From Research to UX Practice: Evaluation Approaches and Tools for Realizing Human-Centered AI Goals

Meena Devii Muralikumar

A dissertation

submitted in partial fulfillment of the

requirements for the degree of

Doctor of Philosophy

University of Washington

2025

Reading Committee:

David W. McDonald, Chair

Gary Hsieh

Amy X. Zhang

Program Authorized to Offer Degree:

Human centered Design and Engineering

University of Washington

Abstract

From Research to UX Practice: Evaluation Approaches and Tools for Realizing Human-Centered AI Goals

Meena Devii Muralikumar

Chair of the Supervisory Committee:

Professor David W. McDonald

Human centered Design and Engineering

The integration of AI models in various products and services poses unique challenges for design and UX practice. Unlike other technologies, AI poses distinct challenges due to its probabilistic nature, technical complexity, and lack of precedent. While collaboration with AI practitioners can help alleviate these challenges, there are significant communication, process, and knowledge barriers to overcome. In this dissertation, I present formative, design, and evaluative research to better support UX practice of AI and UX-AI collaborations, which in turn can support human-centered design of AI products. First, I conduct a qualitative study to understand key challenges that UX practitioners face in model comprehension and reinforcing a human-centered lens on model evaluation. Based on these findings, I design and test visualization-based methods that enable UX practitioners to infer insights about human-AI alignment. I also examine how practitioner tools are currently designed to support interdisciplinary collaborations, such as UX-AI collaborations, through a design space analysis. Based on these three studies, I derive implications for designing evaluation tools for AI/LLM

applications that can better support UX practitioners' needs and improve HCI/UX and AI collaborations.

Table of Contents

List of Figures

List of Tables

Acknowledgements

First and foremost, I would like to thank David W. McDonald for all his support and mentoring throughout the years. You were always receptive of my ideas and discussing research with you has helped me a better researcher. You also enabled me to develop my own voice when it comes to writing and communicating my research. Thank you for being such a supportive and encouraging advisor, even as I navigated the highs and lows of graduate school and personal life. Thank you to my committee members – Gary, Amy, and Leilani. Your feedback was instrumental in helping me complete this dissertation research. I would also like to thank Sean, Julie, David Ribes, Kate, Jennifer, Tyler, and Mark for all the conversations and learnings over the years. My sincere thanks to Kathleen, Allen, and Jane for making life as a HCDE PhD student much easier.

Thank you to my husband, Adithya Srinivasan, for listening to all things related to research and academia, for always helping in whatever way you could, and for sharing the emotional labor of doing a PhD. To my parents, Jeyabharathi and Muralikumar – thank you for believing in me and your unwavering support. Jeyavaishnavi – thank you for being the voice of reason and helping me pick myself up again. Thank you, grandpa, for the kind of person you were and for continuing to inspire. Thank you to my in-laws – Rajalakshmi and Srinivasan, for understanding what this entailed and encouraging me. Thank you to all my friends who took an interest in my work and believed in me. Thank you, Vishal, for being my PhD and conference buddy.

Chapter 1: Introduction

Products and services that incorporate Artificial Intelligence (AI) have proliferated and become part of our everyday lives. AI capabilities are leveraged to power various use cases in education, health, finance, journalism, online communities, software development and so on. While AI provides the promise of enabling novel, intelligent interactions that can improve people's quality of life and work, it also runs many risks. At an individual level, there are risks of misunderstanding the uncertainty associated with AI outputs, being misinformed, or being denied a service based on automated decisions [13,118,159]. AI systems can also have disproportionate impacts for certain groups based on their demographics or other attributes, thus creating a digital divide [14,41,63,159]. At the societal level, AI models and their applications present potential risks of job loss, privacy violations, information manipulation, over-reliance, and adverse environmental impact [21,82,90,127]. AI systems also open a can of worms with respect to accountability, regulation, and ethics that requires participation from multiple interdisciplinary experts to address [120,142,158].

Such challenges at varying levels make it even more important that a human-centered lens is reinforced on the design and development of AI products and services. However, there are distinct challenges to designing with AI. Compared to traditional technologies, AI is probabilistic in nature and technically complex. Algorithms and learning mechanisms have evolved in the past decade to now give rise to large, homogenized, pre-trained models that are capable of a wide range of tasks and can work with multiple modalities like text, video, audio, images [15].These capabilities are incredible and novel, which means there is not much precedent for how to design effectively and ensure the development of human-centered, ethical, and responsible AI systems every step of the way. AI-related design challenges are relevant to both Human-Computer Interaction (HCI) research and practice. HCI research has focused on different aspects of human-centered AI, such as explainable and interpretable AI, ethical AI,

human-AI teaming, and developing human-centered approaches for designing and evaluating AI [22].

To address any research-practice gaps, it is important to i) study actual UX and design practice of AI in order to 'bubble up' insights and situated knowledge and ii) develop design concepts, methods, and tools that can be adapted and used by practitioners [27,58,59,133]. Thus, my first objective was to better understand the challenges of UX practitioners in designing human-centered AI applications, the strategies used by them to overcome these challenges, and needs for new resources and tools that can help resolve these challenges. Since collaboration between UX practitioners and AI practitioners are necessary for designing human-centered AI applications [56,147], I also focus on understanding barriers to collaboration, if any, and what kinds of artifacts and tools support these collaborations.

Thus, I first ask the following research question:

> RQ1. What are the key conceptual, process-related, and collaborative challenges faced by UX practitioners designing for AI?

By conducting a qualitative study, I uncovered various challenges related to gaining contextual understanding of the AI model's capabilities, prototyping and evaluating AI applications in ways that inform human-centered design, as well as factors that challenge interdisciplinary, UX-AI collaborations. Based on these findings, I pose and answer the following questions as well:

> RQ2. How do tools used by practitioners support interdisciplinary, collaborative work in AI design and development?

RQ3. How might we draw from Human-centered AI (HCAI) research and methods to design evaluation approaches that reinforce human needs and behaviors alongside model-centric measures?

1.2 Document Overview

The rest of the dissertation is structured as follows. I will provide an overview of related work at the intersection of UX practice and AI and the literature backdrop against which I designed and conducted my inquiries (Chapter 2).

In Chapter 3, I will describe the interview study I conducted to understand the challenges, needs, and processes of 14 UX practitioners designing with AI (RQ1). This study proved to be a formative investigation that informed the rest of the dissertation research. UX practitioners (UXP) surfaced challenges related to model comprehension - contributed by a lack of contextual model information, lack of visibility into model development processes, and a lack of support to incorporate the model directly in design activities. UXP also discussed difficulties in reinforcing a human-centered lens when they are looped in at later stages, when major design decisions have been made, and when there is more emphasis on model-centric metrics of performance. UXP raised the need for a sandbox type of tool (*model sandbox**)*** that can support hands-on interactions with the model, a more tangible understanding of model capabilities and limitations, prototyping with real model behaviors and outputs, and collaborative testing of different use cases and scenarios. Based on these findings, I derive design implications for designing model sandbox types of tools to support model integration in UX activities and propose using groupware design principles to support collaborations between UX and AI practitioners.

In Chapter 4, I survey tools (n = 18) used by different practitioners to tackle the complexities and quirks of developing AI models and the products that incorporate them. I was motivated to analyze whether groupware principles were used at all to build these tools, and if

not, how else they supported collaborative activities. This analysis resulted in a construction of a design space of how tools support interdisciplinary collaborations in AI design and development (RQ2). Through this design space analysis, I uncovered that a few tools did use groupware design principles support collaborations (between domain experts & data scientists and among a team of data scientists) but none of them focused on UX-AI collaborations.

Revisiting the needs raised in the formative study, I focused on translating the methods used in conducting human-centered evaluations of AI models [67,106] to make them usable and useful for UX practitioners (RQ3). In Chapter 5, I describe how a class of statistical methods - *Maximum Likelihood Estimation (MLE)* [139] - are particularly useful for analyzing how user and model judgements align across various sociotechnical factors. I design interactive visualizations and test how these help UX practitioners i) gain a conceptual and practical understanding of MLE results without having to learn about the underlying statistics and ii) inform human-centered design principles for a given AI-based use case. In Chapter 6, I draw across the results of these studies to derive design implications for AI/LLM evaluation tools.

1.3 Summary of Contributions

My dissertation work makes the following contributions:

a. An empirical understanding of the challenges, strategies, and needs of UX practitioners when designing with AI and collaborating with AI practitioners (Chapter 3).
b. A design space that describes the collaborative potential of tools used in AI development, helps compare different tools, and inspires creation of new kinds of collaborative tools to better support interdisciplinary collaborations in AI application development (Chapter 4).
c. An illustrative example of a human-centered evaluation of an AI model that leverages Maximum Likelihood Estimation (MLE) methods to examine alignment between user and model judgements and analyze factors that affect this alignment (Chapter 5).

d. Alternate ways to visualize results of human-centered evaluations that use MLE methods and the utility of these visualizations in enabling UX practitioners to inform human-centered design of AI (Chapter 5).

Chapter 2: Background and Related Work

2.1 The Changing Landscape of AI

Before delving into research at the intersection of UX/Design practice and AI, it is important to note that the field of Artificial Intelligence (AI) has undergone significant changes in the past few decades. Three eras are particularly relevant and important for this dissertation - machine learning, deep learning, and foundation models [15]. Machine learning algorithms enable development of models that make future predictions based on historical data. The availability of GPUs led to leveraging deep learning techniques over larger datasets to recognize more complex patterns. However, these deep learning models are black-boxes inherently, which makes it harder to peek inside and interpret their predictions. Foundation models are much larger, pre-trained models trained on huge datasets to achieve a wide range of capabilities. Examples of text-based models, also known as Large Language Models (LLMs) include GPT, Gemini, Claude, and Llama. These models exhibit advanced capabilities of natural language understanding and generation.

The related work described in this section covers models of all three eras. Very rarely does this work focus on one paradigm of models (example: [94]) Most of the work focuses on the range of model capabilities (example: [153]), a particular domain (e.g., enterprise [152]), or a particular stage of the design process (e.g., ideation [94,153], prototyping [137]). In my dissertation work, I interview UX designers and researchers working on a range of AI/ML applications - prediction models, natural-language based recommenders, text classification, chatbots, fraud detection, clinical decision-making, and so on (Chapter 3). In the design space (Chapter 4), most tools were meant for working with deep learning or pre-trained text or image models. The evaluation approach discussed in Chapter 5 is applicable for any text-based model,

but Chapter 6 focuses more on a specific NLP task - Question Answering, which is being increasingly powered by LLMs.

2.2 Challenges in Designerly Understanding of AI

The idea of *'designerly ways of knowing'* was put forward by Nigel Cross [31] to support recognition of design as a distinct discipline, different from both the sciences and the arts. He theorized designerly ways of knowing by looking at both design processes as well as design outcomes. Key to designerly ways of knowing is the idea of translation where designers use internal codes to map individual and societal needs to artifacts. This system of codes embodies designerly knowledge and enables the designer to bring about a translation.

While Cross's theory describes design activity generically, its implications for designing with AI/ML have been explored in more detail in the past decade. First came the recognition of 'intelligence' as a new material to design with, which necessitated taking stock of its distinct opportunities and challenges [68]. Yang et al. [149] examined what makes AI/ML difficult and different to design compared to other technologies. They ascribed it to AI's capability uncertainty and output complexity. Capability uncertainty refers to how a designer cannot foresee or predetermine how an AI model is going to behave and evolve over time. Output complexity refers to the inherent complexity associated with an AI model's infinite range of possible outputs that cannot be realistically simulated. Also, not all AI based systems might be equal in terms of the challenges posed [149]. Some might have a fixed set of outputs whereas others might be more complex and adaptive with an infinite possible set of outputs.

Benjamin et al. [11] reframe uncertainty of ML models from something that hinders design to something that can be understood and wielded as a design material. By applying a post-phenomenological lens and through four case studies, they put forward three concepts (*'thingly uncertainty', 'pattern leakage', 'futures creep'*) to demonstrate how ML uncertainty provides design scope to viewing the world as a continuum, projecting learned patterns into the

world, and to influence the present to mediate interactions in the future. Other work has followed similar lines of inquiry to reframe AI uncertainty and errors as a source for designing creative and novel interactions [49,98].

Liao et al. 's [94] work focuses specifically on the more recent advent of large, pre-trained models to identify what designerly understanding of these models constitutes. They describe four key goals - supporting divergent and convergent design thinking, implementing conditional designs, providing transparency for end-users, and collaborating with technical members of the team, and describe corresponding model information requested by designers to meet them. They also propose XAI-based, user-centric, model interrogation in lieu of static documentation to reduce barriers for designers to ideate with pre-trained models.

My first empirical study examines situated practices of UX practitioners understanding the AI model themselves before they translate it in terms of the design (Chapter 4). I also describe how this process varies depending on the stage in the model development process and how a lack of information about training data and details challenges their model comprehension and translation process.

2.2 Empirical Studies with Design/UX Practitioners

Prior work has also investigated how design and UX practitioners work with AI/ML to build an understanding of what strategies work and where there are still challenges. Dove et al. [44] noted that designers found it challenging to recognize how AI/ML models represented a form of intelligence, distinct from human intelligence. Prototyping efforts were also challenged by the dynamic nature of these models and data dependencies. Both these challenges made it harder for designers to reinforce a human-centered lens, as they felt they were falling behind the engineers and data scientists who were building these models.

More experienced designers used certain strategies to try and overcome such challenges [147,152]. They abstracted the model details and focused more on the capabilities offered by the

model using designerly abstractions and examples, to better grasp the AI or ML's material properties [147]. Close collaborations with data scientists and adopting data-driven methods were deemed as important as well. Yildirim et al. [152] find that designers working in the enterprise space are able to innovate when they are not limited to the user interface aspects of the AI project. Designers' co-location and close collaboration with data scientists, and adaptation of artifacts to serve as boundary objects were also instrumental in effectively leveraging AI as a design material.

Zdanowska and Taylor [155] have focused on the enterprise space as well, and find that UX practitioners are able to design enterprise AI/ML systems and that UX practitioners' skillset supports much needed interdisciplinary collaborations among different stakeholders. Participants reported splitting the model design work from the product/application (that incorporated the model) design work, as an important change in the way things are usually done. Processes and workflows were still a work in progress, which motivated high levels of collaboration with AI/ML practitioners. However Zdanowska and Taylor [155] raise a few important areas where the predominant design thinking processes may fall short - designing for dynamic models post deployment, accounting for issues of fairness, accountability, transparency, ethics, and an equal focus on what is going on under the hood rather than at the interface level.

Windl et al. [144] also observed similar approaches in their study. They found that designers are increasingly participating in data and model-related activities to figure out the UX and are taking on new responsibilities such as mediating between experts and non-experts of AI. They observed four main approaches to designing AI systems- two temporal based approaches and two process-oriented approaches. The former category includes the a priori (model before interface), and the post-hoc (model after interface) approach. The latter category includes a model-centric approach, where model is placed at the center of the product design requiring close collaborations between designers and AI engineers, and a competence-centric approach,

where individual team members' unique skills were leveraged resulting in divergent processes, well-suited for more open-ended projects.

Most of these studies focus on either the enterprise context or on more experienced designers. My interviews with practitioners from varied domains and years of experience working with AI led to additional insights and needs regarding *model sandbox* type of tools and how that can support UXP's model comprehension, UX and design activities, and better coordination and collaboration with AI practitioners (Chapter 4).

2.2.1 Collaborations between UX and AI Practitioners

Much of the prior work described has noted that close collaborations between data scientists and UX practitioners are vital for designing human-centered AI systems. However, a lack of mutual understanding of each other's practices and goals creates challenges for collaboration [56,81]. According to Girardin and Lathia [56] bridging the field of data science and design requires deliberate translation efforts as the fundamental epistemologies between the two vary a lot.

Subramonyam et al. [136] put forward a process model for how designers and AI engineers can co-create the form and function of an AI application using user data as probes, however the reality can be quite different. In a different study, Subramonyam et al. [135] identified '*leaky abstractions*' - ad-hoc artifacts that show more low-level details, that both UX and AI practitioners used to share knowledge across boundaries. Most software is developed by following a separation-of-concerns principle [87], where modularization and object-oriented design help to divide things up and tackle them separately. However, AI systems are non-deterministic and evolve over time, invalidating this principle. An analysis of guidelines for ensuring human-centered design of AI systems also showed that they cover multiple components (user interface, user mental models, training data, models) [135]. Artifacts that communicate the 'user-side of things' with respect to the training data, model outputs, AI-

integrated interfaces and artifacts that communicate data structures, model details, and model behavior were shared by UX and AI practitioners respectively. These '*leaky abstractions*' helped puncture boundaries established by the separation-of-concerns principle and build more effective teams [135].

Bruun et al. [17] echo similar findings from their examination of coordination mechanisms between UX and AI practitioners. They propose for teams to coordinate work through '*mutual adjustment'* than through '*standardization of outputs'*. In the latter, each disciplinary group of practitioners produces outputs that are standard, expected, and irrespective of what other groups are doing. But in the former case, at the individual level, practitioners adapt and adjust to what others are doing through communication and allow for collective knowledge to evolve over time.

Other work has focused on cross-functional collaboration with the goal of understanding how it might support Responsible AI (RAI) efforts such as fairness. Deng et al. [37] find that practitioners focusing on RAI had to put in a lot of effort in bridging concepts and evaluations across different disciplines or stakeholders, identify collaborators and existing efforts that they could piggyback on to gain traction, and deal with the invisible labor involved in rallying cross-functional collaborations.

This growing body of research, along with my empirical study (Chapter 4), suggests that boundary objects [132] might not fully satisfy collaborative requirements in HCI-AI collaborations. This insight led to the examination of interactive tools and how they support communication and collaborations in interdisciplinary teams (Chapter 5).

2.3 Resources and Tools for UX practitioners

Different resources and tools have been put forward by researchers to help address design challenges associated with AI. The first category is design guidelines. Amershi et al. [6] proposed 18 design guidelines for designing human-AI interactions, spread over different stages

of the user experience - initial stages, active use, failure cases, and longitudinal interactions. These guidelines provide practitioners heuristics for verifying their designs and inform design of features required for human-AI interactions. Google's PAIR guidebook also provides designers guidance on how to approach designing with AI by detailing various aspects of AI experiences - trust, mental models, errors, feedback, control, and so on [160]. Liao et al. [93] developed a question bank to represent questions users have about the model, which can in turn inform the design space of incorporating explainable AI (XAI) techniques in an application. Thus, it is a more specific design resource for designing XAI features.

The second category is design resources for ideating with AI. When categorizing and understanding models as supervised or unsupervised, it is harder for designers to understand and wield it as a design material [136,153]. Jansen and Colombo [74] created a mix and match toolkit with tangible tokens that had different options for data (video, text, time series) and model capabilities (categorize, cluster, recommend etc.). Yildirim et al. [153] created a similar design resource to support the ideation of AI concepts. From 40 different examples, they abstracted 8 different capabilities. The design kit also provides support for prioritizing between different concepts and exploring the design space through both an impact-effort matrix and a task expertise-AI performance matrix.

The third category is prototyping tools. Subramonyam et al. [137] developed a model-informed prototyping approach, realized through a tool called ProtoAI. Designers can specify inputs to the model while designing the interface and access model outputs as well. Thus, by designing for data instances across different scenarios, they are able to design interfaces that better account for unpredictable or undesirable behaviors. Canvil – a Figma widget, enables designers to explore and adapt LLMs as a design material within the Figma environment and to prototype LLM-powered experiences [50]. PromptInfuser [116] is a Figma plugin that offers similar functionality and is aimed at helping designers prototype the AI and UX components together. Such tools help improve communication of design ideas, support UX-AI

collaborations, and iterate quickly with AI design concepts. Feng and McDonald's [51] work shows how the interactive machine learning technique can be leveraged to address some of the human-centered design challenges faced by designers and support both the ideation and prototyping stages of the design process.

The final category is evaluation tools. Moore et al. [104] designed a tool - fAIlureNotes, supports UX practitioners in exploring the behavior of image models and error patterns from a user-centered perspective. Most model behavior analyses tools are intended for use by the AI practitioner and allow them to identify systematic failures in distinct groups or patterns of data instances in test datasets [19,20]. fAIlureNotes complements these tools by helping UX practitioners leverage scenarios from user research and real-world contexts and compare user expectations with actual model outputs. Another example of an evaluation tool, though not specific to UX practitioners, is the HAX Playbook for surfacing common errors in NLP-based applications [69]. By filling out an initial reflective form about the model use case, input and output types, triggers, subjectivity involved, and how the output is produced, the Playbook will list different types of errors that can happen so that the designer can proactively address them.

These are all valuable tools for UX and design practice of AI. However, to address specific challenges in gathering insights at scale and reconciling user-centered outcomes with model-centric metrics (Chapter 3), I design evaluation methods that enable UX practitioners to examine and make inferences about human-AI alignment (Chapter 5). I envision the tool being useful for evaluating model fit for a particular use context, evaluating end-user acceptance of the model in those scenarios, as well as for more longitudinal evaluations. While a majority of the work described in this section and prior sections focuses on how UX practitioners can inform the design of products or services that incorporate the model, I also focus on how HCI/UX insights can feed back to inform model evaluation approaches (Chapter 6).

2.4 Responsible AI and UX Practice

As AI systems are being leveraged for high stakes scenarios such as clinical decision making, loan approvals, recidivism predictions, researchers recognized that it warranted additional guiding principles compared to other software deployments. Hence, the term 'Responsible AI' (RAI) gained traction to devise and record practices that can guide the ethical and responsible design and development of AI systems [161–164]. RAI can cover a wide range of concerns including accountability, transparency, fairness, privacy, and so on. Though external and third-party audits are important ways to examine AI systems and hold them accountable [119,120,165], internal practices that enable or hinder development of responsible AI systems have also come under scrutiny [16,36,66,70,100].

More recently, researchers have begun identifying the intersections between responsible AI and UX. Liao et al. [95] highlight three areas where UX can play in ensuring Responsible AI design and development. UX practitioners are well-situated to lead '*responsible ideation'* which involves identifying the right thing to build with AI, to address risks and challenges of AI errors failures at the interface level through '*responsible design',* and to surface real-world impact of AI systems through various methods, thus informing '*responsible evaluation'*. While prior work has proposed ways for engaging in responsible ideation and prototyping, I focus on the area of responsible evaluation (Chapters 5 and 6).

Chapter 3: Investigating Challenges UX Practitioners face when Designing with AI/ML

3.1 Motivation

Compared to designing deterministic software applications, designing AI applications requires additional considerations for UX practitioners. Prior work pinpoints exactly what makes AI unique and difficult to design for - capability uncertainty and output complexity [149]. Since AI models are probabilistic, it becomes fuzzier for practitioners to define its capabilities and behaviors upfront for design purposes. With dynamic models that continue to learn and train on new data, the behavior of the model evolves over time making the user experience dynamic as well. Output complexity refers to the inherent complexity associated with an AI model's range of possible outputs that cannot be always realistically simulated. Across different AI systems, the nature and range of possible outputs might differ (for example, a voice assistant versus a pneumonia detecting clinical decision tool).

While Yang et al. 's analysis [149] provides a framework to characterize the unique challenges of designing human-AI interactions, I wanted to understand if and how these challenges manifested in UX practice and whether there are additional challenges as well. Given that there are mismatches between how the HCI research community views design practice and design practice in the wild, leading to research-practice gaps [59,134], I find it prudent to investigate conceptual and organizational challenges that UX practitioners face when designing for AI. Yang et al.'s [149] article also highlights how there are more design challenges than there are design facilitators for human-AI interaction design. By uncovering practitioners'

perspectives, I also hope to inform the design of tools and methods that can address some of these challenges.

3.2 Methods

I conducted semi-structured, qualitative interviews with 14 UX practitioners who were working on various AI/ML applications as designers and researchers. Since participants might not be able to discuss aspects of what they are working on for proprietary reasons, I provided them with the option of talking about either recommender systems or smart features in email systems as an example of designing for AI/ML. I also used design guidelines from Microsoft's Guidelines for Human-AI Interaction (HAX) [6] and associated design examples as an interview probe to i) help participants recall instances of designing or using AI/ML applications that were challenging and ii) elicit more concrete responses. Participants were provided with the HAX guidelines to look through before the interview took place and after they consented to participate. The interview protocol covered questions about what kinds of AI applications participants worked on, their design process, instances of designing or using AI products that posed peculiarities, whether and how they addressed these challenges, and how they collaborated with different stakeholders, especially AI practitioners. I conducted two pilot interviews to test the effectiveness of the protocol before conducting actual interviews.

3.2.1 Participant Recruitment

The main criteria for including UX practitioners in the study is that they should be working on or have previously worked on AI applications in a UX capacity. To that end, I prepared a recruitment questionnaire that asked for basic information about demographics and education, experience in the UX industry, and what kind of AI/ML application they were working on. After obtaining IRB approval, calls for recruitment were posted on social media

(LinkedIn, Twitter) and circulated in relevant mailing lists and Slack channels. Based on the responses received, I reached out to potential research participants using the contact email address provided in the questionnaire. Interviews were conducted virtually using Zoom, typically lasted 45-60 minutes, and were recorded for transcription later. Participants were provided with a $25 honorarium for their contribution. All interviews were conducted between September and December of 2022. Participant information can be found in Table 3.1.

Participant Identifier	Role	Years of UX experience	AI Product Area
P1	UI Tech Lead	3-5 years	Detecting Fraud
P2	UX Designer	>10 years	Task Automation
P3	Conversational Designer	3-5 years	Chatbots, Recommend, Predict Intent
P4	UX Researcher	3-5 years	Smart assistants
P5	Product Designer	3-5 years	NLP-based recommendation, classification and summarization
P6	UX Researcher	1-3 years	NLP applications
P7	UX Researcher	> 10 years	Text Classification
P8	UX Designer	1-3 years	Decision-support systems
P9	UX Researcher	3-5 years	Personalized recommendations
P10	UX Researcher	5-10 years	Classifying musical genres from audio files
P11	UX Generalist	3-5 years	AI applications for Code Generation
P12	UX Designer	5-10 years	Automating Clinical Ultrasound assessments
P13	Product Designer	5-10 years	Predictive Analytics

P14	User Researcher	1-3 years	Enterprise AI

Table 3.1 Participant Information at the time of interviewing

3.2.2 Interview Protocol

To avoid risking participants' non-disclosure agreements and make them feel comfortable sharing data, I designed the interview protocol around common examples of AI systems - recommender systems and smart features in email. Guidelines and design examples from the HAX Toolkit were also used as an interview probe to stimulate discussions. The first part of the interview protocol focused on talking through participants' general UX experience, experience working on AI products, which stages of the AI product development process they were involved in, and whether the HAX guidelines resonated with their work. Then I focused the discussion on specific examples of when participants encountered challenges in designing for AI, what was challenging about it, how they tried to address it, and how successful they were in addressing it.

The second part of the interview protocol focused more on collaborations with model developers (AI engineers, Data Scientists, Research Scientists) and needs for addressing said challenges. Participants talked through team structures, how they collaborated with technical stakeholders, tensions in collaborations, and what kind of information, resources, or artifacts could better support such collaborations. Finally, participants also speculated about and discussed hypothetical tools that can address some of the challenges they face currently.

3.2.3 Data Analysis

Qualitative analysis of the interview transcripts was guided by the grounded theory approach [29]. I conducted a round of open coding where I focused on participants' current practices, needs, strategies, stopgap solutions, and organizational factors. I then grouped codes into relevant categories and subcategories. In the second round of coding, I did axial coding,

which allowed me to better understand the relationship between different coding categories. During this stage, I identified the core phenomena as design practices when working with AI, and additionally uncovered intervening and contextual factors, strategies, and outcomes. The compare and contrast technique also enabled me to understand how participant accounts differed as a function of organizational norms and structures and/or collaborative practices with AI practitioners. In the next section, I report on the key themes from my data analysis.

3.3 Findings

3.3.1 Unpacking the translation process when designing with AI/ML

One of our participants P5, who worked as a product designer, succinctly described what design translation entailed when designing with AI/ML.

