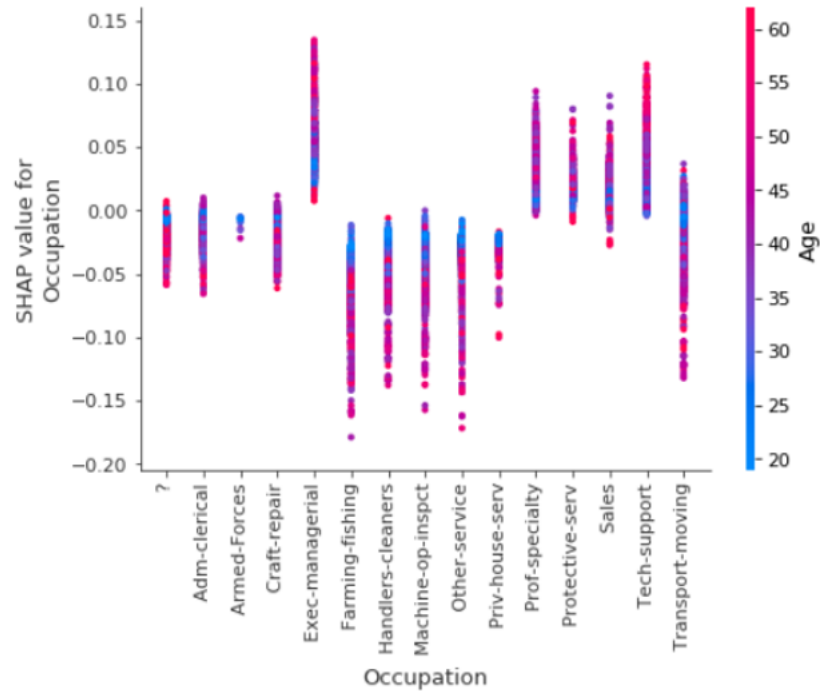
### Global Explanations: What the Model Learned Overall from Training Data

You can also use SHAP's global visualizations to understand the overall importance of each feature. For example, here we can see a graph of all feature importances, ranked from highest to lowest. These are calculated as the mean of the absolute SHAP values for each input feature.



SHAP also provides global trends per feature in the form of dependency plots. For each feature's dependency plot, the x-axis shows the value of that feature for a datapoint ( $x_i$ ), and

the y-axis is the corresponding SHAP value  $\phi_i(x)$ . These plots thus highlight the trend in a feature's importance. For these dependency plots to be dense enough to highlight patterns, each feature is usually plotted against another one, which causes the color of the plotted dots to change. You can observe the pattern when trying to understand the global trends for each feature



## APPENDIX D

### Contextual Inquiry Questions about the Dataset and the Model

Participants were asked to answer the following questions about the dataset and the model. They were given one interpretability tool (either InterpretML's implementation of GAMs or a post-hoc explanation technique, SHAP) to aid their exploration of the data and model.

1. What are the most important features that affect the output Income, according to the explanation above?
2. How does the feature Education affect output Income?
3. How does the feature Age affect output Income?
4. How does the feature EducationNum affect output Income?
5. Can you explain the predicted value of the 5th element in the test set?
6. Below, you can see the local explanations for 20 incorrect predictions. Based on this, can you suggest ways for improving accuracy?
7. How does the feature HoursPerWeek affect output Income?
8. How do some other categorical variables—Occupation and WorkClass—affect output Income?

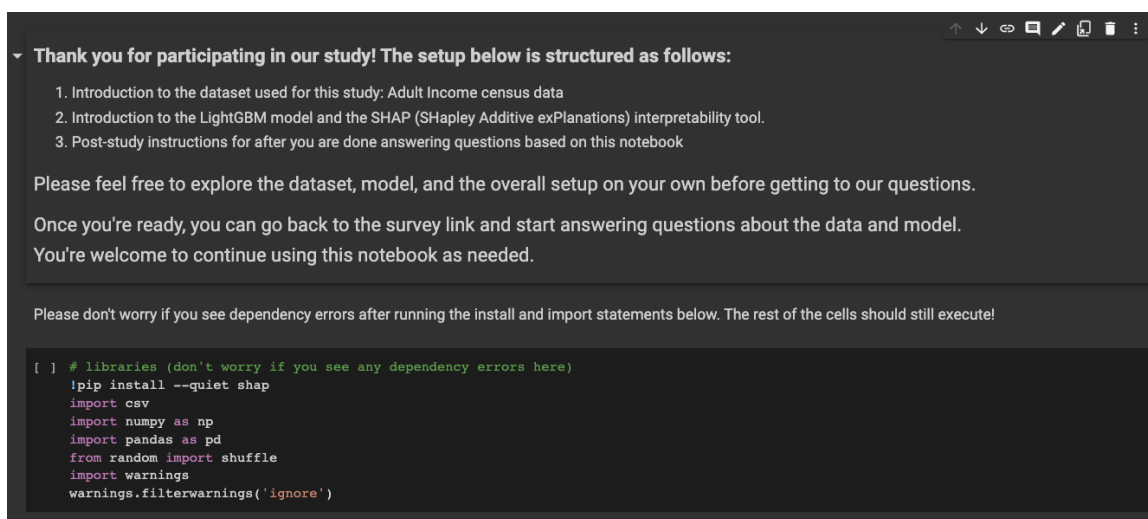
9. If we adjusted for inflation, do you expect this model would do well in predicting income now? Please explain your answer.
10. Are there any other questions you would want to ask based on this data? What information would you hope the explanations would surface?



## APPENDIX E

### Screenshots from a Google Colab Notebook used for the Experiment

The following are screenshots from the Google Colab notebook for one of the interpretability conditions: SHAP, a post-hoc explainer for a blackbox model, and a tool with static visual outputs. These screenshots present how the study setup was introduced to the participants, with information about the dataset, the model, and the capabilities of the interpretability tool for each condition.



The screenshot shows a Google Colab notebook cell with a dark background. At the top right, there are icons for up/down arrows, a refresh icon, a comment icon, a copy icon, a trash icon, and a menu icon. The main content of the cell is as follows:

▼ Thank you for participating in our study! The setup below is structured as follows:

1. Introduction to the dataset used for this study: Adult Income census data
2. Introduction to the LightGBM model and the SHAP (SHapley Additive exPlanations) interpretability tool.
3. Post-study instructions for after you are done answering questions based on this notebook

Please feel free to explore the dataset, model, and the overall setup on your own before getting to our questions.

Once you're ready, you can go back to the survey link and start answering questions about the data and model.  
You're welcome to continue using this notebook as needed.

Please don't worry if you see dependency errors after running the install and import statements below. The rest of the cells should still execute!

```
[ ] # libraries (don't worry if you see any dependency errors here)
    !pip install --quiet shap
    import csv
    import numpy as np
    import pandas as pd
    from random import shuffle
    import warnings
    warnings.filterwarnings('ignore')
```

The Adult Income dataset is a classification dataset: it is used for a prediction task where the goal is to determine whether a person makes over 50k a year. The list of attributes is as follows:

- output variable: **Income**, <=50k and >50k (converted to 0 and 1 respectively)
- input features:
  - **Age**: a continuous number
  - **WorkClass**: a categorical variable that represents different work sectors, including values such as Federal employee, Private, Self-emp-not-inc, Self-emp-inc, Federal-gov, Local-gov, State-gov, Without-pay, Never-worked, Federal worker
  - **Education**: a categorical variable that represents the level of education, including values such as Bachelors, Some-college, 11th, HS-grad, Prof-school, Assoc-acdm, Assoc-voc, 9th, 7th-8th, 12th, Masters, 1st-4th, 10th, Doctorate, 5th-6th, Preschool
  - **MaritalStatus**: a categorical variable with values Married-civ-spouse (Married to a civilian spouse), Divorced, Never-married, Separated, Widowed, Married-spouse-absent, Married-AF-spouse (Married to someone in the Armed Forces)
  - **Occupation**: a categorical variable with values Tech-support, Craft-repair, Other-service, Sales, Exec-managerial, Prof-specialty, Handlers-cleaners, Machine-op-inspct, Adm-clerical, Farming-fishing, Transport-moving, Priv-house-serv, Protective-serv, Armed-Forces
  - **EducationNum**: a continuous value that represents the level of education (0 = Preschool, 16 = Doctorate)
  - **Race**: a categorical variable with values White, Asian-Pac-Islander, Amer-Indian-Eskimo, Other, Black
  - **Gender**: a binary variable, includes Female and Male
  - **CapitalGain**: a continuous number
  - **CapitalLoss**: a continuous number
  - **HoursPerWeek**: a categorical variable including hour ranges of 0-30, 30-60, 60-90, 90+
  - **NativeCountry**: a categorical variable including countries such as United-States, Cambodia, England, Puerto-Rico, Canada, Germany, Outlying-US(Guam-USVI-etc), India, Japan, Greece, South, China, Cuba, Iran, Honduras, Philippines, Italy, Poland, Jamaica, Vietnam, Mexico, Portugal, Ireland, France, Dominican-Republic, Laos, Ecuador, Taiwan, Haiti, Columbia, Hungary, Guatemala, Nicaragua, Scotland, Thailand, Yugoslavia, El-Salvador, Trinidad&Tobago, Peru, Hong, Holand-Netherlands

```
[ ] # Reading and pre-processing training data

csv_columns = [
    "Age", "WorkClass", "fnlwgt", "Education", "Education-Num",
    "Marital-Status", "Occupation", "Relationship", "Race", "Gender",
    "Capital-Gain", "Capital-Loss", "Hours-per-week", "NativeCountry", "Income"]

df = pd.read_csv("https://raw.githubusercontent.com/davisjr/le/satisficing-main-study/main/Data/adult-train.csv", names=csv_columns, sk

cols = ["Age", "WorkClass", "Education", "Education-Num", "Marital-Status", "Occupation", "Race", "Gender",
        "Capital-Gain", "Capital-Loss", "Hours-per-week", "NativeCountry", "Income"]
df = df[cols]

train_cols = df.columns[0:-1]
label = df.columns[-1]
x_df = df[train_cols]
y_df = df[label]
```

```
[ ] X_display = X_df.copy()

#Converting the response / output variable to a binary class
y_df = y_df.apply(lambda x: 0 if x == "<=50K" else 1)

# Converting strings to integers and floats for categorical data
categorical_cols = ["WorkClass", "Education", "Marital-Status", "Occupation", "Race", "Gender", "NativeCountry"]

for col in categorical_cols:
    X_df[col] = pd.Categorical(X_df[col])
    X_df[col] = X_df[col].cat.codes

#Top 5 rows of the full training dataset:
df.head()

[ ] # Reading and pre-processing test data

test_df = pd.read_csv("https://raw.githubusercontent.com/davisjrule/satisficing-main-study/main/Data/adult-test.csv", names=csv_columns)

test_df = test_df[cols]

input_cols = test_df.columns[0:-1]
label = test_df.columns[-1]
test_X_df = test_df[input_cols]
test_y_df = test_df[label]

test_X_display = test_X_df.copy()

#Converting the response / output variable to a binary class
test_y_df = test_y_df.apply(lambda x: 0 if x == "<=50K" else 1)

#Converting strings to integers and floats for categorical data
for col in categorical_cols:
    test_X_df[col] = pd.Categorical(test_X_df[col])
    test_X_df[col] = test_X_df[col].cat.codes

#Top 5 rows of the full test dataset:
test_df.head()
```

