Computing and Communicating Uncertainty for Responsible Human-AI Decision-making

by

Snehal Prabhudesai

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Computer Science and Engineering)
in the University of Michigan
2026

Doctoral Committee:

Associate Professor Nikola Banovic, Co-chair
Associate Professor Xun Huan, Co-chair
Professor Arvind Rao, Co-chair
Professor Richard Gonzalez
Professor Jessica Hullman, Northwestern University
Associate Professor Vera Liao

Snehal Prabhudesai
snehalbp@umich.edu
ORCID iD: 0000-0003-0179-4466

© Snehal Prabhudesai 2026

DEDICATION

To my parents, who raised me to be curious and ask questions about the world. And to Ms. Pranoti, my 7th grade science teacher, who suggested I do a PhD to find their answers.

ACKNOWLEDGEMENTS

This dissertation would not have been possible without the love, support, and immense patience of a whole community of people as they watched me navigate, stumble, shed tears, and celebrate milestones throughout this journey.

First and foremost, thanks to my advisors: Arvind Rao, Xun Huan, and Nikola Banovic, for co-chairing this work and for the very unique perspectives each of you gave me. People are often surprised (and mildly alarmed) when they realize I have three advisors. But I am so lucky and grateful to have a team that trusts me to pick shiny new problems to investigate in their respective fields. Thank you for keeping me focused and grounded when I sometimes wanted to chase butterflies. I deeply appreciate how Arvind balances honesty, practicality, and directness in communication. Thank you for welcoming me into your lab as a fresh Master's student, and for always encouraging me to think beyond the horizon. Xun, thank you for teaching me to be rigorous and detail-oriented. While I may have complained at times, your insistence on getting the math right and refusal to let a hand-wavy argument slide has brought this thesis into fruition. Nikola, thank you for trusting me when I approached you for collaboration. Thank you for encouraging me to say "no" more, for helping me grow into the academic community as an independent researcher, and for encouraging me to explore questions more deeply. The greatest gift that all of you have given me is the ability to build my thesis at the intersection of your fields.

To my committee members: Vera Liao, Jessica Hullman, and Rich Gonzalez, for holding this thesis to the highest standards. Your feedback and thoughtful questions only made it better. Vera, thank you for encouraging me to have faith in myself when I was doubtful, for telling me to "just keep doing good work, and everything will work out". It did! Jessica, I'll admit I was a bit intimidated when I cold-emailed you one day randomly to chat about uncertainty visualization, half expecting to be ignored. You met with me, pointed me to the right direction and shaped this thesis before you were ever formally on my committee. Rich, thank you for asking the questions that came from decades of thinking about judgment and decision-making. They were often the ones I had no good answer to yet, which is exactly why they were so fundamental to this thesis.

I am grateful to the faculty who shaped this work and my growth as a researcher. Emily Mower Provost, the encouragement and sense of safety you create as a mentor is something

I will strive to emulate for my own students. Sabine Loos, our conversations during a conference road trip to Toronto taught me so many lessons about life, academia, and learning to set boundaries. Wes Weimer, thank you for your endless enthusiasm and for never turning me away when I needed help, even when you were clearly busy. Rada Mihalcea, thank you for teaching me the value of bringing science to the community with Friday Night AI.

I'm grateful to my mentors at Microsoft Research, where I spent a formative summer at the Computational Social Science Group. Thanks to Daniel Goldstein, Jake Hofman, and David Rothschild for showing me how to ask sharp empirical questions, identify exciting and important problems, and translate my work into practical applications in the industry.

None of this would have run smoothly without the people who make the department actually work. Thank you to Magda Calvillo, Sarah Snay, Sindhu Kutty, Mithun Chakraborty, Kimberley Diaz, Patricia O'Malley, and Jasmin Stubblefield for your patience with my questions, your mentorship, and for handling the administrative machinery of a PhD. Hence, I rarely had to think about it. A special thank you to Karen Liska, now retired, whose kindness set the tone for how CSE treats its graduate students long before I arrived.

I was lucky to be part of three labs during this PhD. My CompHCI family — Nel Escher, Divya Ramesh, Sumit Asthana, Anindya Das Antar, Tsedeniya Amare, Jaewoong Choi, Ishika Joshi, and Jeff Bilik — thank you for the late-night sessions, venting about frustrations, cupcake birthday celebrations, and celebrating each other's successes. I am so grateful to have found such lifelong friends in you. In RaoLab, I will cherish swapping departmental lore with Santhoshi Krishnan, Avery Maddox, Elisa Warner, Viswes Ravikumar, and Reva Kulkarni. I was devastated when most of you graduated, and realized the mistake of being close friends with seniors who leave you. Jokes aside, I am rooting for you! In UQSciML, Jeremiah Hauth, Aniket Jiwani, Wanggang Shen, and Codie Fiedler Kawaguchi made the complex, hard-to-grasp math fun and exciting. Thank you for your companionship each SIAM, from walks in Amsterdam to balcony evenings over strawberries in Italy.

Thank you to the masters and undergraduate students I had the privilege of working with: Ananya Kasi, Samantha Mazza, Gillian Soekawan, Anmol Mansingh, Rui Nie, Leyao (Hannah) Yang, Vinayak Ahluwalia, and Dingkun Guo, and for the opportunity to learn how to be a supportive mentor.

I have been fortunate to have life-long friendships, both during and before this PhD. To my academic friends in and beyond Ann Arbor — Sarah Whittaker, Meron Demisse, Aymen Ahmed, Sarah Jabbour, Madelyn Gatchel, Alicia Carmichael, Pranjal Srivastava, Aishani Kulshreshtha, and Abhirath Anand, Yasha Iravantchi and Yang-Hsi Hsu, thank you for celebrating wins equally, big and small. To my oldest friends who have been a constant in my life for over a decade (or two): Ketaki Gore, Nayanika Diwadkar, Rutuja Zunzarrao,

Kshitija Mirkale, Darshak Rairakhia, Hrishikesh Belatikar, Vinay Chandrasekhar, Onkar Nabar, Jijo Verghese, Ameya Chitnis, and Rohit Thapliyal. Thanks for reminding me that the machinery of academia manufactures a lot of my problems, and to take it less seriously. To my climbing friends, Jiwon Kim, Hessa Al-Thani, Arlen Dean, and Jaylin Herskovitz, thank you for dragging me up walls, literal and figurative. You taught me to keep on trying the problem until I succeed. To my roommates, Sanjana Eyunni, Tarana Kaovasia, and Agamjyot Sangotra, thank you for making home a place I wanted to come back to.

While I never really owned a pet thanks to this nomadic lifestyle with extensive travel, I am blessed to have enjoyed the presence of my friends' pets. Risking all convention, I thank Smokey, Nilah, Mellon, Chechil, Bernie, and Watson for being unconditionally happy to see me and never once asking how my PhD was going.

At some point I realized I have a very specific set of people I run to when I'm anxious about something. I started calling them, collectively, my panic committee: Yara El-Tawil, Ishika Joshi, and Florian Le Bras. Their sacred duty when something goes wrong is to answer one question: "Should I panic?" They have never once said yes. Thank you for being rays of sunshine on my darkest days, for your immense support and constant encouragement. I love you all.

To my parents and my brother, thank you for a lifetime of support that made this degree possible long before I started it. My parents always encouraged me to be curious and be the best version of myself. My mother, Ruta Prabhudesai, has been incredibly kind, supportive, and patient throughout this journey. My father, Bhalchandra Prabhudesai, taught me to question the world, and that no question is a stupid one. That lesson really led to this dissertation. I am so grateful for my brother Soham Prabhudesai, who handles the grown-up parts of life for me, from servicing my car to wrangling paperwork I'd otherwise ignore. This thesis would not have been possible without their support. To my cousin Saumitra Phatak, thank you for the shared understanding both as family and someone who did his PhD alongside mine. Thanks to Florian Le Bras, my partner and best friend. Your wit, humor, endless imagination, and unwavering sense of justice help me envision a fairer future worth building for all of us.

Thanks to the participants who gave their time and effort towards the studies in this thesis, without whom this thesis would not have existed, quite literally.

And thanks to you, dearest reader, for being curious enough to open this thesis. May you stay curious, always.

TABLE OF CONTENTS

DEDICATION ii

ACKNOWLEDGEMENTS iii

LIST OF FIGURES ix

LIST OF TABLES xi

ABSTRACT xii

CHAPTER

1 Introduction 1

1.1 Thesis Statement 2
1.2 Chapter Organization 2

2 Background and Related Work 4

2.1 Human-AI decision-making 4
2.2 Algorithmic Transparency 5
2.3 Representing Uncertainty in AI Outputs 6
2.3.1 Uncertainty Quantification 6
2.3.2 Uncertainty Communication 8
2.4 Challenges with Understanding Uncertainty 9

3 Quantifying Bayesian Uncertainty of Black-box AI Models 10

3.1 Introduction 10
3.2 Method 12
3.2.1 Step 1: Access a Deterministic DNN Model 13
3.2.2 Step 2: Conduct Sensitivity Analysis to Select a Layer for Bayesian Inference 13
3.2.3 Step 3: Perform Targeted Bayesian Inference to Build pBNN 14
3.3 Findings 14
3.3.1 Illustration: Toy Problem 14
3.3.2 Application: Image Segmentation with U-Net 18
3.3.3 Validation 21
3.4 Discussion 24
3.4.1 Case Study: Uncertainty Communication for Tumor Segmentation . 26

3.5 Summary . . . 28
4 Reflecting on Epistemic Gaps for Uncertainty Communication . . . 29
4.1 Introduction . . . 29
4.2 Method . . . 30
4.2.1 Study Context: Urban Flooding in the Metropolitan Detroit Region 30
4.2.2 Multidisciplinary Research Team . . . 32
4.2.3 Collaborative Autoethnography Workshops . . . 32
4.2.4 Workshop Attendees . . . 34
4.2.5 Data Collection . . . 35
4.2.6 Data Analysis . . . 35
4.3 Findings . . . 36
4.3.1 Producing . . . 36
4.3.2 Transforming . . . 40
4.3.3 Contextualizing . . . 44
4.3.4 Rethinking Design . . . 52
4.4 Discussion . . . 58
4.4.1 RQ1: How do multidisciplinary team dynamics affect design implications for climate resilience solutions? . . . 58
4.4.2 RQ2: What are the design implications that remain hidden or underexplored due to team dynamics? . . . 61
4.5 Summary . . . 62
5 Understanding Uncertainty in Human-AI Decision-making . . . 64
5.1 Introduction . . . 64
5.2 Method . . . 66
5.2.1 Task . . . 66
5.2.2 Data Pre-processing and Algorithms . . . 67
5.2.3 DSS Design Probes . . . 70
5.2.4 Study Software Implementation . . . 73
5.2.5 Tasks and Procedures . . . 74
5.2.6 Participant Recruitment . . . 76
5.2.7 Analysis of Think-aloud Data . . . 77
5.2.8 Limitations . . . 77
5.3 Findings . . . 77
5.3.1 Participants' Lived Experiences Shaped their Sense-making Process and DSS Interactions . . . 78
5.3.2 How Participants Perceived Uncertainty . . . 79
5.3.3 Confusion, Resolution, and Learning Curve . . . 81
5.3.4 Thinking and Deliberation . . . 83
5.3.5 Judgement . . . 86
5.3.6 Reliance and Trust . . . 87
5.4 Discussion . . . 91
5.4.1 Sensemaking from Extracted Cues . . . 91

5.4.2 Uncertainty Forces Decision-makers to Slow Down and Think Analytically . . . 93
5.4.3 Communicating Uncertainty Reduces Undue Over-reliance on ML Algorithms . . . 94
5.5 Summary . . . 95
6 Leveraging Storytelling to Communicate Uncertainty . . . 98
6.1 Introduction . . . 98
6.2 Method . . . 100
6.2.1 Theoretical Underpinning . . . 100
6.2.2 Task . . . 101
6.2.3 Data . . . 101
6.2.4 Model . . . 102
6.2.5 Interfaces . . . 104
6.2.6 Final Study Software . . . 106
6.2.7 Experimental Conditions . . . 106
6.2.8 Participants . . . 107
6.2.9 Procedure . . . 107
6.2.10 Measures . . . 108
6.2.11 Statistical Analysis . . . 109
6.3 Findings . . . 110
6.3.1 Descriptive Statistics . . . 110
6.3.2 Effect of Stories . . . 111
6.3.3 Confidence in Decisions . . . 111
6.3.4 AI Agreement . . . 111
6.3.5 Signal Strength x AI Uncertainty . . . 111
6.4 Discussion . . . 114
6.5 Summary . . . 115
7 Discussion . . . 116
7.1 Representing Uncertainty . . . 117
7.2 Understanding Uncertainty . . . 118
7.3 Decision-making with AI Uncertainty . . . 119
8 Conclusion and Future Work . . . 121
8.1 Thesis Contributions . . . 122
8.2 Future Work . . . 123
8.2.1 Decision-centric Uncertainty Communication . . . 123
8.2.2 Scaffolding User Understanding of Uncertainty . . . 123
8.2.3 AI-Generated Narratives for Uncertainty Communication . . . 124
8.2.4 Human-centered Evaluation of Uncertainty Visualization . . . 124
APPENDICES . . . 125
BIBLIOGRAPHY . . . 125

LIST OF FIGURES

FIGURE

3.1 Flowchart to address scalability challenges of BNN by building pBNN: Step (1) represents a million-parameter deterministic NN. Step (2) Conduct Sensitivity Analysis to identify the layer that has the greatest impact on model output. Step (3) Conduct Bayesian inference on the most sensitive layer to build pBNN. The resulting pBNN combines the strengths of both deterministic and Bayesian NNs, allowing scalability. 12

3.2 SI for all layers in the DNN for the test problem. Layer 9 has the highest SI and is selected for Bayesianization. 16

3.3 Training data (black dots) overlaid with the deterministic DNN prediction (solid orange) and full BNN posterior push-forward uncertainty (solid blue line is the mean, blue shadow is the 95% credible interval). 17

3.4 Normalized KL divergence from the full BNN posterior push-forward to the different pBNNs is highly correlated with the SI that is used to select the layer for Bayesianization (Spearman's correlation $r_s = -0.83$). 18

3.5 U-Net architecture with our naming system of the different tunable layers considered for the pBNN. 19

3.6 SI for all layers in the U-Net used for tumor segmentation. layer_1a has the highest SI and is selected for Bayesianization. 20

3.7 ΔELBO training history for all models. 22

3.8 Normalized KL D_{norm}, for all 19 pBNNs each built with Bayesian inference on one of the layers in the U-Net. 23

3.9 Scatter plot of KL divergence sum versus number of parameters in the Bayesian layer. 24

3.10 Uncertainty map accompanying a prediction made by the pBNN. 26

4.1 Three domains of expertise in the study: physical modeling, human factors and stakeholders, and decision and design. Their integration informs flood risk communication tools. 31

5.1 Base plot split with dotted black line indicating the list price of Alex's house. Line splits the graph into regions where valuation is less (red region) or more (green region) than the list price. Percentages on the x-axis indicate the numerical quantity by which house valuation differs from the list price. 70

5.2 The three DSS design probes: a) Interactive Visualization DSS , b) ML-based DSS, and c) UQ-based DSS. 71

5.3 Examples of list price posterior predictive distributions generated by the Bayesian algorithm for different houses. Each participant had to make a decision for each level of uncertainty - (A) Low, (B) Medium, or (C) High, assigned in random order (Section 5.2.1). For each level of uncertainty, participants were shown only one randomly selected visualization on the web interface out of three types: (I) clear-cut red, (II) split or (III) clear-cut green. 74
5.4 Conceptual map of different cues that participants extract for uncertainty representations of model predictions. Participants start by eyeballing. They do not count when predictions are "clear-cut", but extract more complex cues higher up the pyramid for "split" predictions. 97

6.1 Within-subjects study design with 2x2 conditions. Independent variables are uncertainty and stories. 103
6.2 Model uncertainty (unsure) with signal uncertainty across conditions. 112
6.3 Model agreement with signal uncertainty across conditions. 113

LIST OF TABLES

TABLE

3.1 Architecture of the densely-connected DNN used for the test problem. 15
3.2 Summary of the deterministic U-Net training. 20

5.1 Study participant demographics. We removed data from participants P1 and P5 from analysis due to technical difficulties arising during these study sessions. . . 76
5.2 Summary of the effects of uncertainty communication on different stages of the sensemaking process, with breakdowns in each stage and ways to address them. 91

ABSTRACT

AI-based Decision Support Systems have long been used to support decision-makers in making data-driven decisions. However, lay decision-makers are prone to enchanted determinism, the idea that AI is absolute and its predictions are infallible. As a result, decision-makers find it difficult to identify when to rely on AI and when to exercise their own judgment. The field of Human-centered AI has proposed communicating uncertainty as a form of transparency to mitigate the effects of accepting AI recommendations without questioning.

However, people are naturally averse to uncertainty and probabilistic reasoning. Cognitive biases further hinder people from correctly understanding uncertainty. Factors such as task type, desired outcome, numerical literacy and expertise determine how effective uncertainty communication mechanisms are for making decisions. Decision-makers without prior mathematical or statistical knowledge can struggle to interpret uncertainty and experience a high learning curve.

Existing approaches largely treat uncertainty communication as a *representation* problem, computing accurate uncertainty estimates and encoding them visually while presuming that decision-makers can readily interpret and act on this information. This thesis argues that to prevent blind acceptance of AI recommendations, we must effectively bridge the divide between how AI uncertainty is computed through probabilities and how non-technical decision-makers interpret it through mental shortcuts. This dissertation makes empirical, epistemological, methodological, and design contributions to bridge this divide when communicating AI uncertainty. First, I propose a computational framework based on Bayesian Uncertainty Quantification (UQ) methods that makes it feasible to compute AI uncertainty with lower computational resources. Second, I show how communicating this uncertainty with state-of-the-art visualization methods helps resolve enchanted determinism in lay decision-makers. Third, I unpack the gap between scientific uncertainty communication and decision-makers' cognitive and contextual needs. Finally, I design narrative-based scaffolding and evaluate the effect of uncertainty on AI agreement and decision confidence. Together, this thesis makes a case for considering how decision-makers interpret uncertainty when designing and evaluating uncertainty communication mechanisms.

CHAPTER 1

Introduction

AI-based Decision Support Systems (DSS) promise to support decision-makers by providing accurate predictions based on analyzing patterns in historical data. Such systems are being increasingly adopted by decision-makers who may have domain expertise but lack technical expertise (e.g., clinicians, educators, policymakers) to inform decisions.

However, such decision-makers often have limited access to the model's capabilities, limitations, or confidence. Despite the consequences of wrong decisions being catastrophic in such domains, existing black-box AI-based DSS do not provide insights into the predictions they make. Such black-box AI-based DSS can project a sense of authority [118] and infallibility [45]. This lack of transparency can make it difficult for non-technical decision-makers to determine when to accept AI recommendations and when not to [12].

AI assistance elements provide transparency [64, 212, 43, 130, 124, 85], presuming that additional information about AI's predictions would help decision-makers decide whether to accept, reject, or contest AI. For example, mechanisms such as explainability [177] and interpretability [149] inform decision-makers why and how AI systems made a recommendation. However, prior work [143, 12] has found that such mechanisms are often interpreted as stronger evidence of reliability than is warranted.

As an alternative form of transparency, the HCI community has proposed communicating uncertainty [15, 24] in AI predictions to help decision-makers understand AI's confidence in its recommendations. This has led to the development of uncertainty quantification (UQ) algorithms [79, 153, 218] and visualization approaches [131, 114, 51, 141] that consider uncertainty as a *representation problem*, i.e., providing solutions for how to represent AI uncertainty by quantifying or visualizing it. Similar to other AI assistance elements, uncertainty communication presumes [15] that by providing additional confidence information (i.e., probability of AI's recommendations), decision-makers can better determine when to accept AI recommendations.

However, people are generally averse to probabilistic thinking [216], and ignore probability even if communicated [113]. Understanding probabilistic information also requires prior statistical expertise [46], which non-technical decision-makers may lack. While representing uncertainty may inform decision-makers *what* the risk is, it does not inform them *how* to utilize it. Decision-makers need to engage in the process of sense-making [229] by synthesizing uncertainty information, assigning meaning, and applying it to decide whether to accept AI recommendations.

However, we currently have limited knowledge of how non-technical decision-makers understand and utilize AI uncertainty to make decisions. While prior work on uncertainty quantification and communication frames uncertainty as a *representation problem*, this thesis treats it as an *interpretation problem*. The thesis answers the following research question:

> *How do non-technical decision-makers interpret AI uncertainty representations, and to what extent does this understanding influence acceptance of AI recommendations?*

The key finding of the thesis is that non-technical decision-makers rely on mental shortcuts and heuristics rather than statistical information when interpreting uncertainty in AI predictions. Scaffolding their interpretation leads to analytical thinking and reduces blind acceptance of AI recommendations when in complex scenarios under high AI uncertainty compared to suppressing this information.

1.1 Thesis Statement

The thesis argues that:

> *To prevent blind acceptance of AI recommendations, we must effectively bridge the divide between how AI uncertainty is computed through probabilities and how non-technical decision-makers interpret it through mental shortcuts.*

1.2 Chapter Organization

To investigate how uncertainty affects cognition and decisions, it is first necessary to represent the uncertainty in black-box AI predictions by mathematically quantifying it. Next, this statistical information has to be communicated to decision-makers in a way that facilitates decision-making (e.g., through uncertainty visualization).

Only then can we study how and to what extent AI uncertainty supports responsible human-AI decision-making. Given these aims, this thesis adopts an interdisciplinary and mixed-methods research approach. Based on this rationale, we organize the remainder of the thesis chapter as follows:

Chapter 2 contextualizes the thesis in the space of uncertainty communication, which is an active area of research in uncertainty quantification, uncertainty visualization, and human-AI decision-making. Next, Chapter 3 quantifies the uncertainty in black-box AI predictions by scaling Bayesian uncertainty quantification (UQ) methods to million-dimensional AI systems. Chapter 4 explores how to communicate statistical uncertainty information in decision-centric ways through a collaborative auto-ethnography of an interdisciplinary research team and case study of flood risk communication. Together, these two chapters cover the two necessary conditions (i.e., quantification and communication) for uncertainty representation. Chapter 5 investigates how non-technical decision-makers interpret and make sense of and act upon AI uncertainty in a think-aloud user study. Chapter 6 then quantifies the magnitude of the effects of uncertainty found in the previous chapter through a quantitative user study. Chapter 7 situates the thesis in the larger discourse on uncertainty communication in AI-assisted decision-making, and provides suggestions for the design of such mechanisms. Chapter 8 concludes by inviting the HCI community to adopt a psychological lens to investigate and re-design uncertainty communication for responsible human-AI decision-making and identifies opportunities for future research.

CHAPTER 2

Background and Related Work

2.1 Human-AI decision-making

Decision-making involves cognitive processes to make a selection among alternatives by exercising judgement [53, 83]. Normative and economic theories of decision-making formalize rational choice through expected utility theory [222], which states that a rational decision-maker would choose the action that maximizes the probability-weighted sum of utility across possible outcomes. This framework assumes humans are rational beings [133] who act to maximize an expected utility function [75] in order to make the right decision.

However, descriptive accounts of human judgment show that people systematically depart from this. Rather than computing expected utility, people often rely on biased judgements [216] and make decisions based on perceived outcomes rather than carefully weighing the utility of each alternative [126]. This gap between the normative expected-utility model and people's actual decision-making has motivated critiques of uncertainty communication mechanisms that assume rational users [113, 127]. Decision-making deteriorates further with complexity and stress [169], leading to sub-optimal decisions.

Decision Support Systems (DSS) assist decision-makers in exercising better judgement with the goal of improving the outcomes of their decision-making process [174]. DSS have been historically used for unstructured or semi-structured decision-making [86] (i.e., lacking clear guidelines to inform decisions) to provide data-driven insights to decision-makers. More recently, Machine Learning (ML) has been used in DSS [89] to extract patterns from large volumes of historic data, and make inferences to inform decision-makers.

2.2 Algorithmic Transparency

When ML-based DSS are opaque about the uncertainty in their predictions (including but not limited to the uncertainty arising from the complex, dynamic, and ambiguous nature of the world [82]), they may give a false sense of reliability to users [88], whilst covertly reinforcing harmful societal biases [7] and maintaining the status quo [20].

Transparency mechanisms [61, 65], including uncertainty communication [25], enable decision-makers to understand model limitations and have appropriate trust and reliance on AI [149]. Communicating how uncertain an ML-based DSS is, while providing actionable recommendations, can promote algorithmic transparency. Uncertainty, when communicated appropriately, can help decision-makers better understand algorithms, allowing decision-makers to improve and contest algorithms [159]. Reliable uncertainty estimates may allow decision-makers to determine how much confidence to place in a model's competence [43, 212]. Moreover, an accurate assessment of uncertainty may lead to mitigation of model unfairness [70], improving decision-making [15] and promoting appropriate trust in AI [25, 211].

Explanation mechanisms [178, 193, 62, 4, 227] promote transparency by providing insights into the AI's inner workings and decision-making to end-users, thereby supporting them in evaluating AI decisions. Such mechanisms aim to support end-users in identifying and rejecting AI decisions in situations in which its reasoning was incorrect [190]. However, most existing explanation mechanisms tend to act as evidence of reliability rather than accountability [40, 55], often deceiving end-users into over-relying on AI [143, 116].

Interpretability in AI aims to make model decisions understandable to users, enabling them to trace how inputs lead to specific outputs [61, 181]. Interpretability promotes transparency by revealing model reasoning behind predictions with two methods. First are inherently interpretable glassbox models such as Generalized Additive Models (GAMs) [47] or point systems [123, 237], which provide direct insight into how different features influence predictions. Second are post hoc explanation techniques like feature attribution [151] and counterfactual explanations [224], which analyze trained models to offer interpretability without altering their structure. However, interpretability methods may sometimes oversimplify the explanation, leading to potentially misleading interpretations [160]. Furthermore, some models may be interpretable but still difficult for users to understand or trust, especially when confronted with complex or counterintuitive results [139].

While both explainability and interpretability trace the process by which AI arrived at a prediction, uncertainty communication discloses how confidently the AI made that prediction.

2.3 Representing Uncertainty in AI Outputs

Current black-box AI models do not have the ability to communicate uncertainty by themselves. First, uncertainty has to be measured using uncertainty quantification techniques to obtain numerical estimates. Next, these estimates have to be communicated using suitable techniques that make it easier for non-technical decision-makers to understand. Here, we discuss both these aspects necessary to represent uncertainty in AI outputs.

2.3.1 Uncertainty Quantification

Uncertainty Quantification seeks to understand, characterize, and quantify uncertainty in AI predictions to promote transparency. Consider an AI model G with parameter θ. The model G accepts input x and makes a prediction y. ϵ is the inherent noise in the model. This can be characterized as:

$$y = G(\theta, x) + \epsilon \tag{2.1}$$

Aleatoric uncertainty arises from inherent randomness in the data, while epistemic uncertainty stems from a lack of knowledge or incomplete information about the model or data-generating process. Thus, each component of Equation 2.1 can uniquely contribute to uncertainty. The model G can be a source of uncertainty due to its complexity, assumptions, or architecture. Similarly, model parameter θ can lead to uncertainty if the true value is not known, or due to the optimization process, choice of regularization, etc. The input x can be incomplete, noisy, or not adequately representing the real-world phenomena. Finally, there can be noise ϵ due to the underlying randomness of the process, or factors difficult to control, unobservable variables, or other stochastic processes

Uncertainty Quantification methods aim to explicitly model these uncertainties with computational methods, including Bayesian inference. Bayesian methods aim to update our belief about model parameters θ given some observed data x. A DNN takes input x and predicts output $\hat{y}$; we write $\hat{y} = f(x; w)$ where w represents all tunable model parameters of the DNN (e.g., DNN weights and bias). Given N_T training data points $(x_T, y_T) = x_n, y_n n = 1^{N_T}$, the DNN is typically trained by finding w to minimize a loss function:

$$w^* = \arg\min wL(w, x_T, y_T). \tag{2.2}$$

For example, a popular choice is the least squares loss $\mathcal{L}(w, x_T, y_T) = \frac{1}{N_T} \sum_{n=1}^{N_T} [f(x_n; w) - y_n]^2$. The optimization is often done with stochastic gradient descent [145, 183]. Once Eqn. (2.2) is solved, it produces a deterministic DNN that makes a single-valued prediction for any new input x: $\hat{y} = f(x; w^*)$.

BNN [153, 163, 87, 30, 78], in contrast, treats w as random variables with an associated probability density function (PDF) that represents the uncertainty on w. When training data become available, these PDFs are updated through Bayes' rule:

$$p(w|x_T, y_T) = \frac{p(y_T|x_T, w)p(w)}{p(y_T|x_T)}, \tag{2.3}$$

where $p(w)$ is the prior PDF[1], $p(y_T|x_T, w)$ is the likelihood PDF, $p(w|x_T, y_T)$ is the posterior PDF, and $p(y_T|x_T)$ is the marginal likelihood (a PDF-normalization term). The prior thus represents the uncertainty on w before seeing any training data, and the posterior describes the updated uncertainty after incorporating the training data. Solving the Bayesian inference problem then entails computing the posterior $p(w|x_T, y_T)$.

Conventional Bayesian solutions largely rely on MCMC [97, 184, 6, 37] to *sample* the posterior distribution. While MCMC provably converges to the exact posterior that may be highly non-Gaussian, it converges slowly in high-dimensional settings due to the onset of measure concentration to the so-called *typical set* [23]. Hamiltonian Monte Carlo [164, 105, 23], one of the more scalable MCMC variants, has been used for Bayesian inference for up to hundreds of parameters, but remains orders of magnitude short of the million-parameter BNNs targeted in this paper.

VI [121, 225, 29] provides much better scalability by finding the best approximation to the true posterior from a parametric family of distributions (e.g., all independent Gaussians), thereby turning the sampling task into an optimization one:

$$\theta^* = \arg\min_{\theta} D_{\mathrm{KL}}\big[q(w;\theta), |, p(w|x_T, y_T)\big]. \tag{2.4}$$

where $q(w;\theta)$ is the approximating posterior probability density function (PDF) parameterized by θ, and the Kullback-Leibler (KL) divergence D_{KL} quantifies the farness from the true posterior $p(w|x_T, y_T)$ to $q(w;\theta)$. One can further show that θ^* is also the maximizer of the well-known *evidence lower bound (ELBO)*:

$$\theta^* = \arg\max_{\theta}, \,_q \left[\ln p(y_T|x_T, w) + \ln p(w) - \ln q(w;\theta)\right] = \arg\max_{\theta}, , (\theta) \tag{2.4}$$

[1]Note that $p(w|x_T) = p(w)$, i.e., the prior uncertainty should not change from knowing only the input values of the training data without their output values.

which can be estimated through Monte Carlo that samples $w^{(m)}$ from $q(w;\theta)$:

$$(\theta) \approx \frac{1}{M}\sum_{m=1}^{M}\left[\log p\left(y_T|x_T,w^{(m)}\right)+\log p\left(w^{(m)}\right)-\log q\left(w^{(m)};\theta\right)\right]. \tag{2.5}$$

The optimization can be approached leveraging gradient-based algorithms, where the gradient with respect to θ can be obtained via back-propagation. We adopt the recent Flipout VI algorithm [231] to solve the Bayesian problem. Flipout [231] has been demonstrated to provide substantial computational savings for dense, convolutional, and recurrent neural network architectures. The method injects pseudo-independent weight perturbations in order to decorrelate gradients, thereby achieving drastically decreased variance for the ELBO Monte Carlo estimator. It also offers a vectorized implementation that allows one to take advantage of GPU computations.

However, even with Flipout, the computational and memory requirements for a million parameters are still extremely high, if not outright prohibitive. [140] found that for large DNN architectures with $\mathcal{O}(10^7 - 10^8)$ parameters, MFVI failed to converge altogether. Therefore, one cannot simply take Flipout off-the-shelf to build million-dimensional BNNs; additional algorithmic developments are still needed.

2.3.2 Uncertainty Communication

Once uncertainty is quantified, it can be communicated to end-users in different formats (e.g. numerical [219], verbal [135] or visual [112]) with trade-offs between clarity, decision quality and cognitive effort. Numerical formats such as probability percentages or prediction intervals are more precise than verbal expressions [31], but require prior statistical knowledge to be understood [220]. Despite being more natural, people interpret verbal expressions such as probabilistic phrases and hedges very differently due to ambiguity [41].

Visualizations can communicate more complex information compared to numerical estimates, while being more precise than verbalization. Techniques such as value-suppressing uncertainty palettes [51] and sketchy rendering [233] manipulate visual qualities such as color and style to communicate probabilistic information. Visualization techniques also leverage frequency framing (i.e., converting probabilities to countable units) to account for end-users with lower numeracy skills [103, 99]. For example, quantile dotplots [131] transform probability distributions into dots, a set of discrete countable samples. Prior work has used visualization methods to communicate uncertainty stemming from a range of sources, such as sensor measurements [92] to cartographic data [152]. In this thesis, we use graphical visualization to display probabilistic AI outputs as distributions that quantify uncertainty.

2.4 Challenges with Understanding Uncertainty

Understanding uncertainty is inherently difficult, especially for people without prior training or statistical literacy [46]. People are inherently averse to probabilistic thinking, and cannot reason effectively under uncertainty [216]. Decision-makers rely on mental shortcuts, or heuristics, in order to avoid complex probabilistic reasoning under uncertainty. Failure of these heuristics can lead to severe errors in judgement, adversely affecting the interaction with the ML-based Decision Support System and subsequent decisions. Factors such as decision-makers' prior background and numerical literacy [180] affect how well people are able to understand uncertainty.

Despite various tools [81] and methods [79, 153, 218] to compute and visualize uncertainty in AI outputs, empirical evidence is mixed on how uncertainty improves efficiency [9, 228], quality of decisions [72, 148, 134], trust and reliance [156, 238, 40, 42, 14, 12], and the decision-maker's confidence in decisions [95]. However, there is general consensus that presenting uncertainty leads to better decisions, increased transparency, and reporting.

However, uncertainty visualizations often assume that people are rational actors and make statistically-driven judgements [113]. The effectiveness of uncertainty visualizations depends on contextual factors and user characteristics [209], such as task type and desired decision outcomes [46]. Non-experts may misunderstand graphical uncertainty [199] or fail to realize that the graphic depicts uncertainty, which Joslyn et al. [122] term as "deterministic construal ". Studies in human-AI decision-making find that solely providing uncertainty representations is insufficient to improve decision quality [46], and additional guidance is needed [189]. Work in human-AI decision-making calls for presenting uncertainty aligned to decision-makers' cognitive needs and considering not only *what* information is presented, but *how* to frame it within the broader context of the decision-making task [106].

Prior work [9, 95, 238] has prominently investigated the effects of communicating confidence scores on subjective and objective metrics; few studies have focused on how people perceive, interpret, and make use of uncertainty (e.g., to understand the role of uncertainty communication in helping model creators to understand the capabilities and limitations of their models [188]). While those approaches helped to identify and quantify the effects of uncertainty, they provided limited understanding of cognitive and sense-making aspects [128].

CHAPTER 3

Quantifying Bayesian Uncertainty of Black-box AI Models

3.1 Introduction

Clinical Decision Support Systems (CDSS) based on Artificial Intelligence (AI) are promising to assist clinicians in providing better patient care in high-stakes medical decision-making [179], including for brain tumors. AI-based CDSS have shown potential to accurately segment tumors [137, 162], which could aid in treatment planning for patients with glioma [206], a malignant manifestation of brain tumors [107, 28]. AI models such as Deep Neural Networks (DNNs) can learn patterns from radiologist-annotated multi-modal Magnetic Resonance Imaging (MRI) scans for tumor delineation with high accuracy [186]. These optimistic results have contributed to accelerating approvals from governing agencies such as the U.S. Food and Drug Administration (FDA) [213], with the goal of integrating them into clinical workflows [18].

In the domain of radiology, AI-based CDSS have shown potential to assist in treatment planning, which requires precise segmentation to ensure that the treatment is effective in resecting the bulk of the tumor, without unnecessarily affecting functional brain areas. Radiologists rely on their years of training and practice to estimate the extent of tumor location and spread as precisely as possible. Radiologists delineate the tumor based on its appearance on the MRI, and mark additional boundaries around it to indicate certainty about their segmentation, or lack thereof [205]. Recently, researchers have used DNNs [102, 96, 98, 186] to automate tumor segmentation [129] for a large cohort of patients at a fraction of the time it would take to do it manually.

However, these algorithms do not communicate uncertainty, which remains a key barrier to their responsible adoption in clinical workflows [212]. AI-based CDSS that produce single-

valued (i.e., deterministic) predictions do not reflect the inherent uncertainty in medicine [93], and encourage impressions of "superhuman" ability of AI [45] without mechanisms to contest these claims [15]. This can exacerbate problems with inappropriate trust and reliance on AI-based CDSS [42], with potential harm to patient health [204]. Despite calls for Transparent and Responsible AI [226, 3] to quantify [81, 24], report [176, 159, 8] and design for uncertainty [32], little attention has been dedicated to addressing uncertainty effects in AI models for healthcare [15, 212]. This lack of model transparency remains a key barrier to the responsible adoption of AI in clinical practice [171, 80, 221].

Uncertainty Quantification (UQ) for AI-based CDSS [146] remains challenging due to their high computational costs [117]. For example, Bayesian analysis [21, 22, 197] provides a principled and systematic approach to quantify and update the uncertainty of model parameters, but typically is solved via expensive iterative algorithms such as Markov chain Monte Carlo (MCMC) [97, 184, 6, 37] or variational inference (VI) [121, 225, 29]. While Bayesian methods have been historically used for UQ in medicine [215], they are difficult to scale for handling, e.g., DNNs in AI-based CDSS that often involve millions of parameters [153, 163, 87].

We seek to enable Bayesian UQ for million-parameter DNNs, i.e., to create million-parameter Bayesian neural networks (BNNs) [30, 78]. We achieve this by introducing a low-computation strategy for performing Bayesian inference on a small, strategically chosen portion (layer) of the entire DNN, thereby creating a *partially* BNN (pBNN) that approximates the uncertainty of a full BNN. We use the Flipout [231] VI algorithm to solve the resulting Bayesian problem on the target layer. Unlike related strategies [218, 182, 10] that attempt Bayesian inference only on the last layer of a DNN, which does not necessarily offer the best uncertainty representation [236], our approach allows a justified selection of which part of the DNN to "Bayesianize " guided by sensitivity analysis (SA).

To impart the ability to express uncertainty in tumor segmentation models, we build a pBNN for a state-of-the-art 7.8 million-parameter U-Net [186] proposed for medical image segmentation. To empirically validate our approach, we train a full BNN as a baseline and compute the uncertainty approximation discrepancy of the pBNNs to this full BNN. Our experiment demonstrates that the pBNN based on the SA-selected layer is the most efficient in approximating the full Bayesian uncertainty (per Bayesianized parameter) compared to other layer choices. Our results suggest that the SA-based layer selection offers an effective and inexpensive way to identify the best DNN layer to perform Bayesian inference.

The main contribution of our work is to enable Bayesian UQ for million-dimensional medical DNN models. We propose a novel computational method that uses an SA-based DNN layer selection for targeted Bayesian inference, which is especially useful in scenarios

where practitioners and model developers want to understand the model's uncertainty but do not have the resources for a full Bayesian inference on the entire DNN. Our work advances future interactions for Transparent and Responsible AI that can support the investigation of the uncertainty in DNN models. We illustrate one such interaction, through the use of uncertainty maps, to show how our work can allow clinicians to interpret the uncertainty of the tumor segmentation AI.

3.2 Method

We introduce a novel method to address the scalability challenges of building million-parameter BNNs by performing *targeted* Bayesian inference on a strategically selected portion of the overall DNN—we call the result a partially Bayesian neural network (pBNN). We present the overall procedure via three main steps (see Fig. 3.1): (1) access a deterministic DNN model, (2) conduct SA to select a DNN layer for Bayesian inference, and (3) perform targeted Bayesian inference on the selected layer using Flipout VI to arrive at the final pBNN. We describe each step below in detail.

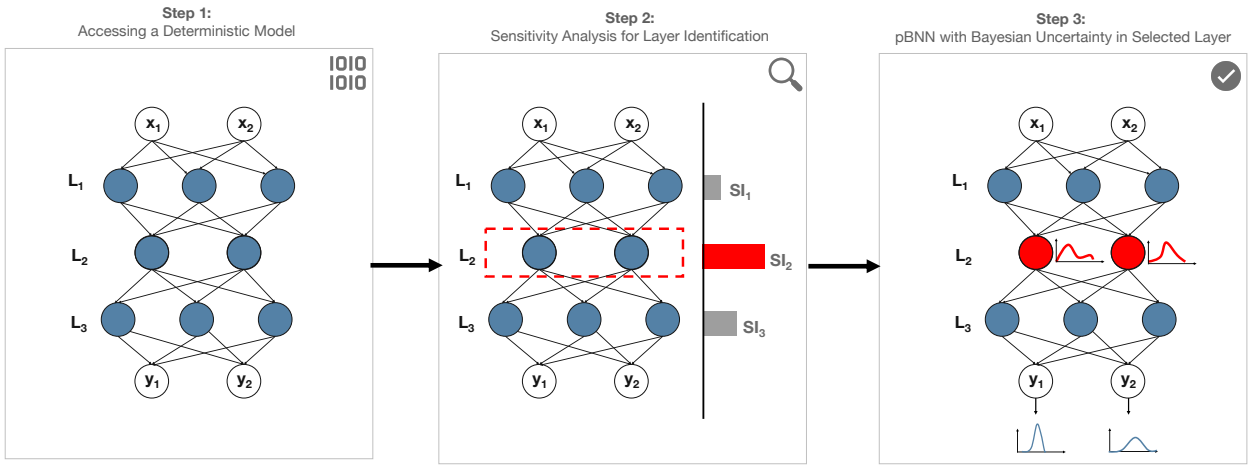


Figure 3.1: Flowchart to address scalability challenges of BNN by building pBNN: Step (1) represents a million-parameter deterministic NN. Step (2) Conduct Sensitivity Analysis to identify the layer that has the greatest impact on model output. Step (3) Conduct Bayesian inference on the most sensitive layer to build pBNN. The resulting pBNN combines the strengths of both deterministic and Bayesian NNs, allowing scalability.