> "*If I don't understand it, as someone who is not an ML researcher, then our users will never understand it. It's important that first I understand it so that I can design something that users will understand as well. So, there's like a translation process there.*" - P5

This translation process is inherent in design practice [31]. Similar to how language translation is more of a semantic process rather than a literal one, a design translation involves understanding the context and establishing common ground for mutual understanding between the system and the end-user [25]. For UX practitioners who are working with AI/ML, understanding the model to some degree is necessary to bring about such a translation. This necessity applies for UX researchers as well as they have to "*know what questions to ask the user and what that might translate to in terms of user facing features*" (P11).

In this subsection, I describe how such design translations require AI model comprehension, but these models evolve continuously through the training process and even post deployment. Thus, UX practitioners (UXPs) also had to iteratively update their understanding. I detail 3 main stages: informing model requirements (pre-training), post-hoc model comprehension (post-training), and continuous improvement through feedback loops (post deployment).

a. Informing Model Requirements

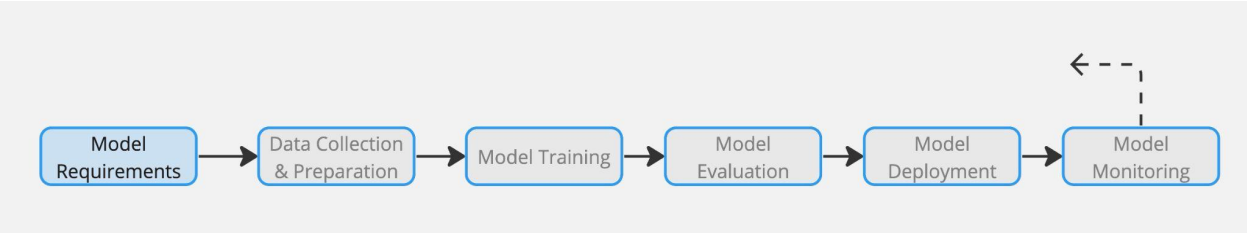


Figure 3.1. A model development pipeline adapted from [5] and showing different stages of the model development process. Only few UXP are involved at the model requirements stage.

Only few participants (P11, P13) discussed this scenario where UXP are involved early on in the model development process (see Figure 1). In such cases, user research is one piece of the puzzle that informs the model development process. P11 recommended a tight coupling between UX and ML practitioners to ensure that findings from user research studies led to features that were possible and feasible to realize through AI/ML capabilities.

> *"It's always tricky to figure out if some of these potential features are achievable. So I think there needs to be a tight coupling between experiment design or study design and kind of consulting with the ML teams. Just to make sure like, should we study this - Is this even possible, right? What are the limitations of types of data we can train on or how much data would you need in order to really see a difference? And then depending*

> *on the technology there could also be cost barriers, right? We need to figure out - Okay, this is like a very limited space in which we can operate with - like testing new features and new experiences. How do I translate that into a study that is not too restrictive and artificial, but could still lead to executable outcomes?*" - P11

Here, P11 is talking about understanding the limits of what would be possible with AI/ML to inform the research study design and obtain actionable insights. The goal here is not just to eliminate blue-sky ideas and design commensurate user research studies, but also work within what would be feasible to build in terms of the type and size of data available and the costs associated with it. However, not many participants discussed translation efforts at this stage except P11 and P13.

b. Post-hoc understanding of the AI/ML model

This stage was the most prevalent among almost all our participants' accounts. In this scenario, the AI/ML model has already been developed by a team of researchers or data scientists (Figure 2). The UX practitioners would then work with the trained model and incorporate it into an existing product and/or design the user experience around it. Unlike the previous case in which UXP are trying to understand the capabilities and limitations of AI/ML in a more generic sense, here, UXP have to understand a specific, trained model insofar as they can design with it.

P14 recalled this example from their user research session -

> "*The data [that the customer provided] was not of the best quality in this case. Sure, we wanted the customers to improve their dataset but we still wanted them to use our predictions. But this customer said to me 'Well, I think if I improve my data, my predictions will be higher quality' and… I realized as a researcher, I wasn't necessarily certain if that was true. If those specific inferences would be higher quality, or if you*

would just get more inferences available. I went to the product manager but we're not sure right? We had to go resolve this with data scientists. And sometimes I don't even realize that I have the questions until the user says this to me, and I'm like, I don't know. I actually don't know what the future would be - how it would evolve if you improved the data." - P14

In this instance, the UX practitioner is trying to understand how the customer data is related to the model output and the model's objective function - whether it optimizes for quantitative or qualitative improvements in its predictions. But P14 also makes a point about how it is not always possible to know how the model evolves in the future with more, better, or different data. P9, a user researcher, who worked as part of the product team to incorporate models developed by the research team in their organization, encountered challenges in mapping outputs of the model to useful information or action items for the end-user.

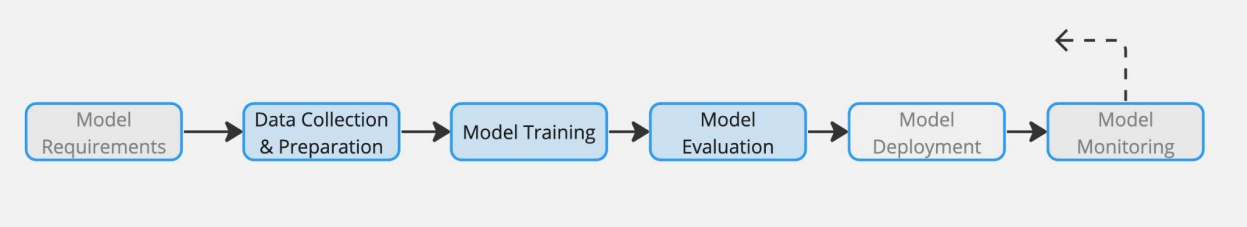


Figure 3.2 Most UXP were discussed these stages when talking about gaining a contextual understanding of the model

"*I didn't have a good understanding of what the thresholds are like... There are different levels of relevance like highly related, maybe just middle, and then low relevance that we maybe shouldn't even show it maybe. But initially, I didn't have good information on what the thresholds are. So, in the prototypes we just show everything and then highlight, like different levels of relevance. And then the users were really*

> *frustrated because a lot of the results were not relevant. If I could kind of understand, "Oh, maybe the lowest threshold is this amount", and then it would be helpful to build the prototype, so that we are not including like, not relevant or unhelpful information, because that would impact users feedback like if it's just too much noise they wouldn't want to use it again. But then it's also kind of biased because we put in too much irrelevant information in the prototype. So yeah, that's an example of what I would want to know uh earlier. But it's also something that I didn't know that I needed to know*" - P9

P9 also added that it would be useful to get insight on whether the model is expected to work well across all user groups or differentially for certain groups of users so that they account for it when executing research studies. P9 also made an interesting comment on how these were examples of things that she didn't know she needed to know. It was a problem of unknown unknowns, from the perspective of the UX practitioner. P13 also reflected how the HAX Guidelines would have been useful for their first AI project as it could have provided concepts that sensitize them or enable them to transform *unknown unknowns* to *known unknowns*.

> "*If I had the Guidelines on my very first project. I feel like I would have been able to do way better, because I was just trying to understand what AI was and what it was capable of, right? I was given this model that already works, and I'm like, "Hey, here's what people could get from it". But I didn't know that I needed to put a 79% confidence until someone told me right. So, there's a lot of different things that I could now have an example of in my head and be able to share with the data scientists like, "hey? This is maybe what we could do", and then they could tweak the model. Or maybe, if we get this other source of data- this would be more accurate, and we could do something more - Yeah, I guess it's like I didn't know what the boundaries are as a designer.*" -

- P13

But as P9 and P13's accounts show, they accrue this knowledge over time and thus form an AI-specific design repertoire over time. One need that was expressed to improve model comprehension and support design translation was the idea of a model sandbox. Noting that model documentation, while important, can only go so far to help make sense of things "*in context*", UXP discussed a hypothetical tool that would allow them to directly interact with the model and explore model outputs for different inputs.

> "*These models are hard to capture in precise documentation. You can say it has this many parameters, and you know the confidence scores or like is this accurate? You could give some metrics, but it's hard to make sense of those metrics in context. Having a sandbox where the UX professionals could actually interact with the model and test like - 'Oh, how would it give me suggestions for smart composition if I give this example versus this example versus this example' right?... So, I think that it's really really important, having internal tools for people to be able to get a sense of model capabilities hands-on and interact and play with it.*" - P11

P5 also expressed the need to see what the model is actually "*spitting out*" as it is different from what they imagine or guess it to be. Echoing the same idea, P7 said, "*We need to improve self-service models for user and user experience researchers. Let the UXR play with the models and understand how the data looks like - the shape of the data - this can help UXR design better research studies. No more black boxes - reveal the underlying assumptions behind the model*" - P7

c. Translation efforts continue post deployment

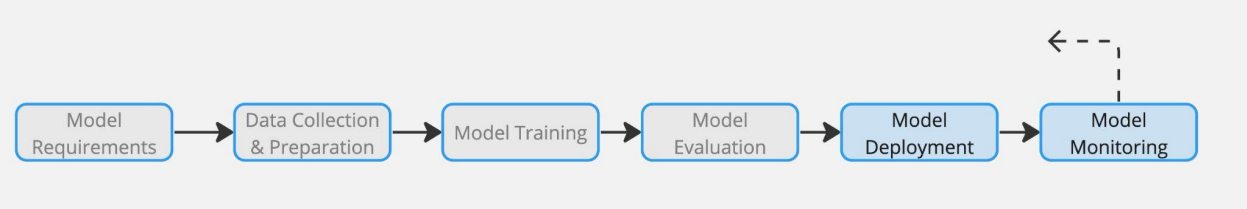


Figure 3.3 UXP continued to understand the model post deployment as it learns through feedback loops

After the model has been deployed, the UXP were analyzing how end-users interacted with the model. With dynamic models especially, UX practitioners had to grow their understanding of the model in terms of feedback loops (Figure 3). For example, P5 described the challenges of working with implicit signals in recommender systems -

> "*One thing I think of, both in my own experience and when I'm designing for our users, it's hard to make corrections if you click on something, and we've read that as an implicit signal that you're interested in this thing. But maybe you accidentally clicked on the wrong thing or you clicked on it and realized it's irrelevant. Now, we're going to learn that that's something that you're interested in, but actually aren't. So how do you override those implicit signals? And maybe, like, keep it from proliferating throughout your recommendations?*" - P5

Though P5 went on to discuss several ways to mitigate misleading implicit signals (using explicit signals only/seamful design, phasing out implicit signals, only using most recent activity), they still acknowledged the challenges of i) aligning with user expectations of what activity should be recorded as implicit signals and ii) encouraging users to provide more explicit feedback to serve useful recommendations. As P5 notes, they had to focus on both how the model evolves as well

as on how users' mental models might co-evolve. Again, not many UX practitioners discussed this stage explicitly. When they did, it was focused more on UX evaluations (detailed in the next subsection 3.2.b) rather than understanding feedback loops per se.

3.3.2 Challenges in Prototyping and Evaluation

In the previous subsection, I focused on UXP's model comprehension efforts based on different stages of the model development process. In this subsection, I focus on the design process and specifically on the stages of prototyping and evaluation.

a. Creating Functional AI Prototypes was Challenging

An important challenge faced by UX practitioners at work is prototyping with real AI/ML functionality.

> "*I think it would have been of limited utility to try and test these concepts on users without the actual AI functioning. I think I can test out some of the user interface aspects of what we want to test. I could test, 'Does the user understand what this button means, for instance, but I wouldn't be able to test - 'Does the user understand when the AI detects something that happens?' And then in real time gives the user feedback like - how would I? How would I mock that up?*" - P12

Wizard-of-Oz studies came up as a useful method in generative or foundational research where a human could mimic AI. Any user feedback obtained from the session could be given as specification of desirable or undesirable model behavior to the engineering team. But since AI/ML will fail in ways that are not easily predicted or imagined by humans, prototyping with working AI/ML was still a necessity. P11 wanted to have the ability to create functional prototypes, test them, and generate insights *before* AI/ML functionality is actually built into the product.

> "*So, I would say that what's missing right now in this field is the tools that allow UX/UI researchers or any other kind of researchers to build rapid prototypes with these technologies. Where could you build very quick prototypes in order to get feedback rather than waiting for the engineering team to put out a product and going and evaluating that?*" - P11

Though few participants shared positive examples of collaborating with engineers to build functional prototypes, this was not the norm and UX practitioners expressed the need to reduce dependency on the engineering team to build prototypes with a working AI/ML component. P9, who worked on a recommendation system, talked through some of the challenges of creating such functional prototypes for recommendation features, specifically the dependency on user data to generate recommendations and obtaining consent for using that data.

In summary, UX surfaced prototyping without functioning AI/ML as a key challenge which limited getting concrete feedback from users. They wanted to have the tools and techniques to independently create functional AI/ML prototypes and generate insights that can inform how AI/ML capabilities should be integrated into the product.

b. Tackling User-centered AI Evaluation

Continuing to the next stage of the design process - testing, UXPs faced challenges that they attributed to i) AI/ML models being probabilistic in nature, which makes it hard to simulate and test thoroughly and ii) reinforcing a user-centered lens on AI/ML outputs. For example, P13 discussed how one of the big barriers they faced as a designer was not being able to test enough use cases. P13 would come up with as many use cases as possible before handing over designs for implementation but not having the support to validate domain-specific or obscure model outputs was a challenge. Additionally, even if the model has an accuracy of 95%, UXP wanted to

investigate what specific outputs are being shown to end-users and how they are perceiving these outputs. But in many AI/ML applications, this can scale up quickly. UXP reported that they did not always have the tools or mechanisms in place to evaluate the outputs of the model in live or production environments, where the model could learn with additional user data. Thus, the scale factor also posed difficulties in evaluating AI/ML models.

> "*One of the challenges I see regularly in my space is we don't have the tools to test the algorithm, the output of the algorithms. Now I can see it has 95% confidence. Right? It is working, but I don't know what people are being shown. I am generalizing it a lot - but we don't always have the tools to see how our model is performing in prod. Because if you are shipping a model that you know, you are affecting 1000, 3000, 10000 people on a daily basis - It's harder to see what the impact is. You either need tools that are going to do it, automated tools which - I don't know how that will be done. Or you gotta take the qualitative approach as someone spot checking it.*" - P3

A strategy that P3 uses is that sometimes they will ask the data scientist specific questions about how they expect the model to work in the product and the impact it will have on the end-user to draw out their assumptions. Qualitative methods are effective in staying close to how a model is performing and ascertain its drawbacks. For example, qualitative analysis of transcripts of interactions with chatbots helped P3 catch instances when the chatbot was wrong in interpreting some information, but there were limits to how many transcripts P3 could read and analyze. P3 added, "*And there was one of me looking at 100 transcripts a day. And we were doing tens of thousands of them a day. That is what I mean by scale*". The scale and output complexity [148] of AI/ML posed difficulties in how UXP gleaned meaningful insights about the user experience and model behaviors. P11 also deduced that satisficing or scrappy research efforts might not

always suffice.

> "*I think we just need to talk to more people. Talking to 5 or 10 people - sometimes it's really not enough, right? So, then the question becomes at a high pace - with tight deadlines - How do you do that? How do you more quickly test or get qualitative and quantitative feedback at scale?*" - P11

P14 also remarked that live testing was less frequent compared to working with prototypes. P14 noted how evaluating AI/ML can get subsumed into other product and business metrics.

> "*We get pulled into other business goals. We might be focused on something simple as whether users understand these 4 steps that they have to do or whether they can simply complete the task right? And yes, there is an AI feature. But are they able to complete it? How long does it take? What do they think the system is doing? Like we might want to make sure that they understand that the system has done something for them. We might want to make sure that they validate a response, if necessary. That they understand how well the AI performed - So from that perspective we are focused on some of these. We want the user to have a positive understanding. But we're just doing a lot of that within the course of our research questions anyway. The business wants to know. Can they complete this task? And do they understand what the system is doing?*" - P14

Building on the idea of the **model sandbox**, UXP continued to brainstorm how it might support prototyping and evaluation efforts. P11 considered how connecting a tool like the model sandbox to Figma or other prototyping software would allow for creating mockups quickly, incorporating real AI/ML functionality in the prototypes, and gathering useful feedback. Other

UXP working as designers such as P5 and P12 echoed similar needs even though they did not call the tool a model sandbox. For example, P5 just wanted an easier way to flow realistic data and model outputs into Figma prototypes. P13 also brought up the idea of a 'model sandbox' to have hands-on interaction with the data and test as many use cases as possible, especially when UX has shorter turnaround times, while noting the following limitation,

> "*I think it also depends on the maturity of your organization. Do we even have time to build that - sometimes on that team of five, four people. We don't have time! But I think if we know that this is going to be a long project, and we're going to benefit a lot from it - putting that infrastructure in place would be great.*" - P13

3.3.3 Cross-cutting Collaborative Challenges

My analysis also uncovered collaborative challenges that UX practitioners faced when designing with AI/ML, particularly with AI practitioners. However, while some challenges are inherent to cross-functional collaborations, others were also a result of organizational processes. I discuss these challenges as cross-cutting collaborative challenges as they can arise during the course of other activities discussed above - model comprehension, design translation, prototyping, and user-centered evaluations.

a. Communication gaps between UX and AI practitioners

Most participants highlighted communication gaps between UX practitioners and AI practitioners. The vocabulary used for describing AI/ML and descriptions of techniques used to build these models were not easy to understand.

> "*There is a huge gulf between the words that the data scientists use and the words that regular people use. How do we talk about these things like I am in my domain, you are*

> *in your domain, right? So, we need to meet in the middle. I think there are a lot of challenges around that.*" - P3

Such gaps were quite obvious in settings such as team meetings and discussions. P3 and P13 noted how it can be difficult to engage, speak up, or contribute when the discussions get too technical.

> "*One time, they went down a rabbit hole about the technology. And this is not a knock on them because they are passionate about it. This is their expertise, right? And this is what they spend all their time trying to think through. But that makes it very hard to engage.*" - P3

Others reported more positive experiences because they worked in tight knit teams with more impromptu conversations (P5), felt comfortable asking basic questions (P5, P12), or leveraged their technical backgrounds to understand things (P6, P8). However, participants did surface needs in areas of communication and documentation. UXP wanted to have more routine, standardized practices of syncing up with AI practitioners, which would give them space to ask questions.

> "*If I don't understand something, I need to go and resolve it, and I often do that with the Product Manager (PM) and sometimes the PMs do have to loop in the data scientists. So, if there was a formal mechanism to maybe make sure that questions are resolved with data scientists before we launch something, that might be helpful.*" - P14

While UXPs do closely collaborate with PMs, in this case, having them as a proxy was not as effective as looping in the data scientist directly. P13, when working with an external consultant, received detailed, helpful documentation which led them to realize the value of having such things in place for alignment.

> "*If I built it internally, I wouldn't make a Powerpoint, or this crazy presentation, because it's like we work together right? But maybe again too - it's like setting standards on how to communicate would be helpful right? If there was a document template that I could give to a developer or data scientist, you know? That would be great*." - P13

P8 also raised a similar need for having documentation that could align the interdisciplinary team on what AI capability is being introduced, what problem it is solving for, and how. Though participants alluded to documentation templates and things useful for designers to know, it can be difficult to articulate a precise list of what should be included, especially in hand-off situations. Some of this uncertainty was because of the inherent nature of AI's capability uncertainty and some of it was because they are unknown unknowns for UXPs (3.3.1).

b. Visibility into AI/ML workflows

Participants also reflected on how the processes and stages of model development remain largely invisible for them. But making them visible and relating it to model behaviors and outputs can help increase AI literacy and support design translation. Being a designer, P12 felt he had to understand how the model gets trained and learns patterns to be able to design for it. P14, a UX Researcher, said knowing what training data was used helped contextualize findings from user research sessions better.

> "*I would like to understand how the data was gathered to train the model, because we often don't have access to it. Data scientists get very crafty with how they're collecting representative data. There are lots of different mechanisms for feeding into the training. As you think about testing it with potential users, they're going to have*

> *different feedback right on its outputs, and that could potentially, I don't know, be dependent on where the data came from and how it was trained. If you're using the synthetic data generator, it probably has its own limitations relative to a different data source.*" - P14

Additionally, only when P14 came to understand the different complexities that AI practitioners deal with as part of internal research, they realized that some of these details would be useful for UX practitioners to know as well. For example, they did not know how many models get tested before one is selected for deployment but knowing factors that led to its selection would be useful to know. P5 also liked to know timelines for developing the model and when it is deemed ready so that it aligns with design iterations and/or user research studies.

> "*The model was not really ready anyway. I started seeing way more errors than I expected to see and super low confidence scores. And the model hadn't really been calibrated yet. So I ended up changing my user study that I was planning and shaped the questions to be more exploratory. I think they were aware of the problems already. So it was just me realizing the extent of the problem.*" - P5

P3 felt that the lack of visibility also led to problems in setting realistic expectations and curating commensurate datasets. P3 noted how stakeholders expected AI practitioners to develop models on the fly when provided a problem or a dataset but the reality was far different. They also speculated about tools that can provide visibility into the process.

> "*Thinking about where the model is built. A lot of times it is just built in like whatever program the AI practitioner is using. And UX-ers are not even gonna know what that is. What would a UI look like that made a model visible, as it is being built? You know*

> *where they [UXP] can say 'I see where you are right now, what are you thinking here?' I don't know if that would be doable. . . That could help.*" - P3

The need to make AI development workflows visible to the extent that it can set the right expectations, provide contextual details, and coordinate UX efforts was highlighted. This provision can also help make more implicit information and assumptions explicit.

c. Synergy between UX and AI

UXPs discussed the challenge of i) reconciling user-centered outcomes with model-centric metrics and ii) having AI developers understand end-user perspectives. When integrating models developed by in-house research teams, the product team had to work on why and how it added value to the end-users. But since the goals and motivations can differ, it introduced challenges in collaborations and evaluations.

> "*From the product side, if we're testing a model, even if the results are like our users don't find this helpful at all, maybe they don't see themselves using it as part of the product - It's fine because we're just trying to build whatever is most useful, and if this doesn't work for them, then we shouldn't put it on the product. But I guess, for the research side, it is more like validating it - Is this much better than the previous version, but not trying to answer if we want to add this to the product, what is the best place? Where would this fit into the existing workflow... If we look at it by itself, it works better. So I feel like the considerations are just a bit different.*" - P9

P5, who worked in a similar setting, noted that one of the questions she finds useful in surfacing was - "*Which ones are really cool for science and academia and which parts are actually going*

to be useful to our users?". However other participants who worked with dedicated AI practitioners still surfaced a variant of this issue.

> "*This is actually more of a personal opinion but they are looking at the model performance. They are not looking at the qualitative side of the impact. And so there is a big gulf there. Just because your model is performing correctly doesn't mean that users are, you know, actually enjoying the thing that you are showing them, right? It is looking at outcomes versus outputs.* "- P3

This divergent lens through which model performance is viewed was also a source of tension in collaborations between UX and AI practitioners. Participants reported trying to involve AI practitioners in user research so that they can gain a better understanding of user perceptions and behaviors with respect to the model. Specifically, participants hoped to show differences between users' mental models of the AI/ML model, how they expect it to work, and how the model might actually function. P13 also hoped to inspire human-centered approaches to model development by involving AI developers in user research.

> "*For them to understand what our users are saying. There's a very broad spectrum of user understanding when it comes to the model performance and calibrating users to that right? Beyond research report-outs. . . sometimes, and they [users] say. 'Oh. 99%, that's awesome, but why is it red? Why is it showing red?' Because that's like an overfitting example. And there are users that understand that perfectly - that 99% is bad. And there are users that think something might be wrong as the model can't be this accurate. And so I also want data scientists to have that awareness of how users think about these things."* - P14

> "*Yeah, I think it would be for them to understand who's using it right, and really getting all these folks engaged more so, into actual use cases. So then we can start building and thinking towards that. I think it's getting them involved in user research at the beginning, understanding what data we have.*" - P13

Having technical stakeholders involved in research studies was also helpful for participants like P9 as she did not always have the right answers to end-users' questions about the model. P11 had a practice of showing video highlights of users interacting with the model or its outputs to challenge assumptions or misconceptions that AI developers might have of the end-users. P3 reported experiences of being involved in feature engineering as they knew the product and customers well, but this was not an organization-wide practice.

d. Establishing a process for collaboration

The final challenge raised by participants was that there was currently no established process for collaborating with AI practitioners and as a result, they came in too late. UXPs acknowledged that it was not easy and time-intensive to figure out what this process should look like, especially when there is pressure to ship features to production.

> "*Having designers work with ML engineers and data scientists. It's easier said than done. And when you look at the traditional structure of all organizations, it is - this is what I do, this is what you do. And to bring someone in knowing when to bring them in, how to bring them in, and you are always up against the battle of 'This is slowing us down!'. A lot of times, what I notice is, they just want to get a model out. Ship it, prove it works. And that is a different incentive, than say, are we getting it right for our customers.*" - P3

P12 acknowledged that being engineering and/or product driven and involving the UX / Design discipline too late into the development cycle was not unique to AI/ML applications. But he still believed there was value in getting a seat at the table early on as that was "*when real design work was being done*". P3 also believed in bringing UX practitioners early into the model development process, ideally during planning stages, to focus discussions on "*what it means to be impacted by these systems rather than building a model that performs well*".

Other participants highlighted similar tensions in following product-first vs model-first workflows. In the former approach, the product team had opportunities to report user needs or problems that required AI/ML-based solutions. In the latter approach, the team had to figure out how to make a model that had already been developed fit into the product. P13, who worked in the enterprise space, has done both, as it varied on a per project basis. He also aspired to explore co-design methods that would allow better collaborations between AI and UX practitioners, but was not sure of its feasibility given the technical details, data dependencies, and different disciplinary norms. P5 and P9 worked for organizations that had a research team in charge of developing AI/ML models. P5 accepted that it made sense that only some of these models might make its way into the product, but P9 wished for closer collaborations to develop models that actually meet user needs and to reduce rework.

3.4 Discussion

Yang et al. [148] attribute AI's design difficulty to capability uncertainty and output complexity. While these factors do contribute majorly to the challenges practitioners face, collaborative and organizational factors are also at play. In UXP's perspective, technology readiness to support individual and collaborative workflows with AI was also lacking. More experienced UX practitioners were able to speak to many aspects of designing for AI compared to less experienced UX practitioners who needed support to tackle unknown unknowns.

However, both groups raised the need for tools that would enable them to have a tangible and contextual understanding of the model's capabilities and limitations, create functional AI prototypes, and evaluate the model from a user-centered lens. Participants named such a tool as a 'model sandbox', which I adopt and unpack further in this section. Below, I explain i) why and how a model sandbox can support design translation for UX practitioners and ii) how model sandboxes help rethink UX-AI collaborations.

3.4.1 Supporting Design Translation and Model Comprehension for UX Practitioners

Participants in our study described different instances of engaging in design translation between AI/ML capabilities and end-user needs, and their efforts to understand the model to some degree. Across the examples discussed in the findings (3.3), a few insights emerge. Unlike a machine learning practitioner who mainly assesses a model through metrics, UX practitioners' assessment of the model and its behavior was connected to mapping implications for the end-user. Rather than a causal effect of the former leading to the latter, these two aspects are co-constructed, especially over time. Instead of a rigid 1:1 mapping, these translation efforts can be perceived as going both ways before a connection has been learnt. To that end, UX practitioners presented the idea of the model sandbox - a hypothetical tool that would allow them to have more hands-on experience with the model, prototype with real AI functionality, design commensurate user research studies, and test different use cases. Access to training data and distributions can also provide insight to UXP about "*where things are coming from*" and compare it against anticipated user behaviors.

A key challenge in this process that was subtly mentioned by UX practitioners was they deal with unknown unknowns. The ability to ask questions is dependent on having some level of

pre-existing knowledge on the topic i.e., one must know enough to know what is not known [103]. In the data, we also see how i) accumulating designerly ways of knowing [31] and ii) reflective practice helps UX practitioners in this aspect [166]. For example, a few practitioners attested to how working in AI/ML helps them recognize it in other products and use them in ways they did not previously. This is an example of UX practitioners developing codes to read design knowledge embodied in other AI products [31]. When probed for what details P9 would like to know about the model, she recalled cases when the model did not perform as well for some users and added that she would like to know if the model worked better for certain groups of users. Thus, building a '*repertoire of examples, cases, and techniques'* [166] over time does help UXP become more fluent in asking questions of the model and examining design tradeoffs in AI/ML applications.