## 2. The Model: LightGBM, with explanation tool SHAP

LightGBM is a gradient boosting framework developed by Microsoft that uses tree based learning algorithms. For more details, please see [this documentation](#).

SHAP, or SHapley Additive exPlanations, is a game theoretic approach to explain the output of any machine learning model. For more details, please see [this documentation](#).

```
[ ] import shap
import lightgbm as lgb
from sklearn.model_selection import train_test_split

# print the JS visualization code to the notebook
shap.initjs()

d_train = lgb.Dataset(X_df, label=y_df)
d_test = lgb.Dataset(test_X_df, label=test_y_df)

# train a lightgbm for the training dataset
model = lgb.train({}, d_train)

[ ] #Training accuracy
train_pred = model.predict(X_df).tolist()
train_pred = [0 if x<=0.5 else 1 for x in train_pred]
accuracy_train = round(sum(train_pred == y_df) / len(train_pred), 5)

#Test set accuracy
predictions = model.predict(test_X_df).tolist()
predictions = [0 if x <= 0.5 else 1 for x in predictions]

accuracy_test = round(sum(predictions == test_y_df) / len(predictions), 5)

print("The accuracy of the model on the training set is: ", accuracy_train)
print("The accuracy of the model on the test set is: ", accuracy_test)
```

### Global Explanation: what the model learned overall from training data

```
shap.initjs()
model.params['objective'] = 'binary'
explainer = shap.TreeExplainer(model)
shap_values = explainer.shap_values(X_df)
shap.summary_plot(shap_values, X_df, plot_type="bar")
```

▼ Partial Dependence Plots (PDPs): the relationship between a feature and the outcome variable

```
[ ] # Visualizing one feature here
# shap.dependence_plot(X_df.columns[0], shap_values[1], X_df, display_features=X_display)

# VISUALIZE ALL INPUT FEATURES
shap.initjs()
for name in X_df.columns:
    shap.dependence_plot(name, shap_values[1], X_df, display_features=X_display)
```

▼ Local explanations: how an individual prediction was made

```
[ ] #Enter a test index you want to test (i.e., from the test_df)
TEST_IDX = 1
idx = [TEST_IDX]
```

```
[ ] # Force plot using index: TEST_IDX
shap.initjs()
shap.force_plot(explainer.expected_value[1], shap_values[1][TEST_IDX,:], test_X_display.iloc[idx[0],:])
```

```
[ ] # Waterfall plot using index: TEST_IDX
shap.plots._waterfall.waterfall_legacy(explainer.expected_value[1], shap_values[1][TEST_IDX,:], test_X_display.iloc[idx[0],:])
```

### 3. Post-Study Instructions

Once you're done with the study:

1. Please download a copy of this notebook with any changes you made during the study. To do so, go to File -> Download -> Download.ipynb. We will ask you to upload this notebook at the end of the survey.
2. Please return to the survey link for additional questions.

## APPENDIX F

### Multiple-Choice Questions about the Dataset and the Model

A full list of hypotheses and analyses corresponding to the dependent variables was pre-registered on AsPredicted: [https://aspredicted.org/462\\_XKP](https://aspredicted.org/462_XKP) (anonymized link). For the hypothesis for response type (with options being plausible, accurate, or randomly inaccurate), with interactive interpretability tools, I accidentally pre-registered this as people selecting responses that are plausible when using interpretability tools with interactive features. Adding interactive elements to systems often promotes deliberative thinking and engagement. If one consistently applies this reasoning to all the metrics, then this hypothesis would inherently be that people select responses that are accurate (rather than plausible or inaccurate) when using interactive interpretability tools. All other pre-registered hypotheses about interactive interpretability tools consistently follow this reasoning.

The experimental setup included five multiple-choice questions (MCQs) as the main task of the study. This appendix presents these MCQs along with the answer key and hints provided for each of them. The legend for the experiment **conditions**:

- **Control**: Normal ML pipeline sans interpretability
- **GAM**: Visual explanations from an interpretability approach; glassbox model
- **SHAP**: Visual explanations from an interpretability approach; post-hoc explainer for a blackbox model

- **ED:** Interactive interpretability tool; Explanation Dashboard, with a glassbox model
- **WIT:** Interactive interpretability tool; What-if Tool, with a post-hoc SHAP explainer for a blackbox model

### **MCQ 1: Global Feature Importance**

**Question:** If you were forced to remove a feature from this model, which of the following features would you remove?

**Answer Options:**

- **Accurate:** EducationNum
- **Plausible and accurate:** Race
- **Visually plausible but inaccurate:** Gender
- **Heuristically plausible but inaccurate:** Age
- **Inaccurate:** Hours-per-week

**Hints:**

- **Control:** The model outputs a global explanation, i.e., importance values for all features. These can be plotted together to get overall feature importances relative to each other.
- **GAM:** The tool outputs Global Explanations which are based on overall feature importances.
- **SHAP:** The tool outputs Global Explanations which are based on the average impact of each feature on the outcome variable.
- **ED:** There is a Feature Importances section in the tool.
- **WIT:** The tool allows you to sort the global partial dependence plots for each feature by importance. Partial dependence plots can be accessed under the Datapoint editor.

### **MCQ 2: Partial Dependence**

**Question:** Which of the following ranges for Age values has the most likelihood of making a high income?

**Answer Options:**

- **Accurate:** 40–50
- **Plausible and accurate:** 55–60
- **Visually plausible but inaccurate:** 35–40
- **Heuristically plausible but inaccurate:** 73–80
- **Inaccurate:** 80–90

### Hints:

- **Control:** Python’s sklearn module offers partial dependence raw values as well as a partial dependence display for input features in a model. You can see the partial dependence raw values in the notebook. To see these values in a visual format, you can use the *from\_estimator()* function of partial dependence display (as shown in examples here). The input arguments needed for it are the same as the inputs for generating the raw partial dependence values.
- **GAM:** The tool outputs partial dependence plots that describe the relationship between an individual feature and the outcome variable.
- **SHAP:** The tool outputs dependence plots that describe the relationship between an individual feature and SHAP values. SHAP values, in turn, help the model make predictions.
- **ED:** Select aggregate feature importance under the Feature Importances section in the tool. You can click on the feature in the chart to show its dependence plot.
- **WIT:** You can access partial dependence plots from the Datapoint Editor tab. If you were already exploring the data and have a datapoint selected before accessing these plots, make sure you switch on the “Global partial dependence plots” on the left.

### MCQ 3: Predict the Outcome

**Question:** Given the following input feature values and importances, what do you think the model predicted for this individual and why?

This question also included a local explanation plot, with the answer cropped out, for all interpretability conditions. Figure F.1 presents these visuals.

FEATURE	VALUE
Age	43
WorkClass	Federal-gov
Education	Doctorate
Education-Num	16
Marital-Status	Never-married
Occupation	Prof-specialty
Race	White
Gender	Female
Capital-Gain	0
Capital-Loss	0
Hours-per-week	50
Native-Country	United-States

### Answer Options:

- **Accurate:** >50K income because most of the features have a positive influence and it adds up to greater than the negative influence.
- **Plausible and accurate:** >50K because the values of input features for this person correspond to those that have positive influence on income based on the partial dependence values.
- **Visually plausible but inaccurate:** <=50k because the intercept has a significant negative influence.
- **Heuristically plausible but inaccurate:** <=50k because the Marital-Status is “Never-married.”
- **Inaccurate:** <=50K income because the sum of all negative importances is greater than positive importances.

### Hints:

- **Control:** We can get local feature importance values, i.e., the impact of each feature in making an individual prediction, using the community-developed XGBoost Explainer. Based on the values output by this explainer, a plot for this current datapoint is included here (see Figure F.2).
- **GAM:** The predicted outcome depends on aggregated positive and negative feature importances.



- **SHAP:** SHAP’s local plots show how each contributing feature pushes the model output from the base value (the average model output over the training dataset we passed) to the predicted model output.
- **ED:** The predicted outcome depends on the aggregated positive and negative feature importances.
- **WIT:** The predicted outcome depends on the aggregated positive and negative attribution values for the datapoint.

#### MCQ 4: Explain Misclassification

**Question:** The model misclassified this datapoint with the following input feature values.

Why do you think that happened?