3.2.1 Step 1: Access a Deterministic DNN Model

The first step is to access a deterministic DNN model (i.e., one that is trained by minimizing the loss in Eqn. (2.2)). The purpose of this DNN is to provide an inexpensive and meaningful starting point for the upcoming SA and layer selection. One scenario is that such a deterministic DNN is already available from a pre-existing study or another research group, and we can simply inherit it. If it does not exist yet, one may train a new model following Eqn. (2.2), which is relatively (to any BNN) inexpensive.

Since this deterministic DNN will be used to guide the selection of model layers for Bayesian inference, it must have the same architecture as the DNN that will eventually undergo Bayesian inference. However, the precise training setup (e.g., choice of loss function, learning rate) for creating this deterministic DNN is flexible since the goal in the next step is to seek a general sense of the sensitivity behavior. This flexibility is important in practice, for instance in situations where we are only given a pre-existing deterministic model but without information on how it was trained.

3.2.2 Step 2: Conduct Sensitivity Analysis to Select a Layer for Bayesian Inference

Next, we perform SA on the deterministic DNN from Step 1 in order to assess the model prediction behavior as a result of model parameter variation. We adopt gradient-based sensitivity for its simplicity, although other types of sensitivity (e.g., variance-based sensitivity) may also be used. Specifically, we first take the partial gradient of model prediction output with respect to the parameters of the candidate layer being considered (e.g., ℓth layer), which can be calculated easily from a DNN using only a single pass of back-propagation. Since this partial gradient is a vector, we then take its L2-norm to transform it into a scalar (other norms such as the L1-norm may be used as well).

Lastly, to avoid automatic favoring of larger layers due to simply having more parameters (and therefore more entries contributing to the gradient vector), we normalize (divide) by the number of parameters in layer ℓ (N_ℓ) to arrive at the *average sensitivity per parameter*. Our overall Sensitivity Index for layer ℓ is:

$$\mathrm{SI}(\ell) = \frac{1}{N_\ell}\Big|\Big|, \nabla_{w_\ell}\hat{y}, \Big|\Big|2. \tag{3.1}$$

The layer with the highest SI, $\ell^* = \arg\max \ell\mathrm{SI}(\ell)$, would then induce the largest change in prediction $\hat{y}$ when its parameters w_ℓ are perturbed. This layer thus presents as the most critical layer, justifying focusing our resources to capture the Bayesian uncertainty for layer

ℓ^*. Layer ℓ^* will then undergo Bayesian inference in the next step.

We note that the gradient is a local operator and captures sensitivity locally at the optimized parameter values of the deterministic DNN (i.e., from Eqn. (2.2)). To capture sensitivity more globally across a wider range of possible w_ℓ values, one may also consider averaging the gradient from multiple locations of w_ℓ, for instance in estimating the prior expectation of the gradient: $_{w_\ell}\left[\frac{1}{N_\ell}\left|\right|, \nabla_{w_\ell}\hat{y}, \left|\right|_2\right]$. However, these global measures are more expensive to compute.

3.2.3 Step 3: Perform Targeted Bayesian Inference to Build pBNN

In the last step, we perform targeted Bayesian inference on the identified layer ℓ^* to build the pBNN. Formally, we partition w into two subsets: $w = w_B, w_D$ where w_B are the DNN parameters corresponding to layer ℓ^* that will undergo Bayesian inference, and w_D consists of all remaining parameters that are treated deterministically (i.e., not as random variables). This sets up a pBNN where only a strategically chosen portion (layer) is Bayesianized.

At this point, we may simply freeze w_D at its deterministic DNN values w_D^* (from Step 1), and invoke VI to compute the *conditional* posterior:

$$\theta_B^* = \underset{\theta_B}{\arg\min}, , D_{\mathrm{KL}}\left[q(w_B;\theta_B) \parallel p(w_B|w_D^*, x_T, y_T)\right]. \tag{3.2}$$

However, allowing w_D to change may allow additional improvement in approximating the posterior. Therefore, we propose to jointly optimize w_D (the deterministic parameters) and θ_B (parameters of the variational posterior for the Bayesian DNN parameters)

$$\theta_B^*, w_D^* = \underset{\theta_B, w_D}{\arg\min}, , D_{\mathrm{KL}}\left[q(w_B;\theta_B) \parallel p(w_B|w_D, x_T, y_T)\right]. \tag{3.3}$$

Lastly, we solve Eqn. (3.3) numerically using the Flipout VI algorithm introduced in Chapter 2, Section 2.3.1.

3.3 Findings

3.3.1 Illustration: Toy Problem

Before applying the pBNN to the tumor segmentation DNN, we first demonstrate our framework on a test problem using a small densely-connected DNN with synthetic training data in order to bring intuitions and insights.

To set up the problem, we generate 256 synthetic training data points following the example from [30] with some modifications:

$$y_i = x_i + 0.3\sin(2\pi x + \epsilon_i) + 0.3\sin(4\pi x + \epsilon_i) + \epsilon_i \tag{3.4}$$

where $\epsilon_i \sim \mathcal{N}(0, 0.02^2)$.

We fit the data to a DNN comprised of an input layer, 9 hidden dense layers with swish activation, and a dense output layer with linear activation (see Table 3.1). The dense layers varied in width to mimic the varying sizes of convolutional layers encountered in tumor segmentation DNNs. The DNN has a total of $13,385$ trainable parameters.

Layer	**Tunable Parameters**
hline Input	0 Dense 1
4 Dense 2	12 Dense 3
40 Dense 4	144 Dense 5
544 Dense 6	2112 Dense 7
8320 Dense 8	2064 Dense 9
136 Output	9

Table 3.1: Architecture of the densely-connected DNN used for the test problem.

In **Step 1** of the pBNN procedure, we produced a deterministic DNN by minimizing the loss in Eqn. (2.2), with the result plotted as the solid orange line in Fig. 3.3. In **Step 2**, we computed the gradient-based SI in Eqn. 3.1 for each of the 10 candidate layers, shown in Fig. 3.2. The bar graph indicates generally lower SI for the middle layers compared to those near the input or output; the lowest SI is at Layer 7, and the highest SI is at Layer 9. Layer 9 is therefore selected for Bayesianization. In **Step 3**, we construct a pBNN by performing Bayesian inference on Layer 9 using Flipout VI and training all other layers deterministically, following Eqn. 3.3. We provide more details of the pBNN implementation and validation below.

Our pBNN is implemented using TensorFlow Probability (TFP) [60], where Layer 9 is replaced with a DenseFlipout layer, and all other layers remain Dense. Training is then performed simultaneously on θ_B and w_D as in Eqn. (3.3). For the Bayesianized DNN parameters (w_B), we adopt independent Gaussian priors with a rather wide standard deviation to reflect an initial uninformative distribution: $p(w_B) \sim \mathcal{N}(0, 10^2)$. A Gaussian likelihood is also used to depict independent Gaussian observation noise ($\sigma_\epsilon = 0.02$ to be consistent with

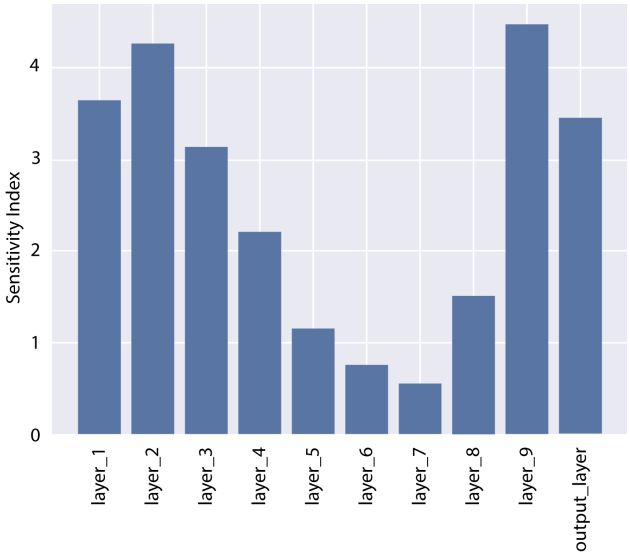


Figure 3.2: SI for all layers in the DNN for the test problem. Layer 9 has the highest SI and is selected for Bayesianization.

the data-generation in Eqn. 3.4) on the y targets:

$$p(y_T|x_T, w_D, w_B) = \prod_{n=1}^{N_T} \frac{1}{\sigma_\epsilon\sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{y_n - f(x_n; w_D, w_B)}{\sigma_\epsilon}\right)^2\right]. \tag{3.5}$$

The variational posteriors are from the family of independent Gaussians $q_i(w_B;\theta_B) \sim \mathcal{N}(\mu_i, \sigma_i^2)$, and so the variational parameters are $\theta = \mu_i, \sigma_i$. We train this pBNN using the Nadam [63] optimizer for 10^4 epochs with a learning rate of 10^{-3} and a batch size of 64.

To assess whether the pBNN constructed with SA-selected Layer 9 provides a good approximation to the uncertainty of the full BNN, we also train a full BNN where all layers are replaced by DenseFlipout layers; the posterior push-forward uncertainty[1] The full BNN is shown in Fig. 3.3. Furthermore, to investigate the performance of pBNNs if a layer different from Layer 9 were selected, we construct separate pBNNs where a different layer is Bayesianized.

Since a good pBNN is one that faithfully approximates the uncertainty in the full BNN, we measure its performance based on how close the pBNN's posterior push-forward is to the full BNN's posterior push-forward via the KL divergence. Following the same reasoning as the SI from Eqn. 3.1, we normalize this quantity by multiplying N_ℓ (since lower KL is

[1]Posterior push-forward is $p(f(x; w_B, w_D)|x, x_T, y_T)$, which differs from the posterior predictive $p(y|x, x_T, y_T)$.

better) to arrive at our validation metric:

$$D_{\text{norm}} = \frac{N_\ell}{N_{tot}}[f(x; w_D^*, w_{\text{partial},B}), ||, f(x; w_{\text{full}}))], \tag{3.6}$$

where $w_{\text{partial},B} \sim q_{\text{partial}}(w_{\text{partial},B}; \theta_B^*)$ are the Bayesianized parameters in the pBNN, and $w_{\text{full}} \sim q_{\text{full}}(w_{\text{full}}, \theta^*)$ are the full set of parameters (all Bayesianized) in the full BNN. The D_{norm} values for each pBNN are plotted against SI in Fig. 3.4, where we observe a Spearman's correlation of $r_s = -0.83$, which is considered to be strongly correlated in the literature [2]. This supports our SA-based layer selection strategy, where the layers with high SI will also lead to low D_{norm}. Lastly, we note that while we perform this brute-force comparison to validate the effectiveness of our approach, in practice one would only train the pBNN on the SA-selected layer as described in Section 3.2.

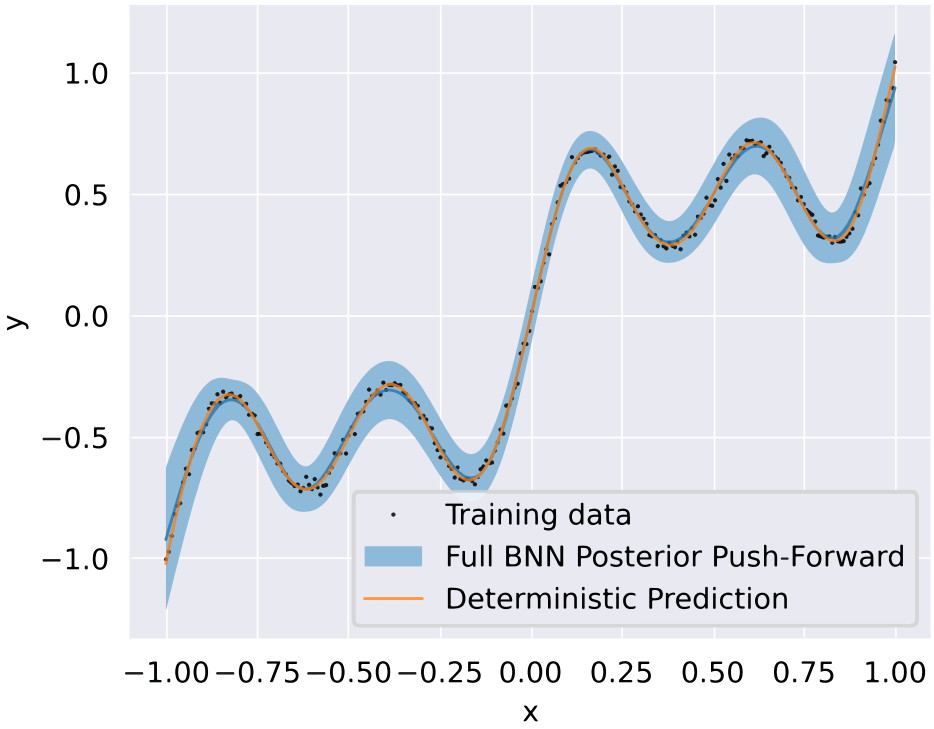


Figure 3.3: Training data (black dots) overlaid with the deterministic DNN prediction (solid orange) and full BNN posterior push-forward uncertainty (solid blue line is the mean, blue shadow is the 95% credible interval).

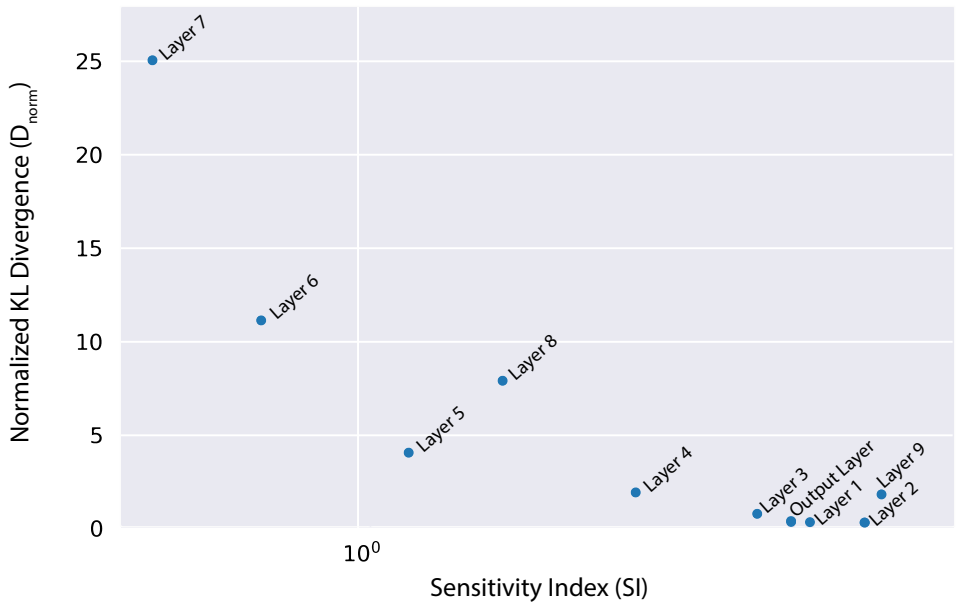


Figure 3.4: Normalized KL divergence from the full BNN posterior push-forward to the different pBNNs is highly correlated with the SI that is used to select the layer for Bayesianization (Spearman's correlation $r_s = -0.83$).

3.3.2 Application: Image Segmentation with U-Net

Here, we present the main demonstration of our pBNN, applying it to a state-of-the-art DNN model used for medical image segmentation: the U-Net [186]. Since our goal is to demonstrate our pBNN framework presented in Section 3.2 and not to develop or justify the choice of architecture, we take the U-Net architecture *as given*.

3.3.2.1 Step 1: Access a Deterministic DNN Model

We begin by training a U-Net deterministically using the architecture and code published by [167]. As shown in Fig. 3.5, the U-Net comprises an encoder, decoder, and skip connections. The encoder module comprises 2D Convolutional layers followed by MaxPooling layers with window size 2×2, and 2D Spatial Dropout layers for regularization. The decoder module is composed of corresponding Conv2DTranspose layers followed by Concatenate layers. Nested between the encoder and decoder is the 2D Upsampling layer with bilinear interpolation. The U-Net has a feature map size set to 32, kernel size 3×3, and dropout rate 0.2. For the 2D Convolutional layers, ReLU activation is used along with He uniform variance scaling initializer, and kernel size is set to 2×2. Overall, this U-Net has a total of 7.8 million tunable parameters.

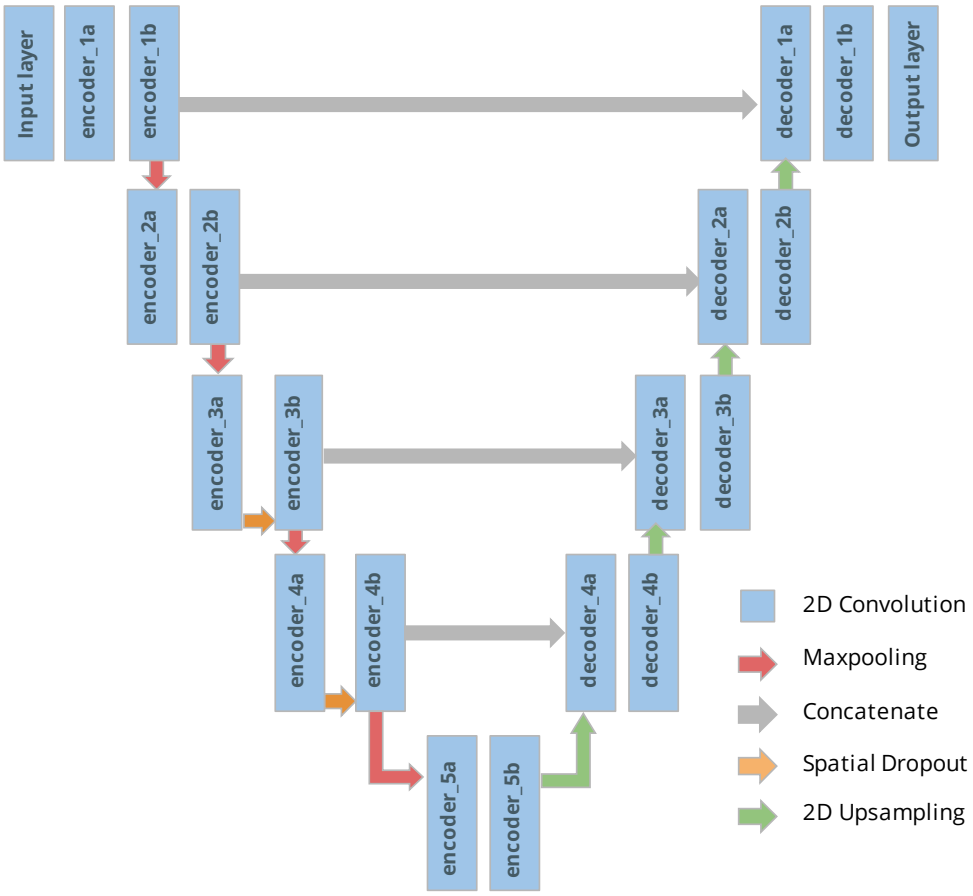


Figure 3.5: U-Net architecture with our naming system of the different tunable layers considered for the pBNN.

We perform the training on a dataset from the Medical Segmentation Decathlon Challenge Task 01 [196] consisting of brain MRI scans of patients with glioma. The training set consists of 58,464 2D images (each sized 144×144) and a separate validation set contains 6,624 images. Data augmentation is employed during training to increase data diversity, including random choices of horizontal and vertical flips, ±45° rotation, and ±2.5° shearing. The ADAM optimizer [136] is run for 30 epochs, using a learning rate of 0.001 and a mini-batch size of 128. The training loss function is a weighted sum of the Dice coefficient [59, 54] and standard Binary Cross Entropy. The training results are summarized in Table 3.2.

3.3.2.2 Step 2: Conduct Sensitivity Analysis to Select a Layer for Bayesian Inference

We computed the gradient-based SI in Eq. 3.1 for each candidate layer, shown in Fig. 3.6. In particular, the layers near the top portion of the encoder carry larger SI. This response

Performance	Value
Training time	5 hrs
Model size	94.3 Mb
Trainable parameters	7.8 million
Validation accuracy	0.9915
Validation Dice coefficient	0.8508

Table 3.2: Summary of the deterministic U-Net training.

gradually decreases as we proceed further into the model, reaching a minimum near the middle, followed by a slight rebound at the decoder layers. The lowest SI occurs at Layer decoder_3a, and the highest SI is at Layer encoder_1a. Layer encoder_1a is therefore selected for Bayesianization.

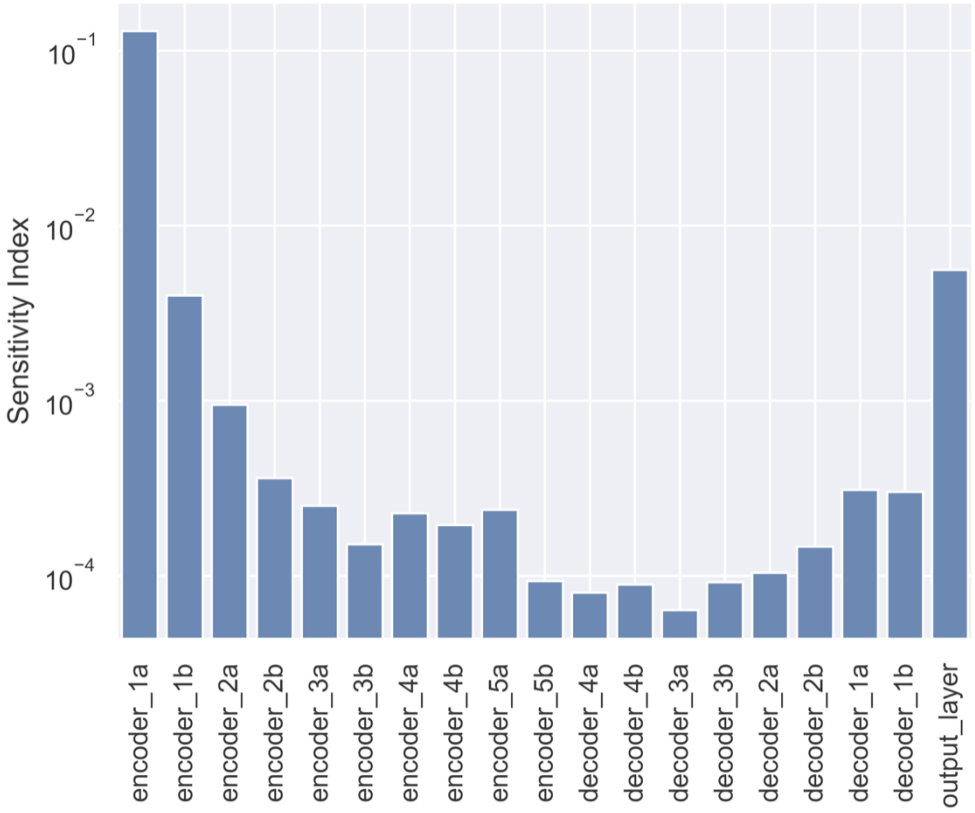


Figure 3.6: SI for all layers in the U-Net used for tumor segmentation. layer_1a has the highest SI and is selected for Bayesianization.

3.3.2.3 Step 3: Perform Targeted Bayesian Inference to Build pBNN

We construct a pBNN by performing Bayesian inference on encoder_1a using Flipout VI and training all other layers deterministically, following Eqn. 3.3. We implement our pBNN using TensorFlow Probability (TFP) [60] by replacing the 2D convolution layer of encoder_1a with `tfp.layers.Convolution2DFlipout`. Training is then performed simultaneously on θ_B and w_D as in Eqn. (3.3). For the Bayesianized DNN parameters (w_B), we adopt independent Gaussian priors with a rather wide standard deviation to reflect an initial uninformative distribution: $p(w_B) \sim \mathcal{N}(0, 10^2)$. Since our data label for each pixel is binary while our model prediction ranges $[0, 1]$, we adopt a likelihood associated with the binary cross-entropy:

$$p(y_T|x_T, w_D, w_B) = \prod_{n=1}^{N_T} \left[y_n f(x_n; w_D, w_B) + (1 - y_n)(1 - f(x_n; w_D, w_B)) \right]. \tag{3.7}$$

The variational posteriors are from the family of independent Gaussians $q_i(w_B; \theta_B) \sim \mathcal{N}(\mu_i, \sigma_i^2)$, and so the variational parameters are $\theta = \mu_i, \sigma_i$. We train this pBNN using the Nadam [63] optimizer for 400 epochs with a learning rate of 10^{-4} and a batch size of 196, using 8 GeForce RTX 2080 Ti GPUs in parallel. The ELBO training history of the pBNN is shown in Fig. 3.7 (plot with encoder_1a emphasized in red). While the plot indicates a small rise around 150 epochs due to numerical instabilities with the optimization algorithm, overall the curve steadies out by 350 epochs (note that the y-axis is logarithmic, which amplifies the lower-value fluctuations) and appears noticeably flatter compared to the beginning of the optimization.

3.3.3 Validation

In this section, we provide validation for the pBNN procedure by illustrating whether the U-Net pBNN constructed with the SA-selected encoder_1a layer provides a good approximation to the uncertainty of the full BNN. To achieve this, we also need to train a full BNN where Bayesian inference is performed on all layers. Such an attempt to obtain the full uncertainty is highly expensive, and is in fact the original motivation for the new pBNN. We clarify that a full BNN is not needed as a part of the regular pBNN procedure described in Section 3.2.

We train the full BNN, where all 19 layers of the U-Net are Bayesianized, using the full extent of our available computational resources: 2000 epochs using the Nadam [63] optimizer with learning rate 10^{-4} and batch size of 196, 94.72 GB of memory, 8 GPUs in parallel, and it required 3 days of continuous computation to arrive at a steadied ELBO. The best-performing model is saved during the training process. Due to the tremendous size of

this BNN, expert experience is also relied upon for randomized initialization and algorithm tuning to arrive at a reasonably converged full BNN. Additionally, we also train the 19 pBNNs, each with one of the 19 layers targeted for Bayesian inference, but with shorter 400-epoch training. The ELBO training history for all models is summarized in Fig. 3.7, with the pBNN using the SA-selected layer (encoder_1a) emphasized in red. While we train the full BNN for significantly longer than the pBNNs, we plot their ELBO to around 350 epochs for better readability and comparison. All training curves gradually decrease and start flattening around 150 epochs, after which they become noisy. The plot suggests that all pBNN models we built have converged reasonably well.

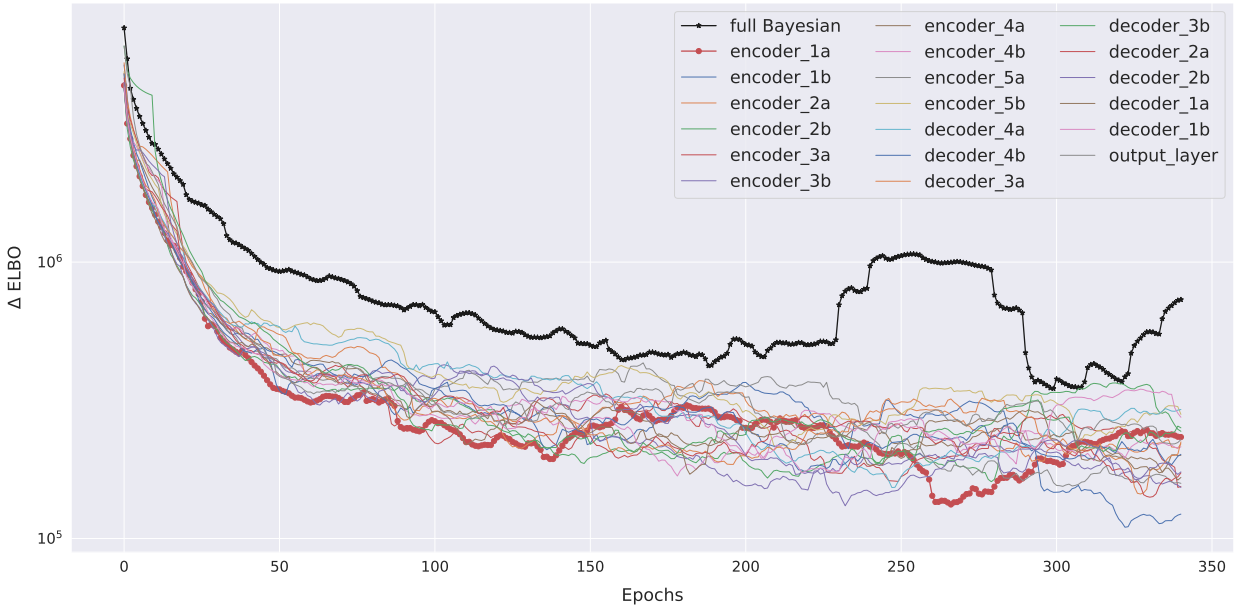


Figure 3.7: ΔELBO training history for all models.

Since a good pBNN is one that faithfully approximates the uncertainty in the full BNN, we measure the pBNN performance using the normalized KL divergence between the pBNN's posterior push-forward and the full BNN's posterior push-forward, D_{norm}, defined in Eqn. 3.6.

With each U-Net output being a 144×144 pixel value, directly computing the KL divergence for the joint 144×144-dimensional posterior push-forward PDF would be very difficult. Therefore, we further simplify the computation of D_{norm} through an independence approximation, breaking the 144×144-dimensional KL into 20,736 pixel-wise one-dimensional KL computations. For a given prediction image, each pixel's KL is estimated by first generating 100 samples of the DNN parameters $w_{\text{partial},B}$ from the optimized variational posterior $q_{\text{partial}}(w_{\text{partial},B}; \theta^* B)$, combine it with the other deterministic parameters and evaluate the corresponding pixel predictions $f(x; w^* D, w\text{partial}, B)$, and for each pixel fit its 100 samples

to a Beta distributions to approximate its posterior push-forward PDF. Beta distributions are often used to model the uncertainty of Bernoulli distributions and are suitable for our problem to ensure pixel-wise prediction values reside between $[0, 1]$. Once the Beta PDFs are extracted, KL divergence can be calculated analytically. The pixel-wise KL divergence values are then summed to obtain the estimate of the joint PDF's KL value (since the KL divergence between two joint PDFs is factored into the sum of KL divergences of individual marginals). The final $D_{\,textnorm}$ values are shown in Fig. 3.8, and they show a correlated, inverse trend compared to the SI values in Fig. 3.6, supporting that SI is a good, inexpensive indicator of the KL performance metric.

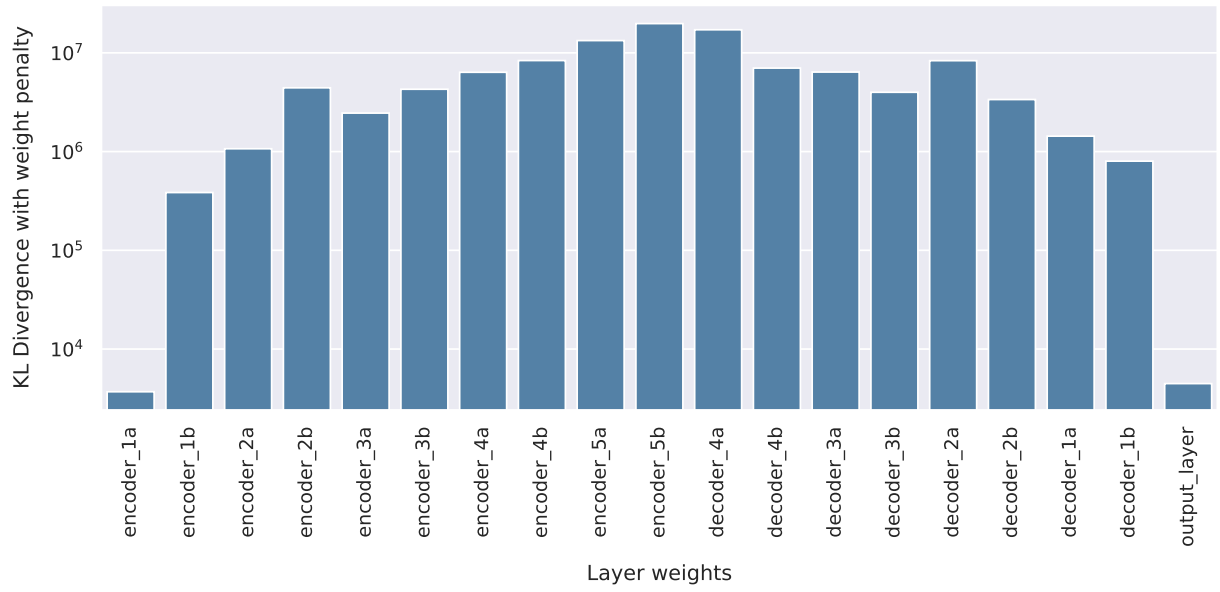


Figure 3.8: Normalized KL D_{norm}, for all 19 pBNNs each built with Bayesian inference on one of the layers in the U-Net.

Figure 3.9 shows a scatter plot for all 19 pBNNs of the KL divergence (not yet normalized by N_ℓ) versus N_ℓ, the number of parameters of the layer targeted for Bayesian inference; the pBNN with the SA-selected layer (encoder_1a) is marked in red. Ideally, we would like to select a layer for Bayesianization that has low KL divergence *and* a small number of layer parameters (i.e., towards the bottom-left corner). However, these two desirable properties generally conflict, and a tradeoff needs to be made. This can be seen by the Pareto optimality front by the plot's lower-left convex hull. In this case, we see our SA criterion indeed identifies one of the Pareto points in layer encoder_1a. The output_layer also resides on the Pareto and would also serve as a reasonable selection for Bayesianization; indeed, the SI values presented in Fig. 3.6 do show the output_layer to be the runner-up. However, we point out

that the nature of how the two layers are good choices is different: encoder_1a achieves a lower KL than output_layer, but output_layer has fewer parameters.

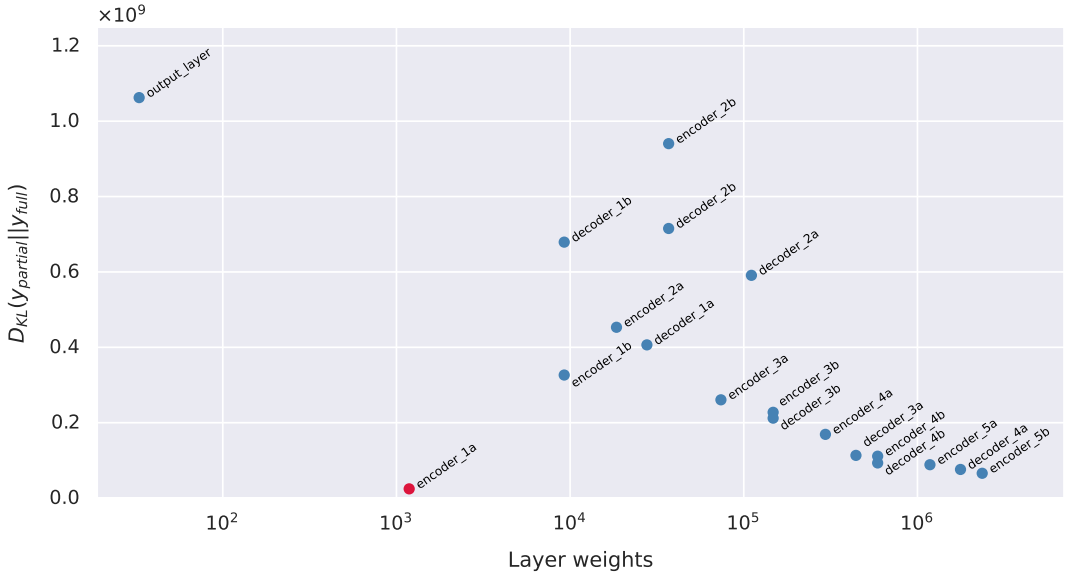


Figure 3.9: Scatter plot of KL divergence sum versus number of parameters in the Bayesian layer.

3.4 Discussion

The results of our empirical validation in Section 3.3.3 illustrate the effectiveness of the pBNN: the pBNN where Bayesian inference is performed on the SA-selected layer achieves the lowest normalized KL divergence—that is, compared to if a different layer were chosen, it provides the best capture of the full BNN uncertainty. Our three-step pBNN methodology in Section 3.2 is set up with implementation considerations in mind. Step 1 allows the leveraging of pre-existing trained DNNs, but also describes the procedure if a new DNN needs to be trained from scratch. Step 2 assesses and identifies the portion of the DNN for Bayesian inference on a layer-by-layer basis (i.e., keeping "layer" as the unit), since accessing and replacing an entire layer (e.g., from a Dense layer to a DenseFlipout layer) is very convenient in programming infrastructure such as TensorFlow Probability. Such modifications do not need the user to program new layers, or to Bayesianize portions of a layer. The gradient-based sensitivity measure adopted is also inexpensive to compute and can be readily extracted from a deterministic DNN. Step 3 takes advantage of existing implementations of Bayesian inference algorithms, such as the Flipout VI, that can be specified and used together with

non-Bayesian layers.

Our method is particularly valuable for high-dimensional problems (e.g., DNNs with millions of parameters), where performing a full Bayesian inference would be extremely expensive. Our work enables principled approximate Bayesian inference to be scaled up to millions of dimensions. While other related *partial* Bayesian strategies compare test error performance [236], we explicitly compute closeness to the fully Bayesian model, with the aim of faithfully capturing the true uncertainty that we are justified to have. Our overarching goal is to provide a transparent quantification of uncertainty of our AI system (not only to optimize accuracy).

We illustrate how our computational framework could be used to visually represent prediction uncertainty in Section 3.4.1. While uncertainty communication is not the primary goal of our paper, we demonstrate how our work can advance an understanding of AI in a healthcare context. We conduct preliminary analysis on these maps to demonstrate how a radiologist might interpret these images. The maps allow one to correlate the model's mistakes and how confidently the model fails. With clinical insights and interpretation, one can also explore the model's behavior on specific use cases. For example, model predictions on out-of-distribution data can be potentially flagged based on uncertainty [170] to caution the radiologist about the model's reliability in those regions. This would be especially useful for healthcare contexts where models are highly domain-dependent and localized [74]. Apart from its potential to enable flagging of unreliable cases, our Bayesian framework can allow for construction of an expert-informed pBNN. It permits a flexible choice for prior and likelihood, potentially allowing practitioners to inject their specialized knowledge. For this, we refer to a branch of Bayesian statistics called expert knowledge elicitation [166].

While our work produced promising results, limitations still exist. While we provided a detailed validation by brute-force computation of the full BNN, one would not do so in practice. Assessing the quality of posterior approximation is a challenge in general for all BNNs, where inexpensive and effective diagnostics are needed to monitor the algorithm's progress, especially for large models. Analyzing ELBO curves is one such approach, but one cannot know when the ELBO is stuck in a local minimum. Furthermore, the magnitude of the ELBO value does not directly indicate how far the variational approximation is from the true posterior, since ELBO differs from the KL divergence by an unknown constant (i.e., one should not expect ELBO to converge towards zero with more epochs, as one would with mean-squared-error). We also have not explored performing Bayesian inference on multiple layers, or portions of layers, which may offer further improvements. Such expansions would require the development of efficient strategies to optimize SI across various combinations of multiple layers, and additional benchmarking and validation to assess the tradeoffs between

their computational cost and ability to approximate the full BNN. Despite these limitations, our work demonstrates the effectiveness of SA-based pBNN in Bayesian UQ of large, million-dimensional DNNs.

3.4.1 Case Study: Uncertainty Communication for Tumor Segmentation

Although the main contribution of our work is to provide a scalable method for Bayesian UQ in large DNNs, here we illustrate one way that our uncertainty information could be used in practice. We introduce uncertainty maps as a tool that communicates how confident (uncertain) a model is in its prediction. Fig. 3.10 illustrates these maps for one such example: the top row displays the 4 modalities of MRI; the bottom row displays the ground truth, prediction, uncertainty, and truth-prediction discrepancy, respectively, for the SA-guided pBNN (Section 3.3.2.3).