The structure of reflection-in-action is as follows. First the practitioner faces a peculiar situation, uniquely different from prior problems that they have tackled [166]. Recognizing these peculiarities, the practitioner reframes their inquiry to make the problem more tractable. Then, they make moves, which can be either exploratory, playful or probing moves or they can be deliberate moves intended to create a desired effect [166]. This move-testing phase can lead to more moves, learnings, confirmations, or unintended effects and further reframing of the original problem [166]. This is also when the situation talks back to the practitioner, enabling a reflective conversation. More importantly, these experiments and reflective conversations happen in a virtual world of practice, where the practitioner can make moves that are reversible and without real-world consequences. The model sandbox, as speculated by many of the participants, should provide a virtual, interactive space for precisely such experimentation and problem reframing. To the extent possible, it should allow for simulating different user actions and discovering errors from a user-centered perspective.

Uncovering unknown unknowns and converting them into known unknowns can also be more of a challenge for novice practitioners, who are new to designing with AI and ML. It can also be viewed as the paradox of reflection-in-action and learning by doing [126,131] where the concepts to understand what needs to be learnt is only discerned by engaging in action. Thus, we should also consider implications for pedagogy by thinking about how we might support these experiences in design education. For example, how might we design a reflective practicum, where we construct the right set of experiences with AI and ML such that students can learn, fail, and reflect-in-action? While Slovák et al. [131] put forward a framework for scaffolding such learning and reflection processes in the context of social-emotional learning studies, extending it to address the quirks AI/ML poses for design curriculum is a worthwhile direction for further research (e.g., [167]).

3.4.2 Supporting UX-AI Collaborations - Revisiting the Idea of Boundary Objects

The idea of boundary objects is commonly used to recognize and describe artifacts that enable knowledge sharing and communication across boundaries of different groups. These boundary objects maintain a common identity across groups but can be used quite differently within each group. Prior work investigating UX practice of AI has largely focused on boundary objects to recognize, uncover, or design boundary objects that support UX-AI collaborations [104,147,152]. However, communication and knowledge-sharing were not the only challenges raised by UXPs. They also brought up more classical CSCW challenges of establishing processes for collaboration and having visibility and awareness into each other's work. Specifically, they brought up the need for a model sandbox to improve both UX practice of AI and collaborations with AI practitioners. Some practitioners viewed it as a self-service tool to minimize dependencies with AI practitioners. Others saw it as an active site for cross-functional

collaboration. But rather than boundary objects, such tools are akin to groupware systems [47]. Thus, I turn to the idea of groupware systems to derive concrete design implications.

Groupware is software for groups. They are systems that support people in achieving a common goal by providing interfaces to a shared environment [47,48,62] and have a rich history in the CSCW literature. Email, video/audio conferencing, digital whiteboards, calendars, collaborative documents are all examples of groupware and are commonplace in today's workplace. Generally, they can support one or more of the following: communication, collaboration, and cooperation (the 3Cs) [47]. For example, email mainly supports communication. A collaborative document editor supports synchronous and asynchronous collaboration. A software project tracking software like JIRA supports coordination of different activities. Groupware can also be considered as a collection of modular functional components, providing a particular set of capabilities, tailored to the group's needs [71].

The groupware's underlying design principle - providing shared access to information, is key to resolving some of the challenges surfaced by participants. There are examples of such AI specific groupware systems in the literature. ZIVA, a tool created by Park et al. [115] improves collaborations between domain experts and data scientists by providing a common interface and shared access to domain knowledge. AIMEE, a tool by Piorkowski et al. [117] enables non-experts of AI to understand and edit an existing model through rules, improving communication and collaboration between data scientists and their clients. What might a groupware system designed to support and improve UX-AI collaborations look like? What components and features might be useful? I offer some design directions below based on the above findings and groupware literature.

Providing ways for UX practitioners to have more visibility into the model development process could be a required component in such groupware. As UXPs noted, it would allow them to increase their AI literacy, plan for UX activities, and map implications for the user experience.

Viewing more contextual information from current and prior stages including what training data was used, what features were selected, and what factors went into selecting a particular version could be particularly helpful. Beyond providing visibility, the groupware system can also incorporate features for better coordination of UX and AI activities. For example, notifications can be sent when the model passes certain evaluation criteria for UXP to start tinkering and prototyping with it. Thus, the set of features can focus on i) providing visibility and awareness and ii) coordinating UX and AI development activities.

Another component of the groupware can focus on collaborative analysis and evaluation of the model. As many of the UXP highlighted the challenges of reconciling user-centric outcomes with model-centric metrics, this set of features can help bridge the divide between i) qualitative and quantitative approaches and ii) model-centrics vs user-centric results. Specifically, it should make it easier for UXP to verify qualitative findings at scale and correlate with model metrics, and for the AI practitioner to understand users' mental models and behaviors with respect to the AI model.

Prior literature helps us learn when and why such groupware systems might fail [61,62]. I draw from these lessons to note additional considerations for building groupware for UX-AI collaborations. For example, user interfaces and actions that can be taken within that interface should be tailored to the needs of each practitioner's role. A particular groupware can offer certain features for AI practitioners themselves, but these may or may not be useful for Product Managers and UX practitioners. If they are, they need to be sufficiently personalized to meet their specific information needs. The action and operations they can undertake should also match their needs and expectations.

While groupware helps support collaboration, it should not hinder individual and single-user workflows. A good example of this principle is the online document editor. It allows people to collaboratively author, edit, review a document but does not hinder individual use either. In this case, UXP should still be able to conduct user-centered model evaluations or prototypes

with real inputs and outputs without having to trigger collaborative workflows. Another well noted reason why such systems fail is the "*disparity between who does the work and who gets the benefit*" [61]. Building an AI model specific groupware system or adapting existing systems requires significant effort. If the onus falls on AI practitioners to develop and maintain such systems, they should get proportionate benefits. Automating certain features in the groupware can reduce the work and effort involved. For example, could we automate the process of creating basic model documentation as well as the process of tailoring it to different practitioner groups? How might systems like Gradio (a Python library allowing creation of front-end/UI components for AI/ML models) be used to reduce the work involved in sharing the models with UXP and getting concrete feedback? Finally, when what the groupware system offers is not aligned with different practitioner groups' expectations, they tend to fail [61].

While I argue that groupware can provide a more dynamic and interactive approach to UX-AI collaborations compared to boundary objects, I also note that they can resolve technical challenges to collaborations rather than address social challenges directly. That is, groupware might be less helpful if the organization does not incentivize collaborations in the first place [62].

3.5 Conclusion

One limitation of this work is that it highlights various collaborative challenges, but only from the perspective of UX practitioners. Interviewing AI practitioners or conducting case studies of what cross-functional challenges arise in the end-to-end process of AI application development can provide a more holistic understanding. Participants in our study worked under different team structures which provided diverse perspectives, but there could be others not accounted for in our study. For example, none of the participants reported enmeshed team structures or working more closely with AI practitioners. Another important limitation of this

work is that it generalizes across different roles UX practitioners take on (such as design, research, content) as well as across different complexities of AI systems. Yang et al. [148] classify AI's design complexity into 4 levels: probabilistic systems, adaptive systems, evolving probabilistic systems, and evolving adaptive systems. Different participants worked on different levels of complexity, with the majority of them working on evolving probabilistic or adaptive systems. However, I did not account for more specific challenges that can arise from working with different input or output modalities (example: text vs voice) and different AI capabilities (predict vs classification vs generation).

Despite these limitations, this study highlights common challenges UX designers and researchers face when designing with AI, particularly with respect to model comprehension, design translation of AI capabilities, prototyping with AI, evaluating human-AI interactions, and collaboration with AI practitioners. Based on the findings, I propose going beyond boundary objects to design groupware-based model sandbox tools that can better enable cross-functional teams and UX practitioners to design human-centered AI experiences.

Chapter 4: Mapping the Design Space of Tools that Support Interdisciplinary AI Application Development

4.1 Introduction

In the previous chapter, I propose leveraging a groupware approach to create tools that can support interdisciplinary collaborations between UX practitioners and AI practitioners. Given that groupware systems host key features to support communication, collaboration, and coordination, I argue that they might be more effective in supporting HCI-AI collaborations than boundary objects. But do such groupware systems exist at all? If so, how do they support AI application development and associated collaborations? If not, what other kinds of software tools support interdisciplinary AI development efforts? To answer such questions, we must be able to analyze the specialized tools used for AI application development. Thus, this chapter focuses on a design space analysis of tools supporting collaborations in AI application development.

Just as a particular design can constrain or enable individual user actions, it can also do so for collaborative actions [55]. Different tools can support different types of collaborations. Johansen categorizes collaborative tools based on two dimensions - time and space, resulting in four categories of collaborative tools [77]. These support collaborations at i) same time/same space, ii) same time/different space, iii) different time/same space, and iv) different time/different space. Thus, they can support collocated or remote and synchronous or asynchronous collaborations.

When analyzing collaborative work and systems, the time and space dimensions have been prevalent in CSCW research. However, my research focuses specifically on the domain of work: developing human-centered and responsible AI systems. While the above dimensions (time, place) are factors in HCI-AI collaborations, the focus is on providing visibility and awareness, translations between two disciplinary groups, establishing processes for collaborations, and synergizing efforts (Chapter 3). This focus better relates to a dimension in Lee and Paine's [89] updated Model of Coordinated Action - number of communities of practice. This dimension can be understood as the number of disciplines involved in achieving coordinated action. Collaborations between practitioners of different disciplines can create discord as there can be differences in vocabulary, artifacts used, and norms and practices followed [89], thus making this an important dimension to consider in a model of coordinated action in addition to the usual suspects (time, place). The question of how practitioner tools might be better designed and developed to support interdisciplinary collaborations hence becomes a relevant and interesting question to explore. Rather than a narrow focus on HCI-AI collaborations, which will reduce the number of samples that can be analyzed, I focus more broadly on interdisciplinary collaborations in AI application development.

Following the results of Study 1 (Chapter 3), one objective of this research is to analyze to what extent tools adopt groupware design principles or not. To do so, I inductively analyze the design space of a set of 18 tools that are built to support individual and collaborative practices of AI application development and have corresponding research publications in HCI-related venues. From this analysis, I derive a design space consisting of seven key design dimensions (Intended Users, Axis of AI work, Tool Architecture, Semantics of Use, Artifact Type, Artifact Availability, Collaboration Goals) and four design spirits of collaboration (Groupware spirit, Core practice & Communication spirit, Community of Practice spirit, and the Visibility & Bridging spirit).

Another objective of this research is to complement empirical studies that surface challenges, strategies, and needs of practitioners working on AI applications as part of AI, engineering, design or product functions [4,37,66,70,107,108,135], including my first investigation. Such studies also provide implications for designing new artifacts that can alleviate practitioners' challenges, among other implications. But we do not know i) whether these findings inform the design of new tools, ii) what problems are being solved when designing these new tools, and iii) what implications they carry for interdisciplinary collaboration. Thus, this design space analysis is a step towards complementing these prior studies and taking stock of what types of tools have been introduced so far. The goal is to go beyond a categorization of discipline-based or problem-based tools.

I will first describe what a design space and design space analysis (4.2) constitute, and the specific methods used to collect and analyze the set of 18 tools (4.3). Next, I describe the design space derived from inductive analysis (4.4 and 4.5), followed by reflections on the design space (4.6) and implications of this design space (4.7).

4.2 Design Space Analysis

A design space represents a set of possibilities [129]. Since the design endeavor involves several decisions and tradeoffs, a design space helps organize and capture them [45,64]. Key properties that can be considered and modified are represented as dimensions, and each dimension is associated with several alternatives, among which one can be chosen. MacLean et al. [99] came up with a Q-O-C representation where Questions (Q) raise important design challenges, Options (O) present potential solutions, and Criteria (C) lists ways to evaluate different options. Their goal was to capture different design decisions in the Q-O-C format as they were designing artifacts (a scroll bar). But the design space also helps with a retrospective understanding of why things were designed a certain way. In fact, design spaces provide descriptive, evaluative, and generative power [10]. A design space can describe a range of systems by representing the key

dimensions that differentiate them. It can also help evaluate and assess multiple alternatives for the same design dimension. Finally, it can support creation of new designs through exploration of new alternatives for the design dimensions.

A design space can be visualized using a Cartesian representation [129]. Different dimensions would correspond to different axes. Options for a specific dimension will occupy different points in that axis. Different artifacts would thus occupy distinct points in this n-dimensional space. Such a representation bodes well with Simon's notion of finding a design solution for a particular problem and for parametric design [130]. But design spaces can also be more conceptual and metaphorical in nature, as represented by design workbooks and annotated portfolios [54,54]. Johansen's 2*2 time-space matrix that categorizes groupware systems serves as a classic example of a design space in the CSCW literature. More recently, Zhang et al. [156] put forward a Form-From model, characterizing the design space of social media systems, with respect to the form and source of its content. But this 2-dimensional design space can be expanded to consider a total of 62 dimensions. Lee et al. [92] contributed a design space of intelligent, interactive writing assistants that is represented by 5 aspects, 4-10 dimensions within each aspect, and several options within each dimension. The focus of constructing a design space can thus be varied but will dictate the selection and number of design dimensions to be analyzed.

In this chapter, I present a design space of how practitioner tools support interdisciplinary collaborations that contribute to various aspects of AI application development. It consists of 7 key dimensions:

- Intended Users
- Tool Architecture
- Axis of AI work
- Semantics of Use
- Artifact type

- Artifact Availability
- Collaboration function/goals

In addition to these design dimensions, I also present four design spirits a tool can embody to support interdisciplinary collaborations.

4.3 Related Work

Considering a broad definition of a design space, there are several prior works that have presented a design space related to AI systems or applications. Morris et al. discuss two design spaces at the intersection of HCI and generative AI (genAI) models [105].One focuses on how HCI can help interface with genAI, and thus focuses on the dimensions of input and output. Each of these have sub-dimensions, with several possible options for each sub-dimension. The second design space focuses on how genAI, as a tool, can support HCI research and practice. Dimensions cover different aspects such as the goal of using genAI, the role genAI plays, the extent to which it retains context, how genAI features in the design lifecycle, media types, and fidelity levels, with corresponding options for each dimension. Such design spaces capture current efforts in these areas, but they also help spur new creations and investigations.

Others have presented a similar set of design possibilities and feature summaries, as part of their research contributions, even though they did not explicitly label it as a design space. For example, apart from interviewing and surveying AI practitioners to evaluate fairness toolkits, Lee and Singh [91] conducted a comparative assessment of 6 fairness toolkits, reviewing features such as tool setup, licensing, classes of models covered, group and individual fairness metrics supported, and techniques for bias mitigation. They also uncovered areas where these tools fall short of meeting practitioner needs for developing real-world fair AI systems. Cabrera et al. [20] present one of the tools included in this design space analysis. But they also present an analysis of how various AI development tools support different stages of the AI model sensemaking process.

Going beyond tool's features that support usability, Wong et al. [145] how tools envision and represent the work of ensuring ethical AI systems. They looked at various aspects of the tool including its source, intended users of the tool, how its motivations and use cases were framed, guidelines it drew on, and its form factor. Through analyzing 27 AI ethics toolkits, they found that most of these toolkits i) focused more on technical aspects rather than the social aspects of doing AI ethics work, ii) failed to provide features that can actually engage interdisciplinary groups and non-AI experts, and iii) emphasized solutionism over restructuring processes, value systems, or business models that incentivize shipping models that can be potentially harmful. My design space analysis is somewhat similar as it identifies how current practitioner tools imagine interdisciplinary collaborative work to be enacted and encode these assumptions into the tool features. However, rather than a pure discourse analysis, I also analyze structural features of the tool.

Other examples of design spaces focus on end-user applications of AI rather than tools that support practitioners. These include a design space of how humans and AI can interact in text generation tasks [24] and a design space of AI-assisted tools for conducting academic research [150]. Design is also a broad term, and the outcomes of the design activity can be new physical products, digital tools, services, processes or methods. If we go beyond technological tools, Lai et al. [85] present a design space of different study designs that aim to understand and evaluate human-AI decision making. They organize and describe the design space of empirical studies in human-AI decision making using three key dimensions - decision tasks, elements of AI assistance provided, and evaluation metrics. Dow et al. [46] put forward a design space that captures important choices in designing evaluations of generative AI models. These include evaluation setting, type of task that is being evaluated, source of model input, mode of interaction with the model, duration of evaluation, metric used, and scoring methods. Though these design spaces do not focus on a tangible artifact, the purpose is similar - to understand

different aspects of what we are designing for, support explorations and alternative choices, and build a more systematic understanding.

4.4 Methods

To conduct a design space analysis, I first had to curate a set of relevant tools. I adopted a criterion-based purposive sampling approach [114], thus identifying instances that are information rich. Specifically, the tool has to support AI model/application development and support collaboration as well. While this could include a range of open-source or proprietary tools, I focused on tools that were published as a technical research contribution, and thus had accompanying research publications. This research publication explained use cases related to communication or collaboration and detailed the tool's functionality, providing necessary information for constructing a design space.

I first searched broadly on the ACM Digital Library for publications that had relevant keywords in the title and/or abstract (*AI, ML, artificial intelligence, machine learning, data science, work, organization, model, practice, design, engineer, collaboration, communication, challenge)*. Out of a total of 1018 results, I selected a subset of 67 articles. But 58 of these were purely empirical work while 9 presented technical contributions. By cross checking for common references (*papers that this paper cites are also cited by)*, I identified 9 additional research articles. The set of 58 articles further informed my understanding of broader challenges associated with working in AI and with interdisciplinary collaborations, not just UX-AI collaborations. All 18 tools in the final sample are associated with a research publication, present a tool as its main contribution, propose tools meant for practitioners working with AI, and situate the tool in an organizational context. This set is listed in Table 4.1.

Tool	Citation	Tool	Citation
T1- ZIVA	[115]	T10 - Gradio	[1]
T2 - ModelLens	[168]	T11 - Angler	[122]
T3 - Symphony	[9]	T12 - Deblinder	[18]
T4- Zeno	[19]	T13 -Interactive Model Cards (IMC)	[30]
T5 - ProtoAI	[137]	T14 - Canvil	[50]
T6 - AIMEE	[117]	T15 - ChainForge	[8]
T7 - AIFinnity	[20]	T16 - AI Playbook	[69]
T8 - DocuML	[12]	T17 - Marcelle	[52]
T9 - fAIlureNotes	[104]	T18 - PromptInfuser	[116]

Table 4.1 List of Tools Examined for Design Space Analysis

4.4.1 Data Analysis

I subscribe to the notion that artifacts embody specific claims about users' tasks, contexts of use, and user behaviors [23] for my analysis of the practitioner tools. Specifically, I analyze design decisions made to solve a particular need and the ways in which these decisions also support or hinder collaboration. If the tool is making claims about individual and collaborative practices that constitute designing AI applications, what are they? Since all tools have an associated research publication, I was able to also analyze i) the formative research and objectives that informed the design of the tool, ii) descriptions of the tool and its

implementation, iii) scenarios of use, and iv) findings from evaluating the tool with practitioners.

I followed an inductive, iterative approach to the analysis. In the first round of analysis, I uncovered as many factors and potential design dimensions as possible. Examples include primary, secondary users and tertiary users, the kinds of AI models the tool focuses on, input and output modalities of the AI model, specific stages of the model pipeline or development lifecycle that the tool supported, features that supported collaboration, and artifacts that could be created, modified, or exported from the tool.

In the second round of analysis, I would iteratively try to select design dimensions. This involves reading and analyzing the design of the tools, coming up with an initial list of design dimensions, noting down what points each tool occupied in the design space based on these design dimensions, and revisiting the tool designs to see if any important and differentiating features were still not captured by the design space. This iterative process allowed me to arrive at the key set of 7 design dimensions. The fit of the design space to the set of tools was verified by ensuring that aspects of the tool that were important for communication or collaboration were captured in the design space.

As part of this design space, I also present four spirits of the tool in supporting collaboration. These can be thought of as higher-level categories that represent distinct combinations of options for the design dimensions. These spirits were uncovered when conducting deeper analyses of tool similarities and differences. For example, how were two tools that catered to the same type of practitioner similar or different? If two tools had similar goals for collaboration, did they leverage the same kind of technical features? Through such compare and contrast processes, I arrived at four archetypes or *spirit of the tool in supporting collaboration*. The term was adopted from DeSanctis and Poole [38] and I provide more contextual details for adopting this specific term in Section 4.5.2.

4.5 The Design Space

First, I will describe each of the 7 key design dimensions. Then, I will describe the four collaboration spirits and justify them using examples, tool descriptions, and the design dimensions.

4.5.1 Design Dimensions

a. **User(s)**: This dimension captures who the intended users of the tool are. Across the tools analyzed, the intended users varied. They were often different practitioner groups involved in the design and development of AI applications. The options included AI Developers, Data Scientists, Domain Experts, UX practitioners, Research Scientists, Product Managers. In a few cases, two different types of practitioners were the intended users. For example, ZIVA (TI) was intended for use by both domain experts and data scientists.

b. **Tool Architecture**: This dimension captures information about the tool's implementation and architecture. The tools were either web applications, plugins, or programming libraries, but all of them supported a front-end component. At the core, the tools either extended an existing application's functionality (e.g., PromptInfuser (T18) - a Figma plugin) or was built as a standalone application (e.g., ZIVA (T1)). When supporting more tightly coupled collaborations, it would make sense to move away from tools that are bespoke to a particular practitioner group and create a new one. This approach would also allow for more flexibility in supporting desired features. On the other hand, in cases of loosely coupled collaborations or to add on AI-specific task functionality, extending the tool that is predominantly used in a practitioner's workflow through plugins or new libraries might be a preferable approach. Thus, this dimension

can help understand collaboration-related goals or constraints that inform the tool's architecture.

c. **Axis of AI work**: This dimension captures the specific domain of AI-related work practices within the broader cross-functional practice. The options or axes can be: Model development, Model application, and Responsible AI. As the name suggests, the first axis refers to work done to develop the AI model and its capabilities. It includes tasks such as data collection, feature engineering, model training or fine-tuning, prompt designing, model evaluations, behavioral analyses, error analyses, model re-learning, and so on. Thus, the Axis of AI work can have a sub-dimension (Task), though it can be quite open-ended. AI practitioners would lead these tasks, but they could also collaborate with domain experts or clients to establish model requirements.

The Model application axis refers to work done to leverage the model capabilities towards a particular use case. This is where software engineers, product managers, and UX practitioners would come into play. These practitioners also need to collaborate with AI practitioners to incorporate the model into a service or a product. The final axis, Responsible AI (RAI), refers to work done by practitioners to ensure that the ultimate artifact - the AI system, follows principles of accountability, fairness, and transparency. Though one could argue that RAI efforts should be conjoined with the development axis, I refrain from doing so for two reasons. Firstly, RAI efforts are multifaceted in nature and often involve varied stakeholders [39]. Secondly, tools to support core model development and tools that support RAI can be quite different in design. Having a distinct axis helps capture differences in tool design and ways in which they seek to support interdisciplinary collaborations.

d. **Semantics of Use**: This dimension captures whether the tool supports single-user semantics, collaboration semantics, or both. With single-user semantics, the practitioner can only see the results of their actions within the tool. With collaboration semantics, the practitioner can also see results of other practitioners' actions within the tool. A Notepad follows single-user semantics, but a Google Doc follows collaboration semantics. Thus, this dimension, adapted from the design space of collaboration architectures [40], directly carries implications for collaborative work. For example, why and when should a tool support single-user vs collaboration semantics? If a tool has single-user semantics, does it support collaborations in other ways? From the set of 18 tools analyzed, 4 of them leveraged collaboration semantics (T1, T2, T10, T14).

e. **Artifact Type and Artifact Availability**: This dimension captures i) the type of artifacts generated within the tool and ii) how the artifact is made available beyond the tool, if at all. The options vary for artifact type, and it is an open-ended category. Some of the artifacts supported by the tools analyzed were model documentation, data documentation, user interface designs, data slices, error reports, visualizations, and model metrics. Some artifacts were static and some interactive. Some tools had features to export even the interactive artifacts in a static format to share with collaborators and stakeholders. Also, some tools were designed with a sole purpose of creating specific artifacts and making it available to collaborators.

Artifacts can support coordination of different activities and articulation work, translate information across boundaries, negotiate these boundaries, and sensemaking in hand-offs and asynchronous collaborations [88,125,128,132].This dimension helps analyze the ways in which these artifacts are created, shared, and distributed to different interdisciplinary practitioners.

f. **Collaboration goals**: This dimension captures the kind of collaborative activity that tool designers intended to support. This was also an open-ended dimension and the options included: knowledge sharing, knowledge alignment, co-creation, coordination mechanism, artifact hand-offs, and so on. At a higher level, the tool can support communication, coordination of different activities, and synchronous or asynchronous collaboration. When analyzing different design decisions, it is also helpful to map them to intended collaboration outcomes. That is, what was the designer's goal for creating this tool in a particular way? When creating new tools, this dimension can also help uncover implicit assumptions of how collaborations are enacted, who is involved in the collaborations, and what challenges are being addressed.

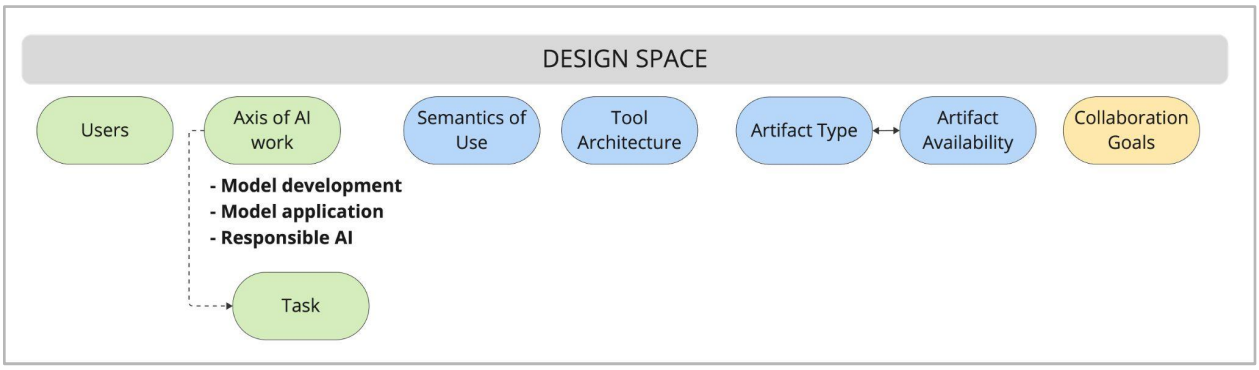


Figure 4.1: The proposed design space and its dimensions. Users & Axis of AI work intended use of the tool, while collaboration goals capture designer's intent for supporting collaborations. The other dimensions capture structural aspects of the tool.

4.5.2 Spirit of the Tool in Supporting Collaboration

Taking stock of the design dimensions, we can see that the design space captures intended use of the tool through dimensions such as Users, Axis of AI work, and Collaboration goal. It also captures structural aspects of the tool through dimensions such as Semantics of Use, Tool

Architecture, Artifact Type, and Artifact Availability. But what do these dimensions mean for designing or evaluating how tools support interdisciplinary collaborations in AI development? For example, could two tools support a similar collaborative effort but through different structural features? Why and when did tools targeting the same type of practitioner implement different collaborative features? Such questions required deeper analyses of the tools selected by comparing the structural features of the tool with what the designer intended the tool to do. While going beyond structural analysis of the tool and its features to map them to intended design objectives and scenarios of use, DeSanctis and Poole's idea of the spirit *of a technology* was useful for framing the analysis [38].