FEATURE	VALUE
Age	50
WorkClass	Local-gov
Education	Bachelors
Education-Num	13
Marital-Status	Married-civ-spouse
Occupation	Prof-specialty
Race	White
Gender	Female
Capital-Gain	0
Capital-Loss	0
Hours-per-week	24
Native-Country	United-States

#### Answer Options:

- **Accurate:** Because the input feature values with positive influence on the income summed up to a value greater than those with negative influence.
- **Plausible and accurate:** Because the model did not take into account that this female is probably in a part-time job situation.
- **Visually plausible but inaccurate:** Because the model assigned too much importance to “Marital-Status.”
- **Heuristically plausible but inaccurate:** Because the model assumed that a married

female would have higher household income by virtue of their spouse, without taking into consideration their own occupation.

- **Inaccurate:** Because the model took “Race = White” into consideration more than it should have.

### Hints:

- **Control:** In addition to the global explanation and the partial dependence plots, you can also access local explanations using the XGBoost Explainer. Relevant details on how to use this explainer are under the Local Explanations section in your Colab notebook.
- **GAM:** You could look at the local explanation for this datapoint. This datapoint is in the test dataframes (test\_X\_df, test\_y\_df). One way to find the index of a specific row in a dataframe is: `Rows = df[(condition1) & (condition2)]`, where a condition can be matching a specific column value like `df[“Age”] == 50`. Be sure to add those parentheses!
- **SHAP:** You could look at the local explanation for this datapoint. This datapoint is in the test dataframes (test\_X\_df, test\_y\_df). One way to find the index of a specific row in a dataframe is: `Rows = df[(condition1) & (condition2)]`, where a condition can be matching a specific column value like `df[“Age”] == 50`. Be sure to add those parentheses!
- **ED:** Individual feature importance tab in the Feature Importances section lists all correctly and incorrectly classified datapoints. When you select a datapoint, you can also see its local feature importance information.
- **WIT:** You can access an individual datapoint under the Datapoint Editor. If you select a datapoint from the scatter plot, its input feature values and attribution values will appear on the left in a table. Datapoint values are editable in the table where they are listed on the left, and you can predict the output for edited values.

### MCQ 5: What-If Question

**Question:** A person with the following input features makes  $\leq 50k$  income. A change to which of the following feature values would cause the prediction to become 1 (i.e.,  $>50K$  income)?

<b>FEATURE</b>	<b>VALUE</b>
Age	30
WorkClass	Federal-gov
Education	Some-college
Education-Num	10
Marital-Status	Married-civ-spouse
Occupation	Adm-clerical
Race	White
Gender	Male
Capital-Gain	0
Capital-Loss	0
Hours-per-week	40
Native-Country	United-States

**Answer Options:**

- **Accurate:** Change Capital-Loss to \$5000
- **Plausible and accurate:** Change Age to 40
- **Visually plausible but inaccurate:** Change Capital-Gain to \$5000
- **Heuristically plausible but inaccurate:** Change Education to “Bachelors”
- **Inaccurate:** Change Occupation to “Exec-Managerial”

**Hints:**

- **Control:** In addition to the global explanation and the partial dependence plots, you can also access local explanations using the XGBoost Explainer. Relevant details on how to use this explainer are under the Local Explanations section in your Colab notebook.
- **GAM:** You could look at the local explanation for this datapoint. This datapoint is in the test dataframes (test\_X\_df, test\_y\_df). One way to find the index of a specific row in a dataframe is: `Rows = df[(condition1) & (condition2)]`, where a condition can be matching a specific column value like `df[“Age”] == 50`. Be sure to add those

parentheses! You can also use `ebm`'s `predict()` function to get predictions for any new datapoints that you generate and want to test.

- **SHAP:** You could look at the local explanation for this datapoint. This datapoint is in the test dataframes (`test_X_df`, `test_y_df`). One way to find the index of a specific row in a dataframe is: `Rows = df[(condition1) & (condition2)]`, where a condition can be matching a specific column value like `df["Age"] == 50`. Be sure to add those parentheses! You can also use a model's `predict()` function to get predictions for any new datapoints that you generate and want to test.
- **ED:** There is a Counterfactuals section in the tool. You can get counterfactuals for individual datapoints and even edit input feature values to test outcomes. New data rows defined based on counterfactuals can appear on the scatter plot with output probabilities.
- **WIT:** You can access an individual datapoint under the Datapoint Editor. Datapoint values are editable in the table where they are listed on the left. You can also automatically generate the nearest counterfactual from the left menu, above the table with feature values.



Figure F.1: Local explanations included with the multiple-choice question about predicting the outcome given input feature values of a datapoint. These were included by default for the interpretability conditions. These explanations are output by: (a) interpretML’s version of GAMs, called Explainable Boosting Machines (EBMs); (b) SHAP applied to a lightGBM model; (c) Explanation Dashboard, using an EBM; and (d) What-If Tool, using SHAP on a lightGBM model. (a) and (b) are examples of static interpretability tools whereas (c) and (d) are from interactive tools. (a) and (c) are glassbox approaches, and the other two are post-hoc explainers for blackbox models.

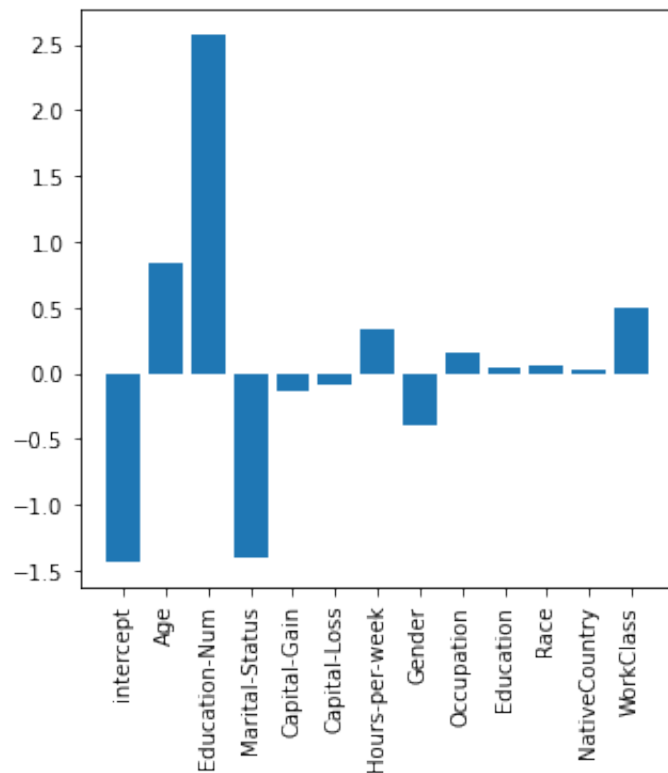


Figure F.2: Local explanation for the multiple-choice question about predicting the outcome given input feature values of a datapoint. This chart is output by the XGBoost Explainer and included as a hint for the control condition participants.

## APPENDIX G

### Descriptive Statistics for All Independent Variables

Figure G.1 presents descriptive statistics for all independent variables. 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 (about the use of data and model in the wild and of accuracy as a key performance indicator) was  $\alpha=0.73$ . This  $\alpha$  value is acceptable for exploratory research [170]. Second, error recognition, since the ratings for the two error-related questions about missing values and redundant features were strongly correlated (Pearson's  $r(115)=0.76$ ,  $p < 0.001$ ).

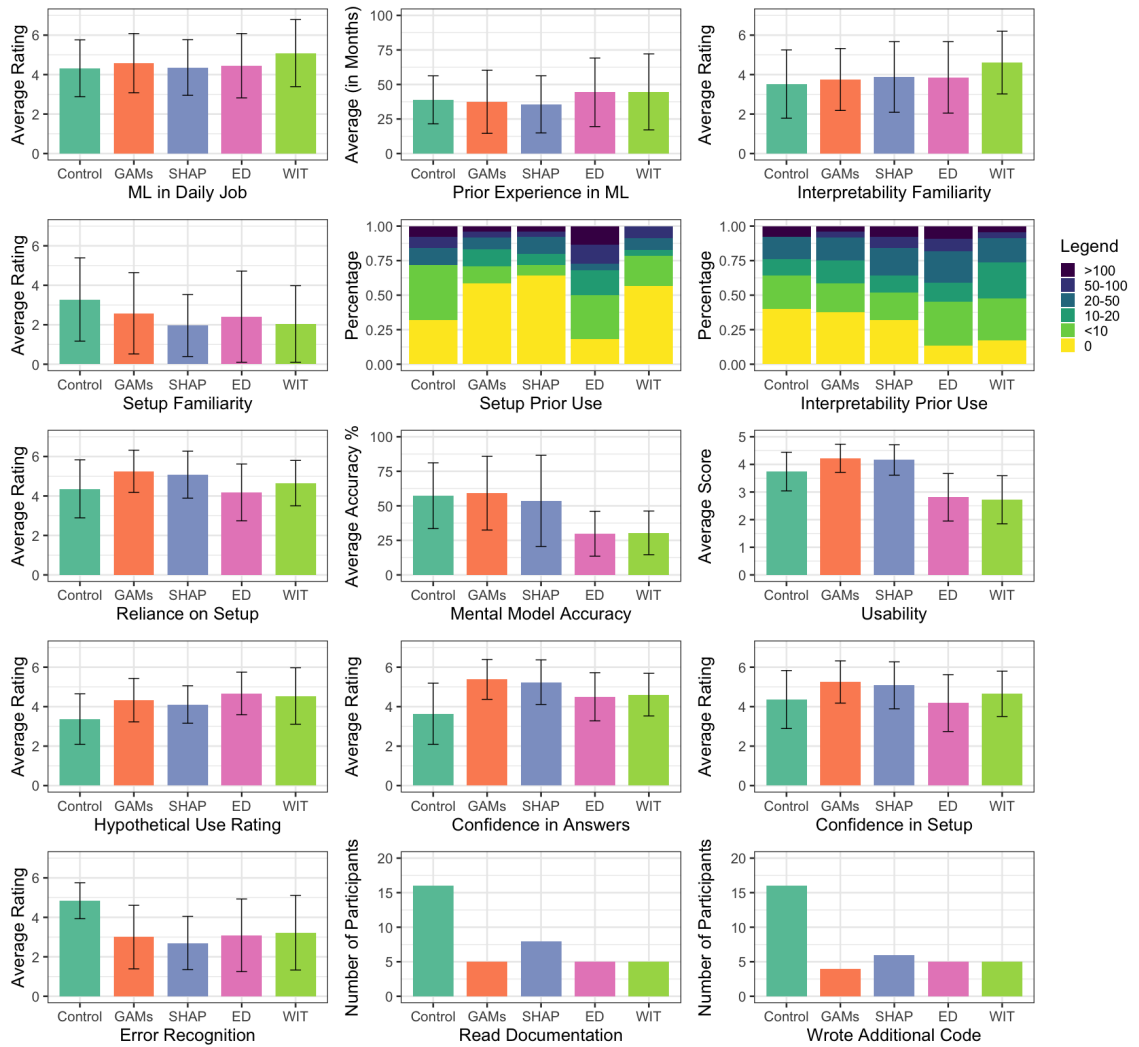


Figure G.1: Descriptive statistics for all the independent variables in the study.