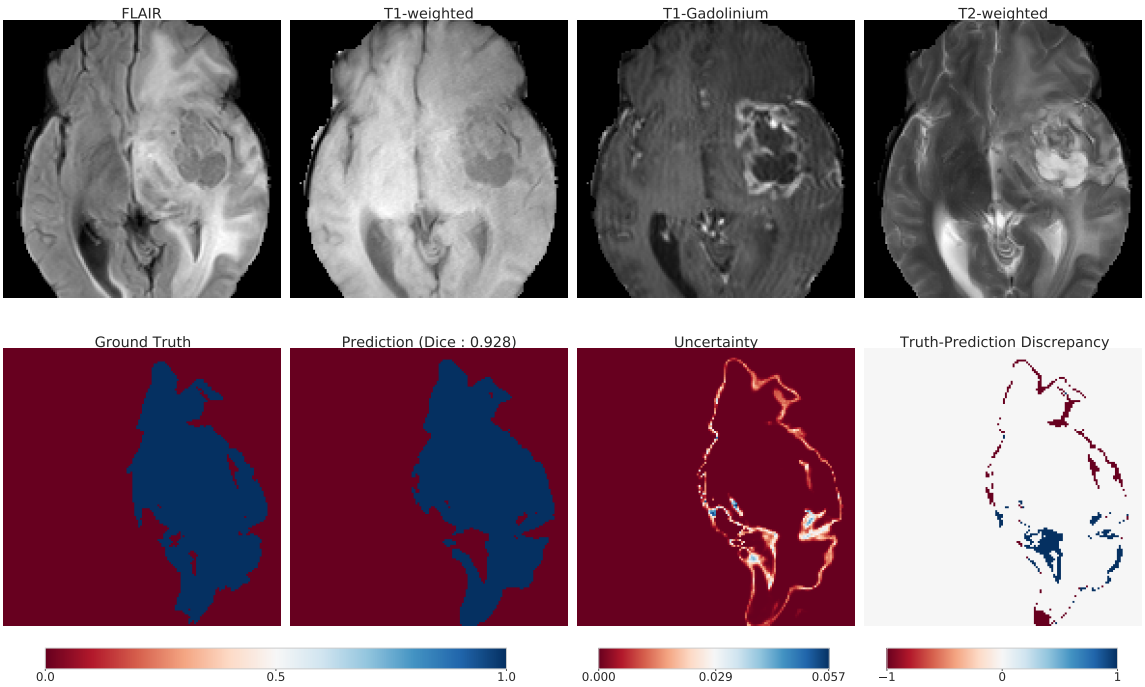


Figure 3.10: Uncertainty map accompanying a prediction made by the pBNN.

3.4.1.1 Visualizing Uncertainty

We describe how we compute and visualize the uncertainty of the pBNN's prediction for a specific use case (Fig. 3.10). We chose this image from our test dataset as an example

because the ground truth indicated a large percentage of the pixels belonging to the tumor class (25

We build the uncertainty map with the same 100 samples by computing within-pixel standard deviation. The range of uncertainty displayed is unique to the image and is not a global uncertainty measure. To visualize the truth-prediction discrepancy map, we subtract the prediction map from the ground truth. The negative extreme value (-1) indicates regions where the model predicts tumor when the tumor is absent in ground truth (false positive, or over-prediction). The positive extreme value ($+1$) indicates regions where the model does not predict a tumor when the tumor is present in the ground truth (false negative, or under-prediction).

3.4.1.2 Interpreting Uncertainty

We show a use-case scenario on how to interpret and analyze uncertainty maps. We asked one of the co-authors, who is a researcher in radiation oncology and a collaborator with routine knowledge of MRI image interpretation, to conduct a preliminary interpretation and analysis of the MRI and the uncertainty map in Fig. 3.10. Note that this interpretation was not performed by board-certified radiologists, so it only serves as a guiding example.

According to their interpretation, it was clear to them that the pBNN was fairly confident in distinguishing the tumor core from the background, but was highly uncertain at the boundary regions of the tumor. This is consistent with manual segmentation. For example, even when a single radiologist performs multiple segmentations on the same image, the greatest differences (intra-rater variability) are often present in the boundary region.

In this particular example, the necrotic, enhancing, and non-enhancing parts of the tumor are surrounded by edema. Since the model predicts the whole tumor and does not perform subclass recognition, it predicts the edematous region of the tumor. However, since edema has a distinct contrast in FLAIR, there is a possibility that this FLAIR modality is contributing largely to the observed uncertainty. This is in line with segmentation by radiologists, which also has the most uncertainty in the FLAIR modality.

Further, comparing the uncertainty map with the truth-prediction discrepancy revealed that the pBNN is highly uncertain where it under-predicts and has low uncertainty where it over-predicts. A possible reason for this is in the context of the model's ability to correctly call out all of the tumor pixels in the ground truth. For over-prediction, the model covered the tumor region and went beyond it (which does not count as a miss), so the uncertainty is low. Whereas for under-prediction, the model failed to correctly call out the tumor (which counts as a miss), thus the uncertainty is high. One such area of under-prediction is in the lateral ventricle of the right hemisphere. This region is darker than the rest of the tumor (in

FLAIR modality), but is still a part of the tumor due to the presence of edema.

One possible explanation of this model behavior is that the model has learned to distinguish tumors based on pixel intensity and contrast. As the algorithm sees the lateral ventricle of the right hemisphere to be significantly darker than the rest of the tumor, it thinks this is not a tumor. While this seems like a reasonable mistake for the model to make, a radiologist would have no problem marking this as a tumor, as they have a firm understanding of anatomy. Example regions of over-prediction are near the middle frontal gyrus and near the necrotic region in the right hemisphere. This region is craggy, and it seems like the model smoothed over these intricacies to get a good enough prediction.

3.5 Summary

In this paper, we proposed an approach to compute the Bayesian uncertainty of million-dimensional DNNs for medical image segmentation. Starting from a deterministic DNN, we used gradient-based sensitivity analysis to identify a layer to perform Bayesian inference, thereby creating a partially Bayesian neural network (pBNN) that is computationally much less expensive to construct than a full BNN. We demonstrated the pBNN method on a state-of-the-art 7.8-million-parameter U-Net for brain tumor segmentation. Our validation indicated that the pBNN based on an SA-selected layer provided the best approximation to the uncertainty from a full BNN, compared to other layer choices.

Our methodology enables model developers and practitioners to compute the Bayesian uncertainty for large deep learning models. In a life-critical domain such as healthcare, deep learning models can have far-reaching impact, but also can make mistakes leading to disastrous consequences. Communicating uncertainty in model predictions helps encourage clinicians to engage with the AI-based DSS with a healthy dose of skepticism and a failure-centric mindset. Additionally, UQ mitigates "super-human" perceptions of AI that can lead to unjustified over-reliance. Thus, quantifying and communicating uncertainty would allow for safer and responsible deployments of AI-based CDSS in clinical workflows.

For future work, we plan to conduct a large-scale evaluation of the impact of uncertainty communication to radiologists for the task of tumor segmentation. In this paper, we conduct preliminary interpretation of these maps and demonstrate the possibility of such a study. We would further like to understand if these maps can be used as a tool for model explanation, and whether these maps are more useful than conventional metrics. It can also be a useful tool in reducing alarm and click fatigue, as the model can only predict when absolutely certain and refrain from predicting otherwise.

CHAPTER 4

Reflecting on Epistemic Gaps for Uncertainty Communication

4.1 Introduction

Climate services aim to support communities in making effective disaster risk management decisions. Communities build disaster resilience in order to anticipate, prepare for, and recover from disasters. Climate services aim to support communities in making effective disaster risk management decisions. Climate services include basic scientific information and data that explain disaster phenomena as well as the design of models, tools, and AI-based decision support systems based on this data. Climate services are produced by different stakeholders, from private consulting firms who work directly with communities, to academic institutions who produce scientific knowledge about disasters and tools that operationalize this knowledge.

However, climate services that are designed and provided to decision-makers are not always actionable or useful, leading to the "climate usability gap". Mismatch between what climate information scientists view as useful and produce, vs what decision-makers can understand, trust, and use to make real-world decisions. Efforts have been made to translate complex science to decision-makers in easy-to-understand ways, such as through visualizations. Despite these efforts, from an epistemic perspective, this gap is persistent as a result of a physics-first, expert-centric scientific knowledge production and tool design paradigm. There is growing recognition in the field to move towards a model of knowledge co-production with multiple disciplines and stakeholders involved.

Prior work in climate services has advocated for multidisciplinary teams to design more actionable and usable climate services. However, simply assembling a multidisciplinary team

does not lead to better design. Different disciplines have different ways of design (e.g., human-centered design process, but that is not the only one). However, team dynamics can lead to one way of design being more valued or dominant over others. This leads to design assumptions, implications, and the eventual designs stemming from dominant discipline. What we don't know is how multidisciplinary teams construct design implications and what design implications remain hidden as a result of dominant ways of thinking.

In this work, we explore the following research questions:

RQ1: How do multidisciplinary team dynamics affect design implications for climate resilience solutions?

RQ2: What are the design implications that remain hidden or under-explored due to team dynamics?

Here, we conducted collaborative autoethnography (CAE) as a multidisciplinary team collaborating over a period of 1.5 years to produce climate services for urban flooding in Detroit. Drawing upon 1.5 years of team progress, ethnographic engagement, and internal discussions, we employ workshop methods to critically reflect on how we produced knowledge across disciplinary and institutional boundaries. To answer our research questions, we adopted a convergence science framework [173] to investigate knowledge production and design of risk communication tools in multidisciplinary teams. We employed workshop methods to scaffold the self-reflection process. In a participatory fashion, all authors collaboratively identified aims, goals, structure, and activities of these workshops. In the first workshop, we focused on individual disciplinary stances and surfacing tensions through the activity of documenting information needs of individual researchers on the team. Next, we engaged in individual self-reflection on questions formulated by identifying common needs through the first workshop. The main purpose was to maintain continuity and prepare for a more focused discussion in the second workshop. Finally, we conducted the second workshop, which focused on achieving alignment and moving towards bridging disciplinary silos.

4.2 Method

4.2.1 Study Context: Urban Flooding in the Metropolitan Detroit Region

We investigated our research questions within the case of flood risk management in the Detroit Metropolitan Area (DMA). The DMA is increasingly vulnerable to urban flash floods

due to aging infrastructure and climate change [214]. In August 2014, the DMA experienced the largest U.S. natural disaster, described as a "1-in-500 year event" by the Federal Emergency Management Agency (FEMA) and National Oceanic and Atmospheric Administration (NOAA) [217]. Four to six inches of rain fell in a four-hour period over this small, yet densely populated area, causing $1.8 billion in direct flood damages to homes, businesses, and infrastructure. This is but one case of the growing flood-related destruction in the US, leading to 1,782 fatalities and damages exceeding $102 billion, affecting 99% of US counties since 2000 [214]. This real-world case study ensures ecological validity of our research findings and allows us to contextualize decision-making within the socio-political landscape, environmental factors, and historical risk patterns of this region.

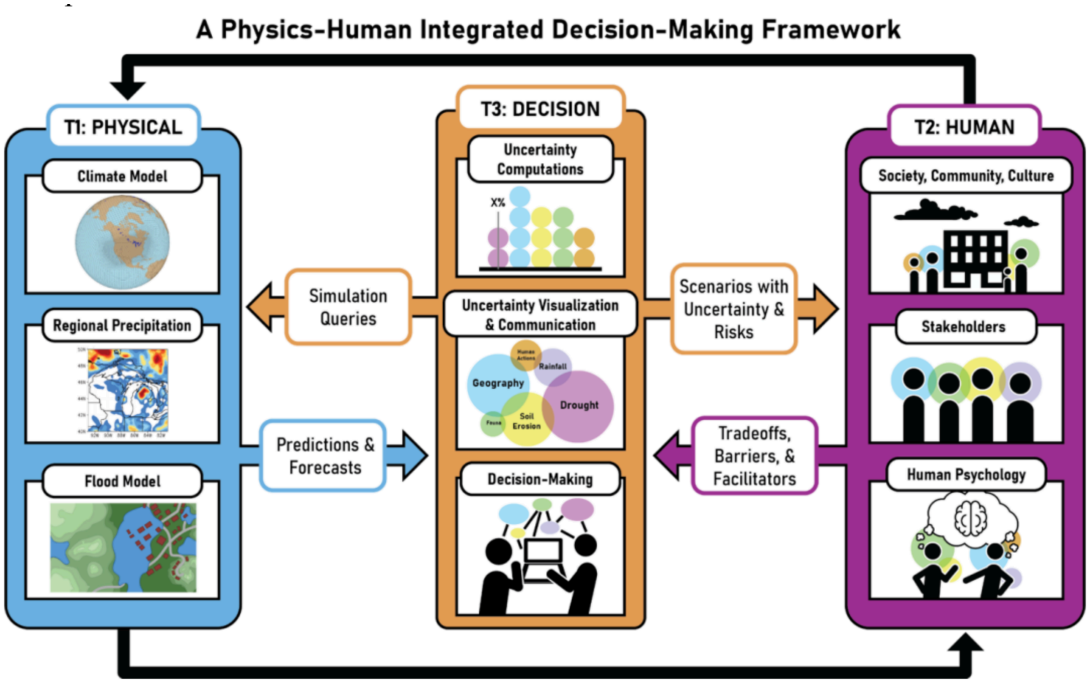


Figure 4.1: Three domains of expertise in the study: physical modeling, human factors and stakeholders, and decision and design. Their integration informs flood risk communication tools.

4.2.2 Multidisciplinary Research Team

In this work, we draw upon a larger federally funded, multi-year, multi-institutional research project to predict and communicate flood risk, which has been ongoing for about two years. The goal of the project was to reduce the human and economic costs of extreme flooding events, which are on the rise due to climate change. The project team included engineers, atmospheric scientists, psychologists and anthropologists to develop better flood predictions and ensure decision-makers can understand them. Together, the team developed new knowledge and methodologies for assessment, analysis and management of climate change impacts on flood risk. The team answered questions such as: *What are the chances, risks and impacts of flooding? What local regions is flooding more likely to occur in? What information do decision-makers need?*.

To answer these questions, the team developed tools and models that examined the causes and impact of flood-producing storms in the region (Section 4.2.1). Climate scientists collected infrastructural data such as local buildings, river channels, stormwater drainage systems, etc., and explored the role of infrastructure in shaping the impact of historical floods. They found that some regions were flooded even when they did not receive intense local rainfall due to aging and poorly designed infrastructure. They further simulated different rainfall conditions and explored their impact on the local landscape. Decision scientists employed survey methods to understand how local stakeholders estimate flood damages, perceive and mitigate flood risk. They also explored the sensitivity of models developed by climate scientists and their influence on decision-making. Anthropologists studied the institutional processes, culture and practices around handling flood risk. Design researchers (HCI) studied local stakeholders' context of use, current tools, and information sharing mechanisms to find implications for design.

4.2.3 Collaborative Autoethnography Workshops

4.2.3.1 Workshop 1

The aim of this phase was to identify the informational needs of participants. The workshop consisted of a series of structured discussions that served as scaffolds for the participants to articulate their informational needs specific to their disciplinary roles (e.g., model building, stakeholder engagement) and the kinds of data required to fulfill those needs. This phase resulted in the identification of a set of core research questions to guide the rest of the study phases. This phase involved a workshop with scaffolding activities to help researchers articulate their information needs. Due to the distributed nature of the team and infeasibility of travel, the workshop was conducted online over Zoom. We closely followed prior work

in participatory methods to support co-creation of research questions [27, 157]. The workshop session lasted about 90 minutes and was led by two facilitators. Facilitators shared a Zoom link with participants and waited for them to join the workshop. After all participants joined, facilitators briefly explained what the role of being a "participant" entailed, set expectations, and asked permission to record the workshop session. Facilitators then briefly explained the purpose of the workshop, what design implications meant, and set the tone for the rest of the workshop. Facilitators then led the participants through a series of structured activities that provided scaffolding for identification of information needs. The workshop was divided into two 25-minute sessions – the first one with researchers (i.e., from disciplines such as climate science, engineering, psychology, etc), while the other was conducted with the ethnography team and stakeholder representatives. In the first session, facilitators posed a probing question to researchers – *"As a researcher, to be able to do your work on this project, what do you need to know?"*. Each researcher elaborated on their most important or pressing informational needs to be able to conduct their work. After this, facilitators encouraged other participants to ask questions and provide comments and feedback. Facilitators proceeded in a round-robin fashion, so all the researchers in the workshop had a chance to answer the probing question. In the next session, the facilitators shifted gears and focused on the users by tailoring the session to ethnographers and stakeholder representatives. The facilitators invited these participants to give a summary or examples of questions that different stakeholders had. Facilitators posed questions to the ethnography team such as – *"Who are the different stakeholders you have encountered? What do they need to know to do their job?"*. Facilitators then asked ethnographers to give a high-level overview of what different stakeholders needed to know. Then, facilitators opened up the discussion to other participants, seeking their comments, thoughts, and follow-up questions. Finally, facilitators wrapped up the workshop by explaining next steps in the study, how research questions will guide further phases, and taking questions from participants. Facilitators thanked all participants for their time and effort and ended the Zoom meeting.

4.2.3.2 Self-reflection Activity

The goal of this phase was to use ethnographic data as boundary objects [203, 147] to support individual reflection guided by the research questions identified in Phase 1. This enabled us to surface individual interpretations that could serve as inputs for the next phase of discussion. To support this, we created role-specific reflection prompts grounded in the research questions identified in Phase 1. This phase lasted two weeks. First, we shared the research questions identified in Phase 1 with the ethnography team, who then identified and shared relevant excerpts from prior field notes, interview transcripts, and memos. Ethnographic excerpts

were then shared with participants along with personalized reflective prompts: what stood out, what was confusing, what resonated with their work, and what gaps they noticed. Participants were then asked to read the ethnographic excerpts, reflect on them, and note detailed responses to these reflective prompts. These reflections then guided the next session to uncover existing gaps and problems in design. The facilitators shared a Google Drive link with the participants to submit their reflections in their preferred format – written notes, audio recordings, pictures, or sketches. The facilitator conducted periodic check-ins with participants for submitting reflections.

4.2.3.3 Workshop 2

The final workshop built on the workshop results from Phase 1 and the individual reflections from Phase 2. The main goal was to achieve convergence across disciplines and between researchers and on-the-ground stakeholders. The session took place four weeks after the previous workshop and lasted about 1.5 hours. Facilitators provided a brief recap of earlier sessions and clarified the workshop focus. Facilitators then encouraged the participants to turn on their cameras if comfortable, in order to create a more engaging space for discussion. Facilitators began the session by focusing on the ethnographic insights shared by the two ethnographers on the team. The facilitator conducted a semi-structured interview with them to bring out key themes from their fieldwork. These questions included elaborating the role of ethnography in the team, specific cases from fieldwork, and reflections on disciplinary knowledge production. Their responses served as a starting point for a broader group discussion. The team explored shared experiences, recurring challenges, and differences in how risk is understood across disciplines and roles. The discussion also touched on how assumptions made by researchers—particularly around expertise—can affect the design of tools and communication practices. After the initial discussion, a second facilitator guided an open-group brainstorming session on how to bridge disciplinary boundaries and engage in knowledge co-production. Facilitators took reflective notes and memos. We also recorded the session video and audio.

4.2.4 Workshop Attendees

Participants had a long history of collaborating on the MURI project, and this project resulted from one such collaboration. Participants were researchers from various disciplines – psychology, ethnography, uncertainty quantification, disaster risk communication. Most participants were in researcher roles such as research manager, postdoc, and professor. In addition, study participants also included community stakeholder representatives.

4.2.5 Data Collection

We collected audio and video recordings of the workshop discussions using Zoom. These recordings captured both individual reflections and group discussions around participants' informational needs. We also took facilitator notes during the session. All data were collected and anonymized. We collected excerpts guided by our research questions from the ethnography team. We collected individual reflection prompts and responses from all participants. We collected audio and video recordings of the session via Zoom, including any chat transcripts and screen captures of the MIRO board. We also gathered all artifacts generated during the session, including sticky notes and facilitator memos.

4.2.6 Data Analysis

We adopted a reflexive thematic analysis (RTA) [33] methodology to analyze the data from the workshops. We followed the six-step process of RTA outlined by Braun and Clarke [35, 34]. We followed prior work in applying RTA to workshop data [94].

In order to become familiar with the data, the co-authors conducted several debrief meetings after the workshops. The first author then engaged in deep reading of the data to become fully immersed and familiar with it. The first author took notes of her reflections, assumptions she was relying on when analyzing data, and her own positionality as an HCI researcher. Notes included insights related to individual data, as well as in relation to the entire dataset. The first author avoided transcribing the data with Rev, but transcribed it with Zoom [1] and fixed errors to get a deeper understanding of the data. The first author then labeled the data with codes that captured these reflections as well as participants' verbalizations. The data were coded in multiple rounds, by iteratively discussing with co-authors. The first author then examined codes and performed affinity diagramming to collate the codes into broader potential themes. In this phase, we began to draw upon theories of interdisciplinarity [173], collaboration [52], and knowledge production [11, 119]. We connected these to our positionality as interdisciplinary authors. The first author then checked the candidate themes against the complete dataset and shared the narrative with co-authors. Themes were then further developed by splitting and combining in a way that was faithful to the core underlying shared concept. The first author then shared the themes with co-authors and further refined them, and noted down the story of each emergent theme. In the final phase, the first author wove together the analytic narrative and the data, and contextualized the analysis in relation to existing literature. The writing of this manuscript is the final stage of reflexive thematic analysis.

[1] https://zoom.us/

4.3 Findings

4.3.1 Producing

Epistemic pluralism led to divergent understanding of terminologies and goals of the project within the research team. Despite valuing interdisciplinary research, knowledge production was in the realm of STEM disciplines. Models were created with the assumption that "scientists know best", pursuing accuracy as a goal. Despite human-centered thinking of first understanding the context of use, modelers "did not want to be pigeonholed". Rather than changing thinking based on context of use, the context and stakeholder needs were quantified into design. Engaging with stakeholders was done by human-centered teams, which took time, effort, and delicate handling, especially in a shifting political environment. However, as models were already built based on assumptions, stakeholders were engaged to determine parameters rather than what gets built. Finally, the team sought to convince the stakeholders to see value in their approach, despite the stakeholders wanting tangible and actionable information to make decisions.

4.3.1.1 Epistemic Pluralism

Epistemic pluralism emphasizes the need for diverse ways of knowing, analyzing, and interpreting climate change by drawing insights from the sciences, social sciences, and humanities. Recognizing this, the team was composed of several disciplines such as hydrometeorology, uncertainty quantification, human-centered design, psychology, and anthropology. Various disciplines have their own norms of what problems counted as important, how to solve them, and what determined success. For example, climate scientists focused on *"correctly capturing an extreme flooding scenario [with] ensemble models"*(P02). Disaster resilience scientists *"followed a human-centered design process, [starting] with identifying who the stakeholders are "* (P15). HCI researchers identified existing decision-making practices and *"current decision-making support tools available to different decision-makers"*(P06), *"the breakdowns with current systems"*(P06) to then *"develop solutions to address these gaps"*(P15). Anthropologists' *"role [was] to describe a society, how things work in that society."*(P12). While these different approaches to investigate the phenomena of climate change existed and were valued by the team, that alone did not guarantee an integrated approach:

> *"We're all experts in our field [and] together we can contribute so much more as this interdisciplinary team. And yet it continues to be a difficult thing to step outside of disciplinary silos." – P06*

Thus, while disciplinary diversity was valued, challenges needed to be overcome before epistemic pluralism was achieved. These challenges included differences in understanding of shared goals and vocabulary. For example, some researchers thought that the *"broader goal [was] really about representing uncertainty and risks."*(P12), while others thought that *"the goal of the project [is] designing something for the stakeholders as well."*(P11). Disciplinary diversity also meant that the same vocabulary held different meanings: *"I struggle with the word uncertainty, because it means so many different things to different disciplines"*(P10). One researcher explained these differences as being due to "taking the path of least resistance":

> *"We tend to take the path of least resistance [as] it's easier to work in our own thing. But there has to be a concerted effort to cross disciplinary boundaries." – P02*

Thus, while it is easy to value multidisciplinary, deep integration work must be undertaken to change the knowledge production model of "working on our own thing" to conducting research together as a team. Others agreed: *"I think as a team, we need to figure out what we mean by what decisions people are making. Which ones do we want to know about, and what sorts of uncertainties are we concerned about knowing more about."*(P11). Therefore, simply having multiple disciplines on a team does not lead to epistemic pluralism, and effort must be undertaken to integrate different disciplines for this model to be effective.

4.3.1.2 Scientists Know Best

Disciplines also viewed knowledge produced in relation to the decision context differently. For example, modelers and climate scientists held the belief that *"uncertainty is important toward decision-making, and uncertainty needs to be computed or captured accurately ... for making a good decision"*(P07). However, they acknowledged not having the knowledge about what decisions stakeholders are making. The below quote illustrates this dichotomy:

> *"If we can calculate the various uncertainties that are attached to the different predictions and forecasts, what uncertainties matter to the decision-making? which further raises the question what decisions are being made?" – (P07)*

Thus, decision-making came into the picture to figure out *which*, rather than *whether* uncertainty matters to decision-makers. On the contrary, Human-centered design researchers acknowledged that *"none of this exists in a vacuum [and first wanted to understand] what decision makers are already doing"*(P06). Scientific knowledge production was prioritized over the design of tools as the perspectives of some disciplines were more dominant than others:

"there are differences in understanding of risk across disciplines, and some are considered more dominant than others."(P15).

Despite *"getting a lot of hints"*(P07) from stakeholders to produce tools that fit in the existing context of use or infrastructure, researchers were *"nervous that [they were] thinking so specific [and] too small [and feared] being pigeonholed"*(P02). Some acknowledged the need to *"find balance and alignment between fundamental research and tangible, concrete tools [so that] it's [more of a] win-win situation"*(P07), yet most felt that *"building into their current infrastructure ... limits what the goal and outcome of our research is"*(P08). Researchers reinforced their stance on advancing "basic science" by invoking the project proposal: *"I would point out also that the proposal here to say that we are doing basic science. We are not inventing gadgets necessarily."*(P12).

Thus, despite disciplinary diversity, scientific knowledge production was prioritized over the design of tools due to fear of being pigeonholed. There was also dissonance in the belief that uncertainty was important for making decisions, but researchers also lacked knowledge about what decisions are being made. To resolve this, researchers made assumptions about what decisions stakeholders are making, relying on intuition to inform the design of models:

> *"My intuition tells me that specific thresholds or specific values that are really driving a lot of these [disaster management] decisions. And this is the information that decision makers are looking for. [For example] when an area is going to have 3 inches of rain or below. So ... our models need to give better, more accurate updates once we start hitting these specific thresholds." – (P08)*

Here, the modelers' intuition that decisions are triggered based on certain thresholds drove design choices. Similarly: *"many stakeholders have binary decisions to make, [such as] if personnel should come to work, or stay away"*(P02) and information needs to be provided to support these binary decisions. Scientific knowledge production was solely in the domain of modelers' expertise, while *"how uncertainty, advance notices and probabilistic information should be communicated to downstream emergency managers and individual groups"*(P02) was a secondary design problem. Thus, with the "scientists know best" design philosophy, scientific information was generated first, and stakeholders and decision context were "downstream communication" problems to be solved by human-centered disciplines.

4.3.1.3 Political Forces Shape Stakeholder Relationships

Establishing trusting relationships with stakeholders required significant time and effort. As one anthropologist notes, *"building connection to the community takes time [as] you have to build relationships with different people"*(P12). Over the period of a year and a half,

ethnographers steadily built trusting relationships and gained access to spaces, up to the point they could *"walk into the base, walk in the hallway and be recognized as someone who just won't leave them alone (laughs) ... that's a good place to be in"*(P11). However, as the political administration changed [2], there was increased sensitivity and scrutiny dealing with topics that were even remotely anti-government, such as climate change. This severely impacted researchers' ability to continue their work, as federal workers were afraid to lose their jobs:

> *"Within that [military] environment, one has to be quite respectful. People are afraid of stepping outside of their swim lane, and as a result, losing their job. So what you say, and how you engage with these people ... requires a lot of delicacy, being really careful and respectful of the world that they're living in." – (P11)*

Preserving stakeholder relationships when navigating a changing political environment required a lot of tact and sensitivity. Another anthropologist agreed: *"the political environment right now makes our job harder. If [someone] is supporting us, is that gonna be seen as a negative by their commander?"*(P12). The changing political environment made it very difficult for the rest of the team to establish new relationships and gain access to stakeholders' spaces. Ethnographers remained the only channel of communication between stakeholders and the rest of the team: *"having firsthand communication with these stakeholders does give me some insight into things, you know, [and] that could be useful [as otherwise] you all don't have the opportunity to get that information."*(P11). Shaped by these conditions, ethnographers took on new and non-traditional roles in the multidisciplinary team:

> *"The role of the ethnographer is to describe society. Our role wasn't and isn't, historically or traditionally, to figure out what people need, and we didn't set out on this project as that. It's come out of this work. We've taken on this role of trying to figure out the [information] needs of the other teams, and establishing these relationships." – (P11)*

Thus, establishing stakeholder relationships took significant effort, and was shaped by political forces outside the teams' control.

4.3.1.4 New Questions, Old Data

As the project evolved and preliminary models were created, new information needs emerged. Researchers wished to engage stakeholders to validate their modeling assumptions and results. For example, climate scientists and uncertainty modelers wanted to know *"specific*

[2] Federal workers were being fired under the Trump administration.

risk thresholds"(P02), *"spatio-temporal resolution"*(P03), *"timescales"*(P07) that stakeholders wanted in their models. The following quote summarizes the rationale for engaging stakeholders:

> *"We [can] run the physical base model and the open flood model [but] it takes a lot of time to run. Depending on the capacity of the computer, we can produce the results at a resolution of 15 min or an hour interval. But the most important question is what information does the decision-maker need to make their decision. It's very important to make sure that our team can provide the results at that resolution." – (P04)*

Once the models were built and run, the information about what specific resolution decision-makers need becomes relevant to make sure that the models can be run for that specific resolution. The team turned to ethnographers and their data in order to seek answers to these information needs. This was like answering new questions with old data:

> *"We had certain objectives for these interviews [which] I conducted over the last year and a half. We didn't even know all the questions at the beginning; we weren't thinking about them in the same way. So we can't have answers to some of those questions based on the interviews that I've done ... that's not what our original research question was. [I will] try to see what I could possibly speak to with the information I gathered. [But] I want to be able to do justice to my data and to the research that I've done, and I don't feel comfortable speaking on behalf of all these people and look for answers to your questions" – (P11)*

The above quote reflects how new information needs for modeling emerged with the evolution of the project, and ethnographers were approached to fill these needs.

4.3.2 Transforming

Once scientific knowledge is produced, it was then "transformed" for downstream users and applications through various modes of knowledge integration. Here, the majority of the transformation work was done by human-centered design teams as they were expected to take objective knowledge and transform it to explain to stakeholders in a way that helped them. However, stakeholders did not think the same way as researchers, which led to the whole basic science vs applied science debate.

4.3.2.1 Convincing stakeholders to see value

Once the scientific information was produced and validated, researchers had to convince stakeholders how it was an improvement over their current ways of doing things. Fundamentally, researchers believed that the current way of doing things was not enough, and that they needed to create a paradigm shift:

> *"We need to be thinking about changing the way that people think about predictability, uncertainty and decision making in this context of floods. We have to have kind of a paradigm shift in terms of the way that things are currently done and what we are proposing." – (P02)*

Thus, researchers believed that something was fundamentally lacking in the current way that stakeholders think about uncertainty, and the communication of scientific information would lead to a paradigm shift. However, this was not easily accepted by stakeholders, and researchers thought it necessary to persuade them to see the value in a basic science approach:

> *"We may have to persuade them of the value of [what] basic science might be able to offer. But [the ethnographers] were very explicit that it will be very difficult to change their minds. So knowing how their attitudes and beliefs differ from our own understanding ... of not what they need, but what they should want. How do we meet them halfway? Do we also begin to incorporate stakeholders' way of thinking [to] persuade them what would be valuable to them in the future?" – (P06)*

While researchers recognized that it is very difficult to change stakeholders' minds and beliefs, they nonetheless wanted stakeholders to accept the basic science approach. To do this, researchers wanted to create something *"tangible [representative] of all of our work together"*(P02), such as *"a product, or a demo, or prototype"*(P06). Other proposed: *"We should work together on building a dashboard where every component of the project can show their result as a collective platform so that it can act as an early product."*(P05). Researchers mentioned how the demo *"has to show pretty clearly how it's an improvement over what they already have in place"*(P01). Thus, researchers believed stakeholders' current approaches were inadequate and that they needed to be persuaded to accept a basic science approach to create a paradigm shift and improve how they make decisions.

4.3.2.2 Translating Complex Science to Stakeholders

During this phase, the design goal was to translate complex science to stakeholders in a way that was useful to them in making decisions. After scientific knowledge was produced and validated, it had to be *"communicated to downstream users and stakeholders"*(P02). Here, the *"language of communication [became important as] more scientific or technical messages might be less understood"*(P02) by stakeholders. For example, *" [complex equations and] formulas might not be useful to people immediately."*(P07) if they did not have *"the expertise to read mathematical formulas"*(P12) and information might have to be delivered *"in plain language"*(P12) instead. For this translation process, human-centered disciplines and *"ethnographic findings [became] increasingly relevant after some of the preliminary model design had been completed."*(P11) while previously they were *"definitely not the center of attention"*(P12). Unlike basic sciences which produced scientific information, these disciplines were valued due to their applied orientation:

> *"There's applied anthropology where the anthropologists are more involved in developing solutions for problems and work with communities in this sort of applied way. And a great grandchild of applied anthropology is this user experience research which is closely aligned with human computer interaction. " – (P12)*

Thus, disciplines such as anthropology, human-computer interaction and user experience research were assigned the task of translating complex science concepts to stakeholders. Despite the expectation to explain science to stakeholders, such disciplines did not have insight into how scientific information was produced and viewed it as "magic":

> *"My understanding is that the modelers are going to do their magic by modeling precipitation [and] the hydrology. Then they're going to give us the results of their modeling and simulation efforts. And then we need to add the human component to it and the cultural values. [But] I'm just not clear on [what] information is going to be sent over to the behavioral team, like our team, to be able to even begin to think about how to communicate that work in a manner that would facilitate decision-making." –(P10human)*

This highlights a fragmented design process, where objective scientific information was first produced, and the human components and values were "added in" later in the process.

4.3.2.3 Lost in Translation

While the research team valued basic science, stakeholders wanted tangible, actionable information that would help them make decisions. However, things were lost in translating

complex science to stakeholders:

> *"We know way more about the weather and the environment than we are using effectively. And that comes down to how all that information is provided, how it's shared, how the warnings go out. There's lots of systems in place [but] a lot of it's lost in the translation." – (P16)*

The above quote from a stakeholder representative captures how stakeholders currently have access to more information than what is usable or actionable. Stakeholders were already aware of their current problems, both through internal large-scale studies and their own lived experience. However, awareness did not automatically lead to action due to human psychology and *"institutional culture"*(P07). The following quote summarizes this:

> *"The [stakeholders] know what their problems are. The issue is, how do you deal with them? How do you develop an emergency system that has both accurate information coming into it, and is listened to by people and responds in the way that it needs to? For example, soldiers feel that they can make it through flooded waters because they've been trained to trust their tactical vehicle. [But] how do you keep soldiers from driving into flooded waters? How do you take that data about storms and flooding and then translate it into an emergency system that is effective?" – (P11)*

Therefore, for effective decision-making, data needs to be both accurate and listened to and responded to by people under existing institutional culture. Researchers noted: *"[Stakeholders] were saying that modeling was actually not of interest to them. Rather they were interested in having rain gauges, because those things are things you can see."*(P07) such as *"tools for engineers to look at information"*(P11).

4.3.2.4 There is no universal number

Researchers wanted to know what specific values of parameters would stakeholders prefer. But they were treating stakeholders like a homogeneous group. Furthermore, it is not a question of stakeholder preference, but what they are going to actually use the information for. The below quote by a stakeholder representative explains how the numbers that the modeling team is looking for are context- and decision-specific:

> *"You can predict atmospherically what is going on. The question is what things could that [information] impact. For example, it can be [given before] 5 days, 3 days, 2 days. [But it should be] enough that we know we can start evacuating*

> *people, or we need to make some big decisions. So the number that you're looking for, how many days, that may change depending on who the stakeholder is." – (P14)*

Thus, things can be predicted atmospherically, but the number you are looking for depends on impact and who the stakeholder is. Others agree that *"it's guaranteed to change depending on the stakeholder, and also their use cases. Like if you have a longer lead time, you might be able to get vulnerable equipment out [of risky areas] or get additional help and resources. Maybe that information is useful 7 days [in advance] as opposed to 12 or 24 hours before [flooding] where you're limited in terms of what you can do."*(P01). Current socio-political systems, such as inter-county coordination, also dictate what information is needed and its level of actionability. *"For example in Southeast Michigan, we knew there was a storm that was gonna hit. [The region spans] multiple counties, which allow for coordination between different stakeholders who have control over the flood gates and stormwater systems. So something like that would be helpful a couple days out, or maybe even 24 hours in advance. That allows for a response at this [county] level for these different people."*(P14)

4.3.3 Contextualizing

Organizational and institutional forces come into play when situating the tools produced with "objectivity" in the "subjective" context of use. People are aware of where problems exist and manage disasters collectively. However, organizational forces constrain the ability of people to make decisions solely based on knowledge. This renders the tools produced ultimately powerless in the face of these organizational forces.

4.3.3.1 People know where problems are

Tools produced with scientific knowledge provide information on how and where problems occur. For example, an equation that provides an understanding of atmospheric forces and a tool that indicates where flooding occurs on the base. However, engineers and master planners have years of lived experience and already know where problems occur. Therefore, lack of knowledge is not the thing that prevents acting upon it:

> *"They know on the base where the flooding happens. The engineers know where the water comes up. They know where the problem is, but there are all these constraining factors that prevent them from being able to execute on those as projects move up the chain of command from an engineer who maybe prioritizes certain projects they want to do." – (P11)*

The ethnographer states that engineers know where the problems on the base are, yet institutional factors constrain them. Engineers and master planners collectively manage disaster risk by dividing day-to-day operations and long-term decision-making: *"One is more concerned with day-to-day security of the installation, and the other one is more concerned with dealing with events, such as inclement weather events"*(P11). Most people work in the Department of Public Works, and others are within training and operations on the installation. Engineers are responsible for managing engineering problems on the base. To carry out these tasks, engineers walk around the base and take note of what problems exist. They also notice problems by *" listening to the squeaky wheels on the base. Who's making noise and who has power to see needs for improvement in terms of engineering, for example, parking lots that need paving or stormwater infrastructure"*(P11). The decisions and needs of stakeholders depend on phases of the disaster. Disaster risk management literature acknowledges that stakeholders' *"tasks and the information needs are quite different when it's before an event, during an event or long after an event"*(P15, external). However, some decisions are dictated by standard practice but still affect disaster response. For example, *"Their pipes have lifespans, roads have lifespans. There are specific formulas for calculating when, in a certain time, you have to repair a pipe to be up to code or repair a road to be up to code. And so there's certain things where you just like you have to. You have to make that decision to invest in that that year, because that year that pipe failed or so, or you know it has hit its expiration date."*(P11).