DeSanctis and Poole [38] proposed the Adaptive Structuration Theory (AST) which considers both the structures present in information technologies and structures that emerge as part of group interaction with that technology. Importantly, they looked at both the structural features and its *spirit*. They defined the spirit of a technology by the goal and values that underlie it. Let's consider an example using group decision support systems (GDSS), which were also their objects of study [38]. Different GDS systems could have different ways of implementing the aggregation of final decisions or votes of a group. A system that allows the group to proceed to the next action after majority votes have been received versus a system that does not allow the next action unless a consensus has been reached requires different implementations. However, they also differ in their *spirit* of how a decision should be made by a group [38]. We might also observe different interactions between groups who use the majority-based GDSS and those who use a consensus-based GDSS. DeSanctis and Poole state that "*to the extent that information technologies vary in their spirit and structural features sets, different forms of social interaction are encouraged by the technology*" [38]. Similarly, *to the extent that practitioner tools vary in their spirit and structure, different forms of collaboration are promoted by them.*

I observe four main spirits in this sample of tools. Each spirit is associated with a set of options for the design dimensions. In fact, spirits can also be thought of as higher-level categories that represent distinct combinations of options for certain design dimensions.

a. The Groupware Spirit

Tools using collaborative semantics of use promote a *groupware spirit of collaboration*. Practitioners can view and act upon the results of each other's actions. ZIVA (T1) and ModelLens (T2) are examples of such a tool.

ZIVA (T1) is a tool meant to improve knowledge sharing and collaborations between domain experts and data scientists. In a formative qualitative study, Park et al. uncovered that the process of creating domain-specific NLP models (legal, sports, disaster) were challenged by domain experts' limited time and availability and by a lack of standard workflows for capturing domain knowledge. They created a tool that allowed domain experts to create concepts from a set of examples, label them, and provide rationale for these labels. The data scientist would be able to view these changes and leverage them while creating domain-specific NLP models.

ModelLens (T2) is meant for improving collaborative error analysis among data scientists. Analyzing and keeping track of the different errors a model makes and categorizing the different model errors was proving to be a challenging task as each data scientist had their own observations and additional effort had to be made to sync all the ad-hoc observations. ModelLens provided a view that collected model errors from different sources, enabled a drill-down view that allowed analysis of specific errors, maintained a custom categorization of the type and reason for model error, and allowed data scientists to view each other's updates.

ZIVA (T1) and ModelLens (T2) are different because the former enabled interdisciplinary collaboration (between data scientists and domain experts) and the

latter supported intradisciplinary collaboration (among data scientists). But they are similar in the way they leverage collaboration semantics of use to support tightly coupled collaborations. In T1, it was asynchronous and serial and in T2, it was most likely synchronous. The work also happens in a distinct space, i.e., not in an add-on of existing tools, providing flexibility in implementing collaboration semantics of use. Both T1 and T2 enabled information sharing, visibility, and coordination of work through groupware-like features. However, it is important to note that in both cases the practitioners were already collaborating closely but did not have the right tools to support them. Thus, T1 and T2 were designed intentionally to improve their collaborative efforts by using groupware features, specifically collaboration semantics of use.

b. The Core Practice and Communication Spirit

Since working with AI involved additional complexities and quirks compared to traditional application development, some tools targeted adaptations of core practice to better address them. These tools also allowed export of artifacts created for sharing with other practitioners. The tool often did not support communication or collaboration explicitly, but the artifacts created through the tool did. 8 of the 18 tools analyzed fit the spirit of core practice and communication, namely ProtoAI (T5), AIFinnity (T7), DocuML (T8), fAIlureNotes (T9), Deblinder (T12), Canvil (T14), ChainForge (T15), and PromptInfuser (T18).

ProtoAI (T5), fAIlureNotes (T9), Canvil (T14), and PromptInfuser (T18) are tools intended for use by UX designers and researchers. ProtoAI and fAIlureNotes are web applications, while Canvil and PromptInfuser are built on top of Figma. ProtoAI

implements a model-informed prototyping approach where different designs are explored and prototypes according to different inputs and model outputs. fAIlureNotes (T9) supports error analyses of computer vision models from a user-centered perspective, by allowing UX practitioners to import user scenarios and verify whether the output labels match user needs in those scenarios. At a high level, both Canvil (T14) and PromptInfuser (T18) are built on top of Figma, allowing the designer to explore LLMs as a design material and tightly couple UI design with AI functionality respectively.

Not only do such tools address challenges of designing with AI, they also create artifacts that play an important role in collaborations. Interactive prototypes, failure cases, explanations of AI outputs and error recovery designs, user-centered error analyses are all vital artifacts for cross-functional communication and collaboration. These artifacts demonstrate user-centered considerations to other practitioners and help in attaining a shared understanding between AI and UX practitioners.

Similarly, AIFinnity (T7), DocuML (T8), Deblinder (T12), and ChainForge (T15) are tools for AI practitioners. AIFinnity is a Jupyter widget, DocuML is a JupyterLab extension, Deblinder is a web application, and ChainForge can either be used as a web app or installed locally. AIFinnity supports behavioral analyses of pre-trained text and image models (T7). It enables practitioners to analyze instances of different inputs and corresponding outputs, group them into schemas, form hypotheses of when the AI exhibits undesirable behaviors, and evaluate them accordingly. Deblinder (T12) has similar functionality but it is intended for post-deployment evaluation. By analyzing end-user feedback and cases when the model fails for the end-user, it helps analyze to what extent these are systematic failures. ChainForge (T15) is a visual programming environment for prompt engineering, including features for iteratively refining prompts,

comparing outputs across different prompts and LLMs, and conducting systematic evaluation. These tools produce data slices, visualizations, error analyses, and performance reports. Again, these artifacts are instrumental when communicating about the model and its capabilities to stakeholders, providing recommendations, discussing ideas, and making decisions. These artifacts are also used as inputs to other tools in the AI development workflow.

DocuML (T8) is a Jupyterlab extension that opens up as a panel next to the code for the purposes of creating model documentation and is based on the model card [102]. The tool was intended to work both ways: i) create traceable and contextual model documentation that linked code cells to sections in the document and ii) nudge the practitioner to think about ethical implications as they were coding and navigating the different sections. The purpose of this tool is to create the model card, a transparent documentation supporting collaboration between upstream model builders and downstream developers. But it also supports an activity that is considered as integral to AI practice as the activity of error or behavioral analyses - responsible and ethical sensitivity towards model development [16].

All these tools have single-user semantics (except Canvil which I discuss further in 4.6.1) as the core problem being solved is '*How can a practitioner do X?*'. However, all tools support creation of artifacts which in turn support collaborations, communications, handoffs, and coordination. Sometimes the artifacts are transient and short-lived (e.g., prototypes) and other times they can persist longer (e.g., model cards).

c. The Community of Practice Spirit

3 tools - Angler (T11), Interactive Model Card or IMC (T13), and the AI Playbook (T16) implemented single-user semantics, but they could be used by practitioners of

different disciplines. Angler (T11) is a visual analytics tool for analyzing errors in machine translation. Practitioners from AI, UX, and business functions can use the tool to examine the prevalence and severity of translation errors and prioritize ones that have a high impact on user experience. Angler is an interactive, visualization-based web application and does not require any code to be written for analysis, making it accessible and usable for both experts and non-experts of AI. The tool helps UX and product practitioners better understand the translation model side of things, and the AI practitioners better understand the user side of things, thus providing a shared, mutual understanding.

IMC (T13) is an interactive tool based on the static model card used for model reporting [102]. It helps a practitioner carry out disaggregated evaluations (i.e., evaluations on sub-groups in the data) and compare the training dataset used for model development to real-world datasets. IMC is a no-code tool and leverages interactivity and visualizations to provide lightweight ways to experiment, interrogate, and attain a tangible understanding of the model. The tool is meant for individual sensemaking by varied stakeholders in an organization, irrespective of AI expertise, and aims to induce productive skepticism. Thus, IMC can i) support discussions and decision making across different stakeholders in an organization and ii) extend the role model cards play in record-keeping, conducting audits, and aligning the team through interactive features.

In both cases, the tools offer ways of collective learning and practice that cuts across team and disciplinary boundaries. That is, a new blended community of practice forms as a result of using these tools [143]. Though the tool only offers single-user semantics of use, the tools are designed in ways that allow practitioners who are not technical experts of AI to gain insights. The tools also highlight human-centered and RAI considerations for AI practitioners by surfacing errors encountered by end-users and

results of disaggregated evaluations. Thus, the tools facilitate diverse perspectives to be involved in decision making, team alignment, and shared understanding.

The AI Playbook (T16) is also a lightweight tool to explore and simulate possible error cases in NLP applications. The tool was intended to connect different disciplines on edge cases and model failures early in the design and development cycle, surface different viewpoints, build consensus, and decide on next steps. Apart from cutting across team boundaries and facilitating alignment, the tool also helps avoid groupthink by having practitioners use it individually and cross-checks for personal biases or blind spots.

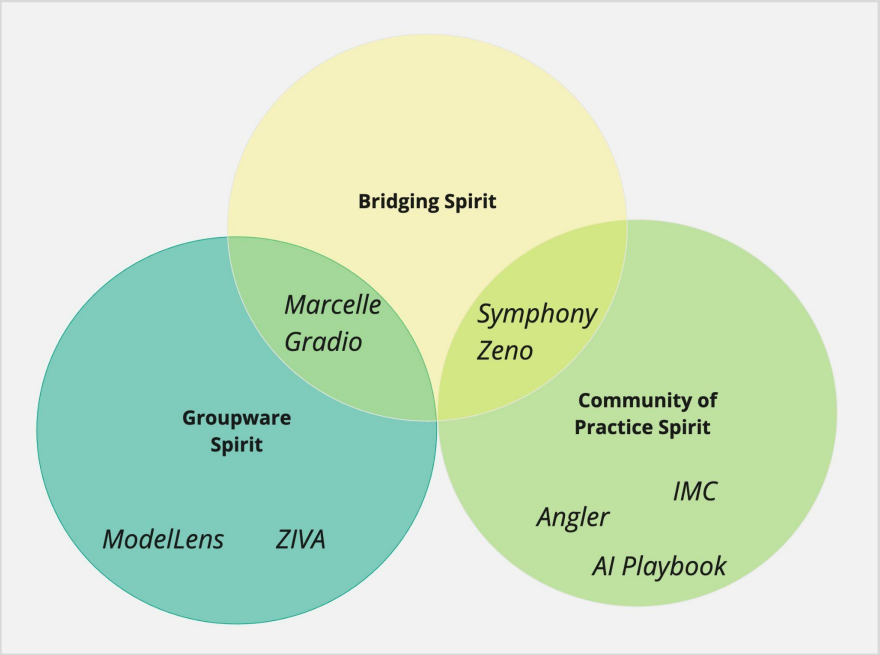


Figure 4.2 Representing the Overlap between Visibility & Bridging Spirit and other spirits.

d. The Visibility & Bridging Spirit

Some of the tools were libraries that aimed to make the model visible and accessible to various collaborators. These packages helped bridge gaps between AI

practitioners and their collaborators or clients, and supported interoperability, extensibility, and reuse as well.

Consider Symphony (T3), a framework for creating task-specific components that are accessible through web dashboards and computational notebooks. It supports reusable components for tasks such as analyzing data duplicates, data outliers, confusion matrices, fairness metrics across subgroups and so on. The goal was to create "*shareable ML interfaces"* and changes can be synced across the computational notebook and web interfaces. By providing this option, Symphony takes the data and model analysis to "*where practitioners work"* [9], fostering collaboration and a shared understanding. Similarly, Zeno (T4) is also a framework consisting of both a Python API and a web interface for conducting behavioral analyses of the model. A practitioner can test and analyze individual instances of model inputs and outputs, recognize common patterns of undesirable behaviors, and test how systematic these behaviors are. It is quite similar in functionality to AIFinnity (T7, a core practice and communication tool), which is a Jupyter widget. But Zeno, like Symphony, is available in both notebooks and as a web application. Practitioners will have to implement the building blocks through code first before the outputs can populate the web application and be made available for other stakeholders. The artifacts made available through Symphony and Zeno are interactive and available through multiple surfaces. In fact, the web applications can function as distinct tools, of a community practice spirit, within an organization (Figure 4.2).

Gradio (T10) is a Python package that makes it easier to share models and demo them with clients, domain experts, and designers. Instead of passing data back and forth, collaborators can easily interact with the model, test different inputs, and flag errors, and all these interactions are available to the AI practitioner for iterating and refining the model. This way, it is not truly collaborative, but the specific AI/ML practitioner can collaborate with different stakeholders better (1: many). Like Gradio, Marcelle (T17) also

enables AI practitioners to share models over the web. One of the components available as part of Marcelle is the shareable, interactive ML application that maintains a common data store to synchronize between the practitioner and the domain-expert. Again, the artifacts created by Gradio and Marcelle are 'tools' themselves. The practitioners can see the changes made by the stakeholders who are using the web-based AI/ML tool. However, I do not categorize these tools under a groupware spirit for the following reasons.

First, some interactive tools created by this set of packages have single-user semantics (e.g., Symphony, Zeno), while others afford collaboration semantics of use (e.g., Gradio, Marcelle). More importantly, these tools are a result of AI practitioners' effort to install the right packages, write code, and set it up. Since models are typically available in computational notebook environments, these packages aim to share it and make it visible for other practitioners through web interfaces. Hence, these tools can be viewed as building bridges from AI practitioners to other groups. But instead of unifying the work environment, it provides different surfaces for practitioners to engage with. I classify it as *Visibility and Bridging* spirit to capture such differences. The question of whether resulting tools of this spirit vs that of *Community of Practice* spirit or *Groupware* spirit play out differently in use contexts is an open question, beyond the scope of this study. However, I revisit the idea in 4.6.2, when I discuss why spirits can be underspecified concept to describe actual tool use.

Figure 4.3 describes how tools belonging to different spirits take on different values for the design dimensions. For example, tools can used by one specific group of practitioners (visibility and bridging or core practice spirit) or multiple practitioner groups (groupware or community of practice spirit). Variations in whether they extend an existing tool or are built as a standalone application (*row*: *Tool Architecture, Figure 4.3)* and in whether they allow for single-user or

collaboration semantics of use (*row*: *Semantics of Use, Figure 4.3*) further give rise to the different spirits discussed above. Figure 4.3 illustrates how the different design dimensions relate to the design spirits by listing the specific values tools of each spirit take.

	GROUPWARE	CORE PRACTICE & COMMUNICATION	COMMUNITY OF PRACTICE	VISIBILITY & BRIDGING
Users	Multiple users from same or different practitioner groups	One practitioner group	Multiple practitioner groups	Model developers
Axis of AI work	1 or more Axes	1 specific Axis	1 or more Axes	Model development /Responsible AI
Tool Architecture	Standalone Tool in most cases - Provides flexibility to implement collaborative semantics	Majorly extension of existing practitioner tools, can be standalone in some cases as well	Standalone/Distinct Tool; Accessible by different groups of practitioners	Programming Libraries, Software packages to support interoperability and extensibility
Semantics of Use	Collaborative Semantics	-	Single-user semantics	Collaborative or Single-user semantics
Artifact Type	-	Create artifacts pertaining to AI-related core practice	-	Creates shareable, interactive interfaces
Artifact Availability	-	Within and outside the tool (static files)	-	Within and outside the tool (interactive)
Spirit Description	Supports tightly coupled collaboration and coordination	Create artifacts, external representations to support communication or handoffs	Forms community of practice around a shared concern	Building bridges from AI/ML practitioners to other stakeholders

Figure 4.3: This image shows what values do tools of different spirits take for the design dimensions described above.

During the description of different spirits, we can see that despite having overlap in functionality and targeting similar problems to be solved, different tools encourage different forms of collaboration. Both AIFinnity and Zeno enable behavioral analyses of models, but differ in terms of the artifact type, artifact availability, and spirits as well. Both Angler and Deblinder support user-centered model evaluations post deployment, but Angler promotes a community of practice spirit and Deblinder promotes the core practice & communication spirit. Such

differences manifest as a result of the tool designer's objectives, who the expected users are, and what axes of AI work it touches upon. Below, I discuss two aspects of the design space that help characterize the nature of the design space better.

4.6 Reflections on The Design Space

4.6.1 Embodying Multiple Spirits

Tools can belong to more than a spirit depending on the collaboration context. Let us consider the examples of Canvil (T14) and PromptInfuser (T18). Both tools are similar in that they allow for designerly exploration of LLM capabilities within the Figma tool. There are other important differences between the two tools in how they combine AI and UI prototyping processes.
The key difference I focus on here is that Canvil is available as a Figma widget, while PromptInfuser is available as a Figma plugin. Figma widgets support collaborative tasks (e.g., voting), but plugins are meant for individual workflows. They are basically different options corresponding to the design dimension semantics of use.

While I categorize them as tools that promote a *Core Practice and Communication* spirit, this classification was based on an organization-wide context. Since Canvil (T14) is a Figma widget, it can support synchronous collaborations among UX and product practitioners (UX/Content Designers, UX Researchers, Product Managers, and so on) but PromptInfuser (T18) cannot. If we consider UX collaborations, Canvil promotes a groupware spirit. The goal here is not to prescribe Canvil over PromptInfuser, but point out that the spirit a tool promotes can actually vary based on the context. In this example, it happens to be a function of one of the design dimensions - Intended Users. Within the larger context of supporting interdisciplinary collaborations in AI design and development, both tools still invoke a *Core Practice and Communication* spirit.

Other tools like Zeno and AIFinnity also highlighted the idea of reusing data slices, schemas, and hypotheses by sharing artifacts with teammates. The tools can promote a certain spirit of collaboration organization-wide, but the implications of the tool for collaborating and distributing work among practitioners within the same discipline can be different.

4.6.2 Design Spirits vs Use Contexts

It is important to note that the *spirits* capture the designers' goals and assumptions about how the tool can support collaborations. These can be derived from explicit rationale, design goals, descriptive use cases as well through the tool's features. The *spirits* describe how a tool presents itself but may or may not account for how the tool is adopted and used by practitioners. The tools can be interpreted and used in ways that are not conceived by the designer [43,109,110].

Consider the tool AIMEE (T6), which was not discussed as part of 4.5.2. AIMEE stands for "*AI Model Explorer and Editor*" and is a standalone tool (T6). It uses decision rules as a way to update and retrain the model by simply editing the rules. These rules are also used to explore and better interpret the model's decision-making process, especially after retraining it. The tool allowed AI practitioners to export new data, model changes, and the rule set as PDF files for the purposes of sharing information, gathering feedback, traceability, and record-keeping. Such features were added after initial rounds of usability testing to better support communication of model capabilities.

While the authors intended AIMEE to bridge knowledge gaps between AI practitioners and clients, the AI practitioners saw the tool as an opportunity to collaborate more closely with clients. Since AIMEE leverages decision rules, which in turn afforded a no-code approach to understanding and editing the model behavior directly, there was an opportunity for clients to co-create models with AI practitioners, provide feedback early on, and iterate much faster. For AI-related work practices, artifacts such as decision rules or other Explainable AI mechanisms can be as instrumental as structural features (e.g., collaboration semantics of use) in supporting

collaborations. The design dimensions Artifact Type and Artifact Availability can help tool designers consider the utility of an artifact for various practitioner groups and which artifacts can support collaborations.

4.7 Discussion - Implications for CSCW Research and Theory

In this section, I discuss implications of the design space for both design and theory at the intersection of HCI, CSCW, and AI.

I believe this design space is also useful in theorizing how technology is designed and used in AI organizations.

I use Orlikowski's *structurational model of technology* [109] to illustrate how. Prior studies presented technology as a neutral and external intervention that could effect change in organizations in expected ways. These ideas were then updated to recognize the human factor involved, how it affects technology-based outcomes and the subjectivity involved. Orlikowski [109], however, argues that both perspectives are valuable and outlines four possible influences between organizations, humans, and technology:

1. Technology as an outcome of human action
2. Technology as a means of human action
3. Organizational context shaping human action with technology
4. Human actions shaping up organization structures through technology

The design space presented here helps analyze these influences. The design space demonstrates how tool designers encode their notions of how collaborations will occur and how they are best supported into the design of the tool (1). But these notions, intentions, and assumptions are influenced by existing organizational contexts. To provide an example from this design space, let us consider Angler and Deblinder as two tools from different organizations. They solve similar

problems of post deployment, user-centered model evaluation. But Deblinder falls under the *Core Practice and Communication* spirit, and Angler under the *Community of Practice* spirit. There are no right or wrong ways to build these tools. But these differences can be tied back to the organizational culture, structures, and norms which can influence the tool designers (3).

It is also important to understand how these tools support and mediate practitioners' actions (2) and whether the use of these tools, over time, influences organizational structures (4). As highlighted above, design spirits may fall short in capturing how the tools are used, appropriated, or even abandoned. Thus, more in-situ and longitudinal research is required to study how practitioners adopt and use these tools. Such research can also help analyze other influences in Orlikowski's model, show the 'other side' of design spaces, and determine to what extent tools help practitioners shift organizational contexts and/or collaborative practices.

The descriptive power of design spaces is also useful in CSCW research for tracing how tools change over time. In recent years, AI has undergone a paradigm shift with the arrival of generative models, and this change is also visible in the tools.

Tools span various tasks and needs such as making use of domain knowledge (T1), conducting behavioral analyses (T4, T7), and tackling tasks like prompt engineering (T15). If we place the tools in a chronological order, we can see that the tools also get updated to reflect newer practices and methods as the nature and capabilities of AI evolves. For example, ChainForge (T15) implements what authors believe are the three important modes for prompt engineering: opportunistic exploration, evaluation, and iterative refinement. Tools also get updated to promote ethical sensitivity and reflexive practices, transparent model documentation, and compliance with responsible AI guidelines (T8). Similarly, design tools are also developed to better support proposed approaches such as 'model-informed prototyping' (T5), 'designerly adaptation' (T14), 'coupling UI and prompt design' (T18).

In prior work, we see that there are multiple approaches to building models and the products that incorporate them [144], including model-first, product-first, or parallel approaches. Most of these cases deal with a bespoke model, trained and developed for a particular use case. But the rise of pre-trained and large language models has led to a homogenous model catering to a wide range of tasks that have to be carefully fine-tuned or adapted to specific, downstream use cases. Consequently, there are higher risks of widening the socio-technical gap [96], despite improved model capabilities. Tools like fAIlureNotes (T9) help address these issues as they enable UX practitioners to investigate errors and mismatches with expected user outputs in multi-modal and generative AI-based scenarios. In part because of the availability of improved AI capabilities out of the box, and in part because of this set of tools, practitioners without technical AI expertise can influence how these models get adapted than wait to work with a bespoke version, in the downstream workflow, by which point design decisions might have already been made. These tools thus help practitioners by being tailored to the tasks they carry out rather than work with native interfaces for foundation models.

4.8 Discussion - Implications for UX Practitioners and AI-UX Collaborations

In Chapter 3, I noted several challenges UX practitioners face in terms of translating AI capabilities to design, prototyping AI-infused concepts, evaluating AI from a user-centered perspective, and collaborating with AI practitioners. Based on these challenges, I propose a groupware-based approach for designing collaborative tools. But what implications does this design space carry considering the results of the previous study? The design space analysis uncovered interesting design dimensions and spirits to consider. Tools under the *Core Practice and Communication* provided functionality to design with real model capabilities and create

functional prototypes. For example, ProtoAI, Canvil, and PromptInfuser are tools that help designers prototype with AI/LLMs and have a better grasp of its capabilities. fAIlureNotes is a tool that enables UX practitioners to evaluate vision models from a user-centered lens. fAIlureNotes (T9) supported user-centered evaluations of image models.

Tools under the *Community of Practice* also provided ways to establish a shared and mutual understanding with AI practitioners. However, the *Groupware* spirit of collaboration did not include tools for HCI-AI collaborations. None of the tools were truly interdisciplinary in nature either, which prior work has also called for (Section 2). A key takeaway is that many of the tools described here can solve a *core set of challenges* for UX practitioners but not the *collaborative challenges*. Also, none of the tools analyzed fit the description of a *model sandbox*, as described by UX practitioners in the prior study. Thus, we can view the tools in terms of the problem it solves for practitioners and in terms of the collaborative features it offers. In some tools the former takes a precedence over the latter, however, designers of tools in the groupware spirit focused on *tackling both* from the start [115,168].

To deliberate further on how the different design spirits might compare and how effective they are in solving collaboration challenges, I leverage a Cartesian representation. This is a common format for visualizing design spaces, where the design dimensions are usually the axes. But I use two aspects of AI work – *Model application* and *Model development* has axes here. I do this so that I can is reduce the dimensionality of the design space to a 2d representation and examine how the spirits fit here. Thus, *Model application* and *Model development* form the x- and y- axes, as shown in Figure 4.4, and we consider UX and AI practitioners as respective representatives of these axes. The pink and blue points thus represent tools that are effective in solving core discipline-specific challenges for either practitioner.

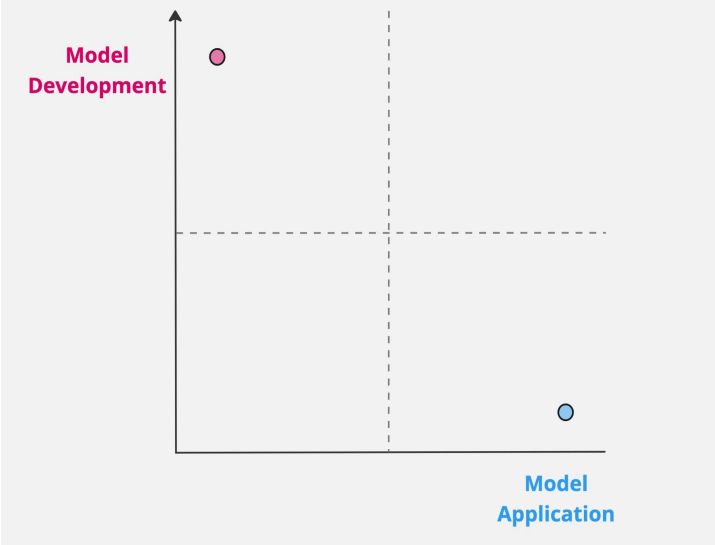


Figure 4.4: Cartesian representation of two axes of AI work. The blue and pink points represent tools that are highly relevant to the respective axis of AI work (Model development or Model application).

Tools that fall under the Community of Practice spirit *could* occupy the lower left quadrant (Figure 4.4). This position is based on Moore et al. 's [104] comparison of a fAIlureNotes (*Core Practice and Communication)* and IMC (*Community of Practice)* and their findings that the former is more useful and usable for UX practitioners. It is unsurprising perhaps that we would lose some of the relevance factor in making the tool more broadly applicable for a range of practitioners. Also, these tools support single-user semantics and are meant to provide a shared understanding of the problem, but they do not support explicitly coordination of work activities or collaborative work.

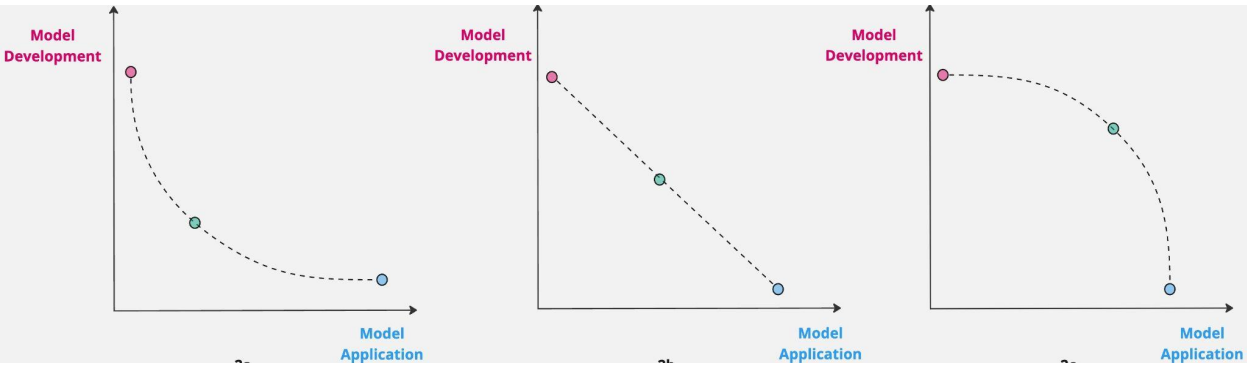


Figure 4.5 Visualizing different points in the 2d-space to speculate what these tools might look like.

The question of why there are no tools currently occupying the upper right quadrant and what such a tool might look like remains an open one. Based on my findings so far, a tool that falls under the Groupware spirit should fit into the upper right quadrant. However, could we map these different design possibilities residing in the 2d space (Figure 4.5), especially in the upper right quadrant (Figure 4.6)? For example, even when we leverage collaboration semantics of use to build groupware style tools, they can vary. The collaboration semantics of use can be implemented to be asynchronous, real-time, serial, or mixed. Tools incorporating these different implementations could reside in the different points mapped out in Figure 4.6.