## **BIBLIOGRAPHY**

## BIBLIOGRAPHY

- [1] Development of nasa-tlx (task load index): Results of empirical and theoretical research. In Peter A. Hancock and Najmedin Meshkati, editors, *Human Mental Workload*, volume 52 of *Advances in Psychology*, pages 139–183. North-Holland, 1988.
- [2] Law enforcement: Predpol law enforcement intelligence led policing software: Predpol law enforcement intelligence led policing software, Sep 2020.
- [3] Hussein Abbass, Axel Bender, Svetoslav Gaidow, and Paul Whitbread. Computational red teaming: Past, present and future. *IEEE Computational Intelligence Magazine*, 6(1):30–42, 2011.
- [4] Ashraf Abdul, Jo Vermeulen, Danding Wang, Brian Y. Lim, and Mohan Kankanhalli. 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*, CHI '18, pages 582:1–582:18, New York, NY, USA, 2018. ACM.
- [5] Mark S Ackerman. The intellectual challenge of cscw: The gap between social requirements and technical feasibility. *Human–Computer Interaction*, 15(2-3):179–203, 2000.
- [6] Ali Alkhatib. To live in their utopia: Why algorithmic systems create absurd outcomes.

In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pages 1–9, 2021.

- [7] Gordon W Allport. The historical background of social psychology (vol. 1). *The handbook of social psychology*, 1985.
- [8] David Alvarez-Melis, Hal Daumé, III, Jennifer Wortman Vaughan, and Hanna Wallach. Weight of evidence as a basis for human-oriented explanations. *arXiv preprint arXiv:1910.13503*, 2019.
- [9] David Alvarez-Melis and Tommi S Jaakkola. 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*, pages 412–421, 2017.
- [10] Saleema Amershi, Maya Cakmak, William Bradley Knox, and Todd Kulesza. Power to the people: The role of humans in interactive machine learning. *AI Magazine*, 35(4):105–120, 2014.
- [11] Saleema Amershi, James Fogarty, and Daniel Weld. 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, pages 21–30, New York, NY, USA, 2012. ACM.
- [12] Saleema Amershi, Dan Weld, Mihaela Vorvoreanu, Adam Fourney, Besmira Nushi, Penny Collisson, Jina Suh, Shamsi Iqbal, Paul N Bennett, Kori Inkpen, et al. Guidelines for human-ai interaction. In *Proceedings of the 2019 chi conference on human factors in computing systems*, pages 1–13, 2019.
- [13] Dario Amodei, Chris Olah, Jacob Steinhardt, Paul Christiano, John Schulman, and Dan Mané. Concrete problems in ai safety. *arXiv preprint arXiv:1606.06565*, 2016.

- [14] Julia Angwin, Jeff Larson, Surya Mattu, and Kirchner Lauren. Machine bias: There's software used across the country to predict future criminals. and it's biased against blacks. *ProPublica, May*, 23:2016, 2016.
- [15] Mary Ann Glynn and Lee Watkiss. Of organizing and sensemaking: from action to meaning and back again in a half-century of weick's theorizing. *Journal of Management Studies*, 57(7):1331–1354, 2020.
- [16] 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. Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai. *Information Fusion*, 58:82–115, 6 2020.
- [17] Pierre Baldi and Laurent Itti. Of bits and wows: A bayesian theory of surprise with applications to attention. *Neural Networks*, 23(5):649–666, 2010.
- [18] Gagan Bansal, Tongshuang Wu, Joyce Zhou, Raymond Fok, Besmira Nushi, Ece Kamar, Marco Tulio Ribeiro, and Daniel Weld. 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*, pages 1–16, 2021.
- [19] Dean C Barnlund. A transactional model of communication. In C. David Mortensen, editor, *Communication theory, Second edition*, pages 47–57. Routledge, 2017.
- [20] Emma Beede, Elizabeth Baylor, Fred Hersch, Anna Iurchenko, Lauren Wilcox, Paisan Ruamviboonsuk, and Laura M Vardoulakis. A human-centered evaluation of a deep learning system deployed in clinics for the detection of diabetic retinopathy. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pages 1–12, 2020.

- [21] Victoria Bellotti and Keith Edwards. Intelligibility and accountability: Human considerations in context-aware systems. *Human-Computer Interaction*, 16(2-4):193–212, 2001.
- [22] Nelly Bencomo and Amel Belaggoun. 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*, pages 460–463, 2014.
- [23] Emily M Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, pages 610–623, 2021.
- [24] Ellen Berscheid. Emotion. In H.H. Kelley, E. Berscheid, A. Christensen, J. Harvey, T. Huston, G. Levinger, E. McClintock, A. Peplau, and D.R. Peterson, editors, *Close Relationships*, pages 110–168. WH Freeman, 1983.
- [25] Astrid Bertrand, Rafik Belloum, James R Eagan, and Winston Maxwell. How cognitive biases affect xai-assisted decision-making: A systematic review. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*, pages 78–91, 2022.
- [26] Herbert Blumer. *Symbolic interactionism*, volume 50. Englewood Cliffs, NJ: Prentice-Hall, 1969.
- [27] Nick Bostrom and Eliezer Yudkowsky. The ethics of artificial intelligence. *The Cambridge handbook of artificial intelligence*, 1:316–334, 2014.
- [28] Clara Bove, Marie-Jeanne Lesot, Charles Albert Tijus, and Marcin Detyniecki. Investigating the intelligibility of plural counterfactual examples for non-expert users: an explanation user interface proposition and user study. In *Proceedings of the 28th International Conference on Intelligent User Interfaces*, pages 188–203, 2023.

- [29] Michelle Brachman, Qian Pan, Hyo Jin Do, Casey Dugan, Arunima Chaudhary, James M Johnson, Priyanshu Rai, Tathagata Chakraborti, Thomas Gschwind, Jim A Laredo, et al. Follow the successful herd: Towards explanations for improved use and mental models of natural language systems. In *Proceedings of the 28th International Conference on Intelligent User Interfaces*, pages 220–239, 2023.
- [30] Virginia Braun and Victoria Clarke. Thematic analysis. In *APA handbook of research methods in psychology, Vol 2: Research designs: Quantitative, qualitative, neuropsychological, and biological.*, APA handbooks in psychology®, pages 57–71. American Psychological Association, Washington, DC, US, 2012.
- [31] Marilyn B Brewer and Wendi Gardner. Who is this "we"? levels of collective identity and self representations. *Journal of personality and social psychology*, 71(1):83, 1996.
- [32] T Bruns and GM Stalker. The management of innovation. *Tavistock, London*, pages 120–122, 1961.
- [33] Zana Buçinca, Maja Barbara Malaya, and Krzysztof Z Gajos. 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):1–21, 2021.
- [34] Andrea Bunt, Matthew Lount, and Catherine Lauzon. Are explanations always important? a study of deployed, low-cost intelligent interactive systems. In *Proceedings of the 2012 ACM international conference on Intelligent User Interfaces*, pages 169–178, 2012.
- [35] Jason W Burton, Mari-Klara Stein, and Tina Blegind Jensen. A systematic review of algorithm aversion in augmented decision making. *Journal of Behavioral Decision Making*, 33(2):220–239, 2020.

- [36] Adrian Bussone, Simone Stumpf, and Dymrna O’Sullivan. The role of explanations on trust and reliance in clinical decision support systems. In *2015 international conference on healthcare informatics*, pages 160–169. IEEE, 2015.
- [37] Carrie J. Cai, Samantha Winter, David Steiner, Lauren Wilcox, and Michael Terry. "hello ai": Uncovering the onboarding needs of medical practitioners for human-ai collaborative decision-making. *Proc. ACM Hum.-Comput. Interact.*, 3(CSCW):104:1–104:24, November 2019.
- [38] Aylin Caliskan, Joanna J. Bryson, and Arvind Narayanan. Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334):183–186, 2017.
- [39] Rich Caruana, Yin Lou, Johannes Gehrke, Paul Koch, Marc Sturm, and Noemie Elhadad. 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, pages 1721–1730, New York, NY, USA, 2015. ACM.
- [40] Federico Maria Cau, Hanna Hauptmann, Lucio Davide Spano, and Nava Tintarev. Supporting high-uncertainty decisions through ai and logic-style explanations. In *Proceedings of the 28th International Conference on Intelligent User Interfaces*, pages 251–263, 2023.
- [41] Matthew Chalmers, Ian MacColl, and Marek Bell. Seamful design: Showing the seams in wearable computing. In *2003 IEE Eurowearable*, pages 11–16. IET, 2003.
- [42] Tianlang Chen, Zhongping Zhang, Quanzeng You, Chen Fang, Zhaowen Wang, Hailin Jin, and Jiebo Luo. “factual”or“emotional”: Stylized image captioning with adaptive learning and attention. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 519–535, 2018.