4.3.3.2 Money, power and authority constrain decision-making

Funding for climate resilience projects is not guaranteed. For decisions to be made (for example, replacing stormwater pipes, paving), it needs money. Some of this money is maintenance and allotted yearly. However, other has to be applied for. It has to be applied to Congress, and Congress has certain priorities. Thus, funding drives decisions more than risk or need. Money availability influences what decisions are made, more than priorities or needs. For example, less important projects can get funded because their budget exactly fits what money is available:

> *"If there's a pot of money that just happens to become available, you might get funded for a project that's not a priority but is within the budget available at the time. So sometimes decisions about what infrastructure to invest in are just made because of what's out there in terms of total pots of money, and using as much of that money up as you can. Money that just becomes quickly available and becomes quickly not available. And you need to be paying attention." – (P11)*

Thus, stakeholders have to be aware of what funding is available in order to make decisions. Not all money is the same – maintenance funding is fixed, disaster funding is not. For yearly maintenance needs of the installation, there are dedicated pots of money. However, *"the resilience funding is extra funding that's obtained from either congressionally appropriated funds or higher levels of the military. Where you have to get it, you have to apply for this funding. It's not your year-to-year operating budget. So some of the needs for stormwater management and infrastructure are just met by that yearly operating budget year after year, incorporated into plans."*(P11). Thus, funding needed for disaster resilience projects or big improvements is harder to get and uncertain. Moreover, money, power, and authority dictate what decisions are made. Institutional hierarchy dictates how much control planners and engineers have in the overall decision-making:

> *"The military is a very regimented, controlled environment; there's a strict chain of command. So there's a shared understanding about how things are done, and limits on what you can do. People operate within those constraints and don't push back." – (P11)*

Authority overrides what decisions are made. For example, if the senior commander wants a coffee shop, then that's what gets funded. Finally, there are some "squeaky wheel" decisions. There are some people who make a lot of noise, and those decisions end up happening:

> *"people who scream a little louder and have a little more money will get the investment as opposed to the people who don't have the power or voice. And so ... some decisions are squeaky wheel decisions, who's making more noise? And that's informing the decisions." – (P11)*

To deal with this, stakeholders need to realign priorities to the senior commander and Congress. Due to the military culture and congressionally appropriated funding, stakeholders need to *"align their decision priorities as best they can with the priorities of the senior commander and the funding goals of the Congress, then make decisions based on that"*(P11). For example, *"For an engineer within the Department of Public Works, for example, they report up to their head of the Department of Public Works. He takes a look at the list, and he modifies it to modifies it, to align with the priorities of the installation at the time as he understands them."*

4.3.3.3 Better accuracy does not mean better decisions

Tools are powerless in the face of institutional constraints. The data provided by tools can be used to provide support for disaster risk management projects, *"data [from tools] could be used*

by people to argue for funding for projects such as the retention basin, or something that's a much higher cost project. But again, they might not be funded if the senior commander wants a coffee shop, which is a true and actual thing that's happening."(P11). Organizational trust is important for credibility and tool adoption. The military relies on trusted relationships with actors and organizations for their disaster risk management needs. For example, the military *"rely a lot on trusted consultants that they've used for a long time."*(P11). The military *"brings in consultants for a particular purpose, which is very well spelled out in a contract with a very specific scope of work and deliverables."*(P16,external). Thus, the military generally chooses to hire familiar consultants, and trust relationships with such actors are built slowly over time. Organizational reputation, such as with national labs, can also facilitate trust. *"[National] Laboratories are very active in the climate risk space, and receive a lot of funding from like DoD, like CERTA, BSTCP, for example. So they're on a lot of those efforts... so that's a trusted relationship."*(P16,external). Thus, the military regularly provides resources, including funding, to such organizations to partner with them. In contrast to these, research teams operate outside these familiar trust mechanisms, and may not be trusted:

> *"The consultants are meeting an immediate need. You're doing basic research, which is kind of a funny thing in this applied space. Right?" – (P16)*

Research teams who do basic research do not meet a clear need, and may not be trusted. This trust also translates to the tools developed by research teams:

> *"I think the biggest resistance we run into is if we propose a technology that that the Army is not familiar with, if you come to them with something that they don't understand, or it's not being done elsewhere, or the Corps of Engineers, or Pacific Northwest laboratories, or all these different other organizations that are doing work for the Army - they haven't seen it or heard about it. Know about it. They're not so, so receptive to it, or they just haven't proven the model right? You know, that's probably the highest level of resistance." – participant from ethnography study*

Thus, there may be institutional resistance to the adoption of new tools if they are unfamiliar. From a psychological perspective, this makes sense:

> *" We all know it's easier to, you know. Sell a box of matches than a spaceship, right? If they don't understand what the spaceship does, you know, and they understand what matches do. They trust groups such as national laboratories to provide them with some new technology, some new solutions, you know, labs*

that they already have a relationship with. They feel labs that already understand them, that speak their language, that understand their needs." – (P06)

. Tools are not trusted by entities if they are very unfamiliar. Therefore, stakeholder representatives thought it necessary to determine how the solutions developed by the research team fit into greater organizational efforts: *"How does this work fit into the projects going on at SEMCOG (Southeast Michigan Council of Governments) and in the greater region? What other universities are doing work similar to ours, and can we collaborate to create a more robust product? What other laboratories (Idaho National Lab / USACE / PNWL) have created or are creating a similar product? How do we demonstrate this can be applicable to non-DoD entities (FEMA / Emergency Management)?*(P14). SEMCOG is a voluntary association of local governments dedicated to fostering regional cooperation and advancing Southeast Michigan. In response, researchers believed they needed better internal coherence and organizational structure similar to a company. One researcher states:

"We should form a company [laughs] ... yeah so we can work like a company. There are some leaders, we have a lot of assignments, and everyone has a responsibility for some specific work. We can have different groups so we can form the idea together and work together. And they must have some input and output to another group. And then we can combine together instead of mostly focusing on the research only. We can work like a company, and then we can have a product. And when we can sell them, I think this could be the best way. " – (P04)

Thus, organizational trust determines tool adoption, and researchers need to adapt to that. Tools currently used do not align with information and decision needs. Stakeholders use tools primarily for two purposes: *"gathering data and making specific engineering decisions"*(Anthropologist, W1). Despite the DoD having its own proprietary tools available, stakeholders prefer using publicly available, open-source tools as *"the open source tools are easier to access and load faster in the internal network versus the DoD tools. That's a big, big reason why."*(P16, W2). They use publicly available tools for weather and climate information, such as NOAA and weather.com. They also use publicly available engineering tools, such as FEMA flood mapping tools. Stakeholders also use tools such as PAVER to make decisions about the infrastructure, such as *"for knowing when you have to repave your roads and parking lots, for installing and replacing pipes."*(Anthropologist, W2). However, there is extensive prior work in disaster risk management which recognizes:

"There's a difference between the tools that are being provided versus what people actually care about. And that's because of the history of weather risk information

tools. A lot of these tools come from a very specific purpose, come from insurance, and have traveled in more of the public sector. And that's shaping how people in the public sector manage disasters." – (Disaster Risk Engineer, W1)

Therefore, the tools that stakeholders use do not necessarily provide them the right information they need to manage disasters and their risk. *"But then the decisions are being informed by people at higher levels of the chain of command. The senior commander right now wants a coffee shop, so the money isn't going to be going to improving stormwater infrastructure unless somebody dies in an accident, for example, on the base; then priorities shift quickly, but they also shift away quickly."*(P11)

4.3.3.4 Resistance and Skepticism

Stakeholders' skepticism towards the benefits of research. Modelers held the belief that users need to adapt their practices to the information given by the models:

" I think we just keep saying, this is the information we would provide to the users on the ground. This is the way we think is the best way to prevent loss of lives. This is the way in which you know emergency managers can benefit from the work that's being proposed. " – (P01)

Thus, scientists held the belief that scientific information and its accuracy lead to better decisions. As a check, they proposed additional data-gathering mechanisms from stakeholders, such as *"creating a real-time feedback system that can receive feedback from users about flood status."*(P04). While researchers wanted to increase flood resilience on the base, *"stakeholders [already] know where flooding is on the base."*(P11), and did not need additional knowledge to tell them that. When engaging with the community, ethnographers had the perception that stakeholders did not really need the research team, and they were competent in their jobs and operating within the space:

"The general feeling is that they don't really feel that they need our knowledge. They are also competent in their jobs. They know how to operate within that space, and we don't. We need to consider that as we continue to engage with these stakeholders." – (P11)

The above quote directly contrasts with the "scientists know best" philosophy of the modeling team, where they treated the models as ultimately the best resource to guide information and stakeholders would benefit from them. An ethnographer mentioned:

> *"People are not convinced really about what we're doing and how that might benefit them. So they don't really see us as very interesting. They don't need us ... [or] perceive that this project was needed necessarily. And [despite] having made a lot of [internal] progress... they have a lot of skepticism about what we're doing." – (P12)*

Towards this contrast, people commented how it was a *"very humbling but an important thing to know how they perceive us, and researchers outside of their structure more broadly"*(P06). Stakeholders were also therefore not convinced what the research team was doing, and how it would benefit them eventually. Stakeholders were engaged with ethnographers, but not the rest of the research team: *"[the stakeholders] kind of know what [the ethnographers] are doing, but they don't really know what the rest of the team is doing."*(P12). Despite this, stakeholders expected to see tangible things in exchange for providing their time towards the project:

> *"You introduce a project to someone, and then they're like, "I'm talking to you, you're clearly getting money to do this. Okay, well, what's the bottom line? What are you gonna deliver?" If I say the project is about flooding on the installation and about weather, and they want to see these things." – (P11)*

Thus, stakeholders were only engaged with a part of the team (the ethnographers), did not perceive the researchers or the project as necessary, and were skeptical about how the project would benefit them. Researchers interpret it as stakeholders thinking about them as a resource: *"That idea of "I gotta deliver" ... it's less about specific results and more about what will those results be? And I think as much as we're learning about them, they're hoping to learn about us, what kind of resource are we really? And I appreciate them thinking about us that way."*(P06).

Moreover, there are differences in how researchers and stakeholders think. Human-centered design researchers recognized that *"the way that we think about these problems... it's not going to be the same as the stakeholders, even those end users that we're designing for"*(P06). While climate scientists and modelers thought about risk in terms of probability, stakeholders did not think of risk in very precise ways:

> *"I don't actually think most of the people I've talked to are thinking so like precisely about uncertainty and quantifying it. There are these biases that affect decision making... and there may be people who will not even consider uncertainty because [they think] it's not real. The storm is gonna come because someone said it is, or someone decided it was gonna be, or you know. the leaves are turned this direction today." – (P11)*

Thus, stakeholders may not even consider climate risk due to cognitive biases or beliefs such as uncertainty is not real. However, stakeholders are not a homogeneous group, *"we're working with a whole community of people on this installation, [which] doesn't all think alike"*(P11). From an anthropological perspective:

> *"From an anthropological understanding, risk is not a probability in isolation, but it's a range of possible outcomes as they affect a decision-maker's well-being or fitness. The same statistical distribution can be "risky" for one individual, but not to another depending on their situation." – (P12)*

Thus, the same "probability distribution" can mean different things to different stakeholders depending on factors such as their own risk threshold or risk tolerance. The following quote from an HCI designer summarizes how different stakeholders think about risk:

> *"From the ethnography excerpts, I learned that stakeholders perceive risk through their lived experiences, responsibilities, and institutional priorities. Local emergency managers and officials focus on immediate safety and operational concerns, such as protecting residents, keeping infrastructure functional, and avoiding liability. Residents perceive risk more personally, based on their memories of past floods, property damage, and the emotional stress of displacement. Engineers and planners often see risk probabilistically, through flood maps, return periods, and long-term mitigation strategies. Policy and funding agencies frame risk in terms of regulatory compliance, political feasibility, and eligibility for grants or disaster recovery programs." – (P09)*

Upon engaging with ethnographic data, researchers articulated the difference between how they think of risk vs how stakeholders think:

> *" Stakeholders often associate risk with the perceived likelihood of project approval, successful completion, or access to funding. In contrast, I view risk more in terms of the vulnerability of an area and the likelihood and severity of damage from a hazard. While this kind of physical or environmental risk is sometimes acknowledged, it's mostly captured under "information/data" uncertainty. Additionally, many stakeholders frame risk as a local concern rather than a regional one. "I can't hold back flow because there's a chance I'll flood my own community if I do."(433-434) This kind of local-first thinking can lead to decision-making that unintentionally puts other areas at higher risk. To mitigate this, risk should be considered at a broader, regional scale, where cooperative action can reduce disproportionate harm across districts. " – (P08)*

4.3.4 Rethinking Design

While the first three themes focused on friction and gaps, this theme focuses on how to design with epistemic pluralism.

4.3.4.1 Questioning Assumptions

To design based on epistemic convergence, it is important to think as a team and question assumptions with critical dialogue. One way to move towards this is to *"identify the small bite-sized [pieces] that we can actually focus on"*(P06) and *"progress ... as a team towards ultimate goals."*(P11). Ethnographers pointed to the *"need to align on some more specific definitions"*(P11) before they can speak of stakeholders and their needs. Outcomes and *"preliminary findings [across] teams ... need to be shared and integrated in a more specific and structured way"*(P11) rather than proceeding alone. Therefore, disciplinary silos first need to be broken in order to effectively think together as a team rather than proceed alone. As it was more comfortable and "path of least resistance" to work solo in interdisciplinary teams, researchers pointed to the need to *"force the interaction [across] different groups ... on a regular basis"*(P02), akin to *"remix of the different [teams] by forcing different angles of cross interests [to] get things rolling."*(P07). Others pointed to making the connections that already exist across teams more explicit:

> *"We need to make the connections between teams [more explicit, and define] what that connection means. For example, when the visualization team is presenting, it needs to be clear where the data is coming from. So clarifying those connections can give us a starting point to make a demo. " – (P09)*

Thus, overall, it was necessary to step outside disciplinary silos and think together as a team.

> *"Learning about these barriers, these different ways in which we ... and our stakeholders are thinking ... is how we break down these barriers, right?" – (P06)*

One of the things that emerged from the workshops was whether we are asking the right questions. How we think of risk, what problems we choose to tackle, and how we propose solutions stem from our disciplinary training. And despite a multidisciplinary approach, we may still choose to tackle the wrong questions. An external participant wonders, *"Is better communication of scientific uncertainty that comes out of these physical models actually addressing the real issues that a lot of the stakeholders are facing?"*(P15). Others acknowledge that *"we may actually have the wrong questions."*(P06). An ethnographer summarizes:

> *"Despite so many advancements and decades of research done on natural hazards ... there are still more deaths, more damage, more vulnerable people. We need to consider as a field if we are asking the right questions that are leading to solutions. It may not be true." – (P11)*

Thus, it is critical to examine if we are asking the right questions and proposing solutions that are truly helpful for stakeholders.

Another way is to rethink the time relevance of design by developing models relevant now and in the future. Just because stakeholders know their problems currently doesn't mean that the problems are going to be the same in the future. Thus, some researchers believe that there needs to be a balance between both current relevance and future-looking, and we need to be building tools that achieve this balance:

> *"The future is going to be different than the present is, and then they're constantly changing that system. So I think we can develop models and produce data that is relevant in the present and will also be relevant in the future. We need to recognize that they know their problems, but that doesn't mean there isn't space for us. Translating [that] to solutions is a bit harder." – (P11)*

Thus, tools have to be simultaneously relevant in the present, but also robust enough to handle the rapidly rising future of climate change.

4.3.4.2 Designing for Communities

We need to design for the community as a whole, and not just for individual stakeholders. While the ethnographic team was engaged with the broader community, only a few specific stakeholder representatives were in constant contact with the team developing the designs. Ethnographers thus cautioned the need to design for the whole community, not just the representatives who are more involved: *"as we design these tools, it's very important that we bring the voices of all these different people together, not just the voice of a key informant that we have easy access to."*(P11)

One way to bridge the gap between basic and applied science is with knowledge co-production. To resolve the differences between basic and applied science, researchers looked to the concept of knowledge co-production. Ethnographers mentioned how this is what the field is moving towards:

> *"I was at this [national] natural hazards workshop, and the field is moving in the direction of co-production ... changing the way people think about predictability and uncertainty. This move towards co-production and trying to work with the*

people who are on the ground facing these hazards on a day-to-day basis is one way forward that maybe can address this gap." –(P11)

The natural hazards field is increasingly recognizing the importance of knowledge co-production and moving towards it. In co-production, researchers prioritize *"engaging with the stakeholder and stakeholder needs from the get-go "* (P11), and *"the stakeholders work with researchers"*(P02) in a collaborative way. Such a method does not assume a linear relationship between basic and applied science, but knowledge is produced by *"translating research into practice and practice into research"*(P11). Researchers also saw this as an opportunity to get stakeholders on board with their approach:

"We need to involve different stakeholders in developing a demo [as] a good way to meet them halfway. Because if they're involved in developing the demo, then they have a say in what it looks like and what it does. And they also may be then more invested in being an advocate for the approach that we're taking, which I think is really the broader goal." – (P12)

Adding to this, the ethnographer also remembered how they conducted online knowledge co-production workshops with other communities of stakeholders, such as reindeer herders:

" [during] COVID-19, all travel was shut down. We ended up doing the workshop online, which created some difficulties, but also some new opportunities. We came up with research threads by mixing [different] teams together. We broke up into different mixed-composition groups, presented to each other and discussed, and then reported back to the whole team. And then we broke down again on different themes. We did it over the period of a week. We [also] did something like that in Finland with reindeer herders. So I think that it works pretty well. " – (P12)

Thus, instead of the "hand-off" process or relying on traditional definitions of collaboration, the team looked towards knowledge co-production as a way to bridge the gap between basic and applied science by being more involved with stakeholders.

It is also important to engage with stakeholders as a team. While so far engagement with stakeholders was primarily done by ethnographers, they also felt the burden of being a primary communication channel between the research team and stakeholders. It was necessary to *"engage with stakeholders as a team"*(P11) rather than only one discipline. Ethnographers believed that having *"a better understanding of the society that [stakeholders] are operating in, and the factors influencing their decision making ... [makes the team] well positioned to engage with stakeholders"*(P11). However, researchers should not assume authority but be *"very respectful of the military culture and subjects, respecting of the chain*

of command"(P11). The team also needed to be on the same page so ethnographers can report to stakeholders with one voice:

> *"As I speak with people and represent our projects in different settings, it would certainly be useful for me to have more information about where every team is in terms of their process and work towards goals ... [to avoid] failure in communication [with stakeholders]. [It is important to] get everybody on the same page perhaps with more structured and comprehensive presentation of findings." – (P11)*

Researchers who did not have disciplinary training for stakeholder engagement were also learning how to do it on the job: *"this is also [a] part of the research that we are learning, how to better engage [stakeholders]. "*(P07). Thus, rather than the burden of stakeholder engagement falling only on one part of the team, it needs to be more equitably distributed.

One way to better connect with stakeholders in the community is by creating personas to bridge barriers. Currently, the team is encoding stakeholder requirements into design parameters such as spatio-temporal resolution, etc. However, another way is to change design thinking by creating personas. Personas are a very popular tool that makes hypothetical stakeholders more realistic, and this can be used as a boundary object between different teams:

> *"Thinking about using like an archetype, like a persona of a decision maker or specific role as a way to demonstrate what we're doing. That could be a starting point. For example, pick one role [and] think, what if I was the person that has to tell people, "hey, don't come into work today. It's gonna be flooded on the Interstate". What would they need to make that decision?" – (P03)*

Despite not having a formal design background, P03 recognized the importance of personas to facilitate thinking about what information different people need to make their decisions. Having an idea of personas would be helpful to know *"what would it be like if I were this role, and how would I be making each one of these decisions"*(P08). Researchers outlined the reasons why they wanted personas:

> *"If we understand exactly what they are currently doing, that might make it easier for us to shape what research we're doing, what goals we're going for. So it would be useful to get information about their current workflow, how a decision maker does their processes in their day-to-day work, how they use these different apps [and tools such as] PAVER, and how do these all integrate together? How do they use these to make decisions, and like what gaps are in their current workflow,*

what things would be easy to replace, that would be well accepted? Or can we supplement or add on to?' – (P08)

4.3.4.3 Mapping Tool to Organization

There needs to be better alignment of the information we deliver to stakeholders to military culture and hierarchy. Rather than scientific information and uncertainty communication disrupting the status quo, it needs to be highly contextualized to the military hierarchy, priorities, and culture. A stakeholder representative illustrates this:

> *"In addition to the basic science uncertainty, we really have to consider [people's] behaviors or biases. There are elected officials who [believe] that people are controlling the weather. I'm not sure how we approach somebody whose attitude is "Jesus take the wheel" ... in a way that doesn't seem confrontational. But you can't ignore it, right?" – (P14)*

Here, a stakeholder representative mentions that information needs to be delivered to officials in a way that does not seem confrontational, maintains hierarchy, and does not challenge their worldviews and beliefs even if it is wrong. The priorities and language of the military adapts to the priorities of the existing administration. This also determines how stakeholders think about certain subjects, such as climate change. An ethnographer recounted how she observed a switch in the language pre- and post-administration change:

> *"Looking at the quotes, [people] use different language [when] having to communicate about extreme weather from pre-administration change and post-administration change. Pre-administration change, [people] were saying, "Oh, yeah, everybody on board understands that climate change is an issue, and there are a lot of efforts to move forward on this front". Post administration change, I haven't heard anything of the sort; [people are] a lot less vocal about that." – (P11)*

Thus, the military context makes it necessary to consider sociopolitical dimensions of uncertainty, especially when communicating results to stakeholders. Thus, it is necessary to *"create a product that can be fluid enough to align with changing administration priorities"*(P14). There are various factors that affect how stakeholders think of uncertainty:

> *"They are doing their job 9 to 5, and maybe not thinking about it [uncertainty] too much outside of the job. There are different types of people at the installation with different belief systems that will affect the way they think about the*

information that we present them and the way they think about disasters. So, for example, belief or not in climate change, support or not for the current administration. There are career military who've been very steeped in the military culture for a long time. The focus on the mission is gonna determine some of the decisions that they make. So goals. Values are kind of shaped by the mission of the installation. I think risk sensitivity in the current environment is something we need to consider as people propose ideas to higher up the chain of command, perhaps, or respond to orders from higher up the chain of command. And then, again, going back to the schools and values as they share knowledge and information with other stakeholders, understanding again the context that they're operating in right now, and what they're having to communicate in order to get funding for the specific sorts of projects they want to do and how they're having to communicate that." – (P11)

Since there are various different stakeholders in the chain of command, we need to make sure that the information is delivered to the people on the ground, so that they can use it to make decisions: *"Some of them might be getting this weather information lower down the chain of command. [We need to think about] how do you get it to the soldier on the base more effectively, or to the commanding officer, or to someone who can use it or apply it in a more effective way."*(P11). Moreover, it is important to *"have a really tight package when you're proposing a project up to your supervisor, and then, as they pass it up the chain of command to get funding for it, and that means having really good documentation, having really accurate reporting."*(P11)

To do so, the tools need to be mapped to the organizational hierarchy. Researchers mentioned how the information needs to be tailored to the political context and different levels in the organization. However, researchers recognized that different knowledge generation approaches (basic vs applied science) can be mapped to different levels in the organization:

"I'm a little concerned about how we can appeal to an organization that has people at the high level, and at the lowest level who are making the decisions and disseminating the information. It seems like if you're going to argue for basic research, then that has to be at a pretty high level. But if you're talking to a forecaster or somebody who's making immediate risk assessment decisions. They may need more of the tool in front of them." – (P01)

Here, the participant speculates how to appeal to the organization as a whole with different approaches. They state that higher-level organizational management might advocate for a

basic research approach, but decision-makers on the ground might need direct and actionable information to make decisions.

Tools also need to align with existing structures in place. The funding agency thought it was important for the solution the team proposes to fit into existing systems already in place, so stakeholders can use it immediately: *"It is important to align whatever you want to change or do, or a decision that needs to be made ... with the systems that are already in place. Like, with whatever requirements the facility already has on it for planning, maintenance, and codes. That really helps when you can offer them something immediate that they can use for something that's being required of them, anyway. Right?"*(P16, External). This sentiment of the funding agency also aligned with what ethnographers heard from stakeholders: *"I'm seeing direct support for this just in the things that people are saying. And this is a hurdle for us to get over if we try to propose new tech or new tools. How does it fit into existing structures? And we have a better sense of those structures now, I think, than we did at the beginning of the project."*(P11)

We need to rethink the purpose of these tools. Tools can be designed to communicate information in a way that resonates with communities and supports advocacy for problems that they want to solve. Ethnographers have mentioned how stakeholders are already aware of where the problems are at the base. The main bottleneck is not lack of knowledge, but funding and political constraints. Thus, tools can specifically target these constraints. For example:

> *" Our data could be used by people to argue for these, you know, for funding for projects such as the, you know, like a retention basin, or something that's a much higher cost project. " – (P11)*

Thus, tools do not need to tell people where the problems are; they already know that. Instead, tools should fit into existing workflows and culture. For example, using the data generated by tools to apply for congressionally allocated funding.

4.4 Discussion

4.4.1 RQ1: How do multidisciplinary team dynamics affect design implications for climate resilience solutions?

While in HCI it is widely accepted that design is an iterative, highly non-linear process, this may not be the case on interdisciplinary teams, where multiple disciplines work together to produce knowledge. In the absence of constant co-ordination and negotiation of boundaries,

it is easy for design to be more linear than iterative, starting from knowledge production (Section 4.3.1), translation of complex science into understandable terms (Section 4.3.2) and finally dissemination to end-users and stakeholders (Section 4.3.3). Here, we unpack how disciplinary dynamics affect each of these points in the design process.

4.4.1.1 Knowledge production

From team science, we know that STEM disciplines tend to dominate over social sciences in collaborative research. This affects the design process – while epistemic pluralism is valued normatively, what gets designed is influenced by dominant disciplines. For example, the human-centered design process starts with stakeholder identification, understanding needs, pain points, and current use of technology, and designing solutions to remedy these. In contrast, "design" in physics-oriented disciplines starts with collecting data of atmospheric phenomena, quantifying it, and making inferences to uncover patterns and trends. While both approaches have significant startup costs—to collect data and establish stakeholder relationships —both are met with challenges. Data collection requires going into the field, taking photographs and placing sensors. Establishing relationships with stakeholders is often fraught with challenges such as establishing trust and gaining access. While challenges exist in both approaches, the start-up cost is different and often unequal. For example, unexpected political challenges can impact trust in stakeholder relationships, resulting in challenges when interviewing people. Practices can have overlap, such as when gaining stakeholder access can benefit modeling by asking stakeholders about their preferences. However, this burden is easy to default to disciplines who are trained in establishing stakeholder relationships, such as anthropologists.

Because of the belief that "scientists know best", the well-documented dynamics of STEM disciplines dominating social sciences also impact what gets designed. This can lead to divides in choosing between two approaches—basic science vs. applied research. Thus, the design process first starts by acknowledging that modeling is important and often a better paradigm than stakeholders' current way of doing things. Climate services are then created by scientists determining what is important to design, what modeling approaches to take, and how to still incorporate stakeholder needs in modeling by way of making parameter and threshold choices.

4.4.1.2 Translating Complex Science into Simpler Representations

After knowledge is produced first, it is then translated for "downstream" stakeholders and use cases. For example, modeling can start by capturing how flood water levels are impacted

by rainfall in a given region and then explained to stakeholders in easy-to-understand and intuitive terms such as through charts, graphs, and maps. Thus, the information presented remains the same at its core (and is determined by scientists), while the presentation changes for public understanding. Here, the underlying assumption is that if things are easy to understand, then they are easy to apply and use to make decisions. However, things fall apart as stakeholders acknowledge they have much more information than what is useful. Understandability does not imply utility.

4.4.1.3 Objective tools in subjective contexts

Whether or not information is useful depends on who is using it, what they are using it for, and what the end goal is. For example, an engineer using flood water level maps can have the end goal of maintaining systems. In contrast, a planning manager can use the same information for determining whether to convert a parking lot into green space for better water absorption. Therefore, communicating water levels in a way that is easy to understand does not immediately allow stakeholders to connect them to relevant decisions, and they have to do some mental calculations to make the information more usable.

Stakeholders' information needs might be completely independent from what scientists assume is important. Stakeholders operate in an institutional culture with well-established norms and values. There may be constraining factors out of stakeholders' control, such as money, politics, and institutional hierarchy, that hinder the usability of information. Therefore, better model accuracy does not lead to "better" information usability. This also indicates a need to think beyond the single stakeholder who is using the climate service. For example, our empirical findings indicated that even though the information may be actionable for the immediate stakeholder, if "the commander wants a coffee shop", there is very little that the stakeholder can do to effectively leverage this information to make decisions. Therefore, tools need to be mapped onto the organization and information flows need to be carefully considered.

4.4.1.4 Designing with Epistemic Pluralism

Disciplinary hierarchies lead to the design of climate services with a physics-first orientation even when stakeholders are involved. We need to move to models of knowledge co-production, and question how we can change our own ways of thinking by being flexible to multiple different perspectives and interpretations.

4.4.2 RQ2: What are the design implications that remain hidden or under-explored due to team dynamics?

Even on interdisciplinary teams, a physics-first orientation seeps through and impacts the design process. Epistemic pluralism needs to not only be normatively valued, but integrated into design thinking. This can be done by (Section 4.3.4) questioning the assumptions we implicitly make when determining what to design, designing for actionability rather than understandability as a proxy, and moving beyond a single stakeholder to considering organizations when designing climate services. Here, we unpack each of these design implications:

4.4.2.1 DI1: Interrogate whose epistemology is encoded in design

Scientists and stakeholders do not think in the same way. Scientists think of risk and uncertainty in the form of numbers and statistics, but users experience uncertainty through their lived experiences, prior knowledge, and mental models. Being scientists, we apply methods to produce knowledge. But once the knowledge is produced, we also determine how to present that to stakeholders. Here, the hidden assumption is that authority to determine what gets designed is an extension of authority to produce knowledge. We communicate in ways that we are used to looking at things (in the form of charts, numbers). And then we simplify them for people to understand. But we need to examine what the purpose of design is beyond simply information presentation. *Is the goal of design to present all possible information? To present information with scientific fidelity? Or to present information that can help users make a very concrete decision?*

4.4.2.2 DI2: Accuracy is not equal to decision actionability

Accurate tools are powerless in the face of political constraints, and therefore accuracy does not lead to better decisions. For climate services to be usable, actionability needs to be defined and grounded in the decision context. Tools can be accurate but can tell users what they already know. A stakeholder representative stated that they have more information than they know what to do with. However, despite this availability of information, stakeholders still struggle to make decisions due to constraints. Therefore, tools can be designed for what stakeholders need to do with the information (e.g., advocate for funding a disaster management project) rather than reiterating what stakeholders already know (e.g., the risk of disaster being high in an area).

4.4.2.3 DI3: Design should not assume rational information processing

What uncertainty means and where it arises from has to be clearly explained. Assumptions made need to be clearly articulated and cross-checked by engaging with diverse perspectives. Stakeholders' sense-making needs to be a key consideration in design. Uncertainty communication should be universally relatable, and tools need to be flexible so that they are understood by different organizational stakeholders rather than a single hypothetical user.

4.4.2.4 DI4: Consider flows and barriers in making information usable

Sometimes it's not a lack of knowledge, but barriers to making the knowledge useful to inform decisions. These barriers need to be mapped and designed for . Designers need to consider organizational information flows, such as where the information gets delivered and who it gets passed on to.

4.4.2.5 DI5: Design uncertainty communication for how it's going to be consumed, not how it's quantified

End users don't consider uncertainty in the same way that scientists do. Anthropologist's insight: people don't think about uncertainty in daily life (not in the same way).

4.4.2.6 DI6: Invite speculation through design

. Design should make plausible futures visible. Design of climate services needs to inform the community not just the current state of things, but also how things will change in the future. These designs need to be speculative and invite stakeholders to think about their implications for decision-making. Stakeholders think about past events in the context of uncertainty.

4.5 Summary

Taken together, these three implications reframe the design problem for uncertainty communication. DI1 suggests that the form of uncertainty communication must be anchored to decision context, not to statistical constructs. DI2 suggests that individual comprehension is an insufficient target, and that uncertainty communication should support the social and institutional processes through which decisions are made. DI3 suggests that uncertainty communication should convey the range of possible futures and contingencies, not just a

snapshot of current risk. These implications collectively motivate the shift from graphical to other forms of uncertainty that resonate with decision-makers, including stories and narratives.

CHAPTER 5

Understanding Uncertainty in Human-AI Decision-making

5.1 Introduction

Decision Support Systems (DSS)

aim to reduce decision-makers' cognitive burden, their decision time, and overall errors in decision outcomes [142]. Recently, ML-based DSS [174] have seen deployment in high-stakes decision-making scenarios, such as production [84], policy [104], welfare [67], and healthcare [194]. Such DSS often aim to support lay decision-makers—decision-makers who are not math-savvy computer scientists or data scientists. Lay decision-makers include end-users with domain expertise (e.g., policy makers, judges, clinicians) and without

(e.g., an end-user inspecting housing price trends to determine the valuation of a house they want to purchase).

However, the lack of transparency into ML-based DSS and their capabilities and limitations could give a false sense of reliability to decision-makers [15]. For example, despite learning from inherently uncertain knowledge of the world [158], existing ML-based DSS are not transparent about uncertainty in their predictions [26]. Suppressing uncertainty information wrongfully paints ML-based DSS as infallible, "superhuman" arbiters [45] that provide absolute judgement without the possibility of doubt [198, 155].

Additionally, suppressing uncertainty grossly oversimplifies the complex, dynamic and ambiguous nature of the world [82], which can lead to over-reliance of decision-makers on inaccurate DSS predictions that may be rooted in harmful historical biases [7, 17, 165, 71, 168, 20].

Thus, the growing demand for AI transparency [64, 212, 43, 130, 124, 85] has led to development of not only uncertainty quantification (UQ) algorithms [79, 153, 218], but also

uncertainty communication mechanisms [131, 73, 114, 81], transparent model reporting practices [176, 159, 8], and calls for embracing uncertainty in the design of interactive computer technology [19, 201, 91, 191, 32]. Yet, except for a limited account of quantified improvements that uncertainty communication could have on decision outcomes [238], there is little knowledge about *how* and *why* lay decision-makers utilize uncertainty when interacting with ML-based DSS.

Here, we conducted a qualitative user study using a think-aloud protocol to understand how lay decision-makers perceive, interpret, and make use of uncertainty when making decisions. In our study, 17 participants took on the role of a decision-maker to help a hypothetical friend decide if the list price of a house matches the true house valuation. The participants interacted with three different DSS to determine the true house valuation: 1) Viz-based DSS, which provided interactive visualization of historical house sales in the region, 2) ML-based DSS that predicted house valuation, but *without* uncertainty information, and 3) UQ-based DSS that made the same predictions, but *with* uncertainty information. All participants experienced varying levels of uncertainty (low, medium, high) when interacting with the UQ-based DSS in randomized order.

We conducted interpretative qualitative analysis on the screen capture and transcripts of each think-aloud with a focus on understanding how participants made decisions. We found that compared to ML-based DSS, the presence of uncertainty information in the UQ-based DSS induced productive skepticism in participants, prompted them to think analytically, and reduced their over-reliance on DSS. However, participants had difficulties understanding the abstract concept of uncertainty, and instead used different heuristics to guide their sensemaking and interpretation.

Our main contribution is new empirical knowledge on how unique characteristics of uncertainty in ML models affect Human-AI interaction. Our work motivates future uncertainty communication designs that inspire decision-makers to think analytically and remain vigilant about algorithmic decision-making. Communicating uncertainty is also related to the concepts of seamful and adversarial design, which have been theorized to promote user agency and support effective Human-AI interaction [115, 128, 25]. Our work provides concrete examples and empirical insights for such concepts. We highlight future research opportunities including creation of on-boarding materials to facilitate understanding of uncertainty, design of user-behaviour-aware uncertainty communication interfaces, and value elicitation of communities affected by uncertainty of ML-based technology.

5.2 Method

Here, we study how lay decision-makers make sense of uncertainty information in ML-assisted Decision Making. We conducted interpretative qualitative work to answer two research questions:

1. *How do decision-makers perceive and reason about uncertainty in their decision-making process?*
2. *How does such reasoning affect their decision outcomes?*

5.2.1 Task

We answered our research questions in a user study with a simple, yet realistic house purchasing decision-making scenario. The participants took on a role of decision-makers who had to help their hypothetical friend, "Alex", determine if the house Alex wanted to purchase was a good deal or a bad deal. The house was a good deal, if the list price of the house (i.e., the house seller's asking price) is less than or equal to the true valuation of the house; and a bad deal, if the list price was more than the true valuation. We used the following prompt to engage the participants in this scenario:

> *Your friend Alex wants to settle down in Ames, Iowa! Alex tells you about their troubles finding a decent house, and you are happy to help them out. Alex plans to move close to their office, and has been looking to buy a house soon. They have found a house they like, and asked your opinion to decide whether they should buy it. Alex tells you the purchase price that the Real Estate Agent provided them. They asked you to help them decide:*
>
> 1. *Whether you recommend Alex to buy the house,*
> 2. *Whether you recommend Alex NOT to buy the house,*
> 3. *If you are unsure and cannot decide what to recommend.*

The participants used three different types of DSS to help their determination: 1) interactive visualization, where decision-makers manually investigated historical house prices, with the help of existing visualization methods, 2) ML-based DSS, where the participants saw outputs of an ML model that predicted true house valuation price, but *without* uncertainty information, and (3) UQ-based DSS, where the participants saw the same ML model predictions, but *with* uncertainty information. The participants interacted with the UQ-based DSS

in three scenarios with three different levels of prediction uncertainty: 1) low, 2) medium, and 3) high uncertainty. Thus, each participant experienced five different conditions.

We randomized the order of conditions for each participant.

The participants then had to decide whether to tell Alex that: 1) the list price is a good deal and recommend them to buy the house, 2) the list price is a bad deal and recommend that they do not buy the house, or 3) tell them they were not confident to make a recommendation either way. We designed this scenario to minimize the effect of proxy tasks on the participants' decision-making process [39] by specifically asking them to make a recommendation (as opposed to judging the decisions of the DSS). Also, asking the participants to make a decision for a "friend" helped us reduce the effect of the participants' own preferences (e.g.,preference for a particular type of house, number of bedrooms, presence or lack of amenities) on their decision.

Thus, to focus on objective house valuation, we created a scenario in which we defer subjective opinions about the house to Alex, who has already decided that the houses we showed to the participant are what Alex wants to buy. We did not show the participant the values of different features of the house (e.g., number of rooms, number of bedrooms, amenities). Instead, the participants only saw the list price and the outputs of the different DSS: 1) historical prices in the interactive visualization, 2) ML prediction in ML-based DSS, and 3) UQ prediction in the UQ-based DSS.

We used the think-aloud method [66], where the participants were asked to verbalize their thoughts and reasoning [101] as they interacted with each of the DSS in randomly assigned order. Unlike existing quantitative ML-assisted decision-making studies [72, 148, 134, 95, 238] that focused on the quantitative measures of decisions outcomes (e.g., in our case that would be how much money Alex saved or lost based on the participant's recommendation), we focused on studying how the participants perceived and reasoned about uncertainty information, and how they interpreted and incorporated that information into their final recommendation. We also sought to identify the effects of uncertainty on different aspects of decision-making, such as participants' reliance on the DSS.

5.2.2 Data Pre-processing and Algorithms

Here, we first describe the dataset and the algorithms we used in our DSS.

5.2.2.1 Dataset

We used the Ames Housing Dataset [50] for our decision-making tasks. We followed open source code [208] on GitHub for data analysis. This included data cleaning, data processing,

and selecting the most influential features for the task of predicting the price of a house in the Ames Dataset. This resulted in a total of 22 features, and 2,616 houses in the dataset. The features included: house age, square footage, basement area, number of bedrooms, number of storeys, type of roof, street length, lot area, land flatness, location, proximity to park or greenbelt, proximity to road or railway, zone, fireplace, size of garage, garage availability, central air conditioning, condition of house, quality of construction, quality of space, and kitchen quality.

We then divided this dataset into train-validation-holdout, with a split of 80%-10%-10% respectively (2,092 training instances; 262 validation; 262 holdout). We used the training data to train the ML- and UQ-based DSS, and the validation data to pick the hyperparameters of the algorithms. We validated the algorithms by combining the training and validation sets, and performed 10-fold cross validation (CV) of our algorithms. We used the "unseen" houses from the holdout set in the decision-making task to represent the houses that Alex wants to buy.

5.2.2.2 ML-based House Price Prediction Algorithm

For simplicity of illustration, but without lack of generality, we used a linear ML model to predict true valuation of a house. Our approach applies to other advanced ML architectures (e.g., deep neural networks [235, 234]). Here, our model took the form:

$$f(x;\beta) = x^T\beta \tag{5.1}$$

where x is the vector of features of a house under consideration (see Sec. 5.2.2.1 for the list of features), β is the vector of unknown model parameters, and $f(x;\beta)$ is the model prediction of the selling price of that house. Using our training set of $m = 2,092$ data points $\mathcal{D}_{\text{tr}} = \{x_i, y_i\}$, $i = 1, \ldots, m$ (where y_i is the true selling price for the ith house) obtained from the Ames Dataset [208], we train our linear ML model by computing its parameters using ordinary least squares (OLS) regression:

$$\hat{\beta} = \arg\min_{\beta} \sum_{i=1}^{m} (y_i - f(x_i;\beta))^2. \tag{5.2}$$

The ML-predicted house valuation price of a house x' is then $\hat{y} = f(x';\hat{\beta})$, which is a single-valued prediction without any uncertainty information. We then evaluated the performance of the ML algorithm using cross-validation ($RMSE = 0.318$).

5.2.2.3 UQ-based House Price Prediction Algorithm

Our Bayesian approach for the treatment of probability and uncertainty closely fits with the culture and incentives of the human-centered design [232, 132]. We selected a Bayesian UQ method because it allowed us to control the level of uncertainty through adjusting the model hyper-parameters, as opposed to frequency-based approaches (e.g., confidence intervals [238]) that would require us to modify the training data. Specifically, we apply the Bayesian ridge regression to the same linear model used in the ML-based prediction. The Bayesian ridge regression provides the Bayesian analogue to the OLS regression solution in the ML-based algorithm above, thereby offering a consistent modeling comparison between the two.

The Bayesian UQ specifically seeks to capture the epistemic uncertainty in the model parameters conditioned on the available evidence in training data $\mathcal{D}_{\text{tr}}$:

$$p(\beta|\mathcal{D}_{\text{tr}}) = \frac{p(\mathcal{D}_{\text{tr}}|\beta)p(\beta)}{p(\mathcal{D}_{\text{tr}})}, \tag{5.3}$$

where $p(\beta)$ is the prior distribution of the model parameters, $p(y|\beta)$ is the likelihood function, $p(\beta|\mathcal{D}_{\text{tr}})$ is the posterior distribution, and $p(\mathcal{D}_{\text{tr}})$ is the Bayesian evidence (marginal likelihood) that acts as a normalization constant to ensure the probability distribution is proper. The prior thus describes the initial uncertainty on β before seeing any training data, and the posterior describes the updated uncertainty after incorporating the training data.

Bayesian ridge regression employs a Gaussian prior $\mathcal{N}(0, \lambda^{-1}\mathbb{I})$ and Gaussian likelihood $\mathcal{N}(x^T\beta, \alpha^{-1})$. We used a hierarchical model, where the hyper-parameters α (precision of data label noise) and λ (prior precision of β uncertainty) are assigned Gamma hyperprior distributions. Each of the two Gamma distributions are further controlled by a shape parameter and rate parameter. The `scikit-learn` implementation finds the numerical approximation of the posterior for this Bayesian ridge regression setup per the algorithm from Appendix A of [210].