While my work proposes groupware features as a solution, it is useful to consider other way in which tools can support UX-AI collaborations. For example, are there other novel spirits that we can design for to support collaboration? The goal of these Cartesian representations is to help reflect on a design space and possible gaps here that can be filled with new ideas, which is precisely what it helped do. The Cartesian representations used here (Figures 4.4-4.6) helped consider alternate possibilities – both within the Groupware spirit and beyond it. There are different ways we could implement groupware-style tools to support UX-AI collaborations and hence the tools can occupy different points in the upper right quadrant (Figure 4.6). Designers can also extend this design space by adding new dimensions and spirits, thereby building novel tools that also reside in the upper right quadrant (Figure 4.6).

With respect to my research questions, I believe there is promise in leveraging groupware-style collaboration features to better support HCI-AI collaborations. Through subsequent research studies focused on proposing new methods and tools (Chapters 5 and 6), I will revisit this design space with empirical and grounded insights.

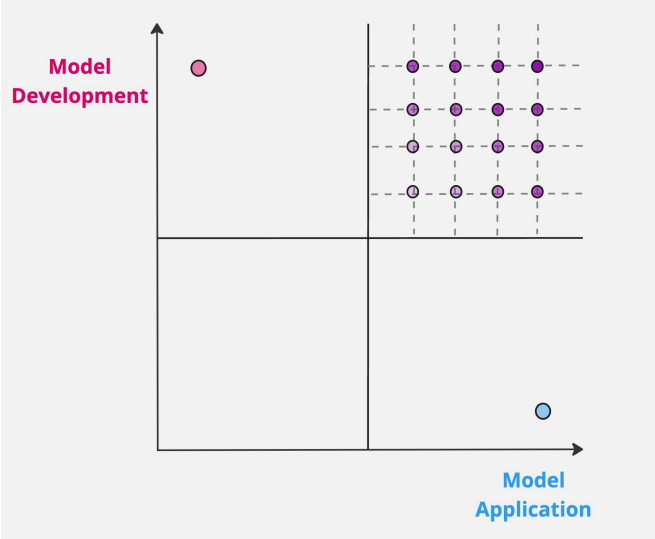


Figure 4.6 Mapping the space of possibilities in the upper right quadrant - where tools are beneficial for model application and development.

These are also interesting questions to consider given the prevalence of pre-trained and generative models. Availability of incredible model capabilities out-of-box has made it easier to generate outputs, prototype, and develop proof of concept for new use cases [76]. This is largely due to the nature of models themselves, along with tools that help access these capabilities. However, evaluation has proved to be the bottleneck. Evaluation challenges are due to the subjective, open-ended nature of outputs, hallucination, lack of consistent and reliable outputs, benchmarks and metrics not transferring to real-world contexts, biased and/or toxic outputs, and misalignment with human needs and objectives [65,138,141,169]. Given these challenges, it would be beneficial to focus new tooling on evaluation of generative AI systems (Chapter 5). I believe this design space can help with reflecting on how to explicitly design to better support

HCI-AI, among other interdisciplinary collaborations, that are necessary for designing human-centered, generative AI applications.

4.9 Limitations and Conclusion

There are a few limitations to note regarding this design space. The first limitation is that this design space is by no means comprehensive. It is an early, emerging design space. As the nature of AI models and work practices evolve, so will the corresponding tools used. It also does not account for all the different types of AI models or the different categories of AI systems. However, being comprehensive is rarely the goal for a design space and might be possible only for more matured areas. The second limitation is that I used tools that were available as technical contributions in HCI-related venues to derive this design space. Tool descriptions and usage scenarios were more readily available as research publications in this case. Though some of the tools analyzed were contributions from industry research labs, I do not have data on what tools are actually used by various industry practitioners in their day-to-day work. Analyzing tools used by practitioners in various organizations can help validate the current design space as well as extend it.

On a similar note, the third limitation is that this design space focuses only on an industry setting where products and services leveraging AI are built for enterprise or consumer use. Though some of the tools are academic research contributions, they also target such industry practitioners. However, there are other contexts where collaborations are happening to design AI systems - open-source, public sector, non-profits, and community organizations [83,97]. Such contexts can also enrich this design space and possibly uncover new dimensions that are not featured here currently.

Despite these limitations, this study contributes an emerging design space of how tools are (and can be) designed to support individual and collaborative practices of interdisciplinary practitioners working on AI application development. In this chapter, I have elaborated on both

the generative and analytical potential of this design space. I believe that this design space can help tool and system designers reflect on alternate options, surface assumptions about collaborations, and design new kinds of tools. With a similar objective, I return to this design space in Chapter 7 to reflect on how tools can be designed based on the evaluation approaches discussed in subsequent chapters.

Chapter 5. Translating Methods for Human-centered AI Evaluations to UX Practice Contexts

5.1 Introduction

In this chapter, I test how interactive visualizations can be leveraged to communicate results of human-centered evaluations of AI models to UX practitioners. I also seek to understand how these methods can inform UX and design practice of AI. I start with describing my prior research study, which was undertaken to evaluate a toxicity detecting model from a human-centered perspective. From there, I discuss how the methods used can be broadly applicable for conducting AI evaluations and how the results can be communicated better using interactive visualizations. Section 5.2 covers related work, and sections 5.3 and 5.4 describe the two investigations I conducted - a formative and an evaluative study, along with the findings. I wrap with implications for using these methods as well as for designing UXP-oriented toolkits (Section 5.5).

5.1.1 Human-centered Evaluation of a Toxicity Detecting Model

Google's Jigsaw team had developed a model called Perspective [170] to support civil and constructive online discussions. This model could detect several attributes of uncivil speech such as toxicity, insults, threats, profanity, and attacks on identity. Perspective was trained on data from various sources including the Wikipedia Talk Pages and the New York Times and annotated via crowdsourcing. The team also released overall model performance metrics (Area under the ROC curve or AUC) as well as AUC metrics for identity-based subgroups [42,146].

With the goal of conducting a human-centered evaluation of Perspective, I set out to understand how human perceptions of toxicity aligned with probabilistic scores of toxicity, generated by Perspective. Such an evaluation can also be considered as a revalidation of the tool's capabilities. However, apart from comparing human and machine ratings of toxicity, I also wanted to i) evaluate how data from different online platforms affected this alignment and ii) to what extent human ratings of other latent attributes corresponded to Perspective's toxicity score.

I selected three different online platforms: news websites, YouTube, and Twitter, but constrained the dataset to news-related comments to make systematic comparisons. Thus, comments collected from YouTube and Twitter were replies made to posts or videos by official news channels. These three platforms not only differ in terms of technical features, but they also present a spectrum of user experiences, commenting styles, and discussion structuring. They are thus distinct sociotechnical spaces, which in turn makes it a diverse dataset to test Perspective's alignment. Apart from toxicity, I also collected human ratings of i) formality of the online comment, ii) how respectful the comment is, and iii) whether the comment presented stereotypes. These attributes were intentionally selected to test Perspective's operationalization of toxicity (*a rude, disrespectful comment that can cause people to leave a discussion*) and check for hidden, latent attributes.

After collecting a set of 300 comments (100 each from each platform), workers in Amazon Mechanical Turk rated them for toxicity, formality, respectfulness, and presence of stereotypes. The same set of comments were also scored by Perspective for toxicity. I had to compare how human ratings of each of these attributes corresponded to Perspective's toxicity score and how this relationship varied across data from different platforms. All human ratings were collected as ordinal data (Likert) to ease the process of making subjective judgements. While the practice analyzing Likert data as interval or ratio data is common in HCI research, I refrained from doing it to better model differences between Likert ratings. For example, the

distance between ‘strongly toxic’ and ‘toxic’ is not the same as the distance between ‘toxic’ and ‘not toxic at all’ [73,121]. This data does not fit the assumptions of a linear regression model. Thus, I use *maximum likelihood estimation (MLE)* techniques to analyze this data [139].

MLE involves estimating parameters of a probability distribution from the observed data such that the observed data would be the most probable occurrence. Imagine if we conducted an experiment where we tossed a dice 1000 times, but we did not know if it was a loaded or unloaded dice. If we recorded the number from every dice throw, we can compute the resulting probability distribution for 1-6. If the resulting probability distribution still looks *uniform* (0.16 for each number in the range 1-6), we can conclude that it is indeed a fair, unloaded dice. The takeaway here is that the probability distribution can tell us what kind of data and phenomenon we are dealing with. Similarly, MLE methods can be used to construct models of binary (logistic regression), categorical (multinomial logistic regression) ordinal (ordered logistic regression), or count data (binomial regression) [139]. Given a dataset of corresponding toxicity ratings from Perspective and humans, I could come up with a probability distribution that shows how the two align based on the below equation:

$$\text{Toxicity}_{\text{Human}} \sim \text{Toxicity}_{\text{Perspective}} + \text{Platform} \quad (1)$$

The resulting probability distribution for (1) is shown in Figure 5.1. In this figure, we can see that as Perspective’s toxicity score increases, the probability of rating a comment as toxic also increases (blue curve). Though Perspective sets a threshold of 0.7 for classifying comments as toxic, even for comments with scores > 0.40, the probability of humans rating these comments as toxic is at least 0.5. Similarly, Perspective sets a maximum threshold of 0.3, below which comments are not toxic. However, for that range of Perspective scores, probability of humans rating a comment as not toxic is only 0.3 – 0.4. However, Perspective does transfer well across these platforms.

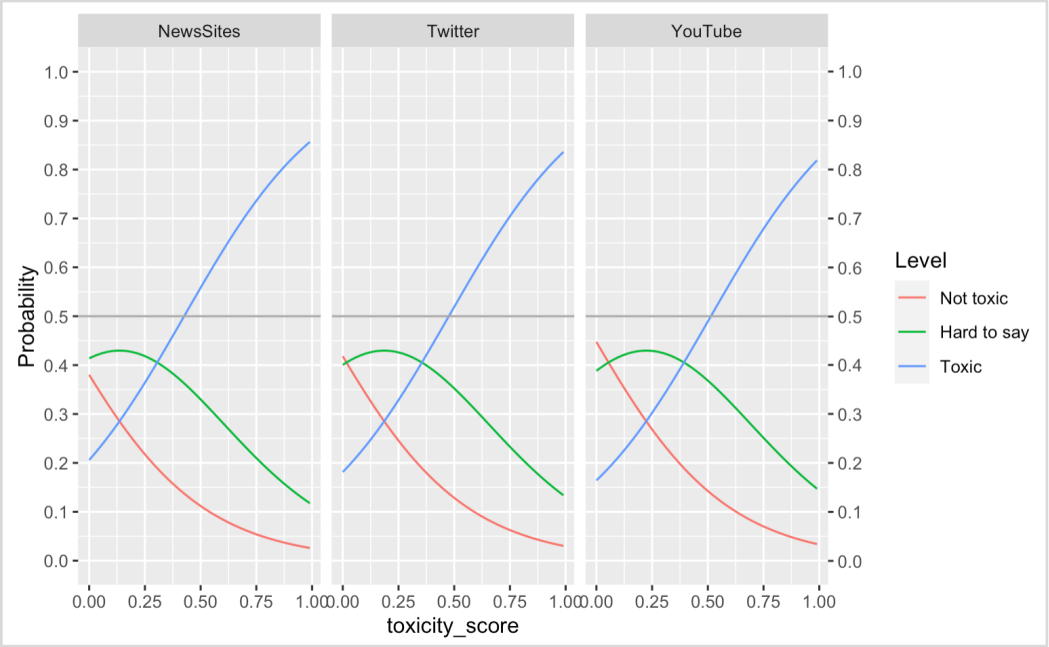


Figure 5.1: Probability of human toxicity ratings with respect to Perspective's toxicity scores (x-axis).

Additional analyses show Perspective's toxicity score is a significant predictor of humans rating a comment as informal, disrespectful, and stereotypical, although the source platform exhibited a significant effect only for the formality attribute. Overall, the analyses helped determine (mis-)alignments between Perspective's toxicity score and human ratings of toxicity across different platforms. While respectfulness is an explicit attribute that is part of Perspective's operationalization of toxicity, formality of speech and presence of stereotypes are not. Yet, as Perspective's toxicity increases, the probability of a human rating the comment as informal and stereotypical is also high. I also fitted an additional model:

$$\text{Toxicity}_{\text{Human}} \sim \text{Formality}_{\text{Human}} + \text{Respectfulness}_{\text{Human}} + \text{Presence of Stereotypes}_{\text{Human.}} \quad (2)$$

With (2), I was able to analyze interesting counterfactuals about how ratings of respectfulness, formality, and stereotypes related to ratings of toxicity. If we switch the rating of

respectfulness from *respectful* to *disrespectful* (the extreme ends of the Likert scale), then the probability of humans rating a comment as *toxic* increases by 75% (Figure 5.2).

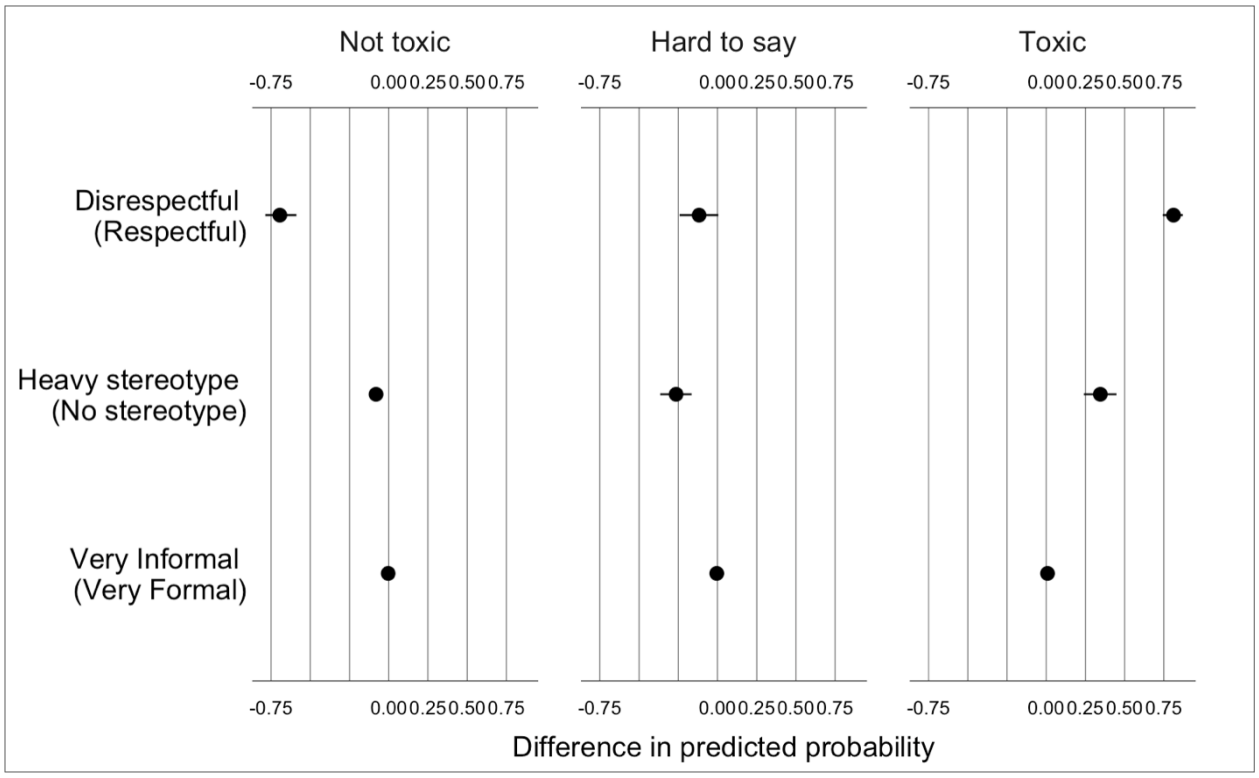


Figure 5.2: Difference in the probability that a human would rate a comment as toxic, simulated by switching ratings for the other attributes: respectfulness, formality of speech, and presence of stereotypes.

If we switch the rating of stereotypes from *no stereotype* to *heavy stereotype* (the extreme ends of the Likert scale), then the probability of humans rating a comment as *toxic* increases by ~30%. But, switching the ratings of formality from *very informal* to *very formal* (the extreme ends of the Likert scale), the probability of humans rating a comment as *toxic* increases by 0.00 (Figure 5.2). Thus, humans did not conflate informal speech with toxicity.

5.1.2 Unpacking MLE methods and Translating to Practice Contexts

The above study was built upon a prior study conducted by Erin Hoffman, David W. McDonald, and Mark Zachry - *Evaluating a Computational Approach to Labeling Politeness: Challenges for the Application of Machine Classification to Social Computing Data* [67]. They attempted to apply a politeness labeling tool, trained on Wikipedia discussions, on their data from Wikipedia Talk Pages. In doing so, they noted discrepancies and revalidated the tool by comparing crowdsourced ratings of politeness to the ML tool's politeness scores using multinomial logistic regression. They found that the tool was somewhat effective at scoring politeness, but ineffective at scoring impoliteness or neutral text. My study focused on the problem of toxicity detection and additionally compares alignment across different online platforms. Together, both studies showcase the utility of using MLE methods (such as ordinal or multinomial regression) for conducting human-centered evaluations of NLP models.

Beyond descriptive statistics and measures of accuracy, inferential statistical methods like *Maximum Likelihood Estimation* (MLE) prove to be useful for such evaluations. It accounts for the data type, constructs appropriate models, and provides a probabilistic framework to make inferences [139]. It provides probabilities for each option in a nominal or ordinal scale i.e., we would be able to infer the probability of users rating 'X' versus 'Y', allowing for nuanced interpretations. Adding covariates (e.g., online platform) allows for examining sociotechnical factors that impact alignment between user expectations and model outputs. We would also be able to estimate uncertainty associated with predictions by computing confidence intervals. Visualizing the expected and actual probability curves, as shown in the above Figure (5.1), is also a compelling way to understand the model in context compared to model metrics.

While measures of accuracy, precision, recall describe model performance on a test dataset, MLE methods can help understand how users will accept model outputs under different conditions. The analysis need not include model outputs explicitly. For example, Goyal et al.

[57] analyzed toxicity ratings (likert data) from rater pools of different identities - a control group, an African-American group, and an LGBTQ group. By fitting the below model, they were able to deduce how likely it is for raters from the control group to rate toxicity as significantly lesser than raters from the African-American or LGBTQ group.

$$\text{Toxicity} \sim \text{Rater\_Pool} \tag{3}$$

These methods are useful for model developers as the results can inform further model improvements and iterations. But they are also useful for practitioners trying to *apply* the model in a specific context. Could they be useful for UX and design practitioners? As discussed in Chapter 3, UX practitioners faced challenges with reinforcing a human-centered lens on AI evaluation especially at scale and with reconciling model-centric metrics with user-centric outcomes. Thus, I wanted to explore how these methods might be useful and comprehensible to UX practitioners, and feature as a potential component of the *Model Sandbox*.

Other work also raises the need for tools that can help UX practitioners test factors that may affect model performance and assess impact on the UX of AI systems [94] and engage in user-centered algorithmic auditing [36]. The recent rise of LLMs due to their incredible natural language understanding and generation capabilities incentivizes their application to power a wide range of use cases. However, using a single, large model for many downstream applications runs the risk of widening the gap between social needs and model capabilities. Liao and Xiao urge rethinking the discipline of LLM evaluation as one that focuses on reducing such socio-technical gaps [96]. The MLE methods, as described above, merits being part of methodological toolkit that blends HCI and NLG evaluations and help addressing the socio-technical gap, and this study will focus on evaluating its effectiveness for UX practitioners. I specifically design the interview protocol and questionnaires to understand how these methods can impact UX and design practice.

5.2 Related Work

5.2.1 Use and Interpretation of MLE Methods in HCI

MLE methods are commonly used in other areas of HCI research as well. However, most studies report the Odds Ratio (OR) as the key effect size [111,112]. Inspired by approaches to interpretations of MLE methods for the social sciences [139], I focus on three other quantities of interest: Expected values/Predicted Probabilities, First Differences, and Relative Risks. Odds ratio is the ratio of odds of an event occurring in a group to the odds of the event occurring in another group. OR is notoriously hard to interpret and most people interpret it as a ratio of probabilities and not odds [34,111]. Let's say the probability of an event occurring within a group A is 0.25 or ¼. Then the odds are 1 to 3 [171].If the probability of an event occurring within a group B is 0.10 or 1/10, then the odds are 1 to 9 [171]. If we compare groups A and B, the odds ratio is (⅓) / (1/9) = 3, but the relative risk or probability ratio is 0.25/0.10 = 2.5. Thus, the event is 2.5 times more likely to occur in group A versus B. But the odds of the event occurring in group A compared to B is 3, which can be quite non-intuitive [171]. When people interpret odds ratio as probabilities, they tend to overestimate the effect [33,34].

Instead of OR, we can use other quantities of interest such as Expected Values/Predicted Probabilities, First differences, and Relative Risks [139]. If we are interested in a particular scenario defined by a set of values for the independent variables, we can compute the associated probability of an event occurring (expected values/predicted probabilities). First or absolute differences are the difference between probabilities of an event occurring for two different scenarios, while Relative risk is the ratio of probabilities. Despite the fact that they can be more intuitive to interpret, odds ratio tends to be the statistic reported by default. The reason could be that it is easier to calculate (exponential of the coefficient) and hence, readily available in

standard statistical libraries. However, open-source libraries offering support to calculate other quantities of interest do exist [3,124].

Because these are non-linear data and we can produce many data points based on the number of scenarios we are dealing with, visualizations can help with interpretation. Kale et al. [79] test how different visualization designs can affect judgment of effect sizes. They uncover that the most optimal visualization design does not necessarily lead to optimal decisions and that users satisfice by using certain heuristics and interpreting visualizations in ways not intended by the visualization designers. By conducting empirical studies with the target population - UX practitioners, I hope to test the effectiveness of visualizations of MLE results (e.g., Figure 5.1) and gain similar insight into their strategies for interpretation.

5.2.2 Human-centered AI Evaluation Tools for UX Practitioners

Given that UX Designers and Researchers play an instrumental role in mapping and assessing how AI models can meet end-users' needs and supporting responsible AI innovation, various tools and techniques have been proposed to better support them.

Yu et al. [154] note how it can be hard for designers to understand details of how the problem is formulated to be solved by an algorithm and related implementation details. Consequently, they may not be able to verify if the model meets design objectives and how inherent tradeoffs can affect different user groups. They designed interfaces with both text and visualizations (confusion matrices) to communicate tradeoffs between i) minimizing false positives versus false negatives and ii) overall accuracy versus sub-group fairness in the context of recidivism prediction. Mixed-method studies showed that these interfaces increased comprehension of algorithmic tradeoffs and provided algorithmic transparency to understand potential consequences. Ye et al. [151] conducted a similar study in the context of Wikipedia, where application designers can explore ORES - a ML tool to classify edits and revisions. The use of interactive visualizations helped designers understand how different model thresholds

affect classification of edits by as good-faith or vandalism for the two groups - experienced editors and newcomers. While these prior studies have used descriptive statistics to communicate model accuracy and sub-group differences, using MLE involves inferential statistics and interpreting probabilities. Using MLE techniques does not necessarily explain the inner workings of a model but rather compares alignment between model and user judgments. It can also help analyze what factors can affect this alignment.

fAIlureNotes developed by Moore et al. [104] also is a relevant example here. The tool fAIlureNotes enables designers to explore computer vision model failures from a user-centered perspective. In a scenario where UX practitioners are testing an object detection model, they can create different user scenarios, upload photos, annotate user expectations of results, track mismatches with model outputs, and so on. While fAIlureNotes helps evaluate user-centric, model failures in computer vision tasks, my work looks at text-based models. Tasks involving a high degree of semantic interpretation, such as language and text interpretation, may not have a single ground truth and can lead to varied, equally valid interpretations [172,173]. Developing narrow, in-house models for such tasks might provide a higher degree of control to align the model with user needs. However, when working with pre-trained and off-the-shelf models, evaluating how model outputs and user perspectives correspond and whether it corresponds across a wide range of domain-specific or contextual data becomes an important prerequisite for applying the model. Prior work has sought to address this problem by proposing audit and evaluation systems for end-users and community stakeholders [35,86], while the focus of this study is on UX practitioners.

5.3 Formative, Qualitative Study

In this subsection, I describe the scenario-based interview study I conducted, with the goal of introducing UX practitioners to visualizations of predicted probabilities (Figures 5.2-5.4) and analyzing how they interpreted it to derive design implications.

5.3.1 Scenario and Data construction

Similar to the case of toxicity detection, I selected sexism detection in online comments as the AI capability and constructed a hypothetical scenario around it. Participants were asked to imagine that they worked for a social media platform and a new AI-powered design feature was being rolled out to detect sexist comments. I used different communities hosted on the platform as an impacting factor to be examined. That is, one could evaluate whether alignment between model and user judgements varies across data from different communities. Sexism detection was binary (yes/no) and hence logistic models were used.

I used the dataset developed by Samory et al. [123], as they collect data annotations using a codebook developed from psychological scales measuring sexism. They also label and train the model on different datasets - tweets that contain the phrase "*call me sexist, but*" tweets displaying hostile sexism, and tweets displaying benevolent sexism [75,123,140]. These datasets were respectively represented as data from different communities A, B, C in the scenario. This might resonate with UX practitioners better as they can draw parallels with different subreddits, Facebook groups, or Discord channels. To avoid biasing participants by characterizing the communities as topic-based (politics, sports) or demographic-based, I kept it nondescript by naming them A, B, and C.

User ratings for this dataset were available as part of the original data's collection and annotation efforts. The same comments were also scored by using the `insult` and `identity_attack` attributes from Perspective and by zero-shot prompting GPT (4-o mini) to score `sexism` directly using the annotation guidelines from Samory et al. [123]. User ratings of sexism were modeled using logistic regression (R), with model scores and source community of the comment as independent variables.

5.3.2 Interactive Visualizations

For this formative study, I focus on visualizing only the predicted probabilities for new data points. That is, given a particular model score of sexism and the community (A/B/C), we can calculate the probability of a human rating the same comment as sexist. For all possible values of the model score of sexism, I generate the probability that humans would rate the comment as sexist for each community (or dataset) (see Figures 5.2-5.4).

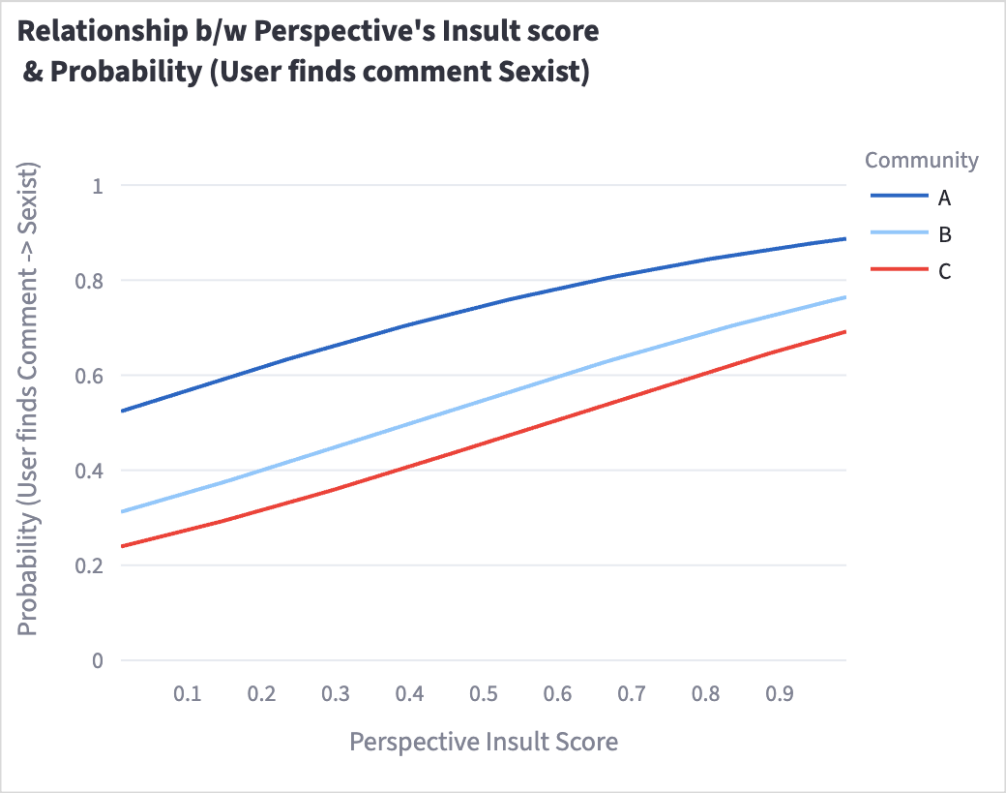


Figure 5.3: On the x-axis, we have Perspective's Insult score which ranges from 0.0 to 0.99. On the y-axis, we have the probability of the user finding a comment sexist. For all three communities, as the Insult score increases, the probability that the user will find the comment sexist also increases but is the highest for community A and lowest for community C.