- [43] Isaac Cho, Ryan Wesslen, Alireza Karduni, Sashank Santhanam, Samira Shaikh, and Wenwen Dou. The anchoring effect in decision-making with visual analytics. In *2017 IEEE Conference on Visual Analytics Science and Technology (VAST)*, pages 116–126. IEEE, 2017.
- [44] Herbert H Clark, Robert Schreuder, and Samuel Buttrick. Common ground at the understanding of demonstrative reference. *Journal of verbal learning and verbal behavior*, 22(2):245–258, 1983.
- [45] Sharolyn Converse, JA Cannon-Bowers, and E Salas. Shared mental models in expert team decision making. *Individual and group decision making: Current issues*, 221:221–46, 1993.
- [46] Juliet M. Corbin and Anselm Strauss. Grounded theory research: Procedures, canons, and evaluative criteria. *Qualitative sociology*, 13(1):3–21, 1990.
- [47] Valdemar Danry, Pat Pataranutaporn, Yaoli Mao, and Pattie Maes. 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*, pages 1–13, 2023.
- [48] Alex J DeGrave, Joseph D Janizek, and Su-In Lee. Ai for radiographic covid-19 detection selects shortcuts over signal. *Nature Machine Intelligence*, pages 1–10, 2021.
- [49] Janez Demšar, Blaž Zupan, Gregor Leban, and Tomaz Curk. Orange: From experimental machine learning to interactive data mining. In Jean-François Boulicaut, Floriana Esposito, Fosca Giannotti, and Dino Pedreschi, editors, *Knowledge Discovery in Databases: PKDD 2004*, pages 537–539, Berlin, Heidelberg, 2004. Springer Berlin Heidelberg.



- [50] Wesley Hanwen Deng, Manish Nagireddy, Michelle Seng Ah Lee, Jatinder Singh, Zhiwei Steven Wu, Kenneth Holstein, and Haiyi Zhu. Exploring how machine learning practitioners (try to) use fairness toolkits. In *2022 ACM Conference on Fairness, Accountability, and Transparency, FAccT '22*, page 473–484, New York, NY, USA, 2022. Association for Computing Machinery.
- [51] Berkeley J Dietvorst, Joseph P Simmons, and Cade Massey. Algorithm aversion: People erroneously avoid algorithms after seeing them err. *Journal of Experimental Psychology: General*, 144(1):114, 2015.
- [52] Wilhelm Dilthey and Frederic Jameson. The rise of hermeneutics. *New literary history*, 3(2):229–244, 1972.
- [53] Carl DiSalvo. *Adversarial design*. Design Thinking, Design Theory, 2015.
- [54] Finale Doshi-Velez and Been Kim. Towards a rigorous science of interpretable machine learning. 2.
- [55] Finale Doshi-Velez, Mason Kortz, Ryan Budish, Chris Bavitz, Sam Gershman, David O’Brien, Stuart Schieber, James Waldo, David Weinberger, and Alexandra Wood. Accountability of ai under the law: The role of explanation. *arXiv preprint arXiv:1711.01134*, 2017.
- [56] Paul Dourish. Algorithms and their others: Algorithmic culture in context. *Big Data & Society*, 3(2):2053951716665128, 2016.
- [57] John J Dudley and Per Ola Kristensson. A review of user interface design for interactive machine learning. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, 8(2):1–37, 2018.
- [58] Rudresh Dwivedi, Devam Dave, Het Naik, Smiti Singhal, Rana Omer, Pankesh Patel,

- Bin Qian, Zhenyu Wen, Tejal Shah, Graham Morgan, et al. Explainable ai (xai): Core ideas, techniques, and solutions. *ACM Computing Surveys*, 55(9):1–33, 2023.
- [59] Mary T Dzindolet, Hall P Beck, Linda G Pierce, and Lloyd A Dawe. A framework of automation use. Technical report, Army Research Lab Aberdeen Proving Ground MD, 2001.
- [60] Upol Ehsan, Q. Vera Liao, Michael Muller, Mark O. Riedl, and Justin D. Weisz. Expanding explainability: Towards social transparency in ai systems. *Conference on Human Factors in Computing Systems - Proceedings*, 19, 5.
- [61] Upol Ehsan, Samir Passi, Q Vera Liao, Larry Chan, I Lee, Michael Muller, Mark O Riedl, et al. The who in explainable ai: How ai background shapes perceptions of ai explanations. *arXiv preprint arXiv:2107.13509*, 2021.
- [62] Miriam Erez, P Christopher Earley, et al. *Culture, self-identity, and work*. Oxford University Press on Demand, 1993.
- [63] Jerry Alan Fails and Dan R. Olsen, Jr. Interactive machine learning. In *Proceedings of the 8th International Conference on Intelligent User Interfaces, IUI '03*, pages 39–45, New York, NY, USA, 2003. ACM.
- [64] Gary Alan Fine and Lazaros Christoforides. Dirty birds, filthy immigrants, and the english sparrow war: Metaphorical linkage in constructing social problems. *Symbolic Interaction*, 14(4):375–393, 1991.
- [65] Susan T Fiske. Thinking is for doing: portraits of social cognition from daguerreotype to laserphoto. *Journal of personality and social psychology*, 63(6):877, 1992.
- [66] Mary Parker Follett. *Creative experience*. Longmans, Green and company, 1924.
- [67] Gerd Gigerenzer. How to make cognitive illusions disappear: Beyond “heuristics and biases”. *European review of social psychology*, 2(1):83–115, 1991.

- [68] Leilani H Gilpin, David Bau, Ben Z Yuan, Ayesha Bajwa, Michael Specter, and Lalana Kagal. Explaining explanations: An overview of interpretability of machine learning. In *2018 IEEE 5th International Conference on Data Science and Advanced Analytics (DSAA)*, pages 80–89. IEEE, 2018.
- [69] Bryce Goodman and Seth Flaxman. European union regulations on algorithmic decision-making and a “right to explanation”. *AI magazine*, 38(3):50–57, 2017.
- [70] Herbert P. Grice. Logic and conversation. pages 41–58, 1975.
- [71] Henning Griethe and Heidrun Schumann. Visualizing uncertainty for improved decision making. In *Proceedings of 4th International Conference on Perspectives in Business Informatics Research (BIR 2005)*, volume 20, 2005.
- [72] David Gunning and David Aha. Darpa’s explainable artificial intelligence (xai) program. *AI Magazine*, 40(2):44–58, 2019.
- [73] Hangzhi Guo and Na Li. Factors impacting k-12 teachers in understanding explanations of machine learning model on students’ performance. 2020.
- [74] Sophia Hadash, Martijn C Willemsen, Chris Snijders, and Wijnand A IJsselsteijn. Improving understandability of feature contributions in model-agnostic explainable ai tools. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, pages 1–9, 2022.
- [75] Moritz Hardt, Eric Price, and Nati Srebro. Equality of opportunity in supervised learning. *Advances in neural information processing systems*, 29:3315–3323, 2016.
- [76] C. Hartshorne. Mind as memory and creative love. In Jordan M. Scher, editor, *Theories of the Mind*, pages 440–463. Free Press of Glencoe, 1962.
- [77] Bertrand K Hassani. Societal bias reinforcement through machine learning: a credit scoring perspective. *AI and Ethics*, 1(3):239–247, 2021.

- [78] Trevor Hastie and Robert Tibshirani. Generalized additive models: Some applications. *Journal of the American Statistical Association*, 82(398):371–386, 1987.
- [79] Pamela R Haunschild and Anne S Miner. Modes of interorganizational imitation: The effects of outcome salience and uncertainty. *Administrative science quarterly*, pages 472–500, 1997.
- [80] Will Douglas Heaven. Predictive policing algorithms are racist. they need to be dismantled. *MIT Technology Review*, 17:2020, 2020.
- [81] Carl G Hempel and Paul Oppenheim. Studies in the logic of explanation. *Philos. Sci.*, 15(2):135–175, 1948.
- [82] E Tory Higgins. Achieving ‘shared reality’ in the communication game: A social action that create; meaning. *Journal of Language and Social Psychology*, 11(3):107–131, 1992.
- [83] Denis J Hilton. Mental models and causal explanation: Judgements of probable cause and explanatory relevance. *Thinking & Reasoning*, 2(4):273–308, 1996.
- [84] Robert R Hoffman, Shane T Mueller, Gary Klein, and Jordan Litman. Metrics for explainable ai: Challenges and prospects. *arXiv preprint arXiv:1812.04608*, 2018.
- [85] Fred Hohman, Andrew Head, Rich Caruana, Robert DeLine, and Steven M. Drucker. 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, pages 579:1–579:13, New York, NY, USA, 2019. ACM.
- [86] Kenneth Holstein, Jennifer Wortman Vaughan, Hal Daumé III, Miro Dudik, and Hanna Wallach. Improving fairness in machine learning systems: What do industry practitioners need? In *Proceedings of the 2019 CHI conference on human factors in computing systems*, pages 1–16, 2019.

- [87] Miroslav Holub. 'brief thoughts on maps', 1977.
- [88] Andreas Holzinger, André Carrington, and Heimo Müller. Measuring the quality of explanations: the system causability scale (scs). *KI-Künstliche Intelligenz*, 34(2):193–198, 2020.
- [89] Andreas Holzinger, Georg Langs, Helmut Denk, Kurt Zatloukal, and Heimo Müller. Causability and explainability of artificial intelligence in medicine. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 9(4):e1312, 2019.
- [90] Lu Hong and Scott Page. The contributions of diversity, accuracy, and group size on collective accuracy. *Accuracy, and Group Size on Collective Accuracy (October 15, 2020)*, 2020.
- [91] Lu Hong and Scott E Page. Groups of diverse problem solvers can outperform groups of high-ability problem solvers. *Proceedings of the National Academy of Sciences*, 101(46):16385–16389, 2004.
- [92] Sungsoo Ray Hong, Jessica Hullman, and Enrico Bertini. Human factors in model interpretability: Industry practices, challenges, and needs. *Proceedings of the ACM on Human-Computer Interaction*, 4(CSCW1):1–26, 2020.
- [93] Jessica Hullman, Xiaoli Qiao, Michael Correll, Alex Kale, and Matthew Kay. In pursuit of error: A survey of uncertainty visualization evaluation. *IEEE transactions on visualization and computer graphics*, 25(1):903–913, 2018.
- [94] E. Hutchins. The social organization of distributed cognition. (Washington, DC), 1991.
- [95] William Ickes and Richard Gonzalez. " social" cognition and social cognition: From the subjective to the intersubjective. *Small group research*, 25(2):294–315, 1994.