To set up scenarios with different levels of prediction uncertainty, we recreated situations of different training data quality. We varied the precision of the noise level of the training data labels through α, while keeping the prior λ constant. In the hierarchical Bayesian setup, α is treated as a random variable (it is not specified directly) and assigned a Gamma hyperprior. Therefore, we vary the shape (α_1) and rate (α_2) parameters of α's Gamma hyperprior to control the entire distribution of α values. Since smaller α values lead to increased uncertainty of the model prediction overall, and larger α values lead to decreased uncertainty, we defined our low uncertainty condition to be $\alpha_1 = 5000$ and $\alpha_2 = 1 \times 10^{-6}$ (i.e., high values of α with mean 5×10^9), medium uncertainty condition with $\alpha_1 = 1 \times 10^{-6}$ and $\alpha_2 = 1 \times 10^{-6}$ (i.e. medium values of α with mean 1), and high uncertainty condition

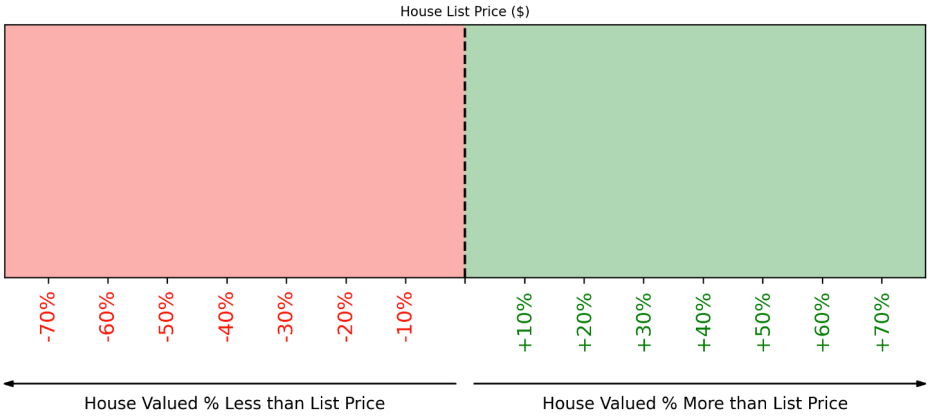


Figure 5.1: Base plot split with dotted black line indicating the list price of Alex's house. Line splits the graph into regions where valuation is less (red region) or more (green region) than the list price. Percentages on the x-axis indicate the numerical quantity by which house valuation differs from the list price.

with $\alpha_1 = 1 \times 10^{-6}$, $\alpha_2 = 200$ (i.e., low values of α with mean 5×10^{-9}). Our cross-validation of the Bayesian algorithm showed that it reached accuracy of $RMSE = 0.344$ for all three scenarios.

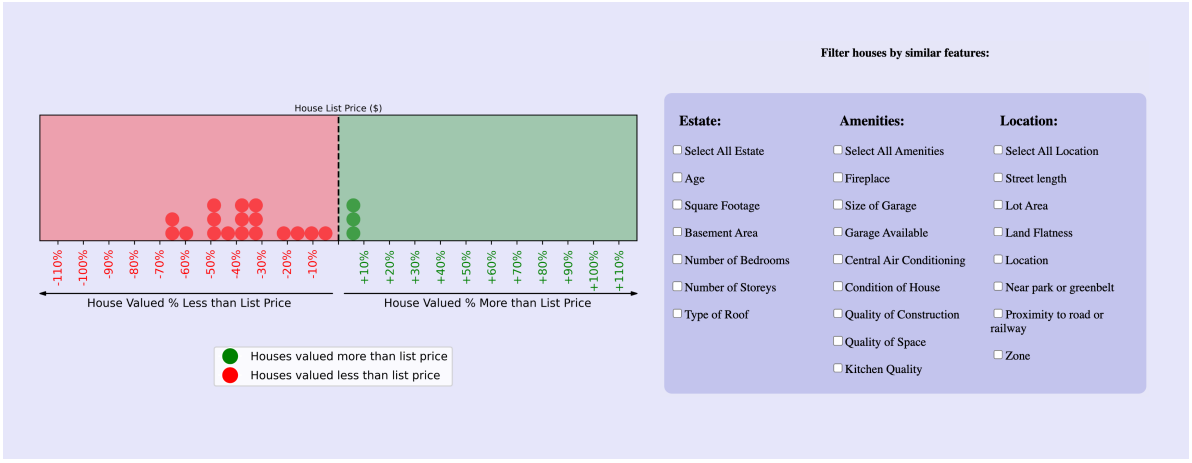
5.2.3 DSS Design Probes

Here, we describe the three DSS design probes that we used in our study. Before we detail each individual DSS, we describe the common foundation that all three DSS shared.

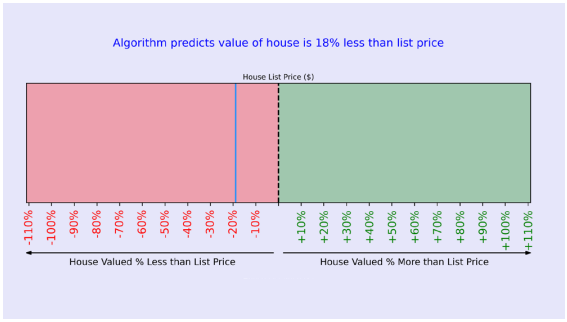
5.2.3.1 Representing Alex's Budget and the House List Price

Each DSS used the same base plot (Fig. 5.1) to display the house list price. The house list price, represented by a dotted black line, split the plot x-axis into two equally-sized house valuation regions. If the true house valuation was on the dotted list price line, that would mean that the house is fairly priced and it is a good deal. If the true house valuation was inside the green-colored region to the right of the list price, that would mean that the house was underpriced (i.e., purchasing the house for the list price would be a good deal). However, if the true house valuation was inside the red-colored region to the left of the list price, that would mean that the house was overpriced (i.e., paying the list price would be a bad deal).

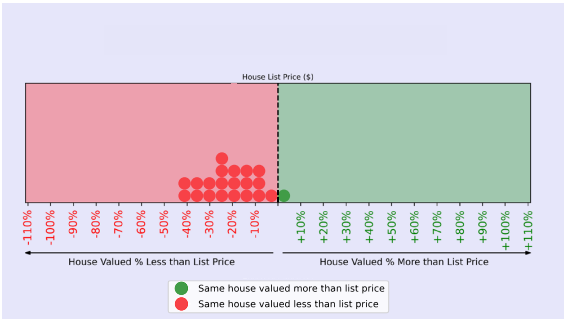
Each DSS then added visual elements on this basic plot to help the participants determine the true valuation of the house plotted on the x-axis. The x-axis displayed house price valuations in percentage relative to list price. We did this because our pilot studies showed that the participants' decisions were swayed by the magnitude of list price when presented with



(a)



(b)



(c)

Figure 5.2: The three DSS design probes: a) Interactive Visualization DSS , b) ML-based DSS, and c) UQ-based DSS.

the actual amount (we refer to this as a "sticker shock" effect). The y-axis was dependent on the type of DSS, which we detail next.

5.2.3.2 Interactive Visualization DSS

The Interactive Visualization DSS (Fig. 5.2a) displayed a dotplot of historical prices of previously sold houses in Ames, Iowa relative to the list price of the house Alex wanted to buy on the x-axis of the base plot. The houses with the same historical prices were stacked; the y-axis of the plot displayed the count of houses corresponding to the price. The dotplot displayed dots corresponding to houses which sold for less than the list price in red, and dots corresponding to houses which sold for amounts equal or more than the list price in green. For clarity, the Interactive Visualization DSS only displayed prices from 20 houses at the time. We randomly sampled those 20 houses from our training and validation data sets.

Participants could further filter houses based on different house features in the dataset,

per these instructions:

> *In this task, you have the chance to examine for what price were houses previously sold in Ames, Iowa. You will first see a graph of previously sold houses in Ames, Iowa. You can find houses similar to Alex's house by selecting Estate, Amenities, or Location-based features. Based on your selection, the graph will update. With the information provided by the graph, you have to decide whether or not you recommend Alex to buy the house at the list price.*

When the participants selected a checkbox corresponding to a house feature, the plot would sample and display 20 new houses from the training and validation sets that had the same selected feature value as the house that Alex is interested in buying. When there were fewer than 20 houses that matched the selected features, the DSS displayed all of those houses. When none of the houses matched the criteria that the participant selected, the DSS displayed the error: "Houses that match criteria not found." This way, the DSS never had to display the actual feature values for any of the houses in line with our design choice that house feature values could add noise to participants' decision-making.

5.2.3.3 ML-based DSS

The ML-based DSS (Fig. 5.2b) used the machine learning algorithm from Section 5.2.2.2 to make a prediction about the true valuation for the house Alex was interested purchasing. The DSS then converted the ML prediction to the percentage of the list price and displayed it as a vertical solid blue line on the x-axis of the base plot. The DSS also indicated the difference between the list price and valuation prediction in text above the plot. If the blue line was in the overpriced, red region, the DSS suggested the list price was a bad deal; otherwise, it indicated a good deal. The DSS listed features the algorithm used to make the recommendation:

> *The tool takes into account different factors such as Age, Square Footage, Basement Area, Number of Bedrooms, Number of Storeys, Type of Roof, Street Length, Lot Area, Land Flatness, Location, Proximity to park or greenbelt, Proximity to road or railway, Zone, Fireplace, Size of Garage, Garage Availability, Central Air Conditioning, Condition of House, Quality of Construction, Quality of space and Kitchen Quality. Based on these factors the algorithm estimates how much a particular house is worth.*

5.2.3.4 UQ-based DSS

The UQ-based DSS (Fig. 5.2c) used the Bayesian UQ algorithm from Section 5.2.2.3 to predict the true valuation for the house Alex was interested in purchasing. We visualized the posterior predictive distribution of the true house valuation (i.e., the uncertainty of the prediction) using quantile dotplots [131]. To generate the dotplots, we modified the quantile dotplot function from the UQ360 Toolkit [81] to sample from the predictive distribution.

The UQ-based DSS displayed its output as dots on the x-axis of the base plot. This is different from the Viz-based DSS because the dots in the UQ-based DSS represented a sample of predicted valuations (and *not* historical house prices). The DSS displayed the dots corresponding to valuations that were less than the list price in red, and those equal or greater than the list price in green. The plot binned, then stacked the dots of similar valuations; the y-axis of the plot displayed the count of valuations in each bin. For clarity, the DSS also only displayed prices from 20 sampled valuations at the time. Figure 5.3 shows example quantile dotplots for the three levels of uncertainty in different regions of the base plot.

The DSS listed features the algorithm used to make the recommendation and provided instructions regarding the dotplots:

> *You will be shown a dotplot, consisting of 20 dots. These dots represent different possible valuations of the same house (that Alex is interested in) made by the decision-support tool. Green dots indicate a good deal. Dots are green if the decision-support tool predicts house value to be higher than the list price. Red dots indicate a bad deal. Dots are red if the decision-support tool predicts house value to be lower than the list price. The more green dots the more likely this house is a good deal; the more red dots the more likely it is a bad deal. With this new information, you have to decide whether the proposed price is reasonable or not, and click on the corresponding button to make your selection.*

5.2.4 Study Software Implementation

We implemented our study software, including the three DSS, as a Django[1] web application running on a dedicated server. We implemented the user interface in HTML, CSS, and JavaScript[2]. We stored the user model, user states, and other user study related data (including logs) in a MySQL[3] database running on a separate server. The participants then

[1] https://www.djangoproject.com/

[2] https://www.w3.org/wiki/The_web_standards_model_-_HTML_CSS_and_JavaScript

[3] https://www.mysql.com/

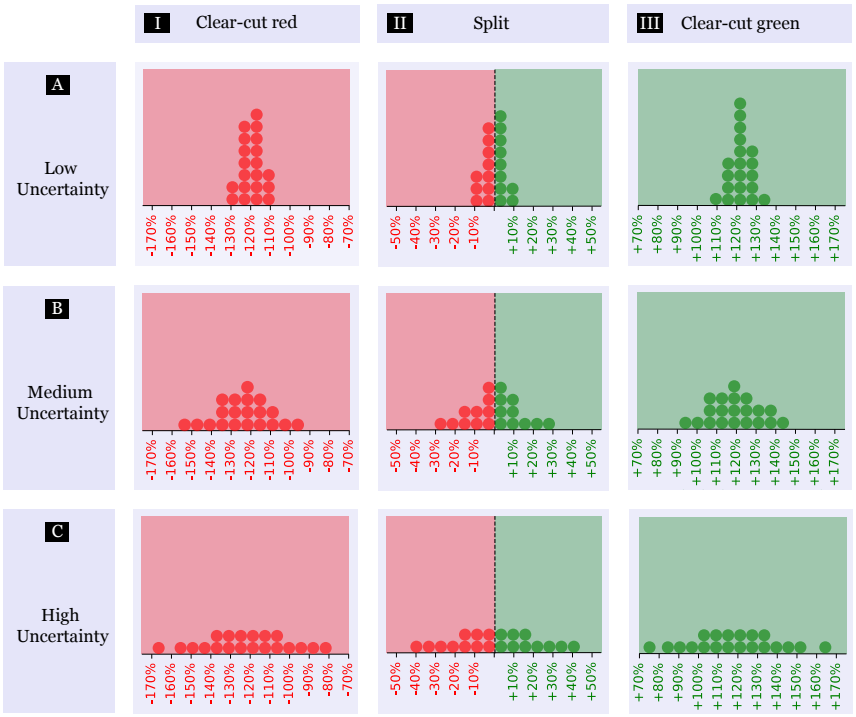


Figure 5.3: Examples of list price posterior predictive distributions generated by the Bayesian algorithm for different houses. Each participant had to make a decision for each level of uncertainty - (A) Low, (B) Medium, or (C) High, assigned in random order (Section 5.2.1). For each level of uncertainty, participants were shown only one randomly selected visualization on the web interface out of three types: (I) clear-cut red, (II) split or (III) clear-cut green.

accessed our study software through their personal computers (e.g., laptops), with a browser of their choice.

5.2.5 Tasks and Procedures

We conducted our user study remotely over videoconferencing software Zoom[4]. After the participant joined the Zoom meeting, the investigator first welcomed them and thanked them for participating. Next, the investigator shared the link to the study website via the Zoom chat, and asked the participant to open the link when ready. The website then displayed a consent form describing the study, expectations, and compensation. The investigator ensured that the participant had reviewed and clearly understood the consent form and clicked on the "I consent" button on the website. Only participants who consented to be part of the study could proceed to the next step.

[4] https://zoom.us/

The investigator then briefed the participant about the think-aloud protocol. After verbally explaining the think-aloud process to the participant, the investigator asked the participant to watch and pay close attention to a short pre-recorded video[5] demonstrating the think-aloud method. This video displayed someone thinking aloud as they engaged in the task of website-exploration. After the video, the investigator asked the participant to share their understanding of the think-aloud process, clarified any misconceptions and answered any questions.

Next, the investigator asked the participant to share their screen and turn off their camera. We audio and screen recorded all sessions for further analysis. The investigator asked the participant to proceed at a natural pace and to start thinking aloud. If the participant fell silent for long periods of time, the investigator prompted them by asking: *"Can you please share what you are thinking right now?"*.

The website then displayed a description of the task and the base plot (Fig. 5.1). After the participant read the description, they proceeded to a quiz, which asked questions about the task and base plot (e.g., *"What does the black line in the graph represent?"*) to ensure that the participant understood both. The participant had to answer all quiz questions correctly to proceed, but could make multiple attempts until correct. There was no time limit for the quiz and participants were allowed to return to previous pages if they wished to see the description again.

After successfully completing the quiz, the participant proceeded to the main task. There, the participant interacted with each of the three DSS, along with the three levels of uncertainty for the UQ-based DSS. Thus, participants had to make decisions for 5 conditions (Sec. 5.2.1), which were randomized in order.

Before interacting with each DSS, the website displayed instructions that were specific to that DSS together with a visual "flyer" of a house that Alex is interested in purchasing. We selected images of visually similar detached houses from a stock photo website[6]. The study website randomized and displayed a different house flyer for each DSS. Note that the flyers did not contain any actual details about the house, and were there to illustrate to the participant that they are dealing with a different house for each DSS.

After completing the main task, the investigator conducted a brief semi-structured interview with the participant. The investigator first asked the participants to give a general description of the tasks they performed. After that, the investigator asked questions based on their notes from the think-aloud.

The participant then completed a brief demographic survey which asked their gender,

[5] https://www.nngroup.com/articles/thinking-aloud-demo-video/
[6] https://www.pexels.com/

age, education, and occupation. To ask about gender, we followed guidelines from [202]. We asked about age using a non-overlapping range. For education and occupation, we provided specific options to select from. To estimate participants' expertise, we asked if they had taken any ML and statistics classes, implemented any ML algorithms, and purchased any property.

At the end of the study, the investigator answered any final questions from the participant. After noting down the payment information, the investigator thanked the participant for their time and the investigator ended the session. The whole study lasted roughly one hour on average. Our organization's Institutional Review Board (IRB) reviewed and deemed the study exempt.

5.2.6 Participant Recruitment

We recruited participants through our organization's participant pool portal, where people can voluntarily indicate interest to participate in research studies. We recruited adult participants who were fluent in English and located in the USA at the time of the study.

We conducted the study remotely because of the COVID-19 pandemic, and requested participants to take part in the study in a quiet setup where they could speak without interruptions.

Nineteen participants took part in our study. We discarded data from two participants due to technical difficulties during the study. We used data collected from the remaining 17 participants in our analysis. We compensated participants $15 per hour for up to two hours. Table 5.1 lists participant demographics from the exit survey.

Table 5.1: Study participant demographics. We removed data from participants P1 and P5 from analysis due to technical difficulties arising during these study sessions.

ID	Gender	Age	Education	Employment	Taken ML class?	Implemented ML algorithm?	Taken stats class?	Purchased property?
P2	Man	35-44	College Degree	Employed	No	No	No	Yes
P3	Man	55-64	College Degree	Retired	Yes	No	Yes	Yes
P4	Woman	25-34	Master's Degree	Employed	No	No	No	Yes
P6	Woman	18-24	College Degree	Part-time	No	No	Yes	No
P7	Man	25-34	Master's Degree	Studying	No	Yes	Yes	No
P8	Man	18-24	Some College	Studying	No	Yes	Yes	No
P9	Demigirl	25-34	Master's Degree	Employed	No	No	No	Yes
P10	Woman	35-44	Master's Degree	Employed	No	No	No	Yes
P11	Woman	18-24	High School	Interning	No	No	Yes	No
P12	Man	65+	College Degree	Employed	Yes	Yes	Yes	No
P13	Woman	25-34	College Degree	Employed	No	No	No	No
P14	Woman	18-24	Some College	Studying	No	No	Yes	No
P15	Man	25-34	College Degree	Unemployed	No	No	Yes	No
P16	Man	55-64	Master's Degree	Retired	No	Yes	No	Yes
P17	Woman	25-34	College Degree	Employed	No	No	Yes	No
P18	Woman	18-24	Master's Degree	Employed	No	Yes	Yes	No
P19	Woman	18-24	College Degree	Employed	No	No	Yes	No

5.2.7 Analysis of Think-aloud Data

To perform interpretative qualitative analysis on the collected data, we first transcribed audio recordings using the *rev*[7] transcription service. We then annotated the transcripts with video details (e.g., details for each condition, any buttons the participants clicked on). Two of the authors independently coded each of the 17 study sessions, and then compared each of the code sets to reach a consensus. We then discussed and refined the codes, before re-coding the sessions together. After reaching consensus, we discussed the interpretations broadly with the study team, and re-coded based on the discussions. We kept detailed records of points of dissent, decisions on merging codes, and code descriptions based on inclusion/exclusion criteria, as well as detailed memos for each participant; both when coding independently and after discussions. We clustered the codes to identify concepts and categories. We performed affinity mapping using a MIRO[8] board to group the codes. We then used the categories to derive the story behind our findings.

5.2.8 Limitations

Although we designed our study with a simple yet realistic housing purchase recommendation task to minimize proxy effects [39], it did not completely eliminate them. Some participants were unfamiliar with the study scenario and doubted their ability to give the right advice. While the scenario avoided effects of subjective opinions in housing choices, participants wanted detailed understanding of Alex's persona to give situated advice.

We asked participants to verbalize their thoughts with a think-aloud protocol, which was difficult to get used to for some of them. To minimize "cold-start" effects, we asked participants to start verbalizing from the page that described task instructions, which made some participants tired at the end.

We used quantile dotplots (Fig. 5.3) to visualize uncertainty for the UQ-based DSS, which introduced a unique structure to reading graphical presentation of uncertainty. Our findings may not translate to other visualization mechanisms, such as frequency intervals or hypothetical outcome plots [114].

5.3 Findings

Here, we present findings of our interpretative qualitative analysis. In addition to exploring general factors that influence the decision-making processes, we focused on different cognitive

[7] https://www.rev.com/
[8] https://miro.com/

processes involved in decision-making when dealing with uncertainty. We used this focus as sensitizing concepts [49], which offered ways of seeing, organizing, and understanding participants' experiences.

5.3.1 Participants' Lived Experiences Shaped their Sense-making Process and DSS Interactions

Unsurprisingly, participants' personal preferences, prior experiences, and knowledge and expertise had an impact on their decision-making process across different DSS.

5.3.1.1 Prior experiences with the housing market shaped participants' decisions

Lived experiences, such as *"[living] in an affluent area where housing prices [were] extremely expensive"* (P6) and *"[working] in housing adjacent with the homeless for a year"* (P19) made participants sensitive about affordability of houses regardless of the DSS they used. Furthermore, P11 believed that *"certain things ... have personal value, if they may not have monetary value"* based on observing their parents buy a home and watching them *"weight the pros and cons of the aspects of the house"*. Of the 17 participants, 6 (35%) had previously purchased a house. Home-owning participants used preconceived notions of how features affected home prices:

> *"My husband and I bought our first home four years ago ... amenities is an important part of purchasing a house ... people list their house for more money if you have certain amenities that catch peoples' eyes." –P4*

Participants with such notions wanted to know what house features the ML- and UQ-based DSS made their predictions based on. When using the Viz-based DSS, many participants carefully examined how the "amenities" features they selected affected historical price trends. However, this also made them forget Alex's preferences: *"Okay, what [features are] important to me?"* (P3). Participants without house-purchasing experience were unsure what features to select using the Viz-based DSS to explore historical prices, and thought that *"maybe it was [understandable] to someone who's [bought a house] before"* (P16).

5.3.1.2 Existing user interface and task expertise aided decision-making

Participants drew analogies between the DSS in the study and other tools they previously used, such as Apartments.com [9], Zillow [10], and Redfin [11]. Participants with such user interface (UI) expertise successfully transferred their knowledge to the DSS UI in our study.

The effects of ML and statistics expertise were most pronounced when interpreting outputs of the ML- and UQ-based DSS. Out of 17 participants, 12 (70%) had taken at least one ML or statistics class, and 5 (29%) had implemented an ML algorithm. Those participants could tell how the uncertainty manifested: *"..the AI is giving 20 different valuations"* (P8). Quantile dotplots reminded participants of normal distributions based on *"[looking] at lots of sets of data, that pattern [the bell curve shape] comes up a lot"* (P9). Participants recognized and interpreted parameters such as mean and variance in quantile dotplots based on their knowledge of statistics:

> *"...[dots] were more like data points from a study ... it creates like, a parietal chart. And the average of that, the median is kind of where it sets. But then the skew is more like how much variation there is." –P3*

The participant P3 could relate the median and variation in UQ-based DSS predictions to their engineering and management job in factory production. When interpreting UQ-based DSS predictions, participants relied on their knowledge of statistics to separate outliers from useful data. For example, P12 did not count a dot towards the extreme end of the distribution as *"this [dot] seems to be more than two standard deviations away"* (P12).

However, participants who did not have such expertise were at time reluctant in making decisions: *"... one of the reasons why I [was] not sure, [was] because I don't know statistics"* (P11). Thus, it was not enough for UQ-based DSS to simply compute uncertainty information; it had to communicate it in a way that the participants could perceive and understand it given their existing knowledge.

5.3.2 How Participants Perceived Uncertainty

In general, participants started their decision-making process by perceiving and extracting high-level visual cues in all of the DSS. They would start with eyeballing the graph to get enough information for their decision as quickly as possible. When eyeballing failed them, they would further engage with the details in the graphs.

[9] https://www.apartments.com/
[10] https://www.zillow.com/
[11] https://www.redfin.com/

5.3.2.1 Participants eyeballed the graphs to get easy cases out of the way

Participants tended to first eyeball the graphs to get a quick impression of house valuations in all DSS. Participants had fewest visual cues to perceive and interpret with the ML-based DSS, resulting in less workload than the other two DSS:

> *"The value of the house is 11% less than the list price, I read blue [caption] first, and I noticed that the blue line, I noticed that it's on the red side." –P4*

The workload for perceiving the dots in the UQ- and Viz-based DSS was similar; but the participants had to perform additional work to generate the output in the Viz-based DSS. When using the Viz- and UQ-based DSS, the participants were *"looking at where the majority of the dots fell"* (P19), *"trying to see if on average, [there] were more green or more red [dots]"* (P17). Eyeballing helped see dots as a whole, rather than having to perceive each individual dot.

5.3.2.2 Participants rushed to deal with the additional workload in Viz-based DSS

To deal with the additional work in the Viz-based DSS, participants rapidly clicked through house features, *"seeing how the spread of dots changes"* (P11), and *"hoping that [the feature they selected] would be more revealing"* (P18). However, they still had to mentally keep track of changes in the distribution of dots across different features. This tedious and time-consuming process made the participants less inclined to conduct detailed data exploration:

> *"... clicking through a lot of variables... and seeing how each one sort of behaves... [was] not the fastest way to do it. But it was [how] I could organize the information in my brain. It's a slower process ... I had to think about it quite a bit more, to make sure I was keeping straight what I was looking at." – P18*

Thus, giving more control to participants in the Viz-based DSS was not always desirable. Instead, delegating that extra work to ML or UQ algorithms helped participants focus on making sense of the visual information.

5.3.2.3 When the graphs were too complex, simply eyeballing them did not work

Participants could not simply eyeball the graphs when the prediction of the UQ-based DSS was "split"; i.e., when *not* all dots were on one side or the other:

"When there's an even spread, I just had to make sure by actually counting ... It's hard to just eyeball." –P14.

When the graph was "split" (i.e., when it was not "clear cut"), participants counted the dots on each side, and considered which side had more and by how much:

"There are more dots on the green side but there's not that many more green dots. So what's that, a difference of five. What percentage is that? 25%? Hmm..." – P11

Participants considered the distance of the dots from the list price along the x-axis; they looked at where the *"mean"* of the distribution (P3, P8, P19) was, and whether the dots were spread *"closer to the [list price] line"* (P13) or not:

"I looked at [percentages] because I could have two dots and positive 10% for green, but I could have two dots at negative 60%, and that would mean, you know, it's a higher percentage negative than the two dots, even though there's still two dots ... on each side." – P16

P16 wanted to know whether the risk associated with different dots was greater despite their count being the same. Still, most participants ignored the percentages on the x-axis in favor of just counting the dots: *"There [were] numbers and the percentages, but the bottom line was you were looking at the number of dots"* (P4).

5.3.3 Confusion, Resolution, and Learning Curve

Participants were confused when initially using the UQ- and Viz-based DSS. This confusion arose from overwhelming details in UQ-based DSS and lack of guidelines for exact feature selection steps with Viz-based DSS. They resolved this confusion by devising and using heuristics in the former case, and developing their own exploration strategies to compensate for the lack of instructions in the latter case.

5.3.3.1 Unlike ML-based predictions, the UQ-based predictions were initially confusing

Participants thought that the UQ-based DSS added *"noise with the dots"* (P14), and found the dots *"just in general, kind of confusing"* (P11), questioning: *"what do [the] 20 dots represent?"* (P6). The participants would blame themselves for such confusion: *"I'm not a crazy math, graph, numbers kind of person, so it was a little confusing to me"* (P4).

Conversely, *"the exact percentage"* (P2) output in the ML-based DSS was *"very, very clear"* (P8).

Participants were confused partly because they were *"not sure what uncertainty information [means]"* (P8). They speculated that the UQ-based DSS predictions were *"not like a sure thing"* (P10) or *"not set-in-stone"* (P13). Even the participants with ML background were confused why the algorithm generated different predictions for the same input:

> *"If it's the same algorithm with the same information, I would assume that it would give out the same output. But if [it predicts the same house to be] 20% undervalued, but also 30% overvalued, what's happening with the AI is a little bit confusing. I'm not really sure if you're running the AI 20 separate times to get different valuations or [were there] 20 separate algorithms to get those 20 dots." –P8*

Some participants were unable to recall the dots' meaning correctly: *"I thought it was green dots were good. (laughs) But they're all on that [higher] side ... uh, I'm confused now."* (P2). This led to further confusion and lack of confidence when initially using UQ-based DSS. However, after reading the instructions again before subsequent use of UQ-based DSS for a different level of uncertainty, participants reflected on and realized their previous mistakes:

> *"And so, the more green dots, the more likely the house... I just realized, I think this is the same as last time. I think I was misinterpreting ... Um, so, a-actually the last house was a slightly better value. Um, so I did make a slight mistake there." –P15*

Thus, our findings show that some of the participants' early confusion about UQ-based DSS was due to the lack of UI expertise. Such confusion cleared up with more practice. However, confusion about the abstract concepts of uncertainty persisted, and participants had to develop their own strategies to deal with it.

5.3.3.2 Participants resolved their confusion about uncertainty by assigning their own meaning to it

To resolve their confusion, participants assigned meaning to the UQ-base DSS representation of uncertainty by interpreting the dots as *"possible scenarios"* (P19), which were *"not, like actual prices that people have paid"* (P10). Even incorrectly assigned meaning, such as assuming dots were *"the different factors"* (P11), enabled participants to reason about the

house valuation: *"all the features ... are telling me that this home's value is less than what it's being asked for"* (P16).

Assigning such meaning enabled participants to think probabilistically: *"a few chances where it might actually be a bad deal"* (P19). They then made use of heuristics such as color of dots to understand valuation:

> *"I think just, remembering this like, green dots is good, red dots is bad, is kind of the way I'm gonna frame my decisions moving forward." –P10*

Our findings indicate assigning meaning, irrespective of its correctness, enabled participants to think probabilistically and devise strategies for further simplification through heuristics. Nevertheless, some participants thought that *"the dots weren't useful to [them] because they weren't assigned any particular unit"* (P6).

5.3.3.3 The Learning Curve

As participants developed their UI and task expertise, they started to think more deliberately; in particular after realizing their previous mistakes: *"reading over to make sure [I was] not missing anything"* (P2). Participants' confidence in their decisions also increased with subsequent use of dotplots: *"Once I understood what was going on like after the first [UQ-based DSS prediction] I think, I became more and more confident in the decisions I was making"* (P11). Eventually, *"things were pretty clear, especially near the end"* (P8).

This learning effect was not restricted to the UQ-based DSS. For example, in Viz-based DSS, participants carefully thought out their own feature selection strategies to resolve lack of detailed, step-by-step instructions. Participants started *"experimenting, and kind of clicking around"* (P10) to *"see what's going on [with the tool]"* (P12). Participants then thought carefully about the house features and made selections to *"update the graph"* (P10). Participants used multiple strategies such as selecting features sequentially, one by one, starting with one and progressively combining more, selecting all features in a group, or by *"picking on in each of the three categories of the estate, amenities, and location"* (P3).

5.3.4 Thinking and Deliberation

After overcoming the learning curve to understand and effectively use uncertainty, participants were able to think critically, and became more confident in their decision-making. Uncertainty also forced participants to consider risk and safety, and alert them to deliberate over plausible reasons behind the predicted valuation.

5.3.4.1 UQ-based DSS encouraged critical thinking and inspired confidence

Participants interpreted "clear-cut" predictions in the UQ-based DSS as *"[reaching] a consensus among the 20 dots"* (P6), and increased participants' confidence in house valuation: *"this house is definitely a much better value"* (P15). Conversely, "split" predictions made participants rightfully vigilant about house valuation: *"I feel like odds are, we'll end up on the red side. Or at least ... potential to."* (P18), evidenced by participants' use of probabilitic language: *"very likely to be, um, valued more than the list price."* (P3). Larger variance in predictions signalled more uncertainty, inaccuracy, and lacking representative data, whereas narrower distributions evoked feelings of confidence, certainty, and prediction accuracy:

> *"If [the prediction] was flatter then, there may not be as many houses in- in that set of data. So it might not be as accurate. But seeing the- the sharp curve, like the big in the middle and ... a large slope down, then [that] showed me there was more data to see, [which] would be more accurate." –P9*

Participants also assumed that the level of uncertainty (i.e., the variance of distributions) was indicative of the market conditions:

> *"If [the dotplot is] wider ... I think it shows that the m- market is not as sure on how to value houses perhaps. So it makes me less sure on the valuation because if they're all tightly grouped together- ... then I think it's pretty clear the market knows what houses are worth." –P15*

Thus, while the overall expression of uncertainty made participants more vigilant and increases their sensitivity to outcomes, the specific degree to which uncertainty was expressed affected their confidence in the competence of UQ-based DSS.

5.3.4.2 UQ-based DSS evoked perceptions of risk and safety

Participants perceived predictions that were "clear-cut" good deals as *"the safest bet"* (P19), while they perceived "split" predictions as having increased risk: *"[Alex] wouldn't be facing the same risk [with all green, as compared to] going for a house that had chances of being over-valued."* (P9).

These initial impressions influenced whether participants engaged in analytical or heuristic-based thinking. When predictions were "clear-cut", participants spent very little time reasoning about the valuation: *"So right away, I see all these red dots, that's a bad deal"* (P2). In such cases, their thinking was very similar to that with ML-based DSS predictions—almost "automatic".

On the contrary, when predictions were "split", participants paused and engaged in slow and deliberate thinking: *"Wait a minute. (laughs) I'm thinking through this now"* (P2). They *"had to kinda think more about what decision [they would] make"* (P9). Despite *"[having] to weigh a lot more yourself"* (P8), similar to Viz-based DSS, participants appreciated the *"benefit of the dots [in] getting more data points"* (P2). In those cases, participants thought that the variance and skew in the UQ-based DSS could alert them to potentially "weird" situations:

> *"[With the ML-based DSS] you're missing a lot of information there that's potentially really, really important. Especially something like, a split situation ... [could mean] that there's something weird going on with that house. Um, but if you just average that and capture it in one line that says that it's equal to the list price, I miss all that information." –P18*

Our findings indicate that uncertainty provides additional context, which in turn, influences whether participants engage in analytical or heuristic-based thinking.

5.3.4.3 UQ-based DSS supported plausible reasoning

When participants paused to deliberate over "split" UQ-based DSS predictions, they also speculated the possible reasons behind the house valuations: *"sometimes just thinking deeper into it, um, people price their house lower sometimes to sell it quicker or if they know something's wrong with it."* (P13). Although participants simply speculated about UQ-based DSS predictions, when using the Viz-based DSS those kinds of participant musings often unveiled their wrong assumptions and backward inferences.

With the Viz-based DSS, participants generally assigned meaning to the historical sales prices for which they had no evidence. For example, P10 inferred that majority red dots in the default graphs represented *"the [bad] situation that we're in in this market"* (P10). Other participants jumped to conclusions about the state of those houses: *"these houses had pretty significant problems that would really lower the true value"* (P6). They would muse that *"maybe [one of the houses] needs some fixing up or something"* (P18). Participants drew inferences they could not substantiate given the DSS functionality, such as the actual age of the houses:

> *"Looking at age [it] looks like it's more on the green side where houses are valued more than the list price. So it's kind of hard, and I'm not really sure, it just says age. I'm assuming older houses are valued more percent-wise than the list price ..." – P9*

This was despite that selecting the "Age" feature did not indicate the age of any of the houses; it only showed the historical prices of houses that had the same age as the house that Alex wanted to purchase. Participants made similar unsubstantiated assumptions when trying to infer what kind of house Alex should look for: *"So, maybe looking at an older home, Alex can, get a little more on the green side"* (P10).

5.3.5 Judgement

The final step in the decision-making process is judgement, where participants had to decide what to recommend to Alex based on their reasoning.

5.3.5.1 Participants withheld decisions when UQ-based DSS predicted good and bad deals were equally likely

In case of clear-cut UQ-based DSS predictions where *"there was [overwhelmingly majority] and they were a larger percentage, it was an easy decision"* (P16), *"or very quick decision on whether or not to buy it."* (P17). Participants made those decisions in a *"split-second"* (P17) with little to no deliberation: *"I mean this is a great value house here right? So yes"* (P7). Decisions about "clear-cut" deals were similarly quick with both the Viz- and ML-based DSS, where at times participants did not even verbalize their reasoning: *"Uh, [4 second pause] no, Alex's not buying"* (P16).

However, *"if [dots were] close to being split in the middle ... [they had] to kind of like make a judgment call"*(P7), and needed *"a little bit more information to make a decision"* (P6). For split UQ-based DSS predictions which *"average [was] about on target for the list price"* (P3), participants thought that *"the [good deal and bad deal] predictions ... cancel each other out"* (P11), which made them unsure and reluctant to make a decision:

> *"[The valuation] looks like it is equal on both sides, so this decision might be a little harder, but let me think through it ... So if the green ones indicate a good deal, and the red ones indicate a bad deal ... we're just literally right in the middle of everything. So I guess this one ... I'm not sure." –P9*

We did not observe this effect in the absence of uncertainty information when ML-based DSS predicted valuation was equal or close to the list price. While participants found such predictions *"not particularly informative"* (P18), they attributed positive properties to the ML-based DSS and made decisions *"based on the tool"* (P10):

> *"So, it's worth that based on, um, this fancier algorithm looking at more information, like the neighborhood ... I'm saying Alex should consider buying this house, so I'm clicking on yes." –P3*

Lacking any information about the uncertainty of the ML-based DSS prediction, participants could have misplaced their confidence in ML algorithm's competency (or lack thereof): *"I felt good about that, especially with knowing that, like, the [ML] algorithm had, had predicted that price, based on, you know, all those things."* (P10).

5.3.5.2 Participants were risk-seeking with ML- and UQ-based DSS, but risk-averse with Viz-based DSS

Participants were willing to recommend bad deals to Alex when the ML- and UQ-based DSS prediction agreed with their personal (often low) risk threshold. Participants determined acceptable risk based on percentages when using ML-based DSS: *"I think a percentage I would probably be comfortable with [recommending] the house [would be] at maybe like 10% or under"* (P11), or proportion of dots when using UQ-based DSS was: *"only slightly above the actual value and meets all of Alex's criteria"* (P15). If the risk of a bad deal was acceptable to participants, they recommended that Alex should purchase the house:

> *"I would say buy the house. It's got all the features [Alex] wants. It's only 1% off. I think that's not an unreasonable percentage to be off by." –P16*

However, when using the Viz-based DSS, participants felt *"safer saying I am not sure rather than making a strong decision in either direction"* (P7). Participants refrained from making a decision as they did not have *"all the information"* (P15)

and did not want to *"make any decision based on (perceived) faulty data"* (P12). Other participants refrained as they did not receive a recommendation from the Viz-based DSS as they expected: *"it doesn't really help me ... about whether Alex should buy that house, or not. So I'm just going to say I'm not sure"* (P3).

5.3.6 Reliance and Trust

Despite little deliberation over clear-cut predictions of ML- and UQ-based DSS during the task, participants expressed distrust in the algorithms during post-task interview. In order for the algorithms to gain their trust, participants deemed transparency and explainability important.

5.3.6.1 Participants did not trust the DSS that did not match their values

When using the ML- and UQ-based DSS, participants had to weigh *"the value of [the] decision support tool against the value of [their] own personal judgment"* (P6). Participants thought that the DSS might miss "intangible" aspects of the house that are important to them: *"the algorithm probably doesn't know, that we're looking for like a certain sized like yard, or how good the garden is in the back, which it's not controlling for"* (P9). While many participants did not question the accuracy of the algorithms, they were skeptical about the usefulness of the two DSS:

> *"I guess the algorithm was probably very accurate, but, I think of the intangibles, [things that] I would consider aesthetically pleasing. I don't care that the algorithm says it's a value, to me [it's] the intangibles, [things that] a computer can't see." – P16*

Participants considered certain aspects of the house, such as aesthetics, to be more important than the price. They *"did not care"* (P16) about the algorithm valuation if they could not see that it aligned with their personal values.

5.3.6.2 Participants did not trust the DSS if it did not conform to their mental models.

Participants developed mental models of how the DSS worked in part through interaction with the DSS.

The feedback they got from extensive interaction with the Viz-based DSS helped build this mental model quickly:

> *"I felt more interactive. I could actually see what would happen when I hit certain functions. The other one was just a generality. It wasn't like I had much to look at." –P16*

Unlike the Viz-based DSS, participants perceived that the ML- and UQ-based DSS were *"doing [the valuation] for me"*. Some participants formed an impression that *"the algorithm [did] not compare with other previously sold houses that are similar."* (P15). This inability to engage further with those DSS, combined with the lack of ML and UQ expertise, for many participants resulted in a disconnect between the algorithms and their mental models.