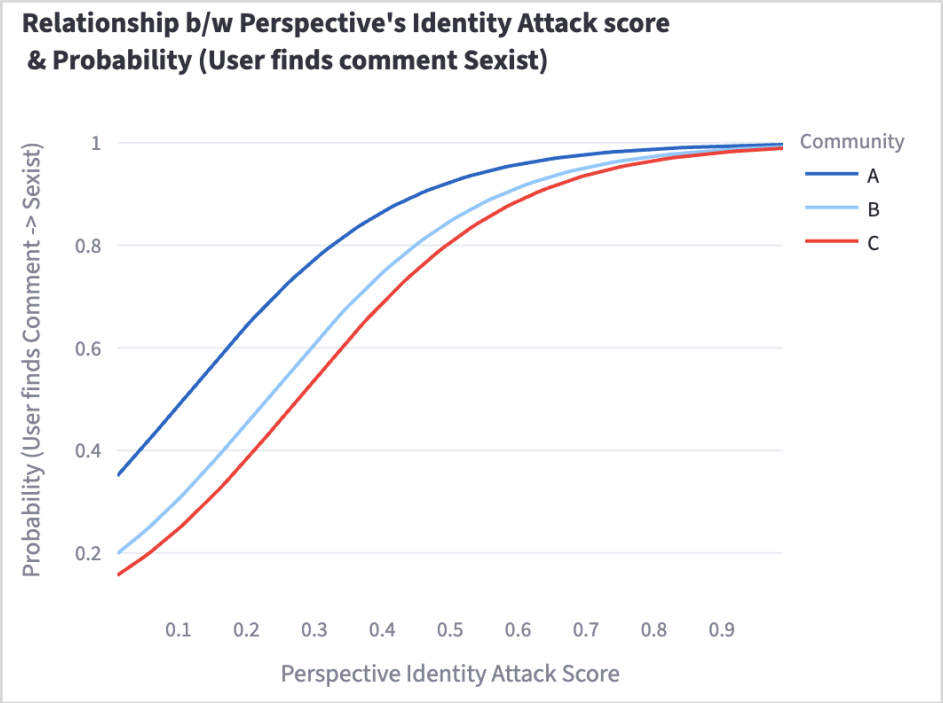


Figure 5.4: On the x-axis, we have Perspective's Identity Attack score which ranges from 0.0 to 0.99. On the y-axis, we have the probability of the user finding a comment sexist. For all three communities, as the Perspective score increases, the probability that the user will find the comment sexist steeply increases. The three curves converge and are close to $y = 1$ for $x > 0.65$.

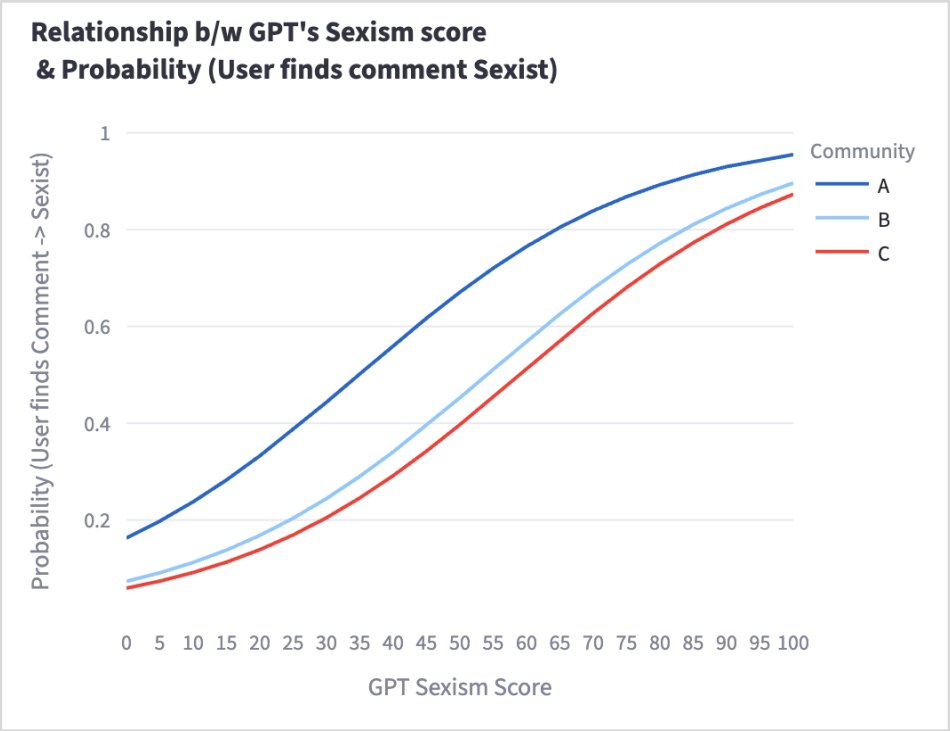


Figure 5.5: On the x-axis, we have sexism scores from prompting GPT, ranging from 0 - 100. On the y-axis, we have the probability of the user finding a comment sexist. The three curves almost form the S-shaped curve that is characteristic of the logistic regression curve. Because GPT was prompted using the same guidelines provided to the data annotators, this logit model (GPT + Community) can better predict the sexism ratings.

Since not all UX practitioners may be familiar with statistics or in dealing with probabilities, I introduced the following interactive features:

1. UXP can choose to view the probabilities for one community at a time. They can select a particular community in the legend, which will toggle the probability curve for that community within the visualization.
2. UXP can also hover on the curves, which will show data from specific points in the visualization. It looks continuous but is actually made of individual data points which can be read using the hover function.

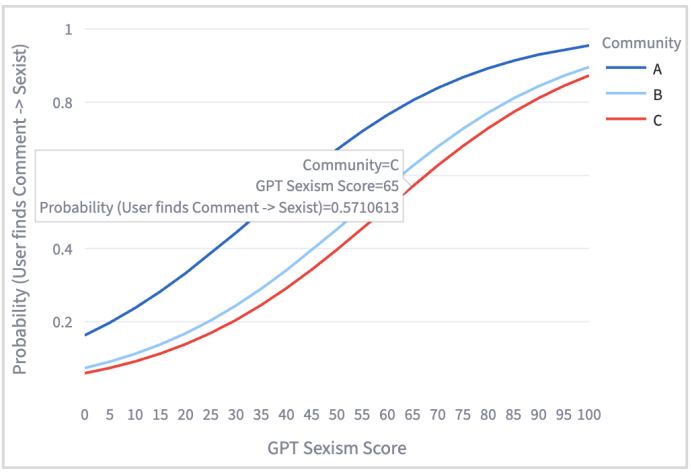


Figure 5.6: UXP can hover and trace the probability curves which will make the tooltip appear.

3. UXP can also drag and draw a box in particular regions of the visualization. For example, they could select a region of x = 0.3 to 0.6 (model score) and y = 0.3-0.6 (Probability of users finding comment sexist). Doing so will bring up a data viewer on the right hand side with i) all comments from A, B, C which have model scores ranging from 0.3 - 0.6, ii) corresponding user ratings of whether the comment is sexist or not (yes/no), and iii) associated user reasoning (*Behavioral Expectations, Denying Inequality, Endorsement of Inequality, Stereotypes and Comparative Opinions, Maybe Sexist - needs context, Not sexist).*

   There might not be data associated with every model score in the range 0.3 - 0.6, but with MLE, we are able to infer the associated probabilities of users finding the comment sexist or not. But the goal with this Data Viewer feature was to give UXP an idea of where the probabilities were coming from, from what data we were making these inferences, and qualitative insight into what kind of comments were relevant to that region.

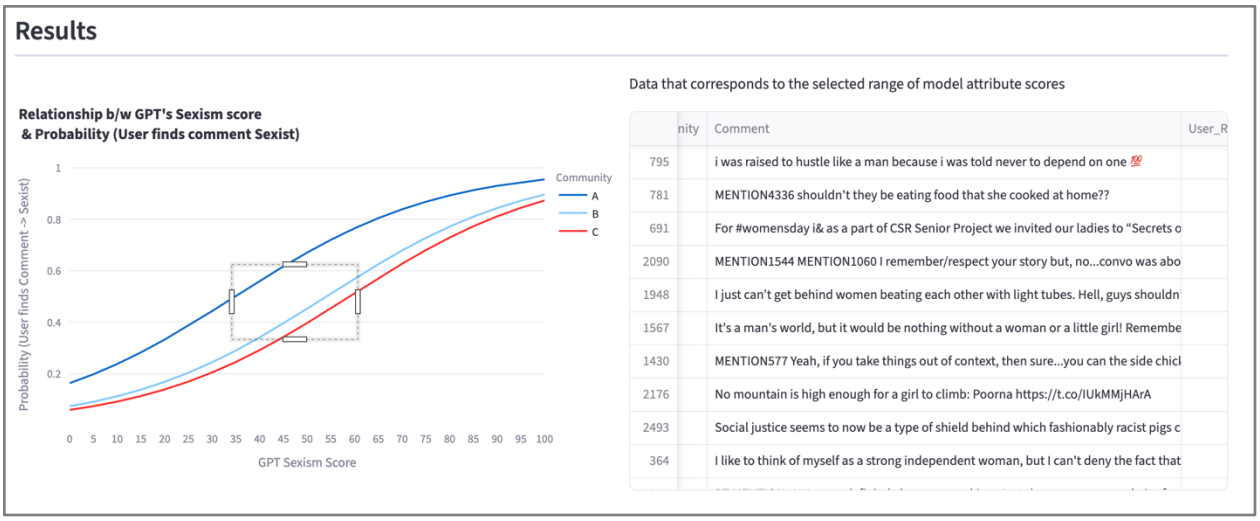


Figure 5.7: Box select feature which brings up the Data Viewer on the right pane.

Additionally, the steps leading up the UXP to the visualizations were also interactive (Figure 5.7).

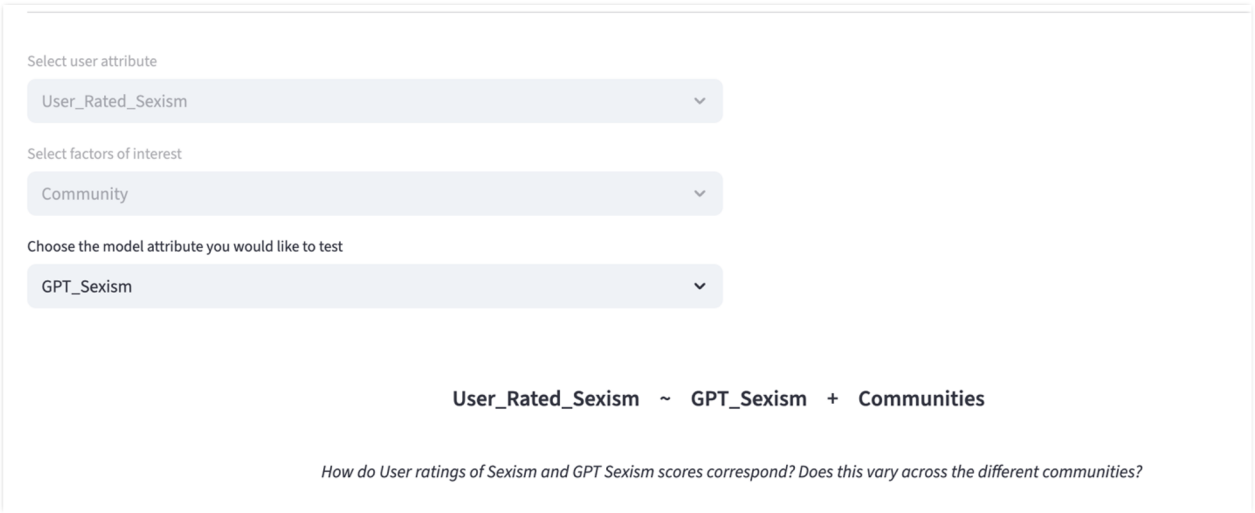


Figure 5.8: Participants can select variables of interest in a drop-down option. The research question and model equation will get updated accordingly.

UXP can view the raw data in a tabular format and filter and sort through them as well. UXP were also told that they could explore how well users' perceptions of sexism correspond to different model judgements, and assess impact of factors as well, in the 'Analytics' page. They could select which Model to focus on - `Perspective_Insult`, `Perspective_Identity_Attack`, `GPT_Sexism` using a dropdown option. Selecting this would update the analysis description, model equation, and the corresponding interactive visualization. I intentionally designed a visual analytics prototype rather than simply hosting interactive visualizations on a web page. This way, I could also gather feedback on using a MLE-based analytics toolkit for UX practice of AI.

Implementation: I developed this mock toolkit in Streamlit, using off the shelf components available as part of Streamlit. All logistic regression analyses were run in R for the original data, after collecting model judgements from Perspective and GPT. The resulting probabilities were generated in R, but visualized using plotly. The Box select feature in plotly was leveraged to build the custom interaction of inspecting the visualization and viewing actual text of the comments side-by-side (Data Viewer).

5.3.3 Study Details

a. Interview Protocol

The interviews were designed to be 1 hour long, wherein I would introduce the scenario, the visualizations, and observe how UX practitioners interacted with them. To ground the scenario better, I also provided four design concepts for how sexism detection would be handled within the communities (Figures 5.7, 5.8, 5.9, 5.10).

After discussing the above design concepts, participants were provided additional information about the different model attributes. Participants were also told that data was

collected from three different communities (A, B, and C) and that an interactive tool was available to explore the data. The participants were asked to share their screens and think-out-loud as they go through different model attributes and interpret the visualizations. After spending considerable time interacting with the visualizations, participants were asked to revisit the design concepts. I noted how participants made new observations, raised questions, and arrived at different design recommendations based on the insights gathered from the visualizations. The design concepts served as 'solution conjectures' that help with an understanding of the problem [32]. In the final stage of the interview, I probed how participants perceived this tool and the visualizations, how they might use it in practice, what UX/design activities would they use it for, what was helpful or challenging about this tool, and whether they thought the tool can support cross-functional collaboration. This study was approved by my institution's IRB.

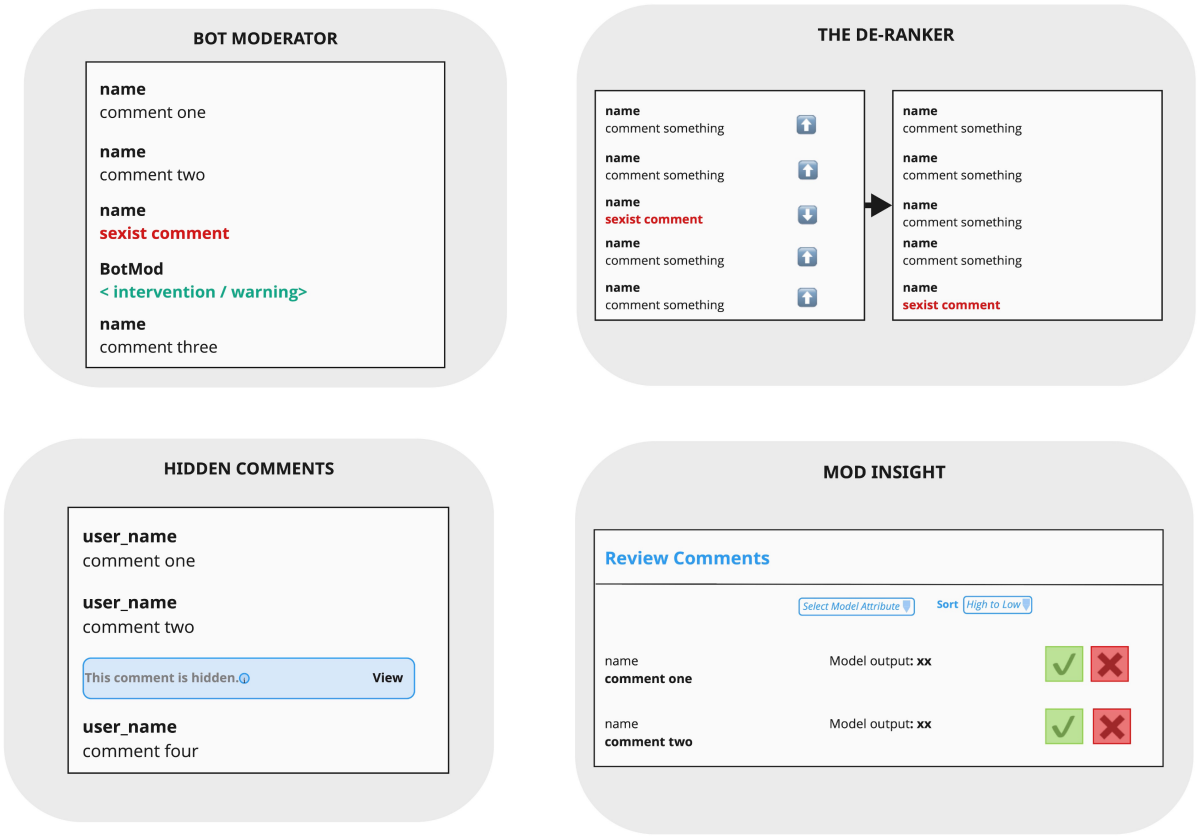


Figure 5.9: BotModerator (top left): A proactive design intervention where a bot intervenes with a warning or message after a sexist comment has been detected. Figure 5.10: De-ranker (top right) A proactive design intervention where sexism scores of comments are used to determine its ranking in the feed. Figure 5.11: Hidden Comments (bottom left) A proactive design intervention where sexist comments are automatically detected and hidden, not removed. Figure 5.12: ModInsight (bottom right): A reactive design intervention where moderators are provided with model judgements to make quicker or better decisions.

b. Participant Recruitment

For this study, I wanted to recruit UX designers and researchers who worked on AI/NLP/LLM infused products and features. I created a recruitment screener that collected information on respondents' current role, area of work, years of experience, educational

background, and their proficiency with statistics, probability, and interpreting information visualizations. I circulated the screener and the call for participation in LinkedIn, Twitter, mailing lists, and Slack-based UX communities. Once a participant was selected based on their responses, I let them know that they could be exposed to toxic, sexist, and profane content as part of this study through consent forms. I conducted interviews with 11 participants - 4 UX Researchers and 7 Designers. All the participants were working on AI or NLP based applications at the time of conducting the study. The interviews were conducted over Zoom, lasting approximately 60 minutes, and participants were provided an honorarium of 30 USD post study completion.

c. Data Analysis

The analysis process was guided by the procedures described by Corbin and Strauss [29]. I conducted a first round of open coding of all the interview data. At this stage, I paid attention to what actions participants took within the tool, what interpretations they arrived at, and what was confusing. Next, I moved to grouping similar codes into categories. At this stage, I analyzed the different stages of visualization interpretation and its outcomes. I also compared different participants' interactions, insights, and attitudes toward the tool to find common patterns. I started seeing repeated themes in the last set of interviews (8-11) which led me to believe that I would not uncover new perspectives unless I made changes to the prototype itself.

5.3.4 Findings

a. Three Stages of Visualization Interpretation and Interaction

Participants went through three stages of understanding visualizations. First, they attempted to understand the visualization at a basic level. At this point, participants used the hover feature to trace along the probability curves to bring up the toolkit and read the data

points. They had clarifying questions about whether the y-axis values implied if there was a higher probability of a comment being sexist or whether more people would find the comment sexist. They also had follow-up questions to clarify the scenario details such as whether the comments were rated by users from different communities or whether the comments themselves were from different communities.

Second, they tried to interpret what the visualization means. Most participants adopted a strategy of doing comparisons to better analyze what the visualization means. For example, P3 established community C as a baseline and compared how values changed for other communities. Participants (P1 P3, P4, P5, P10) also compared how probabilities differed for the lower, middle, and higher range of model scores on the x-axis. Almost all participants switched between visualizations of the three model attributes to compare and assess which was better or worse and what differences between them could imply. They specifically looked at the inflection points and hovered over them to make the tooltip appear.

Comparing model scores which were on a 0-1 scale and the probability of users finding comments sexist which was also on a 0-1 scale, led most participants (P1, P3, P4, P6, P9) to surmise that the relationship between the two should be **linear** and thus the line should have a slope of 1. Participants either stated this explicitly or it was implicit in the way they drew inferences. Using this as a rule of thumb also created confusion on why some curves were more S-shaped (as in the case of GPT) and what that meant. But still, they used it to determine whether model and user judgments align or if the model was over or under rating with respect to the users.

> *If this model was perfectly aligned with user selection, it would be like 45 degrees - like - 1 versus 1, 0.6 versus 0.6, but this is a different shape. The probability is generally lower than the model's prediction.* - P6

Other participants (P8, P11) wondered whether it was a good sign for the probability line to appear flat over a range of x-axis values and what an ideal curve should look like for evaluating model alignment.

> [Figure xx] *So presumably there's some ideal curve where it's 1 to 1 - But I don't know off the top of my head what that is. Maybe this one because it's flatter - The top part of the curve. So there's - you know more of the users have rated these comments as sexist. and the models also rated them as attacking someone's identity. So there seems to be close alignment between what the model is saying and how the users are ranking. And it's interesting that it's closer together here at the upper end of the scale. - P8*

Participants who were not experts in statistics were especially reluctant to make strong sweeping statements, acknowledging that they could be wrong or that they could be bringing in other biases (P1, P7, P11). Erring on the side of caution, they did not want their potentially wrong assumptions to affect design decisions. Along these lines, participants felt it would be helpful to have i) rubrics, templates, and examples to understand and evaluate differently shaped probability curves as well as scaffold comparisons and ii) more cues and information icons to guide them and provide explanations when necessary. Participants P5 and P10 explored comments in different regions of the visualization using the box select feature and to better contextualize what they were observing. P2 and P7 only discerned differences between the communities A, B, and C but were unsure of what the curves implied overall.

In terms of the final stage, 4 participants (P3, P4, P5, P10,) engaged in further sensemaking, going beyond what the probability curves were showing. For example, P4, a UX Researcher, was more interested in what was causing differences between the model score and probability values across the communities A, B, and C. He also selected different regions in the visualization and analyzed the comments, and surfaced questions about the nature and demographics of the communities. He reasoned that the probability of users finding a comment

sexist did not reach 1 because of disagreements in user ratings on what is and is not sexist, and wished to see more information about the labeling process. When analyzing comments in the Data Viewer, both P4 and P10 said it would be more helpful to retain context by situating the comment in the actual conversation.

P5 and P10 had similar goals in terms of analyzing when model and user judgments diverge. They would select an area over the lower or upper range of comments and in the comments viewer, compare examples where users had rated comments as sexist and not sexist for the same score. For example, after analyzing the visualization for GPT-Sexism for score ranges 80-100, 40-55, P5 said:

> [Figure 5.4] *Let's check out a different section [x1=10, x2 =30]. Something that users would rate as sexist but the model did not. So, it's interesting that these examples that users rated as sexist - they deal more with the chivalry side of it. But they lack profanity, but express inequality. That is exactly the challenge of determining things that are inherently sexist, while also determining which things are offensive. These are better examples where you could just De-rank - still leave things visible and leave it up basically like - accept more user feedback - in terms of whether or not it should get flagged. - P5*

P5 used the Data Viewer to look at specific comments and this helped him gain a qualitative understanding of why there might be misalignment between model scores and user ratings of sexism.

b. Deriving Design Implications

Participants used the interactive visualizations to derive implications for design in interesting ways. P3 started simulating the Hidden Design concept (Figure 5.9) with the `GPT_Sexism` visualization data.

> [Figure 5.4] *So in a scenario where we are hiding comments, but we are still allowing users to read at their own discretion. Maybe it's okay to be overly cautious. So maybe we start the cut off point somewhere, like, you know - 50 or 55. So let me like, maybe look at that particular cross section..... If I look at the comments, I can clearly see that even this is not a conservative enough. So what I would do is I would set the threshold very low, and then if people still want to read the comments, they can go ahead and do it and then we can continue to test in production. We could do logging to see, you know, who is clicking to read it, what communities tend to read it. You know, we can get into those kinds of design details.* - P3

Instead of showing the user rating, model score, and the comment in a tabular format, visualizing the data this way in terms of a probability distribution enabled more systematic exploration for participants. Other participants (P1, P5) also used data and insights from the visualization along with the corresponding comments to construct what-if scenarios, realistically simulate and evaluate the designs, and anticipate outcomes. The Data Viewer helped participants dig into the qualitative side of things, surface model behaviors, and match it with the appropriate design concept. For example,

> [`GPT_Sexism`] *Something that users would rate as sexist but the model did not..... So it's interesting that these examples that users rated as sexist - they deal more with the chivalry side of it. But they lack profanity, but express inequality. That is exactly the*

challenge of determining things that are inherently sexist, while also determining which things are offensive. These are better examples where you could just De-rank - still leave things visible. - P5

According to P5, the model rated comments which were both profane and sexist more highly than comments expressing inequality. So, he suggested using De-rank for comments that users would still find sexist but the model might not necessarily score highly. Along these lines, participants started generating conditional designs [84,94]. P5 wanted to leverage multiple design concepts to account for cases when we can expect the model and user perceptions of sexism to align and cases when we expect disagreements. P8 also echoed similar ideas.

In cases where there's the most disagreement between user score and model score will be exactly the time you'd want to flag comments. So I would use ModInsight I guess, on any community that seems to have a lot of things in that kind of - middle - ambiguous zone where you're gonna have disagreements between the model and the user. - P5

I guess one way to handle it could be by assigning - Say comments that are like 0-33 get one design intervention, because we are not really sure. The community isn't as aligned in how it's feeling about this - But then at the other end of the scale, things where there is great agreement between both the model and the community around what is sexist language - Maybe for those kinds of comments you have a more restrictive design intervention. So you are not just using one design intervention. - P8

Participants P2, P6, P9 wanted to know more information on the communities and few others (P7, P11) said they did not see any direct links between the visualizations and the design concepts except in terms of feedback loops.

c. Evaluating overall usability and usefulness

Participants perceived value in being able to visualize different model capabilities this way. P8 said,

> *Understanding how these different AI models perform and how they look at the world and being able to visualize that is really valuable. When you're going to release a feature that could measurably shift how people interact with your platform - one way or another. To be able to visualize that ahead of time and understand if that shift is going to be in the direction you want to go, or in another direction before you invest the time and energy to ship that feature is really valuable.* - P8

P5 felt such tools were helpful when dealing with ambiguous and subjective judgments, particularly to foresee potential harm. Participants (P2, P3, P5, P9) also mentioned using such a tool to gather evidence, get stakeholder buy-in, and influence decision-making. P1 thought of it as a hand-off tool from research to product teams.

P3 felt such tools would be helpful along the entire design cycle. In earlier stages of discovery, P3 could play around with and explore the data more. Later stages of evaluation would require answering specific questions to make decisions. P3 also added,

> *I think design needs to be extremely involved in asking the right questions. So that the analysis that we're doing on the data is driven by a user specific question, right? If it's just about answering questions, yeah I would be happy with a report. But it's not just answering the questions - It's also intuiting what to look for. It is UX that will have perspectives that other functions don't have. So that makes me want to be involved in this process so that we are thinking about it holistically and not just focused on metrics that other functions are focused on.* - P3

On the other hand, P9 - a designer as well, did not see themselves as "*running this*" often but imagined that they could leverage the data points after the visualization has been generated to back up design decisions and as a storytelling device.