- [96] William Ickes, Eric Robertson, William Tooke, and Gary Teng. Naturalistic social cognition: Methodology, assessment, and validation. *Journal of Personality and Social Psychology*, 51(1):66, 1986.
- [97] William Ickes and William Tooke. The observational method: Studying the interaction of minds and bodies. 1988.
- [98] Sarah Inman and David Ribes. " beautiful seams" strategic revelations and concealments. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, pages 1–14, 2019.
- [99] Daniel J Isenberg. The structure and process of understanding: Implications for managerial action. In H.P. Sims Jr. and D.A. Gioia, editors, *The Thinking Organization*, pages 238–262. Jossey Bass, San Fransisco, 1986.
- [100] Laurent Itti and Pierre Baldi. Bayesian surprise attracts human attention. *Vision research*, 49(10):1295–1306, 2009.
- [101] William James. *The principles of psychology*, volume 1. Cosimo, Inc., 2007.
- [102] Marijn Janssen, Martijn Hartog, Ricardo Matheus, Aaron Yi Ding, and George Kuk. Will algorithms blind people? the effect of explainable ai and decision-makers' experience on ai-supported decision-making in government. *Social Science Computer Review*, page 0894439320980118, 2020.
- [103] P Devereaux Jennings and Royston Greenwood. Constructing the iron cage: Institutional theory and enactment. *Debating organization: point-counterpoint in organization studies*, 195, 2003.
- [104] Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12):1–38, 2023.

- [105] Jiun-Yin Jian, Ann M. Bisantz, and Colin G. Drury. Foundations for an empirically determined scale of trust in automated systems. *International Journal of Cognitive Ergonomics*, 4(1):53–71, 2000.
- [106] Jongbin Jung, Connor Concannon, Ravi Shroff, Sharad Goel, and Daniel G. Goldstein. Simple rules for complex decisions. *SSRN Electronic Journal*, 2.
- [107] Mayank Kabra, Alice A Robie, Marta Rivera-Alba, Steven Branson, and Kristin Branson. Jaaba: interactive machine learning for automatic annotation of animal behavior. *Nature methods*, 10(1):64–67, 2013.
- [108] Daniel Kahneman. A perspective on judgment and choice: mapping bounded rationality. *American psychologist*, 58(9):697, 2003.
- [109] Daniel Kahneman. *Thinking, fast and slow*. Macmillan, 2011.
- [110] Daniel Kahneman, Stewart Paul Slovic, Paul Slovic, and Amos Tversky. *Judgment under uncertainty: Heuristics and biases*. Cambridge university press, 1982.
- [111] Harmanpreet Kaur, Harsha Nori, Samuel Jenkins, Rich Caruana, Hanna Wallach, and Jennifer Wortman Vaughan. 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*, pages 1–14, 2020.
- [112] Harmanpreet Kaur, Alex C Williams, Daniel McDuff, Mary Czerwinski, Jaime Teevan, and Shamsi T Iqbal. Optimizing for happiness and productivity: Modeling opportune moments for transitions and breaks at work. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pages 1–15, 2020.
- [113] Matthew Kay, Cynthia Matuszek, and Sean A. Munson. 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*, pages 3819–3828, New York, NY, USA, 2015. ACM.

- [114] Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. Lightgbm: A highly efficient gradient boosting decision tree. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 3146–3154. Curran Associates, Inc., 2017.
- [115] Been Kim, Rajiv Khanna, and Oluwasanmi O Koyejo. Examples are not enough, learn to criticize! criticism for interpretability. In D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems 29*, pages 2280–2288. Curran Associates, Inc., 2016.
- [116] Been Kim, Martin Wattenberg, Justin Gilmer, Carrie Cai, James Wexler, Fernanda Viegas, and Rory Sayres. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (TCAV). In Jennifer Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 2668–2677, Stockholmström, Stockholm Sweden, 10–15 Jul 2018. PMLR.
- [117] Sunnie SY Kim, Elizabeth Anne Watkins, Olga Russakovsky, Ruth Fong, and Andrés Monroy-Hernández. " help me help the ai": Understanding how explainability can support human-ai interaction. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, pages 1–17, 2023.
- [118] Richard Klimoski and Susan Mohammed. Team mental model: Construct or metaphor? *Journal of management*, 20(2):403–437, 1994.
- [119] Karin D. Knorr-Cetina. The micro-sociological challenge of macro-sociology : towards a reconstruction of social theory and methodology. In K. Knorr-Cetina



- and A. V. Cicourel, editors, *Advances in social theory and methodology: toward an integration of micro- and macro-sociologies*, pages 1–47. Routledge & Kegan Paul, Boston, 1981.
- [120] Rafal Kocielnik, Saleema Amershi, and Paul N. Bennett. 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, pages 411:1–411:14, New York, NY, USA, 2019. ACM.
- [121] Charles Kostelnick. The re-emergence of emotional appeals in interactive data visualization. *Technical Communication*, 63(2):116–135, 2016.
- [122] Ravi S Kudesia. Organizational sensemaking. In *Oxford research encyclopedia of psychology*. 2017.
- [123] Todd Kulesza, Margaret Burnett, Weng-Keen Wong, and Simone Stumpf. Principles of explanatory debugging to personalize interactive machine learning. In *Proceedings of the 20th International Conference on Intelligent User Interfaces*, IUI '15, pages 126–137, New York, NY, USA, 2015. ACM.
- [124] Todd Kulesza, Simone Stumpf, Margaret Burnett, Sherry Yang, Irwin Kwan, and Weng-Keen Wong. Too much, too little, or just right? ways explanations impact end users' mental models. In *2013 IEEE Symposium on visual languages and human centric computing*, pages 3–10. IEEE, 2013.
- [125] Isaac Lage, Emily Chen, Jeffrey He, Menaka Narayanan, Sam Gershman, Been Kim, and Finale Doshi-Velez. Human evaluation of models built for interpretability. In *AAAI Conference on Human Computation and Crowdsourcing (HCOMP)*, 2019.
- [126] Vivian Lai, Chacha Chen, Q Vera Liao, Alison Smith-Renner, and Chenhao Tan. Towards a science of human-ai decision making: a survey of empirical studies. *arXiv preprint arXiv:2112.11471*, 2021.

- [127] Himabindu Lakkaraju, Stephen H Bach, and L Jure. 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, 2016.
- [128] Himabindu Lakkaraju and Osbert Bastani. "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, New York, NY, USA, 2020. ACM.
- [129] Ellen J Langer. Minding matters: The consequences of mindlessness–mindfulness. In *Advances in experimental social psychology*, volume 22, pages 137–173. Elsevier, 1989.
- [130] David B. Leake. Goal-based explanation evaluation. *Cognitive Science*, 15(4):509–545, 1991.
- [131] Jaewook Lee, Jaylin Herskovitz, Yi-Hao Peng, and Anhong Guo. 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*, pages 1–15, 2022.
- [132] Yoonjoo Lee, Tae Soo Kim, Sungdong Kim, Yohan Yun, and Juho Kim. 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*, pages 1–22, 2023.
- [133] Kenneth Leiter. *A primer on ethnomethodology*. Oxford University Press, USA, 1980.
- [134] John M Levine. Socially-shared cognition and consensus in small groups. *Current opinion in psychology*, 23:52–56, 2018.

- [135] John M Levine, Lauren B Resnick, and E Tory Higgins. Social foundations of cognition. *Annual review of psychology*, 44(1):585–612, 1993.
- [136] Q Vera Liao, Daniel Gruen, and Sarah Miller. Questioning the ai: informing design practices for explainable ai user experiences. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pages 1–15, 2020.
- [137] Q Vera Liao, Hariharan Subramonyam, Jennifer Wang, and Jennifer Wortman Vaughan. 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*, pages 1–21, 2023.
- [138] Q Vera Liao and Kush R Varshney. Human-centered explainable ai (xai): From algorithms to user experiences. *arXiv preprint arXiv:2110.10790*, 2021.
- [139] Q Vera Liao and Jennifer Wortman Vaughan. Ai transparency in the age of llms: A human-centered research roadmap. *arXiv preprint arXiv:2306.01941*, 2023.
- [140] Brian Y Lim, Qian Yang, Ashraf M Abdul, and Danding Wang. Why these explanations? selecting intelligibility types for explanation goals. In *IUI Workshops*, 2019.
- [141] Peter Lipton. Contrastive explanation. *Royal Institute of Philosophy Supplements*, 27:247–266, 1990.
- [142] Zachary C. Lipton. The mythos of model interpretability. *Communications of the ACM*, 61:36–43, 9.
- [143] Zachary C Lipton. The doctor just won’t accept that! *arXiv preprint arXiv:1711.08037*, 2017.
- [144] Stine Lomborg and Patrick Heiberg Kapsch. Decoding algorithms. *Media, Culture & Society*, 2019.