If the DSS exhibited any behaviour that did not conform with their mental models, participants considered it to be "evidence" that the DSS was incorrect:

> *"[In] my brain, [the dot on the line] can't be green or red. ... It's not plus or minus zero, it's exactly zero ... And based on that, if that part of the algorithm is incorrect, then I don't trust the data." –P12*

There, the participant made a mistake interpreting the visualization. The DSS considered a house valuation that is exactly the same as the list price as a good deal. Nevertheless, such mental model mismatch negatively affected participants' trust.

5.3.6.3 Participants needed assurance about the competence of the algorithms to be able to trust them.

Participants with an ML background watched out for signs that the algorithm was unreliable: *"If the algorithm actually valued the home [too high] it would make me think that the algorithm wasn't reliable"* (P17). They questioned the tool's accuracy: *"There was like no specification from that side. So how do I know that the evaluation provided by that tool was accurate?"* (P7), and wondered *"if [houses were] valued correct"* (P12).

Participants without ML background expressed similar concerns, yet lacked the vocabulary to properly verbalize them. They were concerned that *"maybe- maybe the algorithm is pulling other data ... it may have accidentally pulled in data from like another area. And we don't know that, and then that could influence our decision."* (P9).

5.3.6.4 Participants needed explanations to be able to trust the algorithm.

The participants wanted to know *"why [the houses were] worth more or why they're worth less"* (P16). This was important *"cause anyone could say their house is worth more"* (P2). Yet, participants' needs for explanations were diverse.

Participants with a background in ML needed explanations based on the features of the house that the algorithm considered:

> *"The algorithm might say something like, 'If you have four bedrooms, we'll give it a 10% up, uh, charge. If you have one bedroom, we'll decrease the value by 10%.' So based on all these things.. the algorithm [would] come up with some value." –P12*

Thus, participants with ML background expected an explanation that took the form of a counterfactual [161] or feature attribution [181].

Others suggested *"sort of ... a score or something at the end of like... we rate this property [at] a 4.3 out of 5, [with] breakdown categories, where it's like, well they scored a 5 in quality, a 2 in location."* (P18). Participants also thought it would be helpful *"to show*

[them] an example... seeing a picture helps" (P16) to derive explanations for themselves, including through interaction with the algorithm:

> *"I could perhaps [input] a recently sold [house] and see how the tool priced it compared to the selling price to get an idea for whether or not the tool's algorithm is correct. So I would like to probably use this algorithm, but I would do some, some steps, uh, to verify that I think it's at least somewhat trustworthy." – P15*

Thus, it is unlikely that a single explanation mechanism would be able to satisfy such diverse explanation user needs.

5.3.6.5 Lack of transparency from the ML-based DSS made participants question its credibility and regret trusting it

Participants did not have confidence in the competence of the ML-based DSS: *"It- like it is contradictory what I'm saying, the blue was the easiest to understand, but um, do I believe it? No. (laughs)"* (P2). Although some participants refrained from relying on the ML-based DSS that was *"trying to influence [their] decision"* (P12), those who did rely on it regretted it in hindsight:

> *"I'm, I'm really, really basing all these decisions off a tool that I don't understand very well. So it's putting trust in the tool even though I don't understand it which probably is not the smartest move ... [when] I don't really have transparency into it." –P15*

Other participants had broader questions about the credibility of the tool and demanded "social transparency" [64]:

> *"[I wonder] who made this decision support tool, who financed [its] creation, dispersal and usage ... how many times it has been used, and across what demographics. I'm also thinking about [its] rating from those who may have used it in the past. If [it] was created by, um, a real estate agency ... that has a listing of this house ... that Alex is interested in, there's a, there's a bias there." –P6*

Participants needed to know whether the algorithm was built by *"a bank or some trusted source"* (P2), and whether the *"person that created it is unbiased"* (P15). Others commented that they *"can't always rely on one tool"* (P13) and needed *"a second opinion ... [by having] a contractor look at [the] house"* (P13).

5.4 Discussion

Here, we situate our findings in the larger discourse on transparency in AI-assisted decision-making, and provide suggestions for the design of such mechanisms. We conclude with implications for design of uncertainty communication in ML-based DSS as means to aid algorithmic transparency.

Table 5.2: Summary of the effects of uncertainty communication on different stages of the sensemaking process, with breakdowns in each stage and ways to address them.

Cognitive Process	Effects of Uncertainty	Breakdowns	Implications for Design
Perception	Eyeball proportion first, extract higher-order cues for more complex uncertainty representations	Missing important cues initially due to learning curve	Creation of onboarding material to prime decision-makers for uncertainty; Highlighting different parts of the graph to draw visual attention to specific cues
Reasoning	Think critically and engage analytically when predictions are "split"	Not engaging in critical thinking when dotplot indicates one type of deal	Promote seamfulness and highlight complexity by coloring dots in a bin differently; communicate consequences of a possible wrong decision
Judgement	Perceive uncertainty as "cancelling out" for "split" predictions, refrains from making decision	Rejecting new model behaviour under uncertainty, being paranoid over vigilance	Update mental models of uncertainty; make new behavior with uncertainty consistent with decision-makers' old experiences

5.4.1 Sensemaking from Extracted Cues

Here, we discuss the effects of uncertainty communication on the cognitive process of **perception** as described in Table 5.2.

5.4.1.1 Extraction of complex cues if satisficing is insufficient

Sense-making literature defines cues as "simple, familiar structures that are seeds from which people develop a larger sense of what may be occurring" [229] (Figure 5.4). While extracting visual cues to decode signal in UQ-based DSS predictions, we observed participants *satisficing* [195] with simple heuristics such as eyeballing and glancing rather than finding an optimal strategy. This provides evidence for decision-makers' naturalistic thinking when interpreting graphs, contrary to visualization researchers' assumption of rationality [109]. In addition to satisficing, our findings confirmed that decision-makers use and switch between heterogeneous strategies, such as usage of means, spread of distributions, and frequency or cumulative counts of dots, to extract graphical uncertainty information [127]. Our work expands this further by identifying the patterns of *how*, *why*, and *when* these strategies are employed by decision-makers.

When interpreting uncertainty, decision-makers had to extract cues with higher cognitive effort after an initial pass with low-effort strategies. They combined strategies, such as counting dots to compare at specific percentages, if uncertainty was especially challenging to interpret. To account for and compare with their subjective risk thresholds, decision-makers extracted cues such as percentages and means from uncertainty in predictions. Our work points to *which* nuances in reasoning behaviour are lost when measuring average performance, and strengthens the calls for a new class of behaviour-aware models of uncertainty visualization [127].

5.4.1.2 Visualizations provide scaffolding to think about abstract concept of uncertainty.

We observed that decision-maker participants struggled to understand the meaning of uncertainty during decision-making. This was both due to lack of task experience [128] and a natural aversion to probabilistic thinking [216]. This led them to question the source of the uncertainty, miss important cues, and perceive the representations as noisy and confusing. Although our findings do not dispute that uncertainty may lead to algorithmic transparency [25], we showed that people find the abstract concept of uncertainty opaque and hard to understand. Graphical representations of uncertainty rely on the assumption that users are equipped with knowledge of interpreting charts [16]. However, most participants in our study self-reported that extracting information was challenging due to a lack of statistics background.

Participants devised their own strategies to resolve initial confusion with uncertainty representations. They relied on visual encoding (e.g., color) of quantile dotplots [131] to guide their visual attention, assigned meaning to discrete representations of uncertainty for better recall, and estimated proportions via glancing to determine interpretation effort. With subsequent use, decision-makers were able to reflect on their mistakes and understand the graphs better. With enough exposure to uncertainty representations, decision-makers were able to rely on the quantile dotplots to improve recall [112] and make consistent estimates [131, 127].

5.4.1.3 Streamlining sense-making process during first-time use

To address the breakdowns that we have identified, future designs could use mechanisms such as onboarding [43]. Such mechanisms can be used to fill the learning gap by introducing decision-makers to uncertainty in UQ-based DSS predictions. Onboarding materials could act as cross-functional boundary objects [44] to match visualization researchers' expectations and decision-makers' behavioural strategies, and to facilitate a streamlined meaning-making

process during first-time use [76]. Future designs can further augment this process by assisting decision-makers in directed extraction of visual information. For example, future designs could highlight relevant information (e.g., in case of our DSS the percentages on the x-axis) while reducing emphasis on other visual information (e.g., color of dots can be made gray to indicate un-informativeness [51]).

5.4.2 Uncertainty Forces Decision-makers to Slow Down and Think Analytically

Here, we discuss the effects of uncertainty communication on the cognitive process of **reasoning** as described in Table 5.2.

5.4.2.1 From heuristic-based reasoning to analytical reasoning.

Our work shows further evidence that communicating uncertainty could provide a form of algorithmic transparency [25] and enable decision-makers to have appropriate confidence in and reliance on ML predictions [212, 211]. Our findings further indicate that decision-makers resort to different strategies depending on the uncertainty conditions. They may completely ignore uncertainty in some situations (e.g., clear-cut predictions), and rely on the DSS with much thinking.

However, heuristic-based reasoning can lead to grave errors of judgement [216]. For example, certain configurations of uncertainty (e.g., DSS predicting all-red dotplot when ground truth is higher than list price) led to false notions of confidence in predictions leading decision-makers to make wrong decisions in some scenarios irrespective of the prediction correctness.

To remedy this, decision-makers need to be motivated to account for further uncertainty in uncertainty predictions.

On the other hand, uncertainty increased the demand for cognitive engagement in complex cases, thereby disrupting decision-makers' automatic reasoning. In such cases, decision-makers paused to think, carefully weighted the evidence, and used probabilistic language when making decisions.

Participants questioned what the dots meant, how the algorithm produced the uncertainty representations, and why the algorithm valued it in a particular way. Additionally, uncertainty allowed participants to interpret predictions in failure-centric ways. Decision-makers appreciated having information such as skew and variance, which could potentially alert them to a risky or suspicious situation.

5.4.2.2 Slowing down too much can lead to indecisiveness.

Our findings contest the assumptions that decision-makers will always engage with model predictions rationally when they slow down. For example, "split" uncertainty predictions made participants slow down, but also led to indecisiveness that hurt the decision-making process. When the uncertainty was low, the participants took conscious, yet unnecessary risks and recommended bad deals.

5.4.2.3 Inducing healthy skepticism through seamfulness.

Future designs could remedy unnecessary risk-taking by further promoting seamfulness [48, 100], an approach of using visual elements to highlight the underlying complexities in a graphical representation.

The inherent seamfulness of uncertainty communication may contribute to further decision-makers' analytical engagement and induce healthy skepticism [128]. Thus, future designs could leverage uncertainty communication as a form of adversarial design [128] to promote a wariness and a mindset of doubt, and the consequences of making a wrong decision [4].

5.4.3 Communicating Uncertainty Reduces Undue Over-reliance on ML Algorithms

Here, we discuss the effects of uncertainty communication on the cognitive process of **judgement** as described in Table 5.2.

5.4.3.1 Uncertainty helps to resolve enchanted determinism.

Our findings show further evidence that suppressing uncertainty information can give decision-makers an impression that ML algorithms generate infallible insights, thus promoting *enchanted determinism* [45].

Our findings show that participants recommended houses based on imagined positive properties of ML-based DSS.

However, indicating uncertainty in predictions directed attention away from the underlying algorithm, causing decision-makers to question the algorithm and critically assess the potential outcomes.

5.4.3.2 Difficulties in forming accurate mental models with uncertainty.

Communicating uncertainty could inhibit appropriate reliance on the DSS. For example, when the UQ-based DSS did not align with participants' mental models, they became overly cautious and withheld their decisions.

This indicates the need to promote appropriate mental models that accommodate new model behaviour under uncertainty. Seminal work [120] on mental models shows that decision-makers have to choose between changing their mental models to accommodate the new information, or reject, ignore or reinterpret it. Since uncertainty may generate constant or unexpected new behaviours, re-training users may be costly [13]. Instead, focusing on setting decision-makers' expectations of uncertainty to accommodate new behaviours may be more appropriate [138].

5.4.3.3 Embracing uncertainty as a community.

Our findings also show that future designs could leverage uncertainty representations as cognitive forcing functions [144] to promote vigilance and further reduce over-reliance on ML algorithms [39]. Such future designs should shift the focus from mitigating [230] to embracing uncertainty [36, 115], by incorporating uncertainty in design of interactions with technology [19, 191]. This can be further promoted by engaging domain experts [111, 90] with communities [201, 32] to inform the design of technology by eliciting community-based values and management strategies around uncertainty.

5.5 Summary

In this paper, we investigated how communicating the uncertainty of Decision Support Systems (DSS) affects decision outcomes and cognitive process of lay decision-makers — perception, reasoning, and judgement. Our study revealed that lay decision-makers are able to rely on visualizations even before gaining a conceptual understanding of uncertainty. Our findings showed that the cognitive demands of interpreting uncertainty force the switch from heuristic-based reasoning to analytical reasoning. Finally, model transparency via uncertainty challenges the perception of lay decision-makers that ML is absolute and infallible.

Our findings add to the overarching discussion on the design for uncertainty [69, 199], supporting that it goes beyond the properties of ML systems [32], and is an omnipresent phenomenon [200] that shapes human interaction with technology. Our work opens up opportunities for future research, including the creation and evaluation of novel uncertainty-based designs. We identify how future design work can facilitate better understanding of uncer-

tainty, promote critical reasoning and encourage lay decision-makers' to embrace algorithmic uncertainty. While this work investigates lay decision-makers, it provides a foundation to investigate high-stakes decision-making where errors can have severe consequences. We conclude by inviting researchers to adopt a psychological lens to investigate how human-AI interactions shapes peoples' perception and eventual adoption of AI.

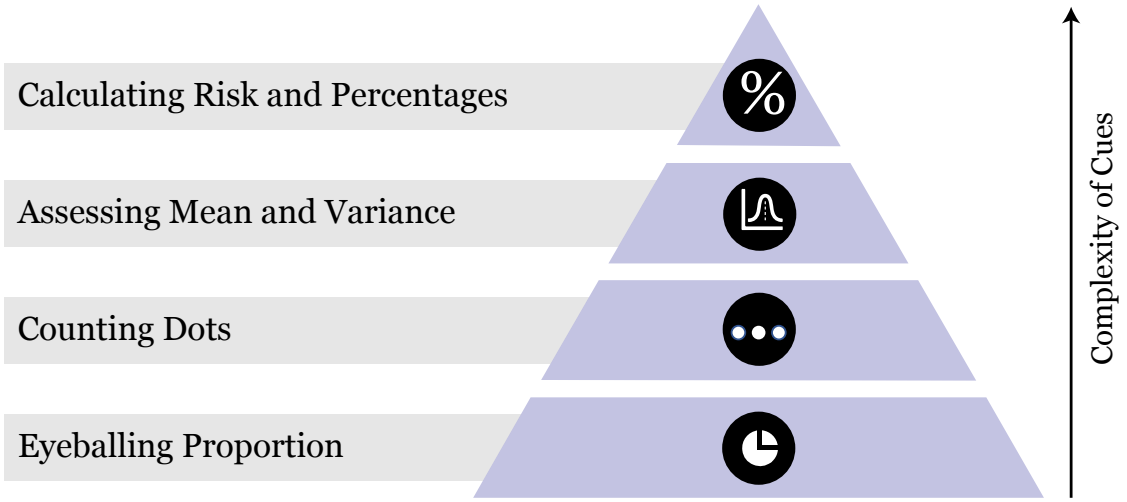


Figure 5.4: Conceptual map of different cues that participants extract for uncertainty representations of model predictions. Participants start by eyeballing. They do not count when predictions are "clear-cut", but extract more complex cues higher up the pyramid for "split" predictions.

CHAPTER 6

Leveraging Storytelling to Communicate Uncertainty

6.1 Introduction

AI-based decision support systems (DSS) promise to augment, rather than replace, human judgement by providing decision recommendations based on algorithmic predictions. Deployed across a number of domains [43], such AI-based DSS perform data analysis from historical data and provide a recommendation, while decision-makers retain oversight and make the final choice. However, a lack of transparency can lead to decision-makers agreeing with AI recommendations even in cases where they should not.

Mechanisms that enable algorithmic transparency, such as explainability [178] and interpretability, aim to provide decision-makers with insight about the AI's inner workings and decision-making. By increasing transparency, such mechanisms aim to help end-users identify and refrain from agreeing with AI recommendations in situations in which the AI's reasoning was incorrect [190]. Yet, such mechanisms can be misleading and can provide decision-makers with a false sense of confidence [12] and tend to act as evidence of reliability rather than accountability [40, 55], often deceiving end-users into over-relying on AI [143, 116].

A large body of work [24, 15, 19, 77] has proposed uncertainty as an alternative mechanism to impart transparency. Uncertainty quantification (UQ) algorithms [79, 153, 218] compute how confident algorithms are when making recommendations. Data visualization mechanisms [114, 81] such as quantile dotplots [131, 73] encode these uncertainty estimates in charts and graphs so that decision-makers can determine situations where it is acceptable to incorporate AI recommendations to inform decisions [238, 159, 43].

Despite improving decision-makers' analytical thinking, graphical uncertainty visualization requires decision-makers to possess a prior mathematical or statistical background to interpret them correctly. As a result, decision-makers without such a background struggle to understand them due to a high learning curve [175]. For example, in the disaster relief domain, lay decision-makers widely report being confused by the cone of uncertainty [199], one of the most popular uncertainty visualization mechanisms.

Prior work has shown that communicating AI uncertainty alone is inadequate to help determine when to accept AI recommendations due to cognitive biases and numerical literacy [46]. Schafer et al. [189] design a capability-focused guidance mechanism in addition to uncertainty representations and find that they improve task performance. A growing body of work in data visualization and HCI has looked at narratives to scaffold novice learning about data [108]. Yet, it remains unknown whether this approach helps people better understand uncertainty, especially in difficult situations, and how it affects their agreement with AI and confidence in decisions.

Here, we conduct a formative quantitative study to explore how AI agreement and decision confidence are affected under cases of varying difficulty levels and uncertainty representation. Our study addresses the following research questions:

- RQ1: How does communicating AI uncertainty impact AI agreement and decision confidence when reviewing algorithmic recommendations?
- RQ2: How does narrative presentation of uncertainty compare to graphical uncertainty in terms of AI agreement and decision confidence?

We contextualize this study in the realistic scenario of recommending house purchasing to a hypothetical friend. Participants assess AI recommendations derived from a Bayesian ridge regression model (from Chapter 5), along with its uncertainty presented in different forms, to determine whether or not to recommend a house to Alex, or they are unsure.

Our results show that stories make decision-makers more unsure and reduce agreement with AI recommendations when uncertainty is high and AI's prediction is ambiguous (i.e., recommendations do not strongly indicate one decision over another). However, stories lead to an increase in AI agreement when uncertainty is suppressed, and AI's predictions are deterministic. Our findings have implications for the design of effective scaffolding mechanisms that help non-technical decision-makers effectively interpret uncertainty.

Our main contribution is empirical knowledge on how uncertainty representation, once it can be interpreted by non-technical decision-makers, influences their decision confidence and agreement with AI recommendations. We present implications for designing mechanisms that can help non-technical decision-makers overcome a high learning curve. Our work motivates

future studies to leverage narrative storytelling to communicate uncertainty to decision-makers without technical or statistical expertise. We highlight future research opportunities, including evaluating the quality of AI-assisted decision-making with the presence of narrative scaffolding.

6.2 Method

6.2.1 Theoretical Underpinning

We draw on dual-process theory [216, 125], which states that when the cost of cognitive processing is high, decision-makers default to heuristic thinking. Chapter 5 empirically demonstrates how the cognitive cost of processing graphical uncertainty information is high for decision-makers without prior statistical or technical knowledge.

When interpreting graphical uncertainty information, such decision-makers need to overcome a high learning curve. They overcome this learning curve by undergoing a learning process to interpret uncertainty. The learning process consists of two main components: declarative (i.e., what is uncertainty as a concept) and procedural (i.e., how to act on uncertainty information) [5]. Decision-makers learn to interpret uncertainty by *procedurally* gaining expertise through repeated exposure (i.e., familiarity and practice with using uncertainty information). They learn to interpret uncertainty *declaratively* by gaining a conceptual understanding of what the dots, the mean, and the spread indicate about the AI's predictions. In the absence of such expertise, decision-makers can agree with AI due to cognitive biases such as enchanted determinism [45] as demonstrated in Chapter 5. Alternatively, they can be unsure about decisions as they cannot interpret uncertainty information effectively, and not because they are unsure about AI's recommendations.

This study evaluates the effects of uncertainty communication on AI agreement and decision confidence, once decision-makers learn to interpret uncertainty. To do so, the study guides decision-makers to overcome the learning curve by addressing both its declarative and procedural components. We operationalize procedural learning by incorporating a training phase in the study design for non-technical decision-makers to gain familiarity and practice using uncertainty information. We operationalize declarative learning by designing narrative-based scaffolding mechanisms that help decision-makers interpret uncertainty probabilistically (based on Chapter 5) rather than through satisficing.

6.2.2 Task

To answer these research questions, we chose a relatable, everyday decision-making scenario of house purchasing. We adopted this task for a number of reasons. First, the task is realistic enough that a lot of people may have experienced making financial decisions before. The housing domain is also where AI models provide a recommendation [207], along with uncertainty in the form of risk scores [172]. We situated participants in the role of making a recommendation to a hypothetical friend (Alex), instead of asking whether to purchase the house themselves to avoid subjectivity in making decisions. To further minimize subjectivity, we only provided AI recommendations and avoided presenting specific house features or images.

Participants took part in a simple yet realistic house purchasing decision-making scenario. The participants took on the role of decision-makers who had to help their hypothetical friend, "Alex", determine if the house Alex wanted to purchase was a good deal or a bad deal. The house was a good deal if the list price of the house (i.e., the house seller's asking price) was less than or equal to the true valuation of the house; and a bad deal if the list price was more than the true valuation. We used the following prompt to engage the participants in this scenario:

Your friend Alex wants to settle down in Ames, Iowa! Alex tells you about their troubles finding a decent house, and you are happy to help them out. Alex plans to move close to their office, and has been looking to buy a house soon. They have found a house they like, and asked your opinion to decide whether they should buy it. Alex tells you the purchase price that the Real Estate Agent provided them. They asked you to help them decide:

1. Whether you recommend Alex to buy the house,
2. Whether you recommend Alex NOT to buy the house,
3. If you are unsure and cannot decide what to recommend.

6.2.3 Data

We used the Ames Housing Dataset [50] for our decision-making tasks. We followed open source code [208] on GitHub for data analysis. This included data cleaning, data processing, and selecting the most influential features for the task of predicting the price of a house in the Ames Dataset. This resulted in a total of 22 features, and 2,616 houses in the dataset. The features included: house age, square footage, basement area, number of bedrooms, number of storeys, type of roof, street length, lot area, land flatness, location, proximity to park or greenbelt, proximity to road or railway, zone, fireplace, size of garage, garage availability,

central air conditioning, condition of house, quality of construction, quality of space, and kitchen quality.

We then divided this dataset into train-validation-holdout, with a split of 80%-10%-10% respectively (2,092 training instances; 262 validation; 262 holdout). We used the training data to train the ML- and UQ-based DSS, and the validation data to pick the hyper-parameters of the algorithms. We validated the algorithms by combining the training and validation sets, and performed 10-fold cross validation (CV) of our algorithms. We used the "unseen" houses from the holdout set in the decision-making task to represent the houses that Alex wants to buy.

6.2.4 Model

Our Bayesian approach for the treatment of probability and uncertainty closely fits with the culture and incentives of the human-centered design [232, 132]. We selected a Bayesian UQ method because it allowed us to control the level of uncertainty through adjusting the model hyper-parameters, as opposed to frequency-based approaches (e.g., confidence intervals [238]) that would require us to modify the training data. Specifically, we apply the Bayesian ridge regression to the same linear model used in the ML-based prediction. The Bayesian ridge regression provides the Bayesian analogue to the OLS regression solution in the ML-based algorithm above, thereby offering a consistent modeling comparison between the two.

The Bayesian UQ specifically seeks to capture the epistemic uncertainty in the model parameters conditioned on the available evidence in training data $\mathcal{D}_{\text{tr}}$:

$$p(\beta|\mathcal{D}_{\text{tr}}) = \frac{p(\mathcal{D}_{\text{tr}}|\beta)p(\beta)}{p(\mathcal{D}_{\text{tr}})}, \tag{6.1}$$

where $p(\beta)$ is the prior distribution of the model parameters, $p(y|\beta)$ is the likelihood function, $p(\beta|\mathcal{D}_{\text{tr}})$ is the posterior distribution, and $p(\mathcal{D}_{\text{tr}})$ is the Bayesian evidence (marginal likelihood) that acts as a normalization constant to ensure the probability distribution is proper. The prior thus describes the initial uncertainty on β before seeing any training data, and the posterior describes the updated uncertainty after incorporating the training data.

Bayesian ridge regression employs a Gaussian prior $\mathcal{N}(0, \lambda^{-1}\mathbb{I})$ and Gaussian likelihood $\mathcal{N}(x^T\beta, \alpha^{-1})$. We used a hierarchical model, where the hyper-parameters α (precision of data label noise) and λ (prior precision of β uncertainty) are assigned Gamma hyperprior distributions. Each of the two Gamma distributions are further controlled by a shape parameter and rate parameter. The `scikit-learn` implementation finds the numerical approximation of the posterior for this Bayesian ridge regression setup per the algorithm from Appendix A

of [210].

To set up scenarios with different levels of prediction uncertainty, we recreated situations of different training data quality. We varied the precision of the noise level of the training data labels through α, while keeping the prior λ constant. In the hierarchical Bayesian setup, α is treated as a random variable (it is not specified directly) and assigned a Gamma hyperprior. Therefore, we vary the shape (α_1) and rate (α_2) parameters of α's Gamma hyperprior to control the entire distribution of α values. Since smaller α values lead to increased uncertainty of the model prediction overall, and larger α values lead to decreased uncertainty, we defined our low uncertainty condition to be $\alpha_1 = 5000$ and $\alpha_2 = 1 \times 10^{-6}$ (i.e., high values of α with mean 5×10^{9}), medium uncertainty condition with $\alpha_1 = 1 \times 10^{-6}$ and $\alpha_2 = 1 \times 10^{-6}$ (i.e. medium values of α with mean 1), and high uncertainty condition with $\alpha_1 = 1 \times 10^{-6}$, $\alpha_2 = 200$ (i.e., low values of α with mean 5×10^{-9}). Our cross-validation of the Bayesian algorithm showed that it reached accuracy of $RMSE = 0.344$ for all three scenarios.

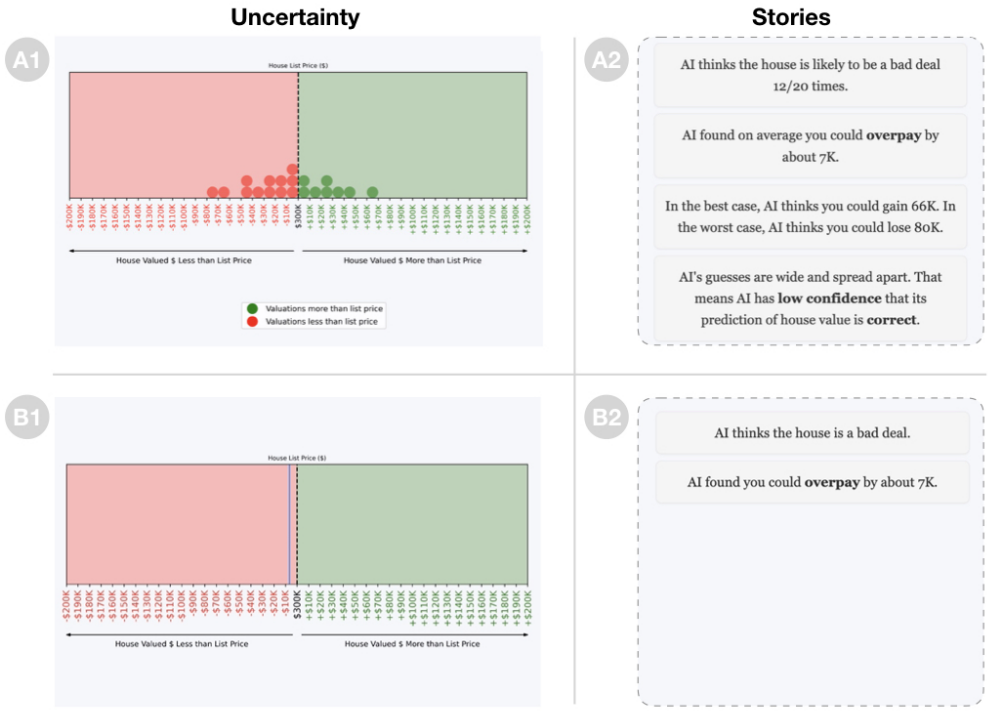


Figure 6.1: Within-subjects study design with 2x2 conditions. Independent variables are uncertainty and stories.

6.2.5 Interfaces

Here we describe how we developed the graphical uncertainty presentations and stories (Figure 6.1) for our study.

6.2.5.1 Uncertainty Visualization

We created graphical representations of the AI's price prediction relative to the list price. Participants viewed a graph where the x-axis was centered around the house's list price. Values to the left indicated the house was worth less than the list price. Values to the right of the list price indicated the house was worth more than the list price. The axis was labeled in dollar increments from -$200K to +$200K.

In *No Uncertainty* conditions, the interface displayed the two-tone chart with a solid blue line indicating the AI's prediction of the house price. In the *With Uncertainty* condition, the software presented the two-tone plot along with a quantile dotplot [131] that represented uncertainty. The spread of the dotplot indicated the level of uncertainty. Dotplots with a narrow spread indicated low uncertainty, while those with a wide spread indicated high uncertainty. We varied the level based on the hyperparameters of the Bayesian ridge regression model (Section 6.2.4), such that the prediction remained unchanged, but only the level of uncertainty varied.

6.2.5.2 Narrative Scaffolding

To develop our narrative-based scaffolding, we consulted data storytelling experts, prior literature, and identified a key set of narrative elements to communicate uncertainty. We created initial designs, exploring scenarios with different uncertainty levels and distributions. We wanted the narratives to be grounded in the visualizations, but also to be able to be read without depending on the visualizations. Through internal team discussions, we identified the final set of criteria to translate uncertainty visualizations into text-based narratives: type of deal, average, extreme scenarios, meaning of uncertainty.

In *With Stories* conditions, the interface displayed the uncertainty visualizations along with textual narrative blocks displayed to the right. We generated the content of these blocks directly based on the outputs of the models. Therefore, the narratives differed between the with uncertainty and no uncertainty conditions. In the **Story-only** condition (Figure 6.1, B2), the narrative consisted of two blocks. The first stated the AI's overall recommendation in plain language. The second presented the direction (i.e., overpay/underpay) and quantified the difference between the AI's prediction and list price. In the *UQ with Story* condition (Figure 6.1, A2), the narrative consisted of four blocks that consisted of direction, average

magnitude, range, and confidence of the AI's uncertainty distribution.

6.2.5.3 Pilot Study and Results

Prior to finalizing the interface designs, we conducted three iterative pilot studies to uncover any remaining issues with the interface. We recruited (N=8) participants through word-of-mouth and personal networks. The pilots served three purposes: to evaluate the comprehensibility of the study, to check whether the visualizations and stories worked as intended, and to identify any remaining usability issues. During each pilot, participants were asked to complete the study, explain their thought process, and provide feedback on the designs. Each pilot took approximately 45 minutes. Several important findings emerged. First, participants wanted to look at house images when cases were more ambiguous, and said that the look of the house would help them identify if they prefer it. To keep the study objective, we avoided providing any photos or other information about the houses apart from AI predictions. We provided an "I am not sure" button in the decision bar as an alternative. Participants also struggled to reason when the x-axis showed percentages. We tested designs with dollar amounts ($) on the x-axis, which participants found much easier to understand without it impacting the decisions they made. We also observed participants making decisions differently when the ML predicted a price close to the list price vs when it was far away. When probed further, participants reported price thresholds above and below within which they were unsure about the ML prediction. In the final study, we picked the signal strength thresholds (less than 10%, greater than 50%) based on these findings.

Participants appreciated the stories as they helped them identify important things to focus on in the uncertainty visualization, and made them more confident that their interpretation was correct. However, they thought the stories were not adding much more to the line plot. Across all pilots, the consistent finding was that narratives were valuable specifically when they could translate complex information to simple, decision-specific terms for participants. Narratives that exactly repeated what the chart said were not especially useful. Narratives that extracted features of the charts that took effort and presented them in terms of consequences (i.e., how much participants could gain or lose) were helpful. Therefore, our designs worked as intended. The final set of pilots revealed no major usability issues, and we proceeded with these designs in the final study software.

6.2.6 Final Study Software

We implemented our study software as a Django[1] web application running on a dedicated server. We implemented the user interface in HTML, CSS, and JavaScript[2]. We stored the user model, user states, and other user study related data (including logs) in a MySQL[3] database running on a separate server. The participants then accessed our study software through their personal computers (e.g., laptops), with a browser of their choice.

6.2.7 Experimental Conditions

We conducted a 2 (uncertainty: absent vs. present) × 2 (narrative: absent vs. present) within-subjects factorial design, yielding four conditions. In the **Baseline** condition, participants viewed a single-line chart showing the AI's deterministic price prediction relative to the list price, without narrative explanation. In the **Story-only** condition, participants viewed the same chart accompanied by a textual narrative summarizing the AI's recommendation and the magnitude of the predicted price discrepancy. In the **UQ-only** condition, participants viewed a quantile dot plot representing 20 samples from the AI's posterior predictive distribution, without narrative explanation. In the **UQ with Story** condition, participants viewed the quantile dot plot accompanied by a narrative explanation describing the direction, magnitude, and spread of the uncertainty distribution. All four conditions included the AI's buy/don't-buy recommendation.

For each condition, we selected 8 houses from the Ames Housing Dataset according to a fully crossed 2 (type of deal: good/bad) × 2 (ML correctness: correct/incorrect) × 2 (signal strength: weak/strong) factorial grid, yielding one house per cell. Type of deal was determined by comparing the list price to the ground truth sale price recorded in the dataset. A house was classified as a good deal when the ground truth sale price exceeded the list price, and a bad deal when the ground truth was below it. This factor determines the objectively correct recommendation — participants should recommend buy for a good deal and don't buy for a bad deal. ML correctness was determined by comparing the model's price prediction to the list price. The model recommended buy when its prediction exceeded the list price and don't buy otherwise. A recommendation was correct when it agreed with the ground truth deal type, and incorrect otherwise. This factor allowed us to construct cases where following the AI led to the right decision and cases where following it led to the wrong one. Signal strength was determined by the magnitude of the discrepancy between the AI's price prediction and the list price, expressed as a proportion of the list price. A discrepancy

[1] https://www.djangoproject.com/
[2] https://www.w3.org/wiki/The_web_standards_model_-_HTML_CSS_and_JavaScript
[3] https://www.mysql.com/

of less than 10% was classified as weak, indicating that the AI's prediction was close to the list price and its directional recommendation was correspondingly uncertain. A discrepancy exceeding 50% was classified as strong, indicating a clear and confident directional prediction. This factor captures the difficulty of the decision: weak-signal cases are objectively harder to evaluate from the visualization alone.

6.2.8 Participants

We recruited participants through Prolific[4] platform. Our inclusion criteria were native English speakers, above age 18, and present in the US. A total of 105 participants took part in the study, out of which 16 participants failed the attention-check or comprehension-check questions, whose data we excluded from our analysis. This resulted in a final sample of 89 participants. On average, participants took approximately 18 minutes to complete the entire study. Participants were compensated at the fixed rate of $15/hr, which is the minimum wage in the US.

6.2.9 Procedure

We conducted an online behavioral experiment on Prolific. To participate in the study, the participant has to access the study website through Prolific. After the participant clicks on the study link, the study software presents them with the consent form describing the study, expectations, and compensation. Only after the participant clicks on the "I consent" button on the website are they allowed to be part of the study and proceed to the next step.

The website then displayed a description of the task and the base plot (Fig. 5.1). After the participant read the description, they proceeded to a quiz, which asked questions about the task and base plot (e.g., *"What does the black line in the graph represent?"*) to ensure that the participant understood both. The participant had to answer all quiz questions correctly to proceed, but could make multiple attempts until correct. There was no time limit for the quiz, and participants were allowed to return to previous pages if they wished to see the description again.

After successfully completing the quiz, the participant proceeded to the main study. The participant completed the main study by sequentially proceeding through the "training" and "testing" phases. In the training phase, the participants engaged in scaffolded learning and completed study conditions in a fixed order of baseline, story only, UQ only, and UQ with story in increasing order of complexity and cognitive load. The training phase mimicked the learning curve experienced by participants when understanding uncertainty [175]. In the

[4] https://www.prolific.com/

testing phase, the participant's understanding of uncertainty was tested in a fully counterbalanced order. The participant was randomly assigned to one of 4! = 24 combinations of 2x2 conditions to cancel out order effects.

Effectively, the participant completed every condition twice over the training and the testing phases. Before each condition, the website displayed instructions that were specific to that condition. Next, the study website displayed the 8 total stimulus cases, which varied by signal strength, correctness of ML recommendation, and type of deal. For each house, the website displayed the AI's recommendation, a graph corresponding to the uncertainty, and stories (present/absent). The study website asked the participant, "Based on the information provided, would you follow the AI's recommendation?". The study website asked the participant to choose how likely they are to follow the AI from a 5-point Likert scale: definitely yes, yes, unsure, no, or definitely no.

After completing the main task, the participant then completed a brief demographic survey which asked their gender, age, education, and occupation. To ask about gender, we followed guidelines from [202]. We asked about age using a non-overlapping range. For education and occupation, we provided specific options to select from. To estimate participants' expertise, we asked if they had taken any ML and statistics classes, implemented any ML algorithms, and purchased any property. The whole study lasted roughly one hour on average. Our organization's Institutional Review Board (IRB) reviewed and approved the study.

6.2.10 Measures

To answer our research questions, we measure whether participants' decisions align with or depart from AI's recommendation (i.e., AI agreement) and whether they feel confident in their decisions. Here we describe our rationale, how we selected the measures, and how we operationalize them based on our study design and interface.

6.2.10.1 Agreement with AI

We measure agreement with AI as a behavioral measure. We make the important distinction that our study design does not allow us to compute *appropriate reliance*, as we do not ask participants to make a decision first and then with AI. We considered agreement with AI based on whether the participant's recommendation matched the AI's stated recommendation for that house.

6.2.10.2 Confidence in Decisions

We measure confidence in decisions as a behavioral measure, as opposed to a self-reported measure. We considered participants to be not confident if they selected "I am not sure" when making a decision. We considered participants to be confident if they selected either "buy" or "don't buy" for the house. Providing a yes/no option and then asking for confidence would have forced participants to pick a side in either direction. Based on the findings from pilots where participants reported being hesitant to pick a side, we include a button for "I am not sure" to measure *decision inconfidence*, so participants can pick a side with confidence.

6.2.11 Statistical Analysis

We analyzed the training and testing phases separately. For each phase, we fit two mixed-effects logistic regression models with participant-level random intercepts. The first model predicted whether a participant selected *unsure*. The second model predicted whether the participant agreed with the AI recommendation; *unsure* responses were coded as non-agreement.

Let i index participants, j index trials, and $p \in \{\text{training}, \text{testing}\}$ index phase. The models were fit separately for each phase p.

6.2.11.1 Unsure response model

$$Y_{ijp}^{\text{unsure}} = \begin{cases} 1, & \text{if participant } i \text{ selected } \textit{unsure} \text{ on trial } j \text{ in phase } p, \\ 0, & \text{otherwise.} \end{cases} \tag{6.2}$$

$$Y_{ijp}^{\text{unsure}} \sim \text{Bernoulli}\left(\pi_{ijp}^{\text{unsure}}\right) \tag{6.3}$$

$$\text{logit}\left(\pi_{ijp}^{\text{unsure}}\right) = \beta_0 + \boldsymbol{\beta}^\top \mathbf{X}_{ijp} + u_i \tag{6.4}$$

$$u_i \sim \mathcal{N}\left(0, \sigma_u^2\right) \tag{6.5}$$

6.2.11.2 AI agreement model

$$Y_{ijp}^{\text{agree}} = \begin{cases} 1, & \text{if participant } i \text{ agreed with the AI recommendation on trial } j \text{ in phase } p, \\ 0, & \text{otherwise.} \end{cases} \tag{6.6}$$

$$Y_{ijp}^{\text{agree}} \sim \text{Bernoulli}\left(\pi_{ijp}^{\text{agree}}\right) \tag{6.7}$$

$$\text{logit}\left(\pi_{ijp}^{\text{agree}}\right) = \gamma_0 + \boldsymbol{\gamma}^\top \mathbf{X}_{ijp} + v_i \tag{6.8}$$

$$v_i \sim \mathcal{N}\left(0, \sigma_v^2\right) \tag{6.9}$$

Here, $\mathbf{X}_{ijp}$ represents the fixed effects for visualization type, story presence, signal strength, AI uncertainty, and their interactions. Visualization type was either line or dot-plot, story presence was either no or yes, signal strength was either weak or strong, and AI uncertainty was either low or high.

The models were estimated separately for the training and testing phases, and we did not estimate direct training-versus-testing contrasts.

6.3 Findings

Here we present the results from 89 participants who successfully completed the study. In the training and testing phase, the study resulted in 2,848 trials from 89 participants each.