Other participants saw themselves using this tool as more of an entry point and in collaborations with model developers, rather than to independently accomplish tasks. For example, P6, P7, P11 wanted to independently explore the tool but check back in with the team's AI engineers to answer questions, verify their understanding, and err on the side of caution. To overcome knowledge gaps, P11 also wondered if there was scope to leverage such visualizations for educational purposes in terms of sensitizing the designer and raising awareness. P8 saw this as a cross-functional tool that can support group decision-making. P9 also thought this tool could provide grounding and a shared language when communicating with AI engineers. Participants suggested features to push the collaborative potential of this tool such as exporting results, sharing snapshots and specific views of the analysis views through links, adding action items, and seeing notes from the team.

P4 saw such a tool being helpful for practitioners who want to partake in responsible and human-centered model development. However, for this particular scenario, he added that it can be hard to talk about diversity and inclusivity because people have different lived experiences, and it can result in emotional labor for people who have experienced marginalization.

> *It certainly helps people who want to be part of that endeavor to have data that they can shop with and say, like, look this algorithm is not performing well as judged by users, and, like this could be an ongoing approach to crowdsourcing smaller samples and checking against GPT for scale, right? Assuming we're all on the same page, that the ground truth is the person and not GPT.* - P4

Participants also felt providing summarized insights would make such tools practically useful. For example, P6 and P9 wanted to directly see the result of comparing user and model judgments and have recommendations that informed the product team. While P6 found it useful to see data specific to different communities and for different model attributes, they still wanted to view translated insights for each community. P6 felt if the conclusions drawn from the visualization are highly accurate, then they could be communicated directly. But for a higher level of uncertainty, the underlying data can be made available for further analysis. Though P5 interacted with the visualization extensively, he wanted to see more actionable insights.

> *I guess a summarized version of what type of insight is available with the data. The table is good for exploring around but I can't make any determinations...Provide what that actionable insight is rather than leave the - I would just say like, don't leave it up to designers to poke around and explore. Summarize it first. It's still good to see the raw data behind it, and just get a better understanding of it once you've tackled the bigger problems first.* - P5

Participants also thought providing summaries and translating the insights can support UX practitioners who may not be as familiar with interpreting such visualizations. Some participants expressed needs and features that can further push the analytical power of such a tool. P3 said,

> *Access to the data is huge, right? And being able to analyze the data is even better. I mean if we can turn this into a Tableau sort of thing, where I can go a little deeper into the data itself, especially because I think this is a cross section of user research as well as the model's accuracy. And so, I would like to slice the data a little bit more to see if we can learn more about how users are thinking about these types of comments.* - P3

Other participants also wanted features to add and examine other impacting factors such as language, region/country, and how long the community has been active. In addition to communities, P5 wanted to analyze how alignment varied according to different reasons users attributed sexism (*Behavioral Expectations, Denying Inequality, Endorsement of Inequality, Stereotypes and Comparative Opinions).*

Some participants (P3, P8, P10) also drew parallels to their own work and spoke about how they would test various factors. They also saw value in having such tools to conduct longitudinal evaluations. Participants P8 and P10 worked with open-source LLMs, which involved comparing human annotations to labels from LLMs and exploring how reliably the LLM could be used. Thus, they saw the direct applicability of such tools for their work.

> *Could we swap in different versions of the dataset to run the analysis and see if anything changes. And if I wanted to evaluate this with a different LLM - could I plug that in and run? That would be super helpful. Like, add a new model. And you plug it in, use your API key and all of that, and it'll connect to the model. And there would be a screen where you could import a data set to analyze… Or if we were working with local models, you know, could we plug those in as well?* - P8

Participants also desired features that further enabled mixed-methods analysis, built on top of the existing Data Viewer (P2, P5, P10). They suggested features to filter and sort data, mark areas of the visualization, annotate, and share/ track insights through a digital whiteboard. Few also mentioned wanting flexibility with how they view and interact with the visualizations. P11 wondered if it would be possible to map specific comments onto the visualization instead of going in the opposite direction. P10 wanted to switch views and see all model attributes in the same visualization for easier comparison.

This is a perfect graph for comparing users and the model. But I would like to compare between models and and users so like, which model is performing the closest to the user - having it in a single graph would be helpful. - P10

5.3.5 Key Takeaways

Participants were able to get a sense of how aligned the model scores and user ratings of sexism were through closer inspection of data points and the analyzing the comments. But most of them established that a line of slope = 1 would represent perfect alignment between model and user judgments and used that as a rule of thumb to evaluate actual data. Upon closer analysis of the interview data, I realized that using such a heuristic did not point to a misunderstanding of what underlying statistical analysis was used (linear regression vs logistic regression). Rather it was their *mental model of how model and user judgments should vary* - linearly and 1:1. Similar to how we would use a model's score to rank comments, set thresholds, and make moderation decisions, the probability of a comment being sexist was treated as a user score.

Given these expectations, the S-shaped curves were quite confusing for the participants. Participants who were not familiar with statistics or probability struggled more. Even when participants were making the right inferences, they were unsure of whether it was the right one. Since this study meant to use a probe to uncover UXP's perspectives, I used the standard visualization for representing probabilities (Figures 5.2-5.4). However, given that they are more common in the discipline of statistics and the challenges in interpreting them, I wanted to redesign them to be more intuitive and comprehensible for practitioners.

The Data Viewer for viewing comments supported participants' evaluation of model alignment, contextualization of sexism ratings, and reasoning of how the model was rating sexism. Some participants thought it was an important feature to have so that UXP can account

for the scale factor as well as empathize with individual user perspectives. Participants wished to have additional information such as the conversation surrounding the comment, disagreements in sexism ratings, and demographics of users who rated the comments. These visualizations are based on predicted probabilities and there might not be actual data for every possible value of x. But seeing a lower number of comments or no comments in narrowly selected boxes raised questions from participants.

Participants explored the design space of moderating sexist comments in tandem with analysis of the different models and attributes. They used the interactive visualizations to simulate different design concepts, anticipate adverse outcomes, and devise conditional designs that would play to the strengths of the model and limit its weaknesses. Upon prompting, they also brought up features that would make the tool more collaborative and features that could make it a better analysis tool. For example, participants asked if there were ways to add and test different variables, select facets of data to view in the visualization, view shifts over time, and make annotations over the visualization. Despite these outcomes, the barrier to entry with this kind of tool and visualizations was still quite high. My next step was to re-design these visualizations and evaluate them in an unmoderated, task-based setting. In this study's probe, the visualizations did not include confidence intervals. I account for this limitation in the re-design and evaluation study.

5.4 Evaluative Study

In this subsection, I describe the follow-up study I conducted to evaluate the redesigned visualizations. This study was also an opportunity to use a scenario that involved a generative model, given the recent prevalence of large language models.

a. Visualization redesign

One of the challenges in the previous visualizations was that it encoded a lot of information in the space of a 2d graph. For each model score on the x-axis, there was a corresponding probability of users finding the comment as sexist. When subtracted from 1, it would give the probability of users not finding the comment sexist. If we used multinomial or ordinal data instead of binary data and additional variables/factors, it could get even more complex.

One way to reduce the complexity might be to show the resulting predicted probabilities for one set of explanatory variable values at any given time. This might also eliminate the need to represent the probabilities as a continuous curve, which in turn led to the idea of using bar charts. If we have a set of covariates $x_1, x_2, \ldots x_n$, then for a set of values for each x, we can represent the probability of 0/1, for each category, or each Likert item as a bar and the associated confidence interval as an error bar or whisker at the top (Figure 5.12 for example).

b. Evaluation Study Rationale

In this study, I aimed to evaluate the redesigned visualization. I also wanted to account for a scenario that involves generative AI (e.g., summary generation or question-answering) rather than predictive or traditional AI (e.g., sexism detection). Hence, I decided to adopt an iterative design and testing approach to evaluate the visualizations, rather than use the visualizations from the formative, qualitative study as a baseline condition. The latter visualizations were particularly frustrating for UX practitioners who did not have a technical and/or statistics related background. Since the generative AI scenario was more complex, the resulting probability curve visualizations could also end up being quite complex. So, I chose to focus on evaluating the redesigned evaluations to ascertain whether they improved understanding through clear post-task measures.

That said, I believe there is still value in using visualizations of predicted probabilities used in the qualitative part (Figures 5.3-5.5). They show the shape of the underlying phenomenon in one shot. However, that visualization format might not be suitable for analyzing multiple factors (which would involve including multiple visualization facets) or for all UX practitioners. Thus, I provide a second set of designs and evaluate them in this study. I believe the visualizations put forward in this evaluative study might serve as a better way to ***introduce*** UX practitioners to MLE results and help them get familiar with its ideas before diving deeper (Section 5.5.3).

For this study, I first describe the scenario and data used for the evaluative study and how I redesigned the visualizations. In the second subsection, I describe how I designed the evaluation task and questionnaires to collect data. In the last subsection, I present the results of the evaluation.

5.4.1 Scenario, Data, and Visualization Construction

I selected the use case of summarizing news articles, where UX practitioners would be asked to imagine that they work as part of a cross-functional team focusing on reimagining digital news consumption. The team has to consider the value in pushing AI-generated news summaries to support engaging with broader topics, receiving updates on stories of interest, providing entry points for a news article, and avoiding feelings of fatigue. Such a use case would require evaluations to ensure that the summaries do not contain factual or false information. Human evaluation will not only complement automatic metrics of evaluation (ROGUE, BertScore, BLEU) but also provide contextual feedback on these summaries.

As part of the scenario, participants were provided the following information about data collection:

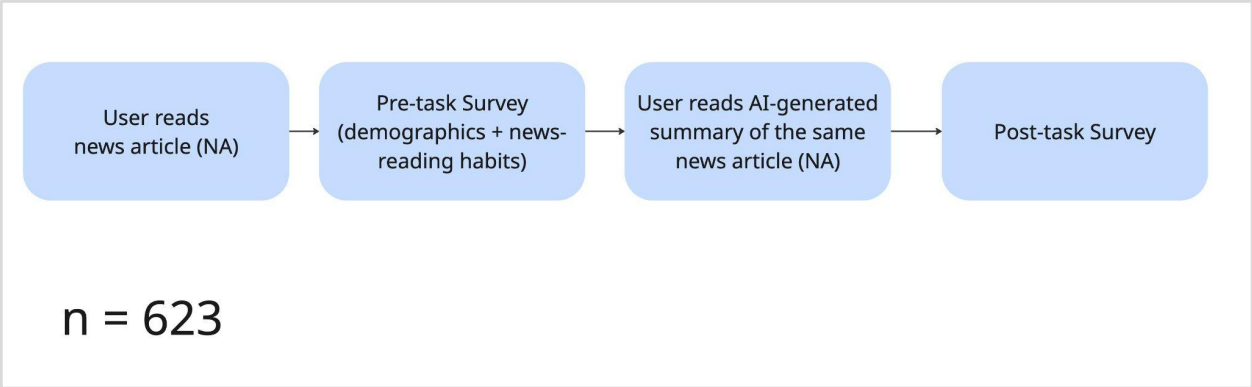


Figure 5.13: The data collection procedure provided to UX practitioners as part of the hypothetical use.

The post-task survey collected Likert ratings (*Strongly Disagree, Disagree, Neutral, Agree, Strongly Agree*) for the following attributes of the news summary:

1. [INF] Informative: The summary covered the key points of the news article.
2. [SURR] Surrogate: The summary captured the article well enough to act as a surrogate or stand-in.
3. [ADD] Additional Context: While the summary provided an overview, engaging with the original source offered valuable additional context.

The scenario is hypothetical but inspired from [28,94,157]. The following explanatory variables were provided:

1. Article Length
2. Article Type
    a. Primary Reports: Objective accounts from first hand investigation
    b. Secondary Reports: Opinion based news, features, commentary etc.
3. Summary Length
4. News Reading Behavior

The scenario description details that a UX researcher had coded the open-ended

questions in the pre-task survey to identify three news reading behaviors, which is included as the final explanatory variable.

a. Tracker: Likes to stay updated; Spends 5-10 mins per day; Uses skimming or scanning techniques to catch up on news
b. Conversationalist: Likes to read news and read the comments section; Often comments and engages in discussion
c. Reviewer: Likes to read in-depth and thoroughly; Spends considerable time reading articles of interest, allocates time for it by possibly saving articles for later

Participants will get to analyze results of running ordinal logistic regression on the three attributes (INF, SURR , ADD). Since the objective of the study was not to understand the effectiveness of news summaries, the data was generated synthetically by reverse engineering from a desired probability distribution. Three different models were fitted using ordinal logistic regression in R.

INF ~ Article_Length + Article_Type + Summary_Length + News_Reading_Behavior (4)

SURR ~ Article_Length + Article_Type + Summary_Length + News_Reading_Behavior (5)

ADD ~ Article_Length + Article_Type + Summary_Length + News_Reading_Behavior (6)

I generated the synthetic data such that only article type and length would significantly impact ratings of INF and only news reading behaviors would significantly impact ratings of SURR and ADD. For INF, UXP would have to analyze the effect of a continuous variable (article length) and a categorical variable (article type). SURR and ADD are inversely related. The higher a news reader rates the summary as a stand-in, the less they might rate the value of

reading the original article. In this case, UXP would examine the effect of news reading behaviors (Conversationalist, Reviewer, Tracker) on ratings of SURR and ADD.

Figure 5.14: Interactive visualization to view probabilities of rating summary informativeness

For INF, I used predicted probabilities (e.g., *What is the probability of a reader agreeing/strongly agreeing that the summary is informative if article length = x and article type = primary report*) and visualized the same using bar charts (Figure 5.14). For SURR, I also use predicted probabilities (e.g., *What is the probability of a Conversationalist disagreeing that the summary serves as a surrogate?)* (Figure 5.15). If we have the probabilities for a particular

attribute rating for Conversationalist and Tracker, we can make pairwise comparisons by calculating the absolute and relative differences in probabilities for the two groups.

For ADD, I present these differences directly instead of showing the actual, underlying probability values (e.g., *How much more likely is it for a Tracker to rate Agree for the given statement, compared to a Conversationalist?*). To visualize absolute and relative differences, I adapt a forest plot which is an established method to visualize effect sizes. Participants can select two behaviors - a reference and a comparator and generate the visualizations seen in Figure 5.16.

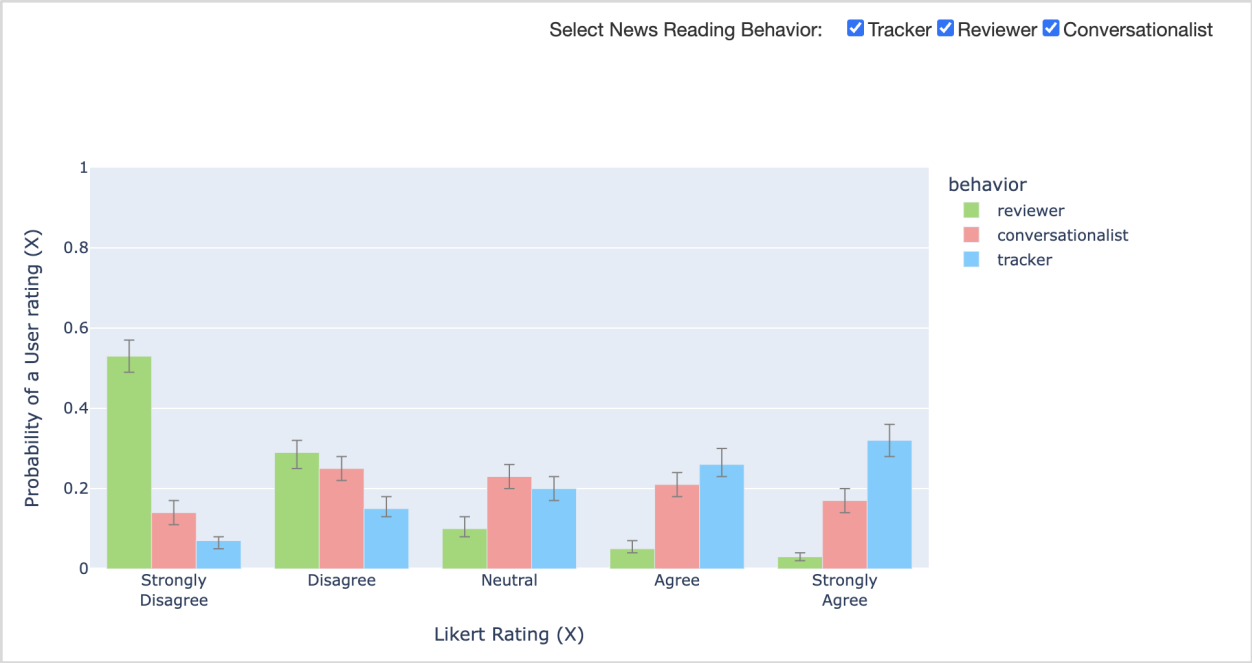


Figure 5.15: Interactive visualization to view probabilities of rating SURR by the three News Reading Behaviors

These three different attributes make up a realistic and sensible use case. They also help test various measures and formats of the visualization. The redesigned visualizations better emphasize an important aspect of analyzing this data - counterfactual and what-if thinking (e.g.,

What if article type was secondary reports instead of primary reports? How did the probability of ratings change for Reviewers vs Trackers?) [101,113,174].

After running the models in R and generating the quantities of interest (predicted probabilities, absolute differences, relative risks), I designed the visualizations and generated them using plotly/Dash.

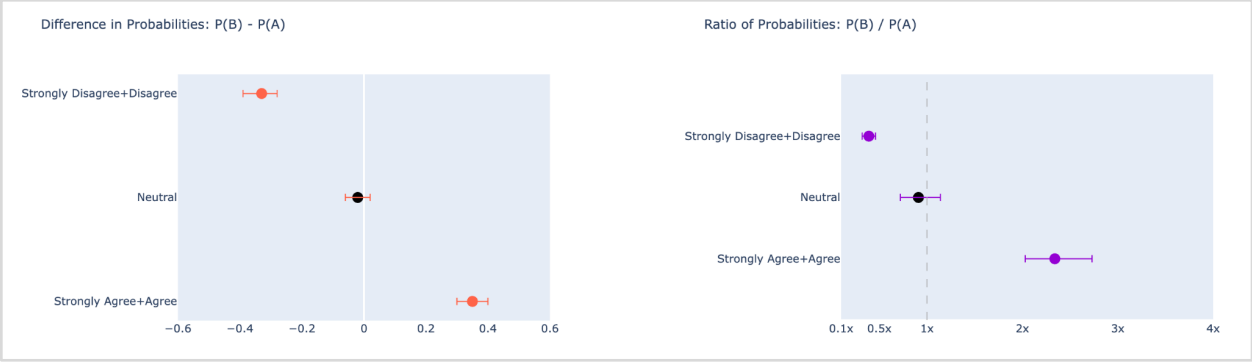


Figure 5.16: Visualization for viewing absolute and relative differences in how different news readers rate the ADD attribute. Immediately we can see that there is a positive absolute difference and a relative difference > 1 for the Likert rating Strongly Agree/Agree.

5.4.2 Study Details

a. Designing the Questionnaire

Friel et al. [53] present three levels of questions that can be asked to test comprehension of a graph: elementary, intermediate, and overall. Elementary level involves extracting basic information from the graph. The intermediate level involves finding relationships between the data in the graph. The final level involves reading beyond the data to extend, assess, predict or make inferences. I adapt these levels to design questions that test comprehension of these probability-based visualizations.

Level 1: Basic Comprehension - Are participants able to extract information from the visualization?

Level 2: Interpretation - Are participants able to interpret the data and understand trends or relationships?

Level 3: Application - Are participants able to apply the insights to reason about a problem or to evaluate solutions?

Attribute	**Level 1**	**Level 2**	**Level 3**
INF	6	2	1
SURR	3	1	1
ADD	4	2	1

Table 5.1 Number of questions of each level for each visualization evaluated

I designed Level 1 questions in a way that participants do not always have to find the probability of an event, but also deduce counterfactual data or outcomes that correspond to certain probabilities. For example, for INF, participants will have to find the probability of a news reader rating *Strongly Agree* for an article length of 950 words and of type Primary Report (Row 1, Table 5.2). But they also need to report which Likert rating has the highest probability of being selected if the article length was 1200 words and of type Primary Report (Row 3, Table 5.2).

Attribute	Article Length	Article Type	Likert Rating	Probability
INF	950	Primary	Strongly Agree	?
INF	1100	Primary	Agree	?
INF	1200	Primary	?	Highest
INF	950	Secondary	Strongly Agree	?
INF	1050	Secondary	?	Highest
INF	1200	Secondary	Strongly Disagree	?

Table 5.2 Design of Questions for Level 1 (INF)

Level 2 questions for INF examined if participants understood how article type and article length affected ratings of AI summary informativeness. Finally, for Level 3, they were presented a case where the team is rolling out AI-generated summaries as push notifications to increase engagement on features and opinion articles (secondary reports) and asked how they would respond.

Attribute	News Reading Behavior	Likert Rating	Probability
SURR	Tracker	Strongly Agree	?
SURR	?	Neutral	Highest
SURR	Reviewer	?	Highest

Table 5.3: Design of Questions for Level 1 (SURR)

Level 2 questions for SURR aimed to understand if the participants are able to rank the three behaviors in the order of who is most likely to use the AI-generated summary as a stand-in / surrogate. Again, for Level 3, an open-ended question was posed, asking how the participant might respond if there were plans to introduce a summary feed in the home page.

Attribute	**News Reading Behavior (reference)**	**News Reading Behavior (comparison)**	**Likert Rating**	**Absolute Differences**	**Relative Differences**
ADD	Tracker	Conversationalist	Strongly Agree/Agree	?	?
ADD	Tracker	Reviewer	Strongly Agree/Agree	?	?

Table 5.4: Design of Questions for Level 1 (ADD)

Similar to the SURR attribute, Level 2 questions for ADD aimed to understand if the participants are able to rank the three behaviors in the order of who is most likely to find additional value in engaging with the original news article. For Level 3, the idea of using adaptive designs to better adapt the design based on who is using it, is proposed. Participants were asked to determine if they would design for three behaviors separately or cluster any of the existing behaviors, or not follow adaptive design at all, and their associated reasoning.

These questions aim to get at how participants understand probabilities of certain events through visualizations of MLE methods and how they are able to apply it in a specific human-AI interaction scenario. Thus, it measures their conceptual understanding. The final page collected more subjective feedback on whether UX practitioners considered the tool to be useful for their practice, what specific design and UX activities they would use it towards, and how this tool, if available, should support them in evaluations of AI systems. I also collected data on their experience working in UX, experience working with AI, major of study, and experience with quantitative methods. The questionnaire was presented using Qualtrics and links to the interactive visualizations (developed using Dash) are available from within Qualtrics.

b. Study Deployment

After obtaining IRB approval, I circulated the call for participation in relevant mailing lists, UX-focused Slack channels, and on professional platforms such as LinkedIn. The inclusion criteria did not change between the prior qualitative study and this one. However, I screened participants to ensure that they had worked with some kind of UX metrics, any quantitative method, and data visualizations before. This prior work experience was defined broadly and the participant could have either conducted analyses themselves or interpreted and talked through the results with other UX practitioners or collaborators. Instead of raffling for a gift card, all participants were provided with a coffee gift card (15 USD).

28 UX practitioners completed the study, out of which 13 were mixed methods UX researchers, 5 were qualitative researchers, 2 were quantitative researchers, 5 were UX Designers, 2 were Product Managers, and 1 was a UX engineer. Considering only the participants who completed the study in a single setting (22), the average time taken to complete was 34.1 minutes.

5.4.3 Results

a. Visualization Evaluation

INF:

For INF, most of the participants got all questions correct. For questions testing basic comprehension (Level 1), 24 out of 28 participants got all answers correct. Based on the answers provided and the fact that they got the Level 2 questions right, it seems likely that the other 4 participants misread the question (2 participants got the same question wrong because they answered with the probability of rating *Strongly Agree* instead of *Agree*. 2 other participants also got a different question wrong because they answered with the probability of rating *Strongly Agree* instead of *Strongly Disagree*.)

However, 26 out of 28 participants answered Level 2 questions correctly. 2 participants stated that the article type (primary and secondary reports) has no effect on ratings of summary effectiveness which was not the case. For the Level 3 question, participants recommended that the 'push notifications with AI-generated summaries' features focus initially on articles that we know have a higher probability of being informative, based on the data. They discussed how they would use insights from the visualization to convince the team to constrain the feature, collect additional analysis on why the AI summaries might fall short, or update the original design.

> *I might recommend this feature should only be used if the secondary reporting is under a certain word count (say, <1000 words), as most users feel the summaries aren't very informative as the article grows in length.* - P11

> *Since AI-generated summaries scored as less informative for secondary reports, I wouldn't want the reader to walk away with only the AI-generated summary, so I would at least want a link to the full article to be included in the push notification.*

> *There could also be cues to the reader that the summary should be taken with a grain of salt, so to say. Ideally, we would source the primary report as well.* - P24

Conversely, a participant wondered if less ratings of informativeness of the summary work well for a push notification because it would prompt the user to find out more.

> *I would recommend a trial period or A/B test of this feature. While AI generated summaries are considered uninformative for secondary reports, this does not mean that users will not respond to the push notifications. For example, the lack of informativeness may prompt them to click on the notification and read more.* - P14

SURR:

For SURR, 26 out of 28 participants got the Level 1 questions right. (One participant got the behavior type wrong when asked which behavior segments had the highest probability of staying neutral about this attribute. Another participant got the value wrong when asked for the probability of a Tracker rating *Strongly Agree*.) However, all participants got the Level 2 question (rank the three behaviors in order of agreement) right.

The Level 3 question asked participants to think about the implications of a summary feed in the news app. Participants responded with i) questions about the design objective of the AI summaries, ii) questions about the size of each reader behavior segment to better inform next steps, ii) how to better cater the design to each of the three behavior segments, and iii) more fine-grained data on how these ratings change for different genres of news article (global news, sports, business).

> *If the summary feed is designed as the "last stop" - e.g. the reader consuming the summary itself is the goal, then focus on trackers, then conversationalists, and avoid*

reviewers. If it's designed to pique reader interest and act as a funnel entry point, then further testing is necessary to see what each group does after seeing the summary. In this case reviewers might be an ideal target if it encourages them to click through to the full article. - P26

We should consider triggering (making summary appear) more frequently based on user segment (e.g., news reading behavior). - P5

[*We should] only push it for users whose reading behavior matches that of the Tracker---doing more of a personalized intervention instead of ruining the experience of Reviewer readers through one size fits all intervention.* - P18

ADD:

For ADD, 20 out of 28 participants got all Level 1 and Level 2 questions right. 3 participants did not get any of the answers right. The others answered ¾ questions correctly. Overall, the number of correct responses was lower for ADD than for INF and SURR. Participants noted that it was their first time seeing this kind of visualization and hence was more unfamiliar and non-intuitive. To understand what was being compared, the direction of comparisons, and the corresponding absolute or relative differences took time and involved a higher task load compared to prior visualizations.

Given the absolute and relative differences between how readers of different behavior segments rate this attribute, the Level 3 question probed how they might leverage adaptive designs. Since both Conversationalists and Reviewers would strongly agree that the original article provides valuable additional information, it would be reasonable to suggest adaptive designs for each of the behaviors, or to cluster Conversationalists and Reviewers together and

treat Trackers separately. Out of the 20 participants who correctly answered Level 1 and Level 2, 7 chose the former, and 13 chose the latter.

> *Conversationalists and Reviewers show similar behavior to a certain extent. Clubbing would allow for an adaptive design for two clusters.* - P17

> *Each of the behaviors place value on different things, and report different levels of value for the AI summaries. It would make the most sense, given that there are only 3 different behaviors, to create adaptive designs catered to each of the behaviors.* - P28

> *The deltas appeared to be significantly more prominent between "trackers" versus the other two, and much less so when comparing "conversationalists" and "reviewers" directly.* - P16

Overall, participants provided the following feedback on the visualizations:

	Visualizations were easy to interpret	*Visualizations were an engaging way to understand the data*	*I am confident in my responses to the questions*
Strongly Agree/Agree	22	25	18
Neutral	2	2	6
Strongly Disagree/Disagree	4	1	4

Table 5.5: Overall feedback from UX practitioners

When asked to select visualizations that were easiest to work with, 17 participants selected both INF and SURR, 5 selected INF, 4 selected SURR, and only 1 selected all three.

b. Application for UX and Design Practice of AI

I also collected participant perspectives on why they thought such MLE-based visualizations were helpful at all, how they might use insights derived from them towards different UX activities, and what parts of the data collection and analysis process they would like to be actively involved in. As shown in Figure 5.17, almost all participants thought the visualizations helped simplify complex information about regression results. For this specific use case, 18 of them considered it useful to relate user feedback on AI outputs to their behaviors. 14 of them found value in simulating different scenarios and doing 'what-if' analyses. 14 of them also thought it was helpful to consider and analyze the effect of multiple factors.