- [145] Tania Lombrozo. The structure and function of explanations. *Trends in Cognitive Sciences*, 10(10):464–470, 2006.
- [146] Tania Lombrozo. Explanation and abductive inference. 2012.
- [147] Alexandra L'heureux, Katarina Grolinger, Hany F Elyamany, and Miriam AM Capretz. Machine learning with big data: Challenges and approaches. *IEEE Access*, 5:7776–7797, 2017.
- [148] Scott M Lundberg, Gabriel G Erion, and Su-In Lee. Consistent individualized feature attribution for tree ensembles. *arXiv preprint arXiv:1802.03888*, 2018.
- [149] Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 4765–4774. Curran Associates, Inc., 2017.
- [150] Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 4765–4774. Curran Associates, Inc., 2017.
- [151] Michael A. Madaio, Luke Stark, Jennifer Wortman Vaughan, and Hanna Wallach. Co-designing checklists to understand organizational challenges and opportunities around fairness in ai. New York, NY, USA, 2020. Association for Computing Machinery.
- [152] Prashan Madumal, Tim Miller, Frank Vetere, and Liz Sonenberg. Towards a grounded dialog model for explainable artificial intelligence. In *First international workshop on socio-cognitive systems at IJCAI 2018*, 2018.
- [153] Sasu Mäkinen, Henrik Skogström, Eero Laaksonen, and Tommi Mikkonen. Who

- needs mlops: What data scientists seek to accomplish and how can mlops help? *arXiv preprint arXiv:2103.08942*, 2021.
- [154] Bertram F Malle. *How the mind explains behavior: Folk explanations, meaning, and social interaction*. Mit Press, 2006.
- [155] George Mandler. *Mind and body: Psychology of emotion and stress*. WW Norton & Company Incorporated, 1984.
- [156] Filip Matějka and Alisdair McKay. Rational inattention to discrete choices: A new foundation for the multinomial logit model. *American Economic Review*, 105(1):272–98, January 2015.
- [157] George Herbert Mead. *Mind, self and society*, volume 111. Chicago University of Chicago Press., 1934.
- [158] David Alvarez Melis, Harmanpreet Kaur, Hal Daumé III, Hanna Wallach, and Jennifer Wortman Vaughan. From human explanation to model interpretability: A framework based on weight of evidence. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, volume 9, pages 35–47, 2021.
- [159] Matthew B Miles and A Michael Huberman. *Qualitative data analysis: An expanded sourcebook*. sage, 1994.
- [160] John Stuart Mill. On the definition of political economy; and on the method of investigation proper to it. *London and Westminster Review*, 4(October):120–164, 1836.
- [161] Tim Miller. Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence*, 267:1–38, 2019.
- [162] Tim Miller. Contrastive explanation: A structural-model approach. *The Knowledge Engineering Review*, 36, 2021.

- [163] Tim Miller, Piers Howe, and Liz Sonenberg. 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)*, 2017.
- [164] Chantal Mouffe. *Agonistics: Thinking the world politically*. Verso Books, 2013.
- [165] Jakob Nielsen. Ten usability heuristics, 2005.
- [166] Richard E Nisbett and Timothy D Wilson. Telling more than we can know: Verbal reports on mental processes. *Psychological review*, 84(3):231, 1977.
- [167] Harsha Nori, Samuel Jenkins, Paul Koch, and Rich Caruana. Interpretml: A unified framework for machine learning interpretability. *arXiv preprint arXiv:1909.09223*, 2019.
- [168] Donald A Norman. Some observations on mental models. In *Mental models*, pages 15–22. Psychology Press, 2014.
- [169] Mahsan Nourani, Chiradeep Roy, Jeremy E Block, Donald R Honeycutt, Tahrira Rahman, Eric Ragan, and Vibhav Gogate. Anchoring bias affects mental model formation and user reliance in explainable ai systems. In *26th International Conference on Intelligent User Interfaces*, pages 340–350, 2021.
- [170] Jum C Nunnally. An overview of psychological measurement. *Clinical diagnosis of mental disorders*, pages 97–146, 1978.
- [171] Jeffrey M; Zhu Pingping; Sommer Marc A.; Ferrari Silvia.; Egner Tobias Oh, Hanna.; Beck. Satisficing in split-second decision making is characterized by strategic cue discounting, 2016.
- [172] Clemens Otte. Safe and interpretable machine learning: a methodological review. *Computational intelligence in intelligent data analysis*, pages 111–122, 2013.

- [173] Heather O'Brien. Theoretical perspectives on user engagement. In *Why engagement matters*, pages 1–26. Springer, 2016.
- [174] Andrés Páez. The pragmatic turn in explainable artificial intelligence (xai). *Minds and Machines*, 29(3):441–459, 2019.
- [175] Cecilia Panigutti, Andrea Beretta, Fosca Giannotti, and Dino Pedreschi. 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*, pages 1–9, 2022.
- [176] Raja Parasuraman, Robert Molloy, and Indramani L. Singh. Performance consequences of automation-induced 'complacency'. *The International Journal of Aviation Psychology*, 3(1):1–23, 1993.
- [177] Kayur Patel, James Fogarty, James A. Landay, and Beverly Harrison. Investigating statistical machine learning as a tool for software development. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '08, pages 667–676, New York, NY, USA, 2008. ACM.
- [178] Charles Sanders Peirce. Illustrations of the logic of science: IV the probability of induction. *Popular Science Monthly*, 12:705–718, April 1878.
- [179] Peter Pirolli and Stuart Card. The sensemaking process and leverage points for analyst technology as identified through cognitive task analysis. In *Proceedings of international conference on intelligence analysis*, volume 5, pages 2–4. McLean, VA, USA, 2005.
- [180] Joseph C Pitt. *Theories of explanation*. Oxford University Press, 1988.
- [181] Louis R Pondy and Ian I Mitroff. Beyond open system models of organization. *Research in organizational behavior*, 1(1):3–39, 1979.

- [182] Jan Pöppel and Stefan Kopp. 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*, AAMAS '18, page 470–478, Richland, SC, 2018. International Foundation for Autonomous Agents and Multiagent Systems.
- [183] Forough Poursabzi-Sangdeh, Daniel G Goldstein, Jake M Hofman, Jennifer Wortman Vaughan, and Hanna Wallach. Manipulating and measuring model interpretability. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pages 1–52, 2021.
- [184] Snehal Prabhudesai, Leyao Yang, Sumit Asthana, Xun Huan, Q Vera Liao, and Nikola Banovic. 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*, pages 379–396, 2023.
- [185] J R Quinlan. Induction of decision trees. *Mach. Learn.*, 1986.
- [186] Gonzalo Ramos, Christopher Meek, Patrice Simard, Jina Suh, and Soroush Ghorashi. Interactive machine teaching: a human-centered approach to building machine-learned models. *Human–Computer Interaction*, 35(5-6):413–451, 2020.
- [187] Charvi Rastogi, Yunfeng Zhang, Dennis Wei, Kush R Varshney, Amit Dhurandhar, and Richard Tomsett. Deciding fast and slow: The role of cognitive biases in ai-assisted decision-making. *Proceedings of the ACM on Human-Computer Interaction*, 6(CSCW1):1–22, 2022.
- [188] James Reason. *Human error*. Cambridge university press, 1990.
- [189] Lauren B Resnick, John M Levine, and Stephanie D Teasley. *Perspectives on socially shared cognition*. Number Washington, DC :. American Psychological Association, 1st ed. edition, 1991.



- [190] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "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, pages 1135–1144, New York, NY, USA, 2016. ACM.
- [191] Hans Peter Rickman. *Dilthey selected writings*. 1979.
- [192] Peter S Ring and Andrew H Van de Ven. Formal and informal dimensions of transactions. *Research on the management of innovation: The Minnesota studies*, 171:192, 1989.
- [193] Karlene H Roberts. Some characteristics of one type of high reliability organization. *Organization Science*, 1(2):160–176, 1990.
- [194] Anna M Rose, Jacob M Rose, Kristian Rotaru, Kerri-Ann Sanderson, and Jay C Thibodeau. Effects of uncertainty visualization on attention, arousal, and judgment. *Behavioral Research in Accounting*, 34(1):113–139, 2022.
- [195] Cynthia Rudin. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5):206–215, 2019.
- [196] Salvatore Ruggieri, Dino Pedreschi, and Franco Turini. Data mining for discrimination discovery. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 4(2):1–40, 2010.
- [197] Daniel M Russell, Mark J Stefik, Peter Pirolli, and Stuart K Card. The cost structure of sensemaking. In *Proceedings of the INTERACT'93 and CHI'93 conference on Human factors in computing systems*, pages 269–276, 1993.
- [198] Wesley C Salmon. *Statistical Explanation and Statistical Relevance*. University of Pittsburgh Press, 1971.

- [199] Jeff Sauro. Supr-q: A comprehensive measure of the quality of the website user experience. *Journal of usability studies*, 10(2), 2015.
- [200] Wolfgang; Newell Ben R. Schulze, Christin; Gaissmaier. Maximizing as satisficing: On pattern matching and probability maximizing in groups and individuals, 2020.
- [201] Alfred Schutz. *The phenomenology of the social world*. Northwestern University Press, 1972.
- [202] Alfred Schutz and Fred Kersten. Fragments on the phenomenology of music. 1976.
- [203] David Sculley, Gary Holt, Daniel Golovin, Eugene Davydov, Todd Phillips, Dietmar Ebner, Vinay Chaudhary, Michael Young, Jean-Francois Crespo, and Dan Dennison. Hidden technical debt in machine learning systems. *Advances in neural information processing systems*, 28:2503–2511, 2015.
- [204] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 618–626, Oct 2017.
- [205] Phoebe Sengers, Kirsten Boehner, Shay David, and Joseph 'Jofish' Kaye. Reflective design. In *Proceedings of the 4th decennial conference on Critical computing: between sense and sensibility*, pages 49–58, 2005.
- [206] Lloyd S Shapley. A value for n-person games. *Classics in game theory*, page 69, 1997.
- [207] Ronald W Shephard and Rolf Färe. The law of diminishing returns. In *Production theory*, pages 287–318. Springer, 1974.
- [208] John Shotter. Duality of structure” and “intentionality” in an ecological psychology. *Journal for the Theory of Social Behaviour*, 13(1):19–44, 1983.