6.3.1 Descriptive Statistics

Of the total 89 participants who completed the study, 53 (59.6%) identified as men, 35 (39.3%) identified as women, and 1 (1.1%) identified as non-binary. Participants ranged in age from 18 to 64, with the largest groups falling in the 25 - 34 (N=30, 33.7%) and 35-44 (N=23, 25.8%) age ranges. Most held at least a four-year college degree (N=38, 42.7%) and were employed full-time (N=51, 57.3%). The majority had no prior ML coursework (N=73, 82%) or experience implementing ML algorithms (N=77, 86.5%), confirming our target population of lay decision-makers. More than half had taken a statistics class (N=58, 65.2%), while about half had prior property purchasing experience (N=44, 49.4%). Participants reported frequent AI use, with half (N=40, 44.9%) using AI tools daily and others (N=33, 37.1%) using AI a few times per week. Mean trust in AI was 4.97 (SD = 1.44) on a 7-point scale (1 = Strongly Disagree, 7 = Strongly Agree), indicating a moderately AI-trusting sample.

6.3.2 Effect of Stories

We fit mixed-effects logistic regression models separately for the training and testing phases, with participant-level random intercepts. For each phase, we modeled two binary outcomes: whether participants selected *unsure*, and whether participants agreed with the AI recommendation. The fixed effects were visualization type, story presence, signal strength, AI uncertainty, and deal type, with planned interactions for visualization × story and signal strength × AI uncertainty. The signal strength × AI uncertainty contrast was estimated at `viz = dotplot`. Planned interaction contrasts used Holm-adjusted p-values.

6.3.3 Confidence in Decisions

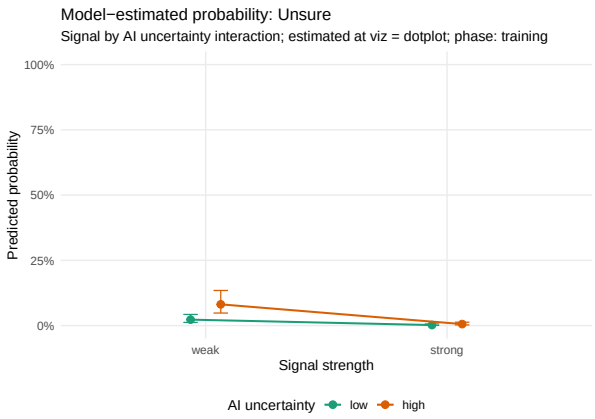
For *unsure* responses, we did not find evidence of a visualization × story interaction in the training phase, OR = 1.58, 95% CI [0.74, 3.36], $z = 1.18$, $p_{\text{Holm}} = .238$. In the testing phase, however, there was evidence of a visualization × story interaction, OR = 2.84, 95% CI [1.16, 6.92], $z = 2.29$, $p_{\text{Holm}} = .0218$. Model-estimated probabilities showed that unsure responses were rare overall, but the dotplot condition was associated with a larger increase in unsure responses when story information was present. In testing, the estimated probability of an unsure response was 0.18% for line + story and 0.72% for dotplot + story, compared with 0.34% for line + no story and 0.47% for dotplot + no story.

6.3.4 AI Agreement

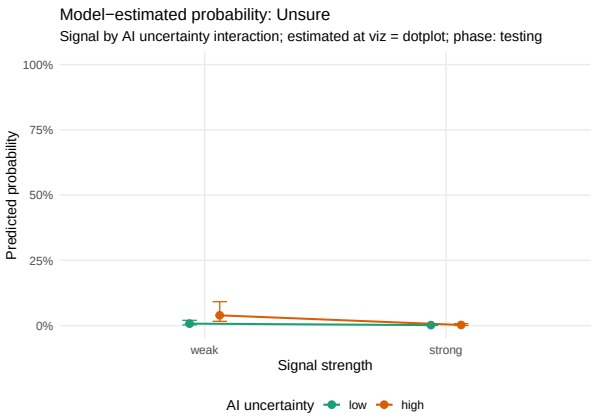
For AI agreement, we found evidence of a visualization × story interaction in both phases. In training, the interaction contrast was OR = 0.54, 95% CI [0.33, 0.89], $z = -2.44$, $p_{\text{Holm}} = .0146$. In testing, the interaction contrast was OR = 0.34, 95% CI [0.18, 0.62], $z = -3.44$, $p_{\text{Holm}} = .000577$. Model-estimated probabilities indicated that agreement was high overall. In training, estimated agreement was 88.35% for line + no story and 93.09% for dotplot + no story, but 95.37% for line + story and 95.21% for dotplot + story. In testing, estimated agreement was 96.57% for line + no story and 97.28% for dotplot + no story, but 98.59% for line + story and 96.77% for dotplot + story. Thus, the relationship between visualization type and AI agreement depended on whether story information was present.

6.3.5 Signal Strength x AI Uncertainty

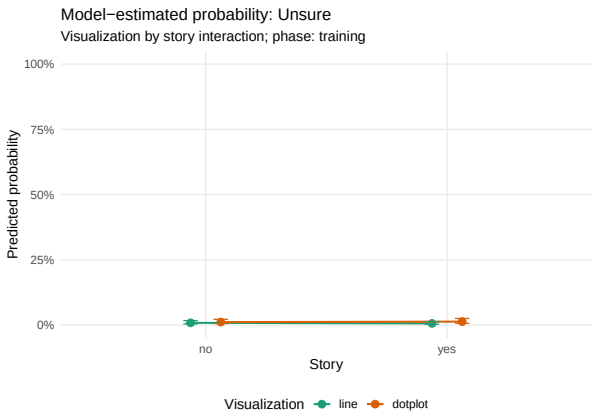
We next examined the planned signal strength × AI uncertainty contrast estimated at `viz = dotplot`. For unsure responses, we did not find evidence of this interaction in training, OR = 0.76, 95% CI [0.22, 2.64], $z = -0.43$, $p_{\text{Holm}} = .665$. In testing, however, there



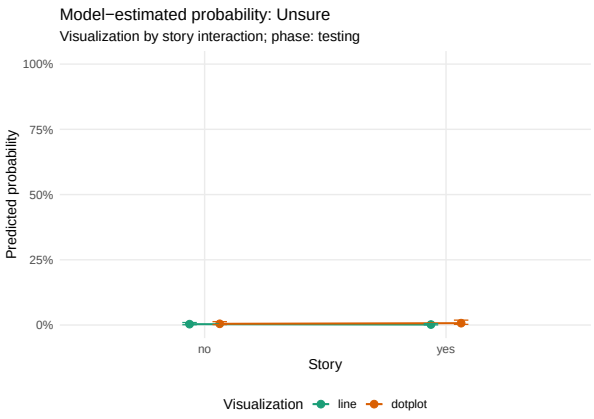
(a) Dotplot, Training



(b) Dotplot, Testing

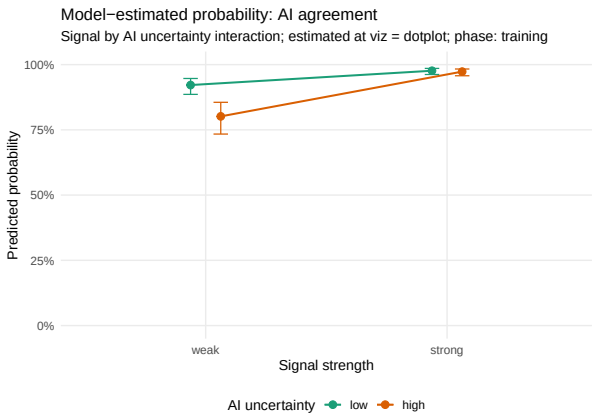


(c) Viz Story, Training

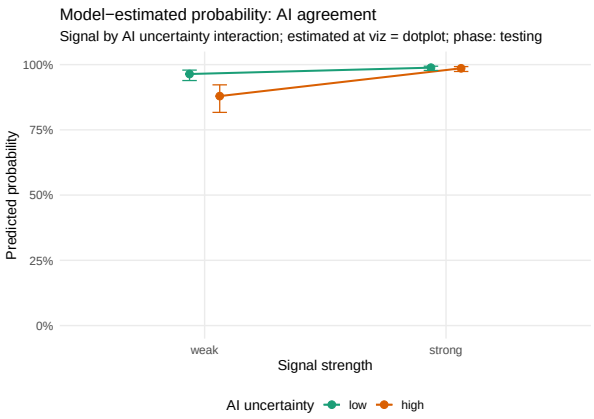


(d) Viz Story, Testing

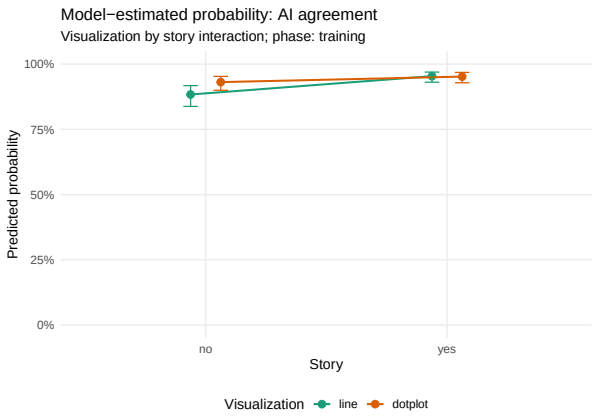
Figure 6.2: Model uncertainty (unsure) with signal uncertainty across conditions.



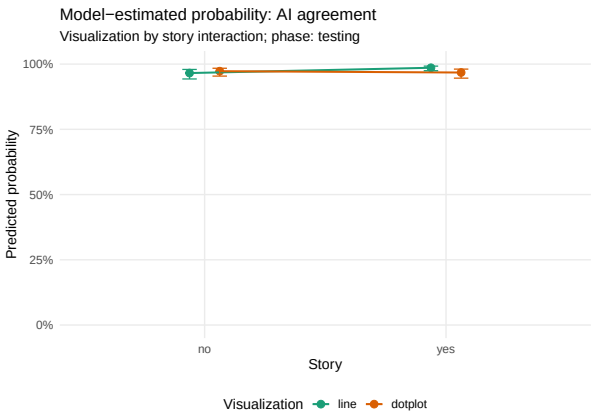
(a) Dotplot, Training



(b) Dotplot, Testing



(c) Viz Story, Training



(d) Viz Story, Testing

Figure 6.3: Model agreement with signal uncertainty across conditions.

was evidence of a signal strength × AI uncertainty interaction, OR = 0.25, 95% CI [0.07, 0.83], $z = -2.26$, $p_{\text{Holm}} = .0238$. Model-estimated probabilities showed that, under dotplot visualization in testing, unsure responses were highest when the signal was weak, and AI uncertainty was high: 3.92%, 95% CI [1.63%, 9.14%], compared with 0.74% for weak signal + low uncertainty, 0.17% for strong signal + low uncertainty, and 0.23% for strong signal + high uncertainty.

For AI agreement under dotplot visualization, we found evidence of a signal strength × AI uncertainty interaction in both phases. In training, the interaction contrast was OR = 2.54, 95% CI [1.48, 4.35], $z = 3.38$, $p_{\text{Holm}} = .000727$. In testing, the interaction contrast was OR = 3.03, 95% CI [1.49, 6.18], $z = 3.05$, $p_{\text{Holm}} = .00227$. Model-estimated probabilities showed that AI agreement was lowest when signal was weak, and AI uncertainty was high. Under dotplot visualization in testing, estimated agreement was 87.93%, 95% CI [81.69%, 92.25%], for weak signal + high uncertainty, compared with 96.40% for weak signal + low uncertainty, 98.83% for strong signal + low uncertainty, and 98.59% for strong signal + high uncertainty.

6.4 Discussion

Our results indicate that participants are more unsure in the story and dotplot condition when they learn how to use the stories (i.e., in the training phase), as stories support the interpretation of uncertainty. Our findings confirm prior work that states simply providing uncertainty representation is not enough, and decision-makers need to be provided additional guidance [189]. Our results indicate that when the uncertainty is high and the signal is weak, participants are more unsure. These effects can be seen across training and testing phases in the study. Therefore, showing uncertainty helps people defer when there is a reason for it. Providing participants with stories in the no uncertainty condition makes them more likely to agree with the AI. Our findings from the pilot study indicate that participants did not find stories helpful when they repeated what the graphs said, but found them much more effective for understanding the dotplots. In the main study, we found stories to have a similar effect to explanations [14, 12]. In this condition, stories may have acted similarly to credibility signals for deterministic AI, and made them agree with the AI, making them more confident.

6.5 Summary

In this work, we studied whether textual explanations of uncertainty visualizations helped non-expert users understand AI uncertainty better than graphical displays alone. In a 2x2 within-subjects study (N=89), participants played the role of helping a hypothetical friend decide whether to buy a house, based on an AI's price prediction. We varied uncertainty (a dotplot vs. a plain line) and stories (present/absent) as our independent variables. Each participant went through a fixed-order "training" phase and a counterbalanced "testing" phase. Our study indicates that narratives paired with the dotplot helped people appropriately defer, specifically in cases when the AI's signal was weak and its uncertainty was high. However, narratives paired with the plain deterministic line chart, which hid uncertainty, increased agreement with the AI, functioning more like a credibility signal rather than a cautionary one. We find that narratives can be a useful "cognitive scaffold" when uncertainty is present, but can mislead decision-makers into relying on AI with high confidence in situations which do not warrant it. Our study calls for further investigation into the design of these narratives.

CHAPTER 7

Discussion

This thesis aimed to investigate *how* non-technical decision-makers interpret AI uncertainty representations, and to *what* extent such understanding influences agreement with AI recommendations.

Our findings indicate that non-technical decision-makers interpret AI uncertainty through mental shortcuts and heuristics rather than as experts represent it: mathematical probabilities. Scaffolding user understanding increases analytical thinking and reduces AI agreement and decision confidence in appropriate situations when uncertainty is communicated, but acts as a credibility signal and increases AI agreement when uncertainty is suppressed.

The above key finding is supported by the thesis's chapters as follows: Chapter 3 illustrates how experts represent uncertainty as mathematical probabilities. Using Bayesian uncertainty quantification (UQ), it is shown that it is possible to compute the uncertainty in AI predictions in the form of probabilistic distributions. Once uncertainty is computed, Chapter 4 investigates the divide between how experts represent AI uncertainty and how this uncertainty is understood and acted upon. Findings show that experts communicate AI uncertainty driven by incentives of precision and accuracy. However, decision-makers need actionability and guidance on how to interpret this information as they lack the same level of statistical and technical expertise. Chapter 5 investigates how AI uncertainty, when quantified and visualized, affects how non-technical decision-makers perceive, understand, and act on this information. Findings indicate that uncertainty can shift decision-makers from passive consumption (heuristic thinking) to more carefully assessing AI outputs (analytical thinking). However, non-technical decision-makers struggle to interpret abstract mathematical probabilities and employ heuristics and mental shortcuts to overcome these challenges. Chapter 6 then investigates how scaffolding mechanisms can offer guidance on uncertainty interpretation and quantifies its effects on decision-making. Findings indicate that such mechanisms reduce AI agreement when its predictions are ambiguous and uncertainty is

high. However, such guidance can act as false credibility when uncertainty is suppressed, increasing agreement with AI regardless of the prediction.

Here, we situate the above findings in the larger discourse on uncertainty communication and responsible human-AI decision-making. We provide suggestions for the design of uncertainty communication in AI as a means to aid algorithmic transparency.

7.1 Representing Uncertainty

Policymakers and advisory bodies recommend communicating uncertainty in AI predictions to decision-makers so that they can know when to act on AI outputs without blindly relying on them. For example, the US Food and Drug Administration (FDA) called for establishing the credibility of AI models by communicating model risks and unintended decision consequences for AI in healthcare [57]. NIST's AI Risk Management Framework [1] recommends that organizations systematically measure AI uncertainty and report it for transparency. However, NIST also states that the problem of inscrutability (i.e., lack of transparency) needs to be tackled first, as AI-based decision-support systems do not inherently communicate uncertainty in their predictions.

Chapter 3 tackles the problem of inscrutability by proposing methods to measure uncertainty of high-dimensional black-box AI models accurately. While uncertainty quantification methods such as Bayesian inference can provide reliable uncertainty estimates [15], they do not scale to million-dimensional AI models. Chapter 3 proposes a computational framework that scales Bayesian UQ to reliably measure uncertainty while keeping computational cost low. Using sensitivity analysis as a heuristic, the framework identifies the layer with the most impact on the output. This layer is then trained with Bayesian inference, while the other layers are trained deterministically. The findings add to the growing body of work on partial Bayesian Neural Networks [56, 192] that make it possible for decision-makers to measure uncertainty given the technical constraints they may have, and obtain more information from black-box AI models.

However, findings from Chapter 4 indicate that availability of accurate uncertainty information does not mean it would be accepted or acted upon in a decision-making context. Factors such as the organizational culture, hierarchy, and authority determine who has the authority to act upon uncertainty information. For example, recommendations driven by uncertainty quantification can be overridden by decisions such as "the commander wants a coffee shop". Funding and institutional priorities also determine whether uncertainty information can be acted upon. Finally, organizational culture sets norms for risk-taking, information flow, and trust in data, which determine whether decision-makers are averse to

uncertainty or receptive to it. Accepting uncertainty depends on who delivers this information and how they do so. For example, organizations trust uncertainty information if delivered by institutions they trust. However, uncertainty estimates computed by unfamiliar institutions would face resistance due to lack of institutional trust. Therefore, despite the availability of uncertainty information, a lot of it is "lost in translation".

Implications: Our findings caution against a "one size fits all" approach to uncertainty communications. Uncertainty representation has to be contextualized within organizational processes and the decision-making task, rather than treated as a downstream problem. Technical and AI experts should consider who will be using the information and what types of decisions will be made based on it before investing resources in computing uncertainty. Uncertainty visualization designers should consider the type of organization that will be adopting the designs, their culture and norms of information sharing. Finally, uncertainty visualization needs to be evaluated not only for understandability but also actionability (i.e., how easily can decision-makers act upon it).

7.2 Understanding Uncertainty

Our findings provide evidence of the expert-user understanding gap in uncertainty communication in the context of human-AI decision-making. Extensive scientific communication literature [58, 223, 185] has described how non-technical decision-makers lack the same degree of statistical literacy as experts and struggle to understand uncertainty. Chapter 5 demonstrates how non-technical decision-makers did not initially understand what uncertainty meant or what the representations signified due to lacking statistical literacy. Even when they overcame a high learning curve by assigning their own meaning to the representations, they did not fully grasp how AI produced the different estimates. These findings indicate that both statistical literacy (i.e., the ability to interpret mathematical information) and AI literacy [150] (i.e., knowledge of capabilities and limitations of AI) are essential to the effectiveness of uncertainty communication for human-AI decision-making.

However, once non-technical decision-makers develop enough statistical literacy through practice, they do not always extract the same information from uncertainty visualization. Our findings demonstrate how people's aversion to probabilistic information [216] translates into how they interpret AI uncertainty information. Chapter 4 finds that decision-makers often do not think of uncertainty at all, or selectively remember worst-case scenarios when managing uncertainty in their daily lives. Building on this, Chapter 5 demonstrates that non-technical decision-makers only extracted higher complexity information (Figure 5.4) when AI predictions were ambiguous. However, they ignored uncertainty information if AI predictions

gave them clear insight into what decision to take. These findings violate the assumption that people are rational actors [222], which many uncertainty visualization mechanisms make [113, 127]. Therefore, visualizing AI uncertainty does not guarantee that decision-makers will make decisions based on the provided information. Rather, non-technical decision-makers make selective use of uncertainty if they cannot resolve the situation with AI predictions alone.

Implications: Our findings suggest that uncertainty communication mechanisms should not assume that the understanding of uncertainty implies its use. Decision-makers may selectively engage with uncertainty, or ignore it despite understanding it. Visualization should aim to communicate uncertainty adaptively, depending on the complexity of the situation in which it is utilized. Evaluation of visualization mechanisms needs to measure how likely decision-makers are to engage with uncertainty in the context of task complexity, cognitive load, and the ambiguity of AI's prediction itself.

7.3 Decision-making with AI Uncertainty

AI developers and technical experts worry that people may ignore or misunderstand uncertainty and often do not report it [111]. The findings of Chapter 5 add evidence to the dangers of suppressing uncertainty information. In complex and uncertain tasks, decision-makers tend to rely more on AI due to recognition of its value combined with lack of confidence in their own ability [187]. This phenomenon is called "enchanted determinism" [45], the belief that AI is superior and its predictions are infallible. When AI uncertainty is suppressed, decision-makers interpret explanations of AI outputs falsely as credibility signals. Chapter 6 shows that explaining deterministic AI outputs (i.e., in the form of stories) leads to an increase in AI agreement in non-technical decision-makers, compared to not explaining the outputs. Our findings are supported by prior work in explainable AI [12, 143], and show how providing scaffolding and suppressing uncertainty can lead to negative effects for decision-making.

Our findings explain why communicating uncertainty helps decision-makers slow down and rely less on AI [135, 106]. Chapter 5 demonstrates how providing uncertainty information fundamentally shifts the way in which decision-makers process information about AI outputs. Instead of heuristic thinking, which leads to agreement with AI recommendations without questioning, uncertainty shifts cognition to analytical thinking, causing decision-makers to take time and think carefully. Uncertainty helps decision-makers estimate when the model is not confident, and exercise their own expertise in making the decision. Chapter 6 shows the magnitude of this effect. Scaffolding that explained graphical uncertainty made people more unsure and override AI recommendations in situations where it was appropriate to do

so (i.e., where uncertainty was high and AI recommendations were ambiguous).

Findings of Chapter 5 and Chapter 6 demonstrate how explanations and uncertainty representations interact to increase or decrease enchanted determinism [45]. Suppressing uncertainty and providing explanations leads to the perception that AI outputs are credible, pushing decision-makers towards accepting them without questioning. However, when uncertainty representations are paired with explanations, they invite decision-makers to scrutinize AI outputs more carefully, reducing AI agreement and decision confidence. Our findings have implications for communicating uncertainty in high-stakes decision-making contexts where explanations are important to contextualize AI outputs. For example, if the diagnosis of a healthcare AI algorithm for patients has high uncertainty, it can explain whether the source of uncertainty is epistemic (e.g., an imaging artifact due to patient movement) or aleatoric (e.g., lack of sufficient training data for a specific patient subpopulation).

Implications: Our findings argue that explanations and uncertainty need to be communicated together. This would assist decision-makers in understanding not only that AI is uncertain, but why and under what conditions. Future work can design explanations of uncertainty and evaluate how it changes decision-makers' cognition and understanding of uncertainty.

CHAPTER 8

Conclusion and Future Work

AI-based Decision Support Systems (DSS) are increasingly being adopted across several domains, from healthcare [117] and climate change [199] to law [68]. Across these domains, they promise to assist decision-makers, especially those without a technical background, in making data-driven decisions. Despite the consequences of wrong decisions being catastrophic in such domains, existing AI-based DSS are black-box and hide the uncertainty in their predictions. Such black-box AI-based DSS can project a sense of authority [118] and infallibility [45]. This lack of transparency can make it difficult for non-technical decision-makers to determine when to accept AI recommendations and when not to [12]. Prior work in HCI has proposed uncertainty as a form of transparency [24], and developed mechanisms to quantify [15] and visualize [113, 131] uncertainty in AI predictions. Such approaches treat uncertainty as a *representation* problem. Such work tackles the problem of computing uncertainty mathematically and communicating it to decision-makers, presuming that decision-makers can make better decisions using this information.

This thesis provides empirical evidence that non-technical decision-makers rely on mental shortcuts and heuristics rather than statistical information when interpreting uncertainty in AI predictions. Scaffolding their interpretation leads to analytical thinking and reduces agreement with AI recommendations without questioning, whereas suppressing this information is more common in complex scenarios under high AI uncertainty. I first computed the uncertainty in black-box AI predictions by scaling Bayesian approaches to neural networks. Next, I investigated existing practices of communicating this uncertainty in numerical format and its limitations. I found that the mechanism of communication needs to translate probabilities into terms that decision-makers understand, such as visualization. Next, I investigated how non-technical decision-makers perceive, reason with, and act on graphical uncertainty representations. I found that uncertainty can guide users to think more carefully, but understanding graphical uncertainty is still difficult. Finally, I designed stories that explain uncertainty visualization and investigated how it impacts decisions. I found

that stories can serve as false credibility signals and increase AI agreement when uncertainty is suppressed. However, when uncertainty is communicated, they can reduce AI agreement and decision confidence in complex situations of high uncertainty. Overall, this thesis offers a complementary perspective to uncertainty representation by focusing on how it is understood and acted upon in the context of responsible human-AI decision-making.

8.1 Thesis Contributions

- **Empirical:** The primary contribution of this thesis is empirical knowledge on how lay decision-makers perceive and use uncertainty visualizations in AI-assisted decision-making. Chapter 5 demonstrates the effect of uncertainty on decision-makers' perception, understanding, and judgement with AI. Chapter 6 quantifies the magnitude of this effect, showing how uncertainty affects AI agreement and decision confidence. Together, findings make a case for why uncertainty communication needs to be designed for user understanding and decision context.

- **Epistemological:** This thesis also contributes an epistemological understanding of how uncertainty is defined, treated, and managed across disciplines of knowledge production. Chapter 4 investigates how social science, human-centered design and basic sciences determine whether uncertainty is conceptualized in terms of accuracy, lived experience, or a mediator of human experience with technology. While disciplinary understanding stems from long-established traditions, interdisciplinary work necessitates shared vocabulary and interrogation of what counts as valid knowledge.

- **Methodological:** The thesis makes a methodological contribution in the form of a computational framework for quantifying AI uncertainty at scale. Chapter 3 demonstrates how Bayesian UQ can be scaled to million-dimensional AI models by partial Bayesian inference informed by sensitivity analysis. This method enables decision-makers with computational resource constraints to compute mathematically rigorous uncertainty estimates for black-box AI models.

- **Design:** The thesis makes design contributions in two ways: First, in Chapter 5 by providing an evaluation of graphical uncertainty representation from a cognitive psychology lens. Second, in Chapter 6 by demonstrating narrative scaffolding as a way to make uncertainty more actionable for nontechnical users. The thesis provides implications for how uncertainty communication mechanisms should be designed and evaluated by considering the interpretation of uncertainty by non-technical decision-makers.

8.2 Future Work

The findings in this dissertation point toward several directions for future research. Here, I outline directions for further exploration.

8.2.1 Decision-centric Uncertainty Communication

Uncertainty communication frameworks need to operationalize abstract mathematical probabilities in terms of decision utility [38]. For example, instead of communicating AI uncertainty as numbers, visualization can translate it into financial or operational costs. Communication mechanisms need to consider the type of decision and its characteristics. For example, operational decisions have tight time frames and very low uncertainty tolerance. In such situations, AI uncertainty can be communicated by simple color-coding or flagging anomalies. Conversely, strategic decisions have long time frames and higher uncertainty tolerance. AI uncertainty can be communicated as what-if scenarios or future possibilities to encourage deeper reflection and discussion. To communicate uncertainty in a way that aligns with organizational culture, training materials and normative guidelines can inform decision-makers how they are supposed to handle situations of differing uncertainty and how much uncertainty to tolerate.

8.2.2 Scaffolding User Understanding of Uncertainty

Our findings indicate that additional support is needed to help decision-makers understand how to interpret uncertainty. Prior work has found that solely providing uncertainty representations is insufficient to improve decision quality [46], and additional guidance and scaffolding are needed. Our work offers scaffolding mechanisms as learner-centered guidance [189] for interpreting uncertainty. In Chapter 6, we designed stories based on how decision-makers extract and reason about uncertainty information from graphs once they gain familiarity (Chapter 5, Figure 5.4). This thesis points to the need to create effective scaffolding mechanisms that improve understanding of uncertainty. For example, mechanisms such as self-explanation and self-reflection [154] can increase engagement with different aspects of graphical uncertainty information. Thus, by designing scaffolding mechanisms that target both statistical and AI literacy, visualizations can guide non-technical decision-makers to interpret uncertainty effectively.

8.2.3 AI-Generated Narratives for Uncertainty Communication

Uncertainty communication is essential for enabling people to make informed decisions with AI systems, yet the dominant paradigm relies on probability distributions and statistical visualizations. Chapter 5 shows how lay decision-makers without a statistical background have a steep learning curve when first using these methods. However, Chapter 4 shows that people make sense of uncertainty through intuitive, lived experiences rather than abstract numbers. This gap between how experts want to communicate uncertainty and how users understand it is an especially big problem in the field of disaster risk communication. Future work can reframe uncertainty communication as a design problem by answering: *how can we create representations that translate complex probabilistic information into forms that people can immediately understand and act upon?* By using large language models (LLMs), AI can transform probability distributions into intuitive, scenario-based descriptions (e.g., narratives, map-based visualizations) as design interventions. Future work can then measure the impact of these design interventions on user comprehension, decision quality, and AI reliance with an online between-subjects study. By leveraging design as a means to translate "numbers" into "stories and pictures", the insights of this thesis can create new methods and principles for intuitive uncertainty communication.

8.2.4 Human-centered Evaluation of Uncertainty Visualization

Chapter 5 shows that people are uncertainty-averse, and tend to avoid complex probabilistic reasoning under uncertainty. Decision-makers seek to reduce complex decision-making to simpler judgments by relying on simple rules-of-thumb, called heuristics. Heuristics save mental effort but can lead to cognitive biases [216], causing severe errors in judgement in certain situations. Prior work evaluates uncertainty visualization by measuring user performance and subjective satisfaction [110]. This thesis adds a complementary perspective on evaluating uncertainty visualization based on user cognition. Instead of subjective metrics and narrow tasks, future work can evaluate uncertainty visualization using a mixed-methods approach that addresses both cognition and its effects on decisions. For example, future work can conduct think-aloud or semi-structured interview methods to identify the cognitive biases when decision-makers interpret visualizations. Using follow-up quantitative methods, future work can capture the degree to which these biases affect decision-making.

BIBLIOGRAPHY

[1] NIST AI. Artificial intelligence risk management framework (ai rmf 1.0). *URL: https://nvlpubs. nist. gov/nistpubs/ai/nist. ai*, pages 100–1, 2023.

[2] Haldun Akoglu. User's guide to correlation coefficients. *Turkish Journal of Emergency Medicine*, 18(3):91–93, September 2018.

[3] Saleema Amershi. Toward responsible ai by planning to fail. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD '20, page 3607, New York, NY, USA, 2020. Association for Computing Machinery.

[4] Saleema Amershi, Dan Weld, Mihaela Vorvoreanu, Adam Fourney, Besmira Nushi, Penny Collisson, Jina Suh, Shamsi Iqbal, Paul N. Bennett, Kori Inkpen, Jaime Teevan, Ruth Kikin-Gil, and Eric Horvitz. Guidelines for human-ai interaction. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, CHI '19, page 1–13, New York, NY, USA, 2019. Association for Computing Machinery.

[5] Lorin W Anderson, David R Krathwohl, Peter W Airasian, Kathleen A Cruikshank, Richard E Mayer, Paul R Pintrich, James Raths, and Merlin C Wittrock. *A taxonomy for learning, teaching, and assessing: A revision of Bloom's taxonomy of educational objectives, complete edition*. Addison Wesley Longman, Inc., New York, 2001.

[6] Christophe Andrieu, Nando de Freitas, Arnaud Doucet, and Michael I. Jordan. An Introduction to MCMC for Machine Learning. *Machine Learning*, 50:5–43, 2003.

[7] Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. Machine bias, May 2016.

[8] Matthew Arnold, Rachel K. E. Bellamy, Michael Hind, Stephanie Houde, Sameep Mehta, Aleksandra Mojsilovic, Ravi Nair, Karthikeyan Natesan Ramamurthy, Darrell Reimer, Alexandra Olteanu, David Piorkowski, Jason Tsay, and Kush R. Varshney. Factsheets: Increasing trust in ai services through supplier's declarations of conformity, 2018.

[9] Syed Z. Arshad, Jianlong Zhou, Constant Bridon, Fang Chen, and Yang Wang. Investigating user confidence for uncertainty presentation in predictive decision making. In *Proceedings of the Annual Meeting of the Australian Special Interest Group for Computer Human Interaction*, OzCHI '15, page 352–360, New York, NY, USA, 2015. Association for Computing Machinery.

[10] Kamyar Azizzadenesheli, Emma Brunskill, and Animashree Anandkumar. Efficient exploration through Bayesian deep Q-networks. In *2018 Information Theory and Applications Workshop, ITA 2018*, 2018.

[11] Justyna Bandola-Gill, Megan Arthur, and Rhodri Ivor Leng. What is co-production? Conceptualising and understanding co-production of knowledge and policy across different theoretical perspectives. *Evidence Policy*, 19(2):275–298, 2023.

[12] Nikola Banovic, Zhuoran Yang, Aditya Ramesh, and Alice Liu. Being trustworthy is not enough: How untrustworthy artificial intelligence (ai) can deceive the end-users and gain their trust. *Proc. ACM Hum.-Comput. Interact.*, 7(CSCW1), apr 2023.

[13] Gagan Bansal, Besmira Nushi, Ece Kamar, Daniel S. Weld, Walter S. Lasecki, and Eric Horvitz. Updates in human-ai teams: Understanding and addressing the performance/compatibility tradeoff. In *Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence and Thirty-First Innovative Applications of Artificial Intelligence Conference and Ninth AAAI Symposium on Educational Advances in Artificial Intelligence*, AAAI'19/IAAI'19/EAAI'19, Honolulu, Hawaii, USA, 2019. AAAI Press.

[14] 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*, CHI '21, New York, NY, USA, 2021. Association for Computing Machinery.

[15] Edmon Begoli, Tanmoy Bhattacharya, and Dimitri Kusnezov. The need for uncertainty quantification in machine-assisted medical decision making. *Nature Machine Intelligence*, 1(1):20–23, jan 2019.

[16] Sarah Belia, Fiona Fidler, Jennifer Williams, and Geoff Cumming. Researchers misunderstand confidence intervals and standard error bars. *Psychological Methods*, 10(4):389–396, 2005.

[17] 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*, FAccT '21, page 610–623, New York, NY, USA, 2021. Association for Computing Machinery.

[18] Stan Benjamens, Pranavsingh Dhunnoo, and Bertalan Meskó. The state of artificial intelligence-based FDA-approved medical devices and algorithms: an online database. *npj Digital Medicine*, 3(1), September 2020.

[19] Jesse Josua Benjamin, Arne Berger, Nick Merrill, and James Pierce. Machine learning uncertainty as a design material: A post-phenomenological inquiry. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, CHI '21, New York, NY, USA, 2021. Association for Computing Machinery.

[20] Ruha Benjamin. *Race After Technology: Abolitionist Tools for the New Jim Code*. Polity Press, Cambridge, UK, July 2019.

[21] James O. Berger. *Statistical Decision Theory and Bayesian Analysis*. Springer Series in Statistics. Springer New York, New York, NY, 1985.

[22] Jose M. Bernardo and Adrian F. M. Smith. *Bayesian Theory*. John Wiley & Sons, New York, NY, 2000.

[23] Michael Betancourt. A Conceptual Introduction to Hamiltonian Monte Carlo. *arXiv preprint arXiv:1701.02434*, 2017.

[24] Umang Bhatt, Javier Antorán, Yunfeng Zhang, Q. Vera Liao, Prasanna Sattigeri, Riccardo Fogliato, Gabrielle Melançon, Ranganath Krishnan, Jason Stanley, Omesh Tickoo, Lama Nachman, Rumi Chunara, Madhulika Srikumar, Adrian Weller, and Alice Xiang. Uncertainty as a form of transparency: Measuring, communicating, and using uncertainty. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '21, page 401–413, New York, NY, USA, 2021. Association for Computing Machinery.

[25] Umang Bhatt, Javier Antorán, Yunfeng Zhang, Q. Vera Liao, Prasanna Sattigeri, Riccardo Fogliato, Gabrielle Melançon, Ranganath Krishnan, Jason Stanley, Omesh Tickoo, Lama Nachman, Rumi Chunara, Madhulika Srikumar, Adrian Weller, and Alice Xiang. Uncertainty as a form of transparency: Measuring, communicating, and using uncertainty. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '21, page 401–413, New York, NY, USA, 2021. Association for Computing Machinery.

[26] Abeba Birhane. The Impossibility of Automating Ambiguity. *Artificial Life*, 27(1):44–61, 06 2021.

[27] Alan F Blackwell. Designing knowledge: An interdisciplinary experiment in research infrastructure for shared description. Technical report, University of Cambridge, Computer Laboratory, 2006.

[28] Fonnet E Bleeker, Remco J Molenaar, and Sieger Leenstra. Recent advances in the molecular understanding of glioblastoma. *Journal of neuro-oncology*, 108(1):11–27, 2012.

[29] David M. Blei, Alp Kucukelbir, and Jon D. McAuliffe. Variational Inference: A Review for Statisticians. *Journal of the American Statistical Association*, 112(518):859–877, 2017.

[30] Charles Blundell, Julien Cornebise, Koray Kavukcuoglu, and Daan Wierstra. Weight Uncertainty in Neural Networks. In *Proceedings of the 32nd International Conference on Machine Learning*, volume 37, pages 1613–1622, 2015.

[31] Robin Bodenberger and Kirsten Thommes. Words or numbers? how framing uncertainties affects risk assessment and decision-making. *Journal of Risk Research*, 29(6):807–830, May 2026.

[32] Ryan David Bowler, Benjamin Bach, and Larissa Pschetz. Exploring uncertainty in digital scheduling, and the wider implications of unrepresented temporalities in hci. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, CHI '22, New York, NY, USA, 2022. Association for Computing Machinery.

[33] Virginia Braun and Victoria Clarke. Using thematic analysis in psychology. *Qualitative Research in Psychology*, 3(2):77–101, 2006.

[34] Virginia Braun, Victoria Clarke, Nikki Hayfield, Louise Davey, and Elizabeth Jenkinson. Doing Reflexive Thematic Analysis. In Sofie Bager-Charleson and Alistair McBeath, editors, *Supporting Research in Counselling and Psychotherapy : Qualitative, Quantitative, and Mixed Methods Research*, pages 19–38. Springer International Publishing, Cham, 2022.

[35] Virginia Braun, Victoria Clarke, Nikki Hayfield, and Gareth Terry. Doing reflexive ta. https://www.thematicanalysis.net/doing-reflexive-ta/, n.d. Website. Accessed August 5, 2025.

[36] Gregor Broll and Steve Benford. Seamful design for location-based mobile games. In Fumio Kishino, Yoshifumi Kitamura, Hirokazu Kato, and Noriko Nagata, editors, *Entertainment Computing - ICEC 2005*, pages 155–166, Berlin, Heidelberg, 2005. Springer Berlin Heidelberg.

[37] Steve Brooks, Andrew Gelman, Galin Jones, and Xiao-Li Meng, editors. *Handbook of Markov Chain Monte Carlo*. Chapman and Hall, 2011.

[38] Stephen B. Broomell, Emily Ho, Daniel M. Benjamin, and David V. Budescu. *Expert Judgment and Communication of Uncertainty*, page 237–246. Springer Nature Switzerland, 2025.

[39] Zana Buçinca, Phoebe Lin, Krzysztof Z. Gajos, and Elena L. Glassman. Proxy tasks and subjective measures can be misleading in evaluating explainable ai systems. In *Proceedings of the 25th International Conference on Intelligent User Interfaces*, IUI '20, page 454–464, New York, NY, USA, 2020. Association for Computing Machinery.

[40] 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. *Proc. ACM Hum.-Comput. Interact.*, 5(CSCW1), apr 2021.

[41] David V. Budescu and Thomas S. Wallsten. *Processing Linguistic Probabilities: General Principles and Empirical Evidence*, page 275–318. Elsevier, 1995.

[42] Adrian Bussone, Simone Stumpf, and Dympna 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, 2015.

[43] 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), nov 2019.

[44] Carrie J Cai, Samantha Winter, David Steiner, Lauren Wilcox, and Michael Terry. Onboarding materials as cross-functional boundary objects for developing ai assistants. In *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems*, CHI EA '21, New York, NY, USA, 2021. Association for Computing Machinery.

[45] Alexander Campolo and Kate Crawford. Enchanted determinism: Power without responsibility in artificial intelligence. *Engaging Science, Technology, and Society*, 6:1–19, January 2020.

[46] Shiye Cao, Anqi Liu, and Chien-Ming Huang. Designing for appropriate reliance: The roles of ai uncertainty presentation, initial user decision, and user demographics in ai-assisted decision-making. *Proc. ACM Hum.-Comput. Interact.*, 8(CSCW1), April 2024.

[47] 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, page 1721–1730, New York, NY, USA, 2015. Association for Computing Machinery.

[48] Matthew Chalmers. Seamful design and ubicomp infrastructure. 2003.

[49] K Charmaz. Grounded theory: Objectivist and constructivist methods, in (eds, denzin, nk and lincoln, ys) grounded theory: Objectivist and constructivist methods, pp249-91, 2003.

[50] Dean De Cock. Ames, iowa: Alternative to the boston housing data as an end of semester regression project. *Journal of Statistics Education*, 19(3):null, 2011.

[51] Michael Correll, Dominik Moritz, and Jeffrey Heer. Value-suppressing uncertainty palettes. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, CHI '18, page 1–11, New York, NY, USA, 2018. Association for Computing Machinery.

[52] National Research Council. *Enhancing the Effectiveness of Team Science*.

[53] Ray Crozier, Rob Ranyard, and Ola Svenson. *Decision making : cognitive models and explanations*. Routledge, 1st edition edition, 1997.

[54] W.R. Crum, Oscar Camara, and D.L.G. Hill. Generalized Overlap Measures for Evaluation and Validation in Medical Image Analysis. *IEEE Transactions on Medical Imaging*, 25(11):1451–1461, nov 2006.

[55] 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*, CHI '23, New York, NY, USA, 2023. Association for Computing Machinery.

[56] Erik Daxberger, Eric Nalisnick, James U Allingham, Javier Antoran, and Jose Miguel Hernandez-Lobato. Bayesian deep learning via subnetwork inference. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 2510–2521. PMLR, 18–24 Jul 2021.

[57] Anjali D Deshmukh and Jennifer K Wagner. Fda draft guidelines for ai and the need for ethical frameworks. *JAMA pediatrics*, 179(9):937–, 2025.

[58] Annal Dhungana, Emma E. H. Doyle, Raj Prasanna, and Garry McDonald. From scientific models to decisions: exploring uncertainty communication gaps between scientists and decision-makers. *Environment Systems and Decisions*, 45(3), 2025.

[59] Lee R. Dice. Measures of the Amount of Ecologic Association Between Species. *Ecology*, 26(3):297–302, jul 1945.

[60] Joshua V. Dillon, Ian Langmore, Dustin Tran, Eugene Brevdo, Srinivas Vasudevan, Dave Moore, Brian Patton, Alex Alemi, Matt Hoffman, and Rif A. Saurous. Tensorflow distributions, 2017.

[61] Finale Doshi-Velez and Been Kim. Towards A Rigorous Science of Interpretable Machine Learning. feb 2017.

[62] Finale Doshi-Velez and Been Kim. Towards a rigorous science of interpretable machine learning, 2017.

[63] Timothy Dozat. Incorporating Nesterov Momentum into Adam. *ICLR Workshop*, 2016.

[64] Upol Ehsan, Q. Vera Liao, Michael Muller, Mark O. Riedl, and Justin D. Weisz. Expanding explainability: Towards social transparency in ai systems. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, CHI '21, New York, NY, USA, 2021. Association for Computing Machinery.

[65] Upol Ehsan, Q. Vera Liao, Michael Muller, Mark O. Riedl, and Justin D. Weisz. Expanding explainability: Towards social transparency in ai systems. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, CHI '21, New York, NY, USA, 2021. Association for Computing Machinery.

[66] K Anders Ericsson and Herbert A Simon. *Protocol analysis: Verbal reports as data.* The MIT Press, Cambridge, MA, US, 1984.

[67] Nel Escher and Nikola Banovic. Exposing error in poverty management technology: A method for auditing government benefits screening tools. *Proc. ACM Hum.-Comput. Interact.*, 4(CSCW1), may 2020.

[68] Nel Escher, Jeffrey Bilik, Nikola Banovic, and Ben Green. Code-ifying the law: How disciplinary divides afflict the development of legal software. *Proc. ACM Hum.-Comput. Interact.*, 8(CSCW2), November 2024.

[69] Motahhare Eslami. Embracing seamfulness and uncertainty in designing around hidden algorithms. 2017.

[70] Kawin Ethayarajh. Is Your Classifier Actually Biased? Measuring Fairness under Uncertainty with Bernstein Bounds. Apr 2020.

[71] Virginia Eubanks. *Automating inequality: How high-tech tools profile, police, and punish the poor*. St. Martin's Press, 2018.

[72] Chi Feng and Youssef M. Marzouk. A layered multiple importance sampling scheme for focused optimal Bayesian experimental design. *arXiv preprint arXiv:1903.11187*, 2019.

[73] Michael Fernandes, Logan Walls, Sean Munson, Jessica Hullman, and Matthew Ka. Uncertainty Displays Using Quantile Dotplots or CDFs Improve Transit Decision-Making. 2018.

[74] Samuel G. Finlayson, Adarsh Subbaswamy, Karandeep Singh, John Bowers, Annabel Kupke, Jonathan Zittrain, Isaac S. Kohane, and Suchi Saria. The clinician and dataset shift in artificial intelligence. *New England Journal of Medicine*, 385(3):283–286, July 2021.

[75] Milton Friedman and L. J. Savage. The utility analysis of choices involving risk. *Journal of Political Economy*, 56, 1948.

[76] Michael Froehlich, Charlotte Kobiella, Albrecht Schmidt, and Florian Alt. Is it better with onboarding? improving first-time cryptocurrency app experiences. In *Designing Interactive Systems Conference 2021*, DIS '21, page 78–89, New York, NY, USA, 2021. Association for Computing Machinery.

[77] Jasmina Gajcin, Jovan Jeromela, Joel Wester, Sarah Schömbs, Styliani Kleanthous, Karthikeyan Natesan Ramamurthy, Hanna Hauptmann, and Rifat Mehreen Amin. Cure2026: Communicating uncertainty to foster realistic expectations via human-centered design. In *Companion Proceedings of the 31st International Conference on Intelligent User Interfaces*, IUI '26 Companion, page 243–247, New York, NY, USA, 2026. Association for Computing Machinery.

[78] Yarin Gal. *Uncertainty in Deep Learning*. PhD thesis, University of Cambridge, 2016.

[79] Yarin Gal and Zoubin Ghahramani. Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning. In *33rd International Conference on Machine Learning, ICML 2016*, volume 3, pages 1651–1660, 2016.

[80] Marzyeh Ghassemi, Tristan Naumann, Peter Schulam, Andrew L. Beam, and Rajesh Ranganath. Opportunities in machine learning for healthcare. *CoRR*, abs/1806.00388, 2018.

[81] Soumya Ghosh, Q. Vera Liao, Karthikeyan Natesan Ramamurthy, Jiri Navratil, Prasanna Sattigeri, Kush Varshney, and Yunfeng Zhang. Uncertainty quantification 360: A hands-on tutorial. In *5th Joint International Conference on Data Science &; Management of Data (9th ACM IKDD CODS and 27th COMAD)*, CODS-COMAD 2022, page 333–335, New York, NY, USA, 2022. Association for Computing Machinery.

[82] Lisa Gitelman. *Raw data is an oxymoron*. MIT press, 2013.

[83] Cleotilde Gonzalez. Decision-making: A cognitive science perspective. In *The Oxford handbook of cognitive science.*, pages 249–263. Oxford University Press, New York, NY, US, 2017.

[84] Germán González Rodríguez, Jose M Gonzalez-Cava, and Juan Albino Méndez Pérez. An intelligent decision support system for production planning based on machine learning. *Journal of Intelligent Manufacturing*, 31(5):1257–1273, 2020.

[85] Mitchell L. Gordon, Kaitlyn Zhou, Kayur Patel, Tatsunori Hashimoto, and Michael S. Bernstein. The disagreement deconvolution: Bringing machine learning performance metrics in line with reality. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, CHI '21, New York, NY, USA, 2021. Association for Computing Machinery.

[86] George Anthony Gorry and Michael S. Scott Morton. A framework for management information systems. Working paper, 1971.

[87] Alex Graves. Practical Variational Inference for Neural Networks. In *Advances in Neural Information Processing Systems 24 (NIPS 2011)*, pages 2348–2356, Granada, Spain, 2011.

[88] Ben Green. The flaws of policies requiring human oversight of government algorithms. *Computer Law & Security Review*, 45:105681, 2022.

[89] Ben Green and Yiling Chen. The principles and limits of algorithm-in-the-loop decision making. *Proc. ACM Hum.-Comput. Interact.*, 3(CSCW), nov 2019.

[90] Miriam Greis, Passant El. Agroudy, Hendrik Schuff, Tonja Machulla, and Albrecht Schmidt. Decision-making under uncertainty: How the amount of presented uncertainty influences user behavior. In *Proceedings of the 9th Nordic Conference on Human-Computer Interaction*, NordiCHI '16, New York, NY, USA, 2016. Association for Computing Machinery.

[91] Miriam Greis, Jessica Hullman, Michael Correll, Matthew Kay, and Orit Shaer. Designing for uncertainty in hci: When does uncertainty help? In *Proceedings of the 2017 CHI Conference Extended Abstracts on Human Factors in Computing Systems*, CHI EA '17, page 593–600, New York, NY, USA, 2017. Association for Computing Machinery.

[92] Miriam Greis, Aditi Joshi, Ken Singer, Albrecht Schmidt, and Tonja Machulla. Uncertainty visualization influences how humans aggregate discrepant information. In *Conference on Human Factors in Computing Systems - Proceedings*, 2018.

[93] Frances Griffiths, Eileen Green, and Maria Tsouroufli. The nature of medical evidence and its inherent uncertainty for the clinical consultation: qualitative study. *BMJ*, 330(7490):511, January 2005.

[94] Sophie Grimme, Susanna Marie Spoerl, Susanne Boll, and Marion Koelle. My data, my choice, my insights: Women's requirements when collecting, interpreting and sharing their personal health data. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, CHI '24, New York, NY, USA, 2024. Association for Computing Machinery.

[95] Shunan Guo, Fan Du, Sana Malik, Eunyee Koh, Sungchul Kim, Zhicheng Liu, Donghyun Kim, Hongyuan Zha, and Nan Cao. Visualizing uncertainty and alternatives in event sequence predictions. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, CHI '19, page 1–12, New York, NY, USA, 2019. Association for Computing Machinery.

[96] Intisar Rizwan I Haque and Jeremiah Neubert. Deep learning approaches to biomedical image segmentation. *Informatics in Medicine Unlocked*, 18:100297, 2020.

[97] WK Hastings. Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, 57(1):97–109, 1970.

[98] Mohammad Havaei, Axel Davy, David Warde-Farley, Antoine Biard, Aaron Courville, Yoshua Bengio, Chris Pal, Pierre-Marc Jodoin, and Hugo Larochelle. Brain tumor segmentation with deep neural networks. *Medical image analysis*, 35:18–31, 2017.

[99] Sarah T. Hawley, Brian Zikmund-Fisher, Peter Ubel, Aleksandra Jancovic, Todd Lucas, and Angela Fagerlin. The impact of the format of graphical presentation on health-related knowledge and treatment choices. *Patient Education and Counseling*, 73(3):448–455, December 2008.

[100] Nicole Hengesbach. Undoing seamlessness: Exploring seams for critical visualization. In *Extended Abstracts of the 2022 CHI Conference on Human Factors in Computing Systems*, CHI EA '22, New York, NY, USA, 2022. Association for Computing Machinery.

[101] Morten Hertzum, Kristin D Hansen, and Hans H K Andersen. Scrutinising usability evaluation: does thinking aloud affect behaviour and mental workload? *Behaviour & Information Technology*, 28(2):165–181, 2009.

[102] Mohammad Hesam Hesamian, Wenjing Jia, Xiangjian He, and Paul Kennedy. Deep learning techniques for medical image segmentation: Achievements and challenges. *Journal of digital imaging*, 32(4):582–596, 2019.

[103] Rebecca Hess, Vivianne H. M. Visschers, and Michael Siegrist. Risk communication with pictographs: The role of numeracy and graph processing. *Judgment and Decision Making*, 6(3):263–274, April 2011.

[104] M Hino, E Benami, and N Brooks. Machine learning for environmental monitoring. *Nature Sustainability*, 1(10):583–588, 2018.

[105] Matthew D Hoffman and Andrew Gelman. The no-u-turn sampler: adaptively setting path lengths in hamiltonian monte carlo. *J. Mach. Learn. Res.*, 15(1):1593–1623, 2014.

[106] Joshua Holstein, Lars Böcking, Philipp Spitzer, Niklas Kühl, Michael Vössing, and Gerhard Satzger. Balancing the unknown: Exploring human reliance on ai advice under aleatoric and epistemic uncertainty. *ACM Trans. Comput.-Hum. Interact.*, 32(6), December 2025.

[107] Ferlay J Shin HR, F Bray, D Forman, C Mathers, and DM Parkin. Estimates of worldwide burden of cancer in 2008: Globocan 2008. *Int J Cancer*, 127(12):2893–917, 2010.

[108] Oliver Huang, Muhammad Fatir, Tianyu Luo, Sangho Suh, Hariharan Subramonyam, and Carolina Nobre. Narrative scaffolding: A narrative-first framework for data-driven sensemaking. In *Proceedings of the 31st International Conference on Intelligent User Interfaces*, IUI '26, page 1112–1140, New York, NY, USA, 2026. Association for Computing Machinery.

[109] Jessica Hullman. Why Authors Don't Visualize Uncertainty.

[110] Jessica Hullman. Why evaluating uncertainty visualization is error prone. In *ACM International Conference Proceeding Series*, 2016.

[111] Jessica Hullman. Why authors don't visualize uncertainty. *IEEE Transactions on Visualization and Computer Graphics*, 26(1):130–139, 2020.

[112] Jessica Hullman, Matthew Kay, Yea-Seul Kim, and Samana Shrestha. Imagining Replications: Graphical Prediction & Discrete Visualizations Improve Recall & Estimation of Effect Uncertainty.

[113] 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, January 2019.

[114] Jessica Hullman, Paul Resnick, and Eytan Adar. Hypothetical outcome plots outperform error bars and violin plots for inferences about reliability of variable ordering. *PLOS ONE*, 10(11), 2015.

[115] Sarah Inman and David Ribes. "beautiful seams": Strategic revelations and concealments. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, CHI '19, page 1–14, New York, NY, USA, 2019. Association for Computing Machinery.

[116] Sarah Jabbour, David Fouhey, Stephanie Shepard, Thomas S. Valley, Ella A. Kazerooni, Nikola Banovic, Jenna Wiens, and Michael W. Sjoding. Measuring the Impact of AI in the Diagnosis of Hospitalized Patients: A Randomized Clinical Vignette Survey Study. *JAMA*, 330(23):2275–2284, 12 2023.

[117] Maia Jacobs, Jeffrey He, Melanie F. Pradier, Barbara Lam, Andrew C. Ahn, Thomas H. McCoy, Roy H. Perlis, Finale Doshi-Velez, and Krzysztof Z. Gajos. Designing ai for trust and collaboration in time-constrained medical decisions: A sociotechnical lens. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, CHI '21, New York, NY, USA, 2021. Association for Computing Machinery.

[118] Maurice Jakesch, Advait Bhat, Daniel Buschek, Lior Zalmanson, and Mor Naaman. Co-writing with opinionated language models affects users' views. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, CHI '23, New York, NY, USA, 2023. Association for Computing Machinery.

[119] S. Jasanoff. *States of Knowledge: The Co-production of Science and the Social Order*. Routledge, 2004.

[120] Philip N. Johnson-Laird. *Mental models : towards a cognitive science of language, inference, and consciousness*. Cambridge, MA: Harvard University Press, 1983. Excerpts available on Google Books.

[121] Michael I. Jordan, Tommi S. Jaakkola, Lawrence K. Saul, and Florham Park. An Introduction to Variational Methods for Graphical Models An Introduction to Variational Methods for Graphical Models. *Machine Learning*, 37:183–233, 1999.

[122] Susan Joslyn and Sonia Savelli. Visualizing uncertainty for non-expert end users: The challenge of the deterministic construal error. *Frontiers in Computer Science*, 2, January 2021.

[123] Jongbin Jung, Connor Concannon, Ravi Shroff, Sharad Goel, and Daniel G. Goldstein. Simple rules for complex decisions. *SSRN Electronic Journal*, February 2017.

[124] Ekaterina Jussupow, Kai Spohrer, Armin Heinzl, and Joshua Gawlitza. Augmenting medical diagnosis decisions? an investigation into physicians' decision-making process with artificial intelligence. *Information Systems Research*, 32(3):713–735, September 2021.

[125] Daniel Kahneman. *Thinking, Fast and Slow*. Farrar, Straus and Giroux, New York, first paperback edition, 2011.

[126] Daniel Kahneman and Amos Tversky. *Prospect theory: An analysis of decision under risk.* Cambridge University Press, New York, NY, US, 1988.

[127] Alex Kale, Matthew Kay, and Jessica Hullman. Visual reasoning strategies for effect size judgments and decisions. *IEEE Transactions on Visualization and Computer Graphics*, 27(2):272–282, 2021.

[128] Harmanpreet Kaur, Eytan Adar, Eric Gilbert, and Cliff Lampe. Sensible ai: Re-imagining interpretability and explainability using sensemaking theory. In *2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22, page 702–714, New York, NY, USA, 2022. Association for Computing Machinery.

[129] Michael R Kaus, Simon K Warfield, Arya Nabavi, Peter M Black, Ferenc A Jolesz, and Ron Kikinis. Automated segmentation of mr images of brain tumors. *Radiology*, 218(2):586–591, 2001.

[130] Anna Kawakami, Venkatesh Sivaraman, Logan Stapleton, Hao-Fei Cheng, Adam Perer, Zhiwei Steven Wu, Haiyi Zhu, and Kenneth Holstein. "why do i care what's similar?" probing challenges in ai-assisted child welfare decision-making through worker-ai interface design concepts. In *Designing Interactive Systems Conference*, DIS '22, page 454–470, New York, NY, USA, 2022. Association for Computing Machinery.

[131] Matthew Kay, Tara Kola, Jessica R. Hullman, and Sean A. Munson. When (ish) is my bus? User-centered visualizations of uncertainty in everyday, mobile predictive systems. In *Conference on Human Factors in Computing Systems - Proceedings*, 2016.

[132] Matthew Kay, Gregory L. Nelson, and Eric B. Hekler. Researcher-centered design of statistics: Why bayesian statistics better fit the culture and incentives of hci. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, CHI '16, page 4521–4532, New York, NY, USA, 2016. Association for Computing Machinery.

[133] Ralph L. Keeney, Howard Raiffa, and David W. Rajala. Decisions with multiple objectives: Preferences and value trade-offs. *IEEE Transactions on Systems, Man, and Cybernetics*, 9:403–403, 1979.

[134] Amirhossein Kiani, Bora Uyumazturk, Pranav Rajpurkar, Alex Wang, Rebecca Gao, Erik Jones, Yifan Yu, Curtis P Langlotz, Robyn L Ball, Thomas J Montine, Brock A Martin, Gerald J Berry, Michael G Ozawa, Florette K Hazard, Ryanne A Brown, Simon B Chen, Mona Wood, Libby S Allard, Lourdes Ylagan, Andrew Y Ng, and Jeanne Shen. Impact of a deep learning assistant on the histopathologic classification of liver cancer. *npj Digital Medicine*, 3(1):23, 2020.

[135] Sunnie S. Y. Kim, Q. Vera Liao, Mihaela Vorvoreanu, Stephanie Ballard, and Jennifer Wortman Vaughan. "i'm not sure, but...": Examining the impact of large language models' uncertainty expression on user reliance and trust. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '24, page 822–835, New York, NY, USA, 2024. Association for Computing Machinery.

[136] Diederik P. Kingma and Jimmy Ba. Adam: A Method for Stochastic Optimization. In *3rd International Conference on Learning Representations, ICLR 2015*, San Diego, CA, 2015.

[137] Martin Kocher, Maximilian I Ruge, Norbert Galldiks, and Philipp Lohmann. Applications of radiomics and machine learning for radiotherapy of malignant brain tumors. *Strahlentherapie und Onkologie : Organ der Deutschen Rontgengesellschaft ... [et al]*, 196(10):856–867, oct 2020.

[138] 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, page 1–14, New York, NY, USA, 2019. Association for Computing Machinery.

[139] Maya Krishnan. Against interpretability: a critical examination of the interpretability problem in machine learning. *Philosophy & Technology*, 33:487–502, 2020.

[140] Ranganath Krishnan, Mahesh Subedar, and Omesh Tickoo. Specifying weight priors in bayesian deep neural networks with empirical bayes. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(04):4477–4484, Apr. 2020.

[141] Michael Krisper, Jürgen Dobaj, and Georg Macher. Patterns for communicating numerical uncertainty. In *Proceedings of the 24th European Conference on Pattern Languages of Programs*, EuroPLop '19, New York, NY, USA, 2019. Association for Computing Machinery.

[142] 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, 2021.

[143] Himabindu Lakkaraju and Osbert Bastani. "how do i fool you?": Manipulating user trust via misleading black box explanations. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, AIES '20, pages 79–85, New York, NY, USA, 2020. Association for Computing Machinery.

[144] Kathryn Ann Lambe, Gary O'Reilly, Brendan D Kelly, and Sarah Curristan. Dual-process cognitive interventions to enhance diagnostic reasoning: a systematic review. *BMJ Qual. Saf.*, 25(10):808–820, October 2016.

[145] Yann A. LeCun, Léon Bottou, Genevieve B. Orr, and Klaus-Robert Müller. Efficient BackProp. In Grégoire Montavon, Geneviève B. Orr, and Klaus-Robert Müller, editors, *Neural Networks: Tricks of the Trade*, pages 9–48. Springer-Verlag Berlin Heidelberg, 2012.

[146] Christian Leibig, Vaneeda Allken, Murat Seçkin Ayhan, Philipp Berens, and Siegfried Wahl. Leveraging uncertainty information from deep neural networks for disease detection. *Scientific Reports*, 7(1), December 2017.

[147] Susan Leigh Star. This is not a boundary object: Reflections on the origin of a concept. *Science, Technology, amp; Human Values*, 35(5):601–617, August 2010.

[148] Zhiyuan Jerry Lin, Jongbin Jung, Sharad Goel, and Jennifer Skeem. The limits of human predictions of recidivism. *Science Advances*, 6(7), 2020.

[149] Zachary C. Lipton. The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery. *Queue*, 16(3):31–57, jun 2018.

[150] Duri Long and Brian Magerko. What is ai literacy? competencies and design considerations. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, CHI '20, page 1–16, New York, NY, USA, 2020. Association for Computing Machinery.

[151] Scott M. Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 4768–4777, Red Hook, NY, USA, 2017. Curran Associates Inc.

[152] A. M. MacEachren, C. A. Brewer, and L. W. Pickle. Visualizing georeferenced data: representing reliability of health statistics. *Environment and Planning A*, 1998.

[153] David J. C. MacKay. A Practical Bayesian Framework for Backpropagation Networks. *Neural Computation*, 4(3):448–472, 1992.

[154] Karen Mann, Jill Gordon, and Anna MacLeod. Reflection and reflective practice in health professions education: a systematic review. *Advances in health sciences education*, 14:595–621, 2009.

[155] Charles F. Manski. The lure of incredible certitude. *Economics and Philosophy*, 36(2):216–245, 2020.

[156] Sean McGrath, Parth Mehta, Alexandra Zytek, Isaac Lage, and Himabindu Lakkaraju. When does uncertainty matter?: Understanding the impact of predictive uncertainty in ML assisted decision making. *arXiv*, 2020.

[157] Roisin McNaney, John Vines, Andy Dow, Harry Robinson, Heather Robinson, Kate McDonald, Leslie Brown, Peter Santer, Don Murray, Janice Murray, David Green, and Peter Wright. Enabling the participation of people with parkinson's and their caregivers in co-inquiry around collectivist health technologies. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, CHI '18, page 1–14, New York, NY, USA, 2018. Association for Computing Machinery.

[158] Sally Engle Merry. *The Seductions of Quantification: Measuring Human Rights, Gender Violence, and Sex Trafficking*, volume 100. Chicago: University of Chicago Press.

[159] Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. Model cards for model reporting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, FAT* '19, page 220–229, New York, NY, USA, 2019. Association for Computing Machinery.

[160] Christoph Molnar. *Interpretable Machine Learning*. 2020.

[161] Ramaravind K. Mothilal, Amit Sharma, and Chenhao Tan. Explaining machine learning classifiers through diverse counterfactual explanations. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, FAT* '20, page 607–617, New York, NY, USA, 2020. Association for Computing Machinery.

[162] Umaira Nazar, Muhammad Attique Khan, Ikram Ullah Lali, Hong Lin, Hashim Ali, Imran Ashraf, and Junaid Tariq. Review of Automated Computerized Methods for Brain Tumor Segmentation and Classification. *Current medical imaging*, 16(7):823–834, nov 2020.

[163] Radford M. Neal. *Bayesian Learning for Neural Networks*. Springer-Verlag New York, New York, NY, 1996.

[164] Radford M. Neal. MCMC Using Hamiltonian Dynamics. In *Handbook of Markov Chain Monte Carlo*, pages 113–162. 2011.

[165] Safiya Umoja Noble. Algorithms of oppression. In *Algorithms of Oppression*. New York University Press, 2018.

[166] Anthony O'Hagan, Caitlin E. Buck, Alireza Daneshkhah, J. Richard Eiser, Paul H. Garthwaite, David J. Jenkinson, Jeremy E. Oakley, and Tim Rakow. *Uncertain Judgements: Eliciting Experts' Probabilities*. John Wiley & Sons, Ltd, Chichester, UK, 2006.

[167] David Ojika, Bhavesh Patel, G. Anthony Reina, Trent Boyer, Chad Martin, and Prashant Shah. Addressing the memory bottleneck in AI model training. arXiv [preprint] Available at: http://arxiv.org/abs/2003.08732, 2020.

[168] Cathy O'neil. *Weapons of math destruction: How big data increases inequality and threatens democracy*. Broadway books, 2016.

[169] Judith Orasanu and Terry Connolly. The reinvention of decision making. In *Decision making in action: Models and methods.*, pages 3–20. Ablex Publishing, Westport, CT, US, 1993.

[170] Yaniv Ovadia, Emily Fertig, Jie Ren, Zachary Nado, D. Sculley, Sebastian Nowozin, Joshua Dillon, Balaji Lakshminarayanan, and Jasper Snoek. Can you trust your model's uncertainty? evaluating predictive uncertainty under dataset shift. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

[171] Nicolas Papernot, Patrick D. McDaniel, and Ian J. Goodfellow. Transferability in machine learning: from phenomena to black-box attacks using adversarial samples. *CoRR*, abs/1605.07277, 2016.

[172] Shreyasha Paudel, Sabine Loos, and Robert Soden. Hype versus historical continuity: Situating the rise of ai in climate and disaster risk modeling. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, CHI '25, New York, NY, USA, 2025. Association for Computing Machinery.

[173] Lori Peek, Jennifer Tobin, Rachel M. Adams, Haorui Wu, and Mason Clay Mathews. A framework for convergence research in the hazards and disaster field: The natural hazards engineering research infrastructure converge facility. *Frontiers in Built Environment*, 6, July 2020.

[174] Daniel J Power. Decision Support Systems: A Historical Overview BT - Handbook on Decision Support Systems 1: Basic Themes. pages 121–140. Springer Berlin Heidelberg, Berlin, Heidelberg, 2008.

[175] 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*, IUI '23, pages 379–396, New York, NY, USA, 2023. Association for Computing Machinery.

[176] Mahima Pushkarna, Andrew Zaldivar, and Oddur Kjartansson. Data cards: Purposeful and transparent dataset documentation for responsible ai, 2022.

[177] Emilee Rader, Kelley Cotter, and Janghee Cho. Explanations as mechanisms for supporting algorithmic transparency. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, CHI '18, page 1–13, New York, NY, USA, 2018. Association for Computing Machinery.

[178] Emilee Rader, Kelley Cotter, and Janghee Cho. Explanations as mechanisms for supporting algorithmic transparency. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, CHI '18, pages 1–13, New York, NY, USA, 2018. Association for Computing Machinery.

[179] Pranav Rajpurkar, Emma Chen, Oishi Banerjee, and Eric J. Topol. AI in health and medicine. *Nature Medicine*, 28(1):31–38, January 2022.

[180] Valerie F. Reyna and Charles J. Brainerd. Numeracy, ratio bias, and denominator neglect in judgments of risk and probability. *Learning and Individual Differences*, 18(1):89–107, 2008.

[181] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. Model-agnostic interpretability of machine learning, 2016.

[182] Carlos Riquelme, George Tucker, Jasper Snoek, and Google Brain. Deep Bayesian Bandits Showdown. *NIPS 2017 Bayesian Deep Learning Workshop*, 2017.

[183] Herbert Robbins and Sutton Monro. A Stochastic Approximation Method. *The Annals of Mathematical Statistics*, 22(3):400–407, 1951.

[184] Christian P. Robert and George Casella. *Monte Carlo Statistical Methods*. Springer New York, New York, NY, 2004.

[185] Claire Roney, Jana Laura Egelhofer, and Sophie Lecheler. The "bread and butter" of science: understanding how scientists communicate uncertainties. *Humanities and Social Sciences Communications*, 13(1), March 2026.

[186] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-Net: Convolutional Networks for Biomedical Image Segmentation. In *Medical Image Computing and Computer-Assisted Intervention - MICCAI 2015 - 18th International Conference Munich, Germany, October 5 - 9, 2015, Proceedings, Part III*, volume 9351 of *Lecture Notes in Computer Science*, pages 234–241. Springer Verlag, may 2015.

[187] Sara Salimzadeh, Gaole He, and Ujwal Gadiraju. Dealing with uncertainty: Understanding the impact of prognostic versus diagnostic tasks on trust and reliance in human-ai decision making. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, CHI '24, New York, NY, USA, 2024. Association for Computing Machinery.

[188] Téo Sanchez, Baptiste Caramiaux, Pierre Thiel, and Wendy E. Mackay. Deep learning uncertainty in machine teaching. In *27th International Conference on Intelligent User Interfaces*, IUI '22, page 173–190, New York, NY, USA, 2022. Association for Computing Machinery.

[189] Ulrike Schäfer, Lars Sipos, and Claudia Müller-Birn. 'the ai is uncertain, so am i. what now?': Navigating shortcomings of uncertainty representations in human-ai collaboration with capability-focused guidance. *Proc. ACM Hum.-Comput. Interact.*, 9(7), October 2025.

[190] Max Schemmer, Niklas Kuehl, Carina Benz, Andrea Bartos, and Gerhard Satzger. Appropriate reliance on ai advice: Conceptualization and the effect of explanations. In *Proceedings of the 28th International Conference on Intelligent User Interfaces*, IUI '23, pages 410–422, New York, NY, USA, 2023. Association for Computing Machinery.

[191] Hugo Scurto, Baptiste Caramiaux, and Frederic Bevilacqua. Prototyping machine learning through diffractive art practice. In *Designing Interactive Systems Conference 2021*, DIS '21, page 2013–2025, New York, NY, USA, 2021. Association for Computing Machinery.

[192] Mrinank Sharma, Sebastian Farquhar, Eric Nalisnick, and Tom Rainforth. Do bayesian neural networks need to be fully stochastic? In Francisco Ruiz, Jennifer Dy, and Jan-Willem van de Meent, editors, *Proceedings of The 26th International Conference on Artificial Intelligence and Statistics*, volume 206 of *Proceedings of Machine Learning Research*, pages 7694–7722. PMLR, 25–27 Apr 2023.

[193] Hua Shen, Chieh-Yang Huang, Tongshuang Wu, and Ting-Hao Kenneth Huang. Convxai: Delivering heterogeneous ai explanations via conversations to support human-ai scientific writing. In *Companion Publication of the 2023 Conference on Computer*

Supported Cooperative Work and Social Computing, CSCW '23 Companion, pages 384–387, New York, NY, USA, 2023. Association for Computing Machinery.

[194] Leonard M Silva, Emilio M Pereira, Paulo GO Salles, Ran Godrich, Rodrigo Ceballos, Jeremy D Kunz, Adam Casson, Julian Viret, Sarat Chandarlapaty, Carlos Gil Ferreira, Bruno Ferrari, Brandon Rothrock, Patricia Raciti, Victor Reuter, Belma Dogdas, George DeMuth, Jillian Sue, Christopher Kanan, Leo Grady, Thomas J Fuchs, and Jorge S Reis-Filho. Independent real-world application of a clinical-grade automated prostate cancer detection system. *The Journal of Pathology*, 254(2):147–158, April 2021.

[195] Herbert A. Simon. Rational choice and the structure of the environment. *Psychological Review*, 63:129–138, 1956.

[196] Amber L. Simpson et al. A large annotated medical image dataset for the development and evaluation of segmentation algorithms. *arXiv preprint arXiv:1902.09063*, 2019.

[197] D. S. Sivia and J. Skilling. *Data Analysis: A Bayesian Tutorial*. Oxford University Press, New York, NY, 2nd edition, 2006.

[198] Meredith Skeels, Bongshin Lee, Greg Smith, and George Robertson. Revealing uncertainty for information visualization. In *Proceedings of the Working Conference on Advanced Visual Interfaces*, AVI '08, page 376–379, New York, NY, USA, 2008. Association for Computing Machinery.

[199] Robert Soden, Lydia Chilton, Scott Miles, Rebecca Bicksler, Kaira Ray Villanueva, and Melissa Bica. Insights and opportunities for hci research into hurricane risk communication. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, CHI '22, New York, NY, USA, 2022. Association for Computing Machinery.

[200] Robert Soden, Laura Devendorf, Richmond Wong, Yoko Akama, and Ann Light. Modes of uncertainty in hci. *Foundations and Trends® in Human–Computer Interaction*, 15(4):317–426, 2022.

[201] Robert Soden, Perrine Hamel, David Lallemant, and James Pierce. The disaster and climate change artathon: Staging art/science collaborations in crisis informatics. In *Proceedings of the 2020 ACM Designing Interactive Systems Conference*, DIS '20, page 1273–1286, New York, NY, USA, 2020. Association for Computing Machinery.

[202] Katta Spiel, Oliver L. Haimson, and Danielle Lottridge. How to do better with gender on surveys: A guide for hci researchers. *Interactions*, 26(4):62–65, jun 2019.

[203] Susan Leigh Star and James R. Griesemer. Institutional ecology, 'translations' and boundary objects: Amateurs and professionals in berkeley's museum of vertebrate zoology, 1907–39. *Social Studies of Science*, 19(3):387–420, 1989.

[204] Eliza Strickland. Ibm watson, heal thyself: How ibm overpromised and underdelivered on ai health care. *IEEE Spectrum*, 56(4):24–31, 2019.

[205] Joep C. Stroom and Ben J.M. Heijmen. Geometrical uncertainties, radiotherapy planning margins, and the ICRU-62 report. *Radiotherapy and Oncology*, 2002.

[206] Roger Stupp, Warren P Mason, Martin J Van Den Bent, Michael Weller, Barbara Fisher, Martin JB Taphoorn, Karl Belanger, Alba A Brandes, Christine Marosi, Ulrich Bogdahn, et al. Radiotherapy plus concomitant and adjuvant temozolomide for glioblastoma. *New England journal of medicine*, 352(10):987–996, 2005.

[207] Padma Susarla, Dexter Purnell, and Ken Scott. Zillow's artificial intelligence failure and its impact on perceived trust in information systems. *Journal of Information Technology Teaching Cases*, 16(1):13–19, 2024.

[208] Alvin T. Tan. Cracking the ames housing dataset with linear regression. https://github.com/at-tan/Cracking_Ames_Housing_OLS, 2019.

[209] Ryan Tennant and Tania Randall. Human performance and perception of uncertainty visualizations in geospatial applications: A scoping review. *IEEE Transactions on Visualization and Computer Graphics*, 31(10):9387–9404, October 2025.

[210] Michael E. Tipping. Sparse bayesian learning and the relevance vector machine. *J. Mach. Learn. Res.*, 1:211–244, sep 2001.

[211] Richard Tomsett, Alun Preece, Dave Braines, Federico Cerutti, Supriyo Chakraborty, Mani Srivastava, Gavin Pearson, and Lance Kaplan. Rapid Trust Calibration through Interpretable and Uncertainty-Aware AI. *Patterns*, 1(4):100049, 2020.

[212] Sana Tonekaboni, Shalmali Joshi, Melissa D. McCradden, and Anna Goldenberg. What clinicians want: Contextualizing explainable machine learning for clinical end use. In Finale Doshi-Velez, Jim Fackler, Ken Jung, David Kale, Rajesh Ranganath, Byron Wallace, and Jenna Wiens, editors, *Proceedings of the 4th Machine Learning for Healthcare Conference*, volume 106 of *Proceedings of Machine Learning Research*, pages 359–380, Ann Arbor, Michigan, 09–10 Aug 2019. PMLR.

[213] Eric J Topol. High-performance medicine: the convergence of human and artificial intelligence. *Nature Medicine*, 25(1):44–56, 2019.

[214] Vinh Ngoc Tran, Valeriy Y Ivanov, Weichen Huang, Kevin Murphy, Fariborz Daneshvar, Jeff H Bednar, G Aaron Alexander, Jongho Kim, and Daniel B Wright. Connectivity in urbanscapes can cause unintended flood impacts from stormwater systems. *Nature Cities*, 1(10):654–664, 2024.

[215] C. Tsagkaris, A.S. Papazoglou, D.V. Moysidis, M. Papadakis, and C. Trompoukis. Bayesian versus frequentist clinical research now and then: Lessons from the greco-roman medical scholarship. *Ethics, Medicine and Public Health*, 23:100805, August 2022.

[216] Amos Tversky and Daniel Kahneman. Judgment under uncertainty: Heuristics and biases. *Science*, 185(4157):1124–1131, 1974.

[217] US Department of Commerce, NOAA. August 11, 2014 historic rainfall. Online, 2014. Retrieved 2024-12-01.

[218] Laurent Valentin Jospin, Wray Buntine, Farid Boussaid, Hamid Laga, and Mohammed Bennamoun. Hands-on Bayesian Neural Networks - a Tutorial for Deep Learning Users, 2020.

[219] Anne Marthe van der Bles, Sander van der Linden, Alexandra L. J. Freeman, James Mitchell, Ana B. Galvao, Lisa Zaval, and David J. Spiegelhalter. Communicating uncertainty about facts, numbers and science. *Royal Society Open Science*, 6(5):181870, May 2019.

[220] Anne Marthe van der Bles, Sander van der Linden, Alexandra L.J. Freeman, and David J. Spiegelhalter. The effects of communicating uncertainty on public trust in facts and numbers. *Proceedings of the National Academy of Sciences of the United States of America*, 117(14):7672–7683, 2020.

[221] Effy Vayena, Alessandro Blasimme, and I. Glenn Cohen. Machine learning in medicine: Addressing ethical challenges. *PLoS Medicine*, 2018.

[222] John von Neumann and Oskar Morgenstern. *Theory of Games and Economic Behavior (60th Anniversary Commemorative Edition)*. Princeton University Press, 2007.

[223] Detlof von Winterfeldt. Bridging the gap between science and decision making. *Proceedings of the National Academy of Sciences*, 110(supplement$_3$) : 14055–14061, *August*2013.

[224] Sandra Wachter, Brent D. Mittelstadt, and Chris Russell. Counterfactual explanations without opening the black box: Automated decisions and the GDPR. *CoRR*, abs/1711.00399, 2017.

[225] Martin J. Wainwright and Michael I. Jordan. Graphical Models, Exponential Families, and Variational Inference. *Foundations and Trends® in Machine Learning*, 1(1–2):1–305, 2007.

[226] Dakuo Wang, Pattie Maes, Xiangshi Ren, Ben Shneiderman, Yuanchun Shi, and Qianying Wang. Designing ai to work with or for people? In *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems*, CHI EA '21, New York, NY, USA, 2021. Association for Computing Machinery.

[227] Zhenlin Wang, Xun Huan, and Krishna Garikipati. Variational system identification of the partial differential equations governing the physics of pattern-formation: Inference under varying fidelity and noise. *Computer Methods in Applied Mechanics and Engineering*, 356(1):44–74, nov 2019.

[228] Hilde J. P. Weerts, Werner van Ipenburg, and Mykola Pechenizkiy. A human-grounded evaluation of shap for alert processing, 2019.

[229] Karl E Weick. *Sensemaking in organizations*, volume 3. Sage, 1995.

[230] Mark Weiser. The computer for the 21st century. *SIGMOBILE Mob. Comput. Commun. Rev.*, 3(3):3–11, jul 1999.

[231] Yeming Wen, Paul Vicol, Jimmy Ba, Dustin Tran, and Roger Grosse. Flipout: Efficient pseudo-independent weight perturbations on mini-batches, 2018.

[232] John H. Williamson, Antti Oulasvirta, Per Ola Kristensson, and Nikola Banovic. *Bayesian Methods for Interaction and Design*. Cambridge University Press, 2022.

[233] Jo Wood, Petra Isenberg, Tobias Isenberg, Jason Dykes, Nadia Boukhelifa, and Aidan Slingsby. Sketchy rendering for information visualization. *IEEE Transactions on Visualization and Computer Graphics*, 18(12):2749–2758, 2012.

[234] Haiping Xu and Amol Gade. Smart real estate assessments using structured deep neural networks. In *2017 IEEE SmartWorld, Ubiquitous Intelligence & Computing, Advanced & Trusted Computed, Scalable Computing & Communications, Cloud & Big Data Computing, Internet of People and Smart City Innovation (SmartWorld/SCALCOM/UIC/ATC/CBDCom/IOP/SCI)*, pages 1–7, 2017.

[235] Jiayi Xu. A novel deep neural network based method for house price prediction. In *2021 International Conference of Social Computing and Digital Economy (ICSCDE)*, pages 12–16, 2021.

[236] Jiaming Zeng, Adam Lesnikowski, and Jose M. Alvarez. The Relevance of Bayesian Layer Positioning to Model Uncertainty in Deep Bayesian Active Learning, 2018.

[237] 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, September 2016.

[238] Yunfeng Zhang, Q. Vera Liao, and Rachel K. E. 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*, FAT* '20, page 295–305, New York, NY, USA, 2020. Association for Computing Machinery.