- [209] Herbert A Simon. A behavioral model of rational choice. *The quarterly journal of economics*, 69(1):99–118, 1955.
- [210] Herbert A. Simon. Rationality as process and as product of thought. *The American Economic Review*, 68(2):1–16, 1978.
- [211] Herbert A. Simon. *Models of Bounded Rationality: Empirically Grounded Economic Reason*. The MIT Press, 07 1997.
- [212] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034*, 2013.
- [213] Venkatesh Sivaraman, Leigh A Bukowski, Joel Levin, Jeremy M Kahn, and Adam Perer. Ignore, trust, or negotiate: understanding clinician acceptance of ai-based treatment recommendations in health care. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, pages 1–18, 2023.
- [214] Ben R Slugoski, Mansur Lalljee, Roger Lamb, and Gerald P Ginsburg. Attribution in conversational context: Effect of mutual knowledge on explanation-giving. *European Journal of Social Psychology*, 23(3):219–238, 1993.
- [215] James F Smith and Thomas Kida. Heuristics and biases: Expertise and task realism in auditing. *Psychological bulletin*, 109(3):472, 1991.
- [216] Mark Snyder and Phyllis White. Moods and memories: Elation, depression, and the remembering of the events of one’s life. *Journal of personality*, 50(2):149–167, 1982.
- [217] Kacper Sokol and Peter A Flach. Glass-box: Explaining ai decisions with counterfactual statements through conversation with a voice-enabled virtual assistant. In *IJCAI*, pages 5868–5870, 2018.

- [218] Thilo Spinner, Udo Schlegel, Hanna Schäfer, and Mennatallah El-Assady. explainer: A visual analytics framework for interactive and explainable machine learning. *IEEE transactions on visualization and computer graphics*, 26(1):1064–1074, 2019.
- [219] Aaron Springer and Steve Whittaker. Progressive disclosure: empirically motivated approaches to designing effective transparency. In *Proceedings of the 24th international conference on intelligent user interfaces*, pages 107–120, 2019.
- [220] William H Starbuck and Frances J Milliken. Executives’ perceptual filters: What they notice and how they make sense. 1988.
- [221] Barry M Staw. Attribution of the “causes” of performance: A general alternative interpretation of cross-sectional research on organizations. *Organizational behavior and human performance*, 13(3):414–432, 1975.
- [222] Erik Štrumbelj and Igor Kononenko. Explaining prediction models and individual predictions with feature contributions. *Knowledge and information systems*, 41(3):647–665, 2014.
- [223] Simone Stumpf, Vidya Rajaram, Lida Li, Weng-Keen Wong, Margaret Burnett, Thomas Dietterich, Erin Sullivan, and Jonathan Herlocker. Interacting meaningfully with machine learning systems: Three experiments. *International Journal of Human-Computer Studies*, 67(8):639–662, 2009.
- [224] Harini Suresh, Steven R Gomez, Kevin K Nam, and Arvind Satyanarayan. Beyond expertise and roles: A framework to characterize the stakeholders of interpretable machine learning and their needs. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pages 1–16, 2021.
- [225] Alistair Sutcliffe. Designing for user engagement: Aesthetic and attractive user interfaces. *Synthesis lectures on human-centered informatics*, 2(1):1–55, 2009.

- [226] Alistair Sutcliffe. Designing for user experience and engagement. In *Why engagement matters*, pages 105–126. Springer, 2016.
- [227] William B Swann. Quest for accuracy in person perception: A matter of pragmatics. *Psychological review*, 91(4):457, 1984.
- [228] Maxwell Szymanski, Martijn Millecamp, and Katrien Verbert. Visual, textual or hybrid: the effect of user expertise on different explanations. In *26th International Conference on Intelligent User Interfaces*, pages 109–119, 2021.
- [229] Sarah Tan, Rich Caruana, Giles Hooker, Paul Koch, and Albert Gordo. Learning global additive explanations for neural nets using model distillation. *arXiv preprint arXiv:1801.08640*, 2018.
- [230] Shelley E Taylor. Asymmetrical effects of positive and negative events: the mobilization-minimization hypothesis. *Psychological bulletin*, 110(1):67, 1991.
- [231] Leigh Thompson and Gary Alan Fine. Socially shared cognition, affect, and behavior: A review and integration. *Personality and Social Psychology Review*, 3(4):278–302, 1999.
- [232] Amos Tversky and Daniel Kahneman. Availability: A heuristic for judging frequency and probability. *Cognitive psychology*, 5(2):207–232, 1973.
- [233] B van Fraassen. The pragmatic theory of explanation. In Joseph C Pitt, editor, *Theories of Explanation*. Oxford University Press, 1988.
- [234] Francisco J Varela, Evan Thompson, and Eleanor Rosch. *The embodied mind: Cognitive science and human experience*. MIT press, 2016.
- [235] Kush R Varshney and Homa Alemzadeh. On the safety of machine learning: Cyber-physical systems, decision sciences, and data products. *Big data*, 5(3):246–255, 2017.

- [236] Helena Vasconcelos, Matthew Jörke, Madeleine Grunde-McLaughlin, Tobias Gerstenberg, Michael S Bernstein, and Ranjay Krishna. Explanations can reduce overreliance on ai systems during decision-making. *Proceedings of the ACM on Human-Computer Interaction*, 7(CSCW1):1–38, 2023.
- [237] Michael Veale, Max Van Kleek, and Reuben Binns. Fairness and accountability design needs for algorithmic support in high-stakes public sector decision-making. In *Proceedings of the 2018 chi conference on human factors in computing systems*, pages 1–14, 2018.
- [238] Himanshu Verma, Jakub Mlynar, Roger Schaer, Julien Reichenbach, Mario Jreige, John Prior, Florian Evéquo, and Adrien Depeursinge. 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*, pages 1–19, 2023.
- [239] Janet Vertesi. Seamful spaces: Heterogeneous infrastructures in interaction. *Science, Technology, & Human Values*, 39(2):264–284, 2014.
- [240] Timothy J Vogus and Kathleen M Sutcliffe. Organizational mindfulness and mindful organizing: A reconciliation and path forward. *Academy of Management Learning & Education*, 11(4):722–735, 2012.
- [241] Sandra Wachter, Brent Mittelstadt, and Chris Russell. Counterfactual explanations without opening the black box: Automated decisions and the gdpr. *Harv. JL & Tech.*, 31:841, 2017.
- [242] Byron C. Wallace, Kevin Small, Carla E. Brodley, Joseph Lau, and Thomas A. Trikalinos. 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*, pages 819–824, New York, NY, USA, 2012. ACM.
- [243] Dakuo Wang, Elizabeth Churchill, Pattie Maes, Xiangmin Fan, Ben Shneiderman, Yuanchun Shi, and Qianying Wang. From human-human collaboration to human-ai collaboration: Designing ai systems that can work together with people. In *Extended abstracts of the 2020 CHI conference on human factors in computing systems*, pages 1–6, 2020.
- [244] Danding Wang, Qian Yang, Ashraf Abdul, and Brian Y. Lim. Designing theory-driven user-centric explainable ai. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems, CHI '19*, pages 601:1–601:15, New York, NY, USA, 2019. ACM.
- [245] Danding Wang, Wencan Zhang, and Brian Y Lim. Show or suppress? managing input uncertainty in machine learning model explanations. *Artificial Intelligence*, 294:103456, 2021.
- [246] Xinru Wang and Ming Yin. Are explanations helpful? a comparative study of the effects of explanations in ai-assisted decision-making. In *26th International Conference on Intelligent User Interfaces*, pages 318–328, 2021.
- [247] Karl E Weick. *Sensemaking in organizations*, volume 3. Sage, 1995.
- [248] Karl E Weick and Kathleen M Sutcliffe. *Managing the unexpected: Sustained performance in a complex world*. John Wiley & Sons, 2015.
- [249] Karl E. Weick, Kathleen M. Sutcliffe, and David Obstfeld. Organizing for high reliability: Processes of collective mindfulness. *Research in Organizational Behaviour*, 21:81–123, 1999.

- [250] Karl E Weick, Kathleen M Sutcliffe, and David Obstfeld. Organizing and the process of sensemaking. *Organization science*, 16(4):409–421, 2005.
- [251] Daniel S Weld and Gagan Bansal. The challenge of crafting intelligible intelligence. *Communications of the ACM*, 62(6):70–79, 2019.
- [252] Mark Wenman. *Agonistic democracy: Constituent power in the era of globalisation*. Cambridge University Press, 2013.
- [253] Terry Winograd, Fernando Flores, and Fernando F Flores. *Understanding computers and cognition: A new foundation for design*. Intellect Books, 1986.
- [254] Qian Yang, Alex Scuito, John Zimmerman, Jodi Forlizzi, and Aaron Steinfeld. Investigating how experienced ux designers effectively work with machine learning. In *Proceedings of the 2018 Designing Interactive Systems Conference, DIS '18*, pages 585–596, New York, NY, USA, 2018. ACM.
- [255] Qian Yang, Aaron Steinfeld, and John Zimmerman. 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*, pages 1–11, 2019.
- [256] Jiaming Zeng, Berk Ustun, and Cynthia Rudin. Interpretable classification models for recidivism prediction. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 180(3):689–722, 2017.
- [257] Wencan Zhang and Brian Y Lim. Towards relatable explainable ai with the perceptual process. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, pages 1–24, 2022.
- [258] Yunfeng Zhang, Q Vera Liao, and Rachel KE Bellamy. 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*, pages 295–305, 2020.

- [259] Yayan Zhao, Mingwei Li, and Matthew Berger. Graphical perception of saliency-based model explanations. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, pages 1–15, 2023.
- [260] Jianlong Zhou, Amir H Gandomi, Fang Chen, and Andreas Holzinger. Evaluating the quality of machine learning explanations: A survey on methods and metrics. *Electronics*, 10(5):593, 2021.
- [261] Jichen Zhu, Antonios Liapis, Sebastian Risi, Rafael Bidarra, and G. Michael Youngblood. Explainable ai for designers: A human-centered perspective on mixed-initiative co-creation. In *2018 IEEE Conference on Computational Intelligence and Games (CIG)*, pages 1–8. IEEE, Aug 2018.