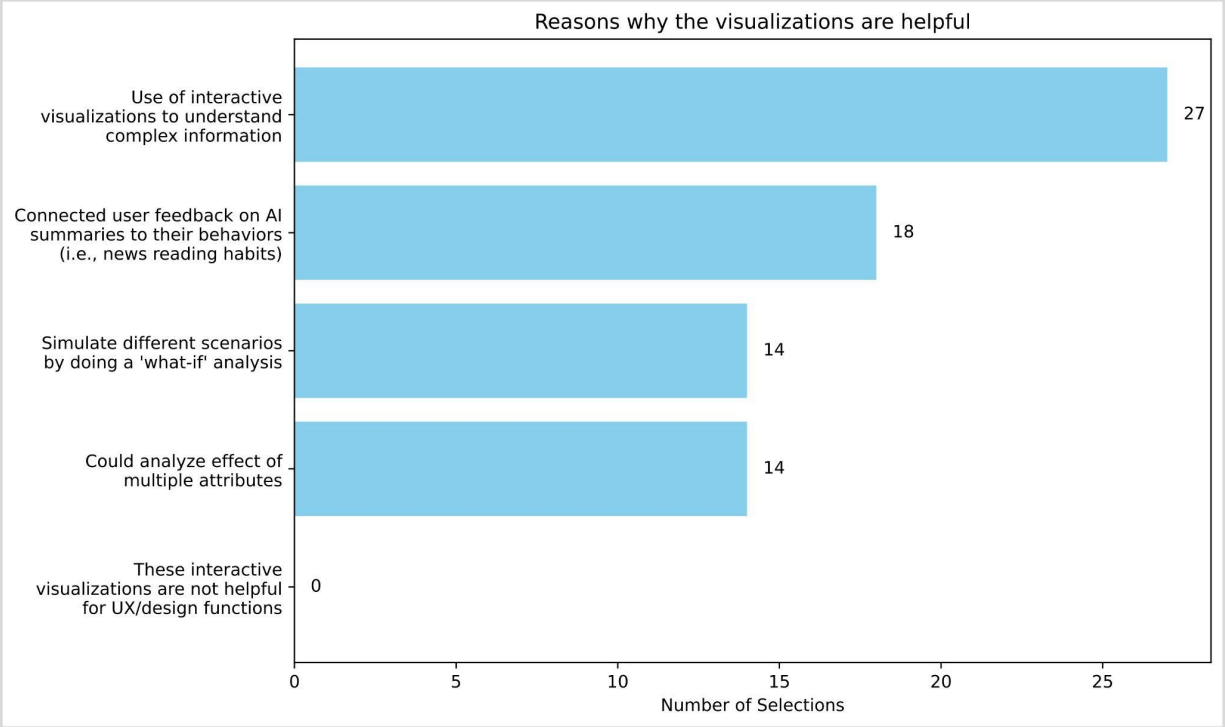
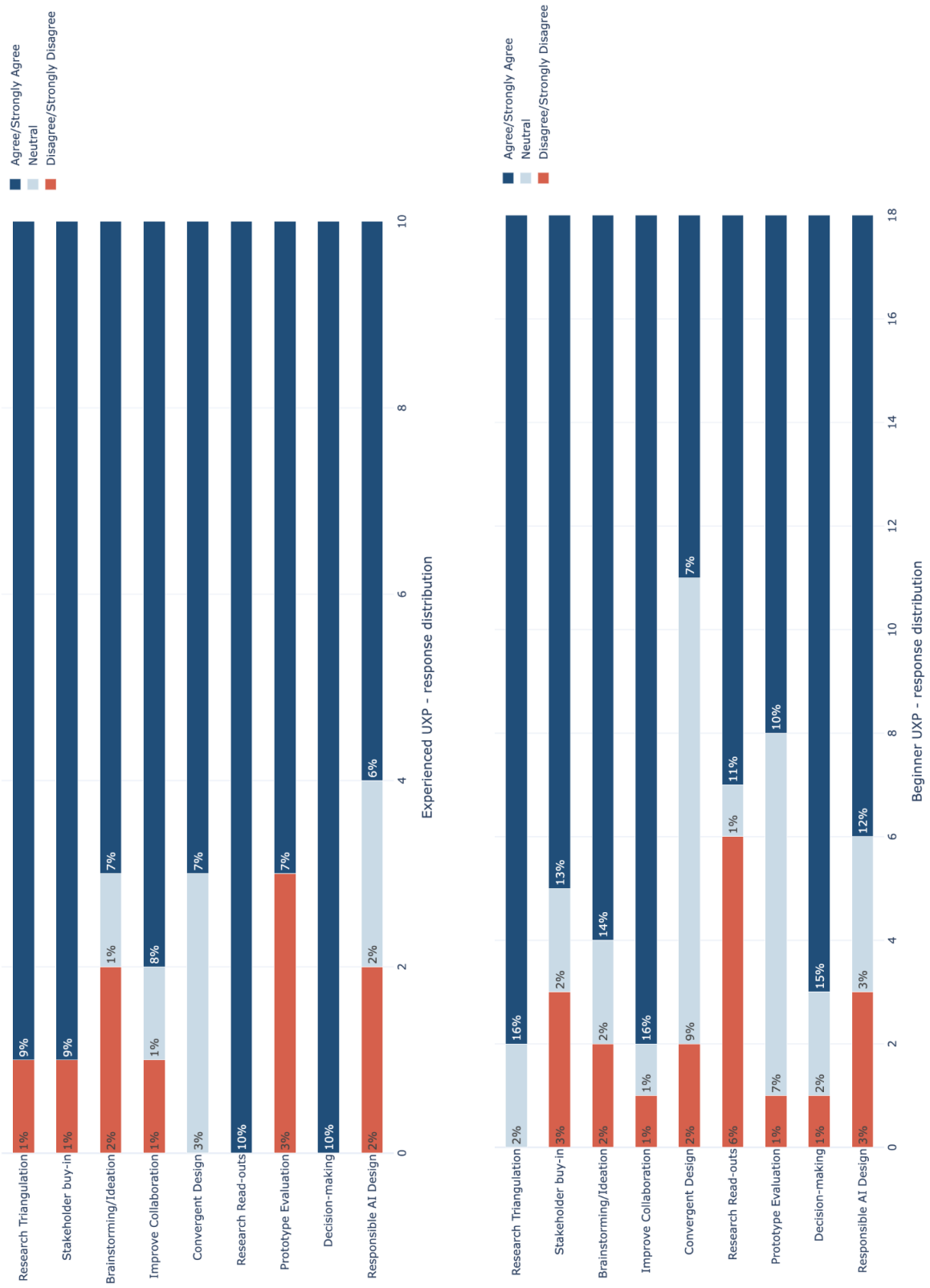


Figure 5.17: Participant perspectives on why the visualizations were helpful

I also wanted to understand how they might use insights from the visualizations towards different design/UX activities when working with AI, if such visualization-based tools were available. For this question, slicing by role (designer vs researcher) did not lead to meaningful differences, but slicing by experience working with AI (beginners vs experts) revealed some differences. Experienced UX practitioners thought such visualizations would be most helpful for research read-outs, decision-making, followed by triangulating with data from other methods and getting stakeholder buy-in (Figure 5.18). More novice UX practitioners thought such visualizations would be most helpful for triangulating other methods and improving collaborations with technical stakeholders, followed by getting stakeholder buy-in and informing decisions (Figure 5.19). It is possible that more experienced UX practitioners were familiar with AI and/or established collaborations with AI practitioners, in comparison to novice UX practitioners.



Figures 5.18, 5.19: Participants' perspectives on which UX/Design activities would be supported by the visualizations

Apart from stating that they would like to test different factors (e.g., news genre), participants thought it was important to be part of the data collection process. As Figure 5.20 shows, not all participants wanted to be involved in running MLE methods themselves. However, they found value in understanding the results of these methods. They also wanted to play an active role in deciding what outcomes are measured, what factors are analyzed, and making decisions based on the results. As P3 said from the prior study, *"it's intuiting what to look for"*.

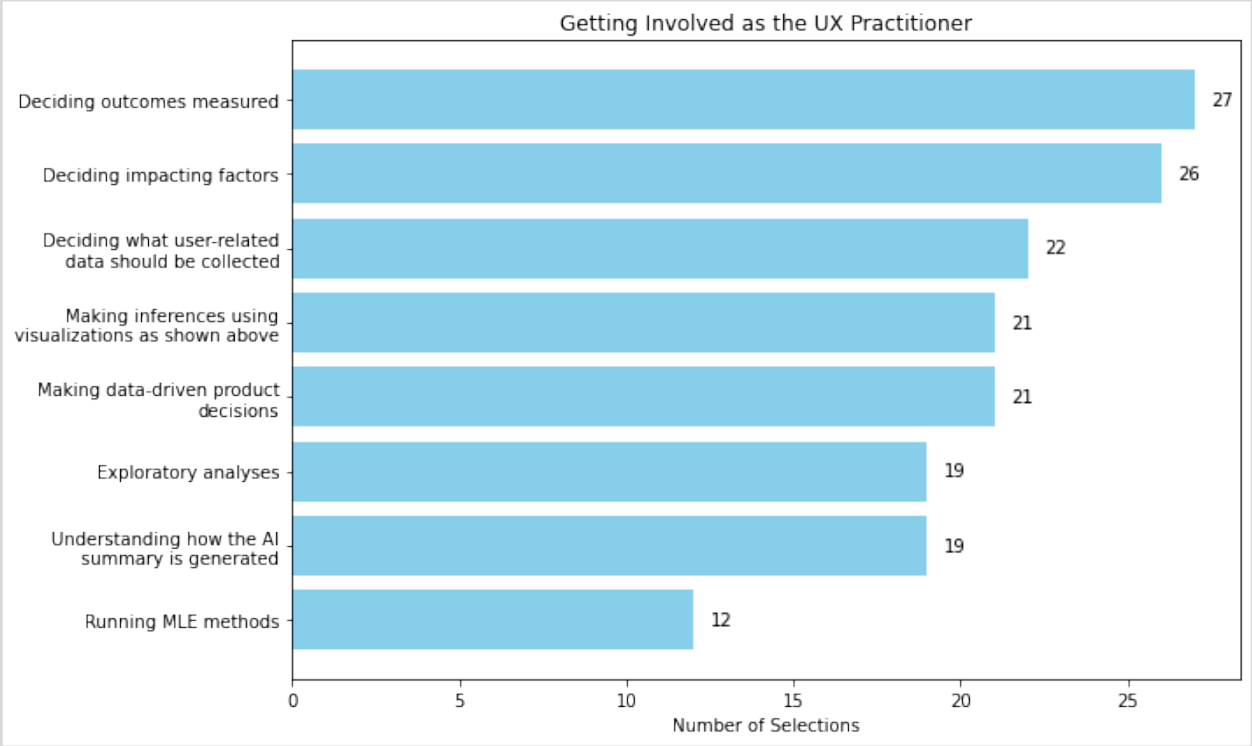


Figure 5.20: Participant perspectives on what aspects of the tool's features they should have access to as the UX Practitioner

16 participants said they would recommend such an interactive, visualization-based tool to other UX practitioners, and 12 participants selected 'Maybe'. In a follow-up, open-ended question, participants provided feedback on improving the interactions with the visualizations

and how to scaffold interpretations of the visualization. Some participants wanted the tool to include tutorials and examples to ensure that their interpretations of absolute and relative differences were correct. This was important for making the tool as accessible as possible which in turn can support collaborations. Participants also thought tutorials and examples would help for engaging with non-technical stakeholders or stakeholders who were not as familiar with statistics. Others mentioned creating static versions of the visualizations for sharing with non-UX practitioners.

Some asked about adding other variables and easy ways to (re-)generate visualizations. Another participant mentioned that the questions served as a way of guiding and scaffolding them through the different visualizations. A few participants suggested making the visualization more reactive to changing the article length on the slider and including options to bucket the word length by 10 or 50 words or manually enter an input.

5.5 Discussion

Through both studies, I demonstrate how prevalent ways of visualizing predicted probabilities can be re-designed to make it easier for non-statisticians to interpret the data (Figures 5.12-5/14) and how to fuse smaller, more qualitative analyses alongside analyses at scale (Data Viewer - Figure 5.6). The redesigned visualizations were easier to interpret compared to what was reported in the formative study (5.4.3). In this section, I discuss how visualizing MLE results inform UX practice of AI and design implications for building practitioner-oriented toolkits.

5.5.1 Utility of Visualizing MLE Results

The findings of the qualitative and the quantitative study demonstrate the value of visualizing results of MLE methods for human-centered AI evaluations. There are two aspects to visualizing results of MLE methods - i) explanations and ii) predictions.

Explanations. Beyond checking p-values to see which factors/independent variables were significant, these visualizations show how the outcome/dependent variable is related to them [80]. In the first study, participants were able to understand how the model and user judgements align with each other. Even if we do not compare user and model judgements directly, these methods help understand how and why user judgements differ based on different demographic and/or sociotechnical factors. For example, how does user evaluation summaries vary based on reading behaviors or how do user judgements of toxicity vary based on personal identity? If there are tangible differences, these factors will continue to be in play when we introduce AI in these tasks. By running these methods and visualizing the results, UX practitioners are able to understand how different factors of interest affect outcomes. This understanding is further supported by what-if and counterfactual thinking [113,174].

In the follow-up study, 25/28 participants agreed that this kind of insight would help triangulate findings from other methods. This explanatory power can also be leveraged in research read-outs (21/28), to get stakeholder buy-in (22/28), while decision-making (25/28), and collaborating with technical stakeholders (24/28).

Predictions. The second aspect is predictions (*What is the probability of a user-related outcome for a particular set of factors?)*, which helps with simulating future scenarios, evaluating design concepts and exploring the design space, and minimizing risk of adverse outcomes. Design evaluation would also be possible to some extent by just using varied inputs and corresponding

outputs but it can be random and unsystematic [94]. In the interviews, designers were able to more systematically explore consequences for low/high probabilities of user ratings or low/high model scores. They were able to determine conditions that would make a design concept effective or unsuitable.

In the follow-up study, there was agreement that these insights would inform decision-making (25/28), design ideation (21/28), and support stakeholder buy-in (22/28). The number of participants who thought it would help during the convergent design phase and with evaluating prototypes was comparatively lower (14 and 17 respectively). One possible explanation could be that the participants did not do related tasks of selecting between two design concepts or evaluating a design idea after interacting with the visualizations. Only participants in the qualitative study engaged in thinking about these aspects (5.3.3).

5.5.2 Judging Effects Between Outcomes and Factors

The study results also raise interesting implications for interpreting expected values / predicted probabilities, absolute differences, and relative differences. Commonly used effect sizes such as Cohen's d have associated guidelines for interpreting it and what values constitute a small, medium, or large effect [26]. But what thresholds might we set for predicted probabilities when making decisions? For absolute differences, a positive difference shows higher probability, while a negative difference shows lower probability. With absolute and relative differences, the confidence interval should not include the 0 and 1 respectively, for the effect to be significant. But there are no precise guidelines for comparing between two significant values of absolute or relative differences.

Ortloff et al. 's [111] study presents an interesting relationship between the scenario and judgement of effect sizes. They presented scenarios related to Security/Privacy and a scenario related to Work/Productivity improvements. Though the authors hypothesized that for the same numerical value, researchers might estimate the effect to be qualitatively larger if the context is

related to Security/Privacy than when it is related to Work/Productivity. However, only some of their participants did so, while others set similar thresholds for effect sizes despite the very different implications of both scenarios. Participants of the study (HCI researchers) leaned more on the study context to judge effect sizes if they were not experts in statistics. For others, judging effects in more areas like medicine or clinical decision making was more critical than the presented contexts. For my own study, if participants had been provided both use cases (sexism detection and AI-generated news summaries) and similar types of effects (actual probabilities or differences), then we could have observed differences in interpretation. The use cases presented can vary on the 'UX' spectrum and/or on the 'Responsible AI' spectrum. Apart from the use case, sample size, confidence intervals, and statistical experience of the practitioner can also influence interpretations and judgements [111]. Future work can help explore these aspects (especially confidence intervals) further and more widely among practitioners working in the AI space. It is also important to focus on how these judgements feature in and support collaborative discussions, negotiations, and decision making in a cross-functional team setting.

In the studies I conducted, I did observe that making comparisons helped participants judge implications of these effects to some extent. For example, Conversationalists and Reviewers are 2 and 3 times more likely than Trackers to find more value in the original news article. But a few participants made comparisons between Conversationalists and Reviewers (1.25 times) and decided to cluster them. However, others who did not report making that comparison also happened to treat them as distinct behaviors. The Data Viewer in the qualitative study also helped to an extent (5.3.3.b). Participants decided on an arbitrary threshold for user probability of finding comments sexist. Then on examination of actual comments that fall under that region in the visualization, they adjusted the thresholds. Sexism can be viewed along many dimensions - profane, hostile, subtle, sarcasm and so on. Based on the design idea they were working with and the kind of sexism they were seeing in the comments, participants refined probability thresholds.

5.5.3 Designing Practitioner-Oriented Toolkits

UX practitioners come from diverse backgrounds and have varying skillsets. In the formative study, UXP displayed varying levels of ease in working with the visualizations (5.3.3), In the evaluative study, though 26/28 participants got interpretation of INF and SURR visualizations correct, 8 participants did not find the ADD visualization (representing absolute and relative differences in a forest plot) as intuitive. Thus, when designing tools for UXPs and other stakeholders, two design principles are important: designing for different fidelities and progressive disclosure. Practitioners who are not as familiar with statistics or probabilities should be able to get high level insights with minimal friction. We cannot eliminate difficulty altogether because this analysis does require them to question, understand, and critique what they are seeing. But we should remove unnecessary challenges in this process. Instead of overwhelming or increasing the workload for practitioners, progressive disclosure will help ease them into the task and support learnability as well. As one participant rightly pointed out, the questions in the evaluative study served to scaffold participants' explorations and enabled them to arrive at key takeaways (direction of the effect between a factor and an outcome). For a practitioner-oriented toolkit, similar approaches can be used for scaffolding and progressing through the analysis, and verification of interpretations.

Once UXP were exposed to visualizations from running certain models, they got curious about other factors and expressed interest in testing them as well. Using open-source packages provides options for computing expected values / predicted probabilities, absolute differences, and relative differences. Thus, the practitioner-oriented toolkit should provide a front-end for triggering new analyses and viewing the results as visualizations (similar to [78]). It might be easier to decide what visualization to use based on a set of if-then rules, which will depend on the number of independent variables and their data types. However, robust evaluation of the fit

of the model (goodness of fit) can be trickier. MLE methods require larger sample sizes and most measures of goodness of fit are relative [139].

Chapter 6: Discussion

Through this dissertation, I make both empirical and design contributions. In my first study, I interviewed 14 UX practitioners to uncover challenges of designing with AI and collaborating as part of cross-functional teams that involved AI researchers and practitioners. This study raised several directions for future investigations and thus was formative in informing this dissertation work. One of these directions was analyzing the practitioner tools designed by academic and industry researchers to better tackle the challenges of developing human-centered and responsible AI systems. Specifically, I analyze who the intended users of these tools are, what aspects of AI work they seek to inform, and how the tool supports interdisciplinary collaborations often required to develop AI systems. This inquiry led to the construction of a design space that captures the important dimensions of tool design and the spirit of the tool in promoting collaborations. Through this work, I argue that tools that better account for and enable interdisciplinary communication and collaborations (such as UX-AI collaborations), *when and as needed*, can help its users (i.e., the practitioners) build human-centered and responsible AI systems. For example, some model training details can be extraneous to a UX professional (example: learning rates, epochs) but some are vital to approaching the design of the product incorporating the model. Information silos through team structures or tool features introduces challenges to downstream processes of designing and developing the AI application.

The second direction of inquiry was related to evaluation approaches that bridge HCI and AI. My objective was to adapt the *Maximum Likelihood Estimation* methods which proved to be useful for comparing user judgements and model outputs in the case of a politeness and a toxicity-detecting classifier [67,106]. To make these methods transfer easily for UX practice, I created interactive, visualization-based tools and tested it to see how UX practitioners

conducted analyses and made inferences (Chapter 5). These visualizations make it easier to understand inferences about how user and model judgements align, which in turn helps the UXP gain insights at scale.

In this discussion section, I connect the threads between the three studies, offer implications for designing collaborative tools based on MLE methods, and conclude with directions for future work.

6.1 How do MLE toolkits help address challenges raised by UX Practitioners in the interview study?

In Chapter 3, we saw UX practitioners bringing up the challenge of reinforcing a user-centered lens when evaluating AI products and tackling output and scale complexities.

> "*One of the challenges I see regularly in my space is we don't have the tools to test the algorithm, the output of the algorithms. Now I can see it has 95% confidence. Right? It is working, but I don't know what people are being shown. I am generalizing it a lot - but we don't always have the tools to see how our model is performing in prod. Because if you are shipping a model that you know, you are affecting 1000, 3000, 10000 people on a daily basis - It's harder to see what the impact is. You either need tools that are going to do it, automated tools which - I don't know how that will be done. Or you have to take the qualitative approach as someone spot checking it.*" - P3 (Chapter 3)

P3 also raises an important point about mixing smaller scale qualitative methods with larger scale quantitative methods. For her own case of evaluating chatbot transcripts, she added, "*And there was one of me looking at 100 transcripts a day. And we were doing tens of thousands of them a day. That is what I mean by scale".* P11 also highlighted the importance of validating

feedback and perspectives at scale ("*How do you more quickly test or get qualitative and quantitative feedback at scale?*") (Chapter 3). Researchers have also raised the need for mixed-methods tools to bridge RAI evaluation efforts across different cross-functional partners [37] and to conduct end-user algorithmic audits [36,37]. Qualitative analyses are also increasingly recognized as vital for AI practitioners to conduct LLM evaluations [7].

The research done in Chapter 5 is a response to these needs and challenges. I designed for communicating MLE results through visualizations in such a way that it would still retain and combine the qualitative aspect (through the Data Viewer - 5.3.2, 5.3.3). Qualitative methods are instrumental to gain a deeper understanding, but UX practitioners also need ways to test insights at scale and get stakeholder buy-in. Future evaluation tools should similarly incorporate features for conducting mixed-method analyses.

Another UX practitioner said the following when asked about what model information she would like to receive,

> "*Maybe just a simple demo, of how the model works if we put in this, this is what we're getting from the model, and then this is the limitation. Also, when we're testing, is there anything that we should be aware of? Sometimes the model doesn't work for all our users. So that is also a constraint that I would want to know, because that would affect the users that I'm recruiting and talking to, and maybe it works better for certain groups of users in this area. So that is something that will be helpful.*" – P9 (Chapter 3)

P9 was concerned about differential impact for end-users and wanted to account for that in her user research sessions. Liao et al. 's work [94] also shows that UX practitioners need to be able to discover and test *impacting factors* that can potentially affect the user experience of an AI system. Only then can they engage in divergent and convergent design stages. That is, the

capabilities of the model can help with ideating different concepts [94]. However, understanding how different factors impact the UX can help select between different concepts or explore conditional designs [94]. The MLE toolkits can help UX practitioners analyze how user judgments of model outputs vary, or how user and model judgements align. They can also inspect how these judgments vary with respect to different sociotechnical factors. In Chapter 5, the formative and evaluation study support Liao et al. 's [94] findings on model information needs for UX practitioners by situating the use of MLE visualizations in design-related tasks.

Similarly, prior work has raised the need to involve end-users in algorithmic auditing efforts as external auditors may be smaller in number and may not share the same background or experiences as the end-users [36]. But a key challenge for UX practitioner is to play the numbers game as most organizations place an emphasis on quantified results. UX practitioners in the interview study also noted challenges in reconciling user-centric outcomes with metrics of model performance. Deng et al. [36] propose leveraging strategies of "tactical quantification" (originally proposed by Irani et al. [72], when presenting Turkopticon) to advocate on behalf of the end-user and get stakeholder buy-in. Though I have not empirically verified it, I believe MLE toolkits have the potential to be used for tactical quantification – a tool through which UX practitioners can advocate for end-users by leveraging numbers (specifically probabilities).

In the formative and qualitative study, I observed that a few practitioners struggled with the visualizations shown. My initial hypothesis was that UX designers struggled more than UX researchers. However, that was not the case. It was rather familiarity with statistics, probability, and whether they came from a technical background. Participants who did not check these boxes requested summaries or takeaways from the visualization before drilling down. Another designer however wanted to be able to *direct what they were looking at*.

Challenge	**Reference**	**MLE Toolkit**
Reinforcing a human-centered lens on AI evaluation	Interview study, Chapter 3	User judgments on AI outputs are the object of analysis.
Tackling scale, while still leveraging qualitative methods	Interview study, Chapter 3 Deng et al. [37]	Supported by the design of the Data Viewer, where UX practitioners could see comments alongside probability curves.
Conducting user-centered algorithmic audits	Deng et al. [36]	When the right factors are selected, the MLE visualizations can also be useful in conducting user-centered algorithmic audits.
Uncovering and testing factors that impact UX	Interview study, Chapter 3 Liao et al. [94]	With MLE methods, factors of interest can be tested by UX practitioners by adding them as explanatory variables.
Understand risks and benefits of different design alternatives	Interview study, Chapter 3 Liao et al. [153] Yildirim et al. [153]9/5/25 11:38:00 AM	UX practitioners simulated design concepts by using probabilities of a user outcome, gleaned from the visualization.

Table 6.1: A summary of how the MLE toolkit helps address different UX/Design challenges raised in Chapter 3 and other prior work.

> "*If it's just about answering questions, yeah, I would be happy with a report. But it's not just answering the questions - It's also intuiting what to look for. It is UX that will have perspectives that other functions don't have. So that makes me want to be involved in this process so that we are thinking about it holistically and not just focused on metrics that other functions are focused on.*" – P3 (Chapter 5)

However, leveraging progressive disclosure and including sensitizing examples and rubrics, can go a long way in helping UX practitioners from different backgrounds find value in the MLE toolkit.

6.2 Revisiting UX-AI Collaborations and The Design Space - The Case of MLE Toolkits

Let us consider the hypothetical use case of AI-generated news summaries to deliberate how practitioner-oriented toolkits can support interdisciplinary collaborations, specifically between UX and AI practitioners.

The first step involves constructing a dataset of news articles, generating AI summaries, and collecting automation and human metrics of the summary's effectiveness. AI practitioners may use human feedback on AI-generated news summaries as a ground truth and measure how correlated these are to automatic metrics (ROGUE, BLEU, BERTScore). Additionally, they might run tests for evaluating performances across different summarization techniques (extractive vs abstractive) and different language models, and for verifying if the summaries are factually grounded. While the underlying data collected can be useful and important for UXP to know, UXP *will focus on different questions* to inform human-centered design of the

application. The scenario used in the study (5.4) provides examples of questions that UXP will primarily be interested in (*What factors influence differences in how AI-generated summaries are rated?)*. How can we provide access to the same data but enable analysis of how various factors impact user feedback? This can be made possible by adopting a groupware design which ensures access to shared information objects. However, because the nature of inquiry is different, we should tailor the interface, interactions, and functionality accordingly [60]. This tailoring should involve a *knowledge recontextualization*, rather than simple, UI-level changes [2,60]. A groupware system that fails to account for an individual or a group's role and skill based requirements and provide commensurate components may also fail to achieve critical mass [60–62]. The other design spirits for collaboration (core practice & communication, community of practice, or visibility & bridging) do not fit here because they do not have provisions for tailoring and personalization in ways that groupware systems have traditionally provided. An interesting line of inquiry here would be to evaluate whether the availability of such MLE methods (and visualizations) addresses communication challenges between UX and AI practitioners and helps build a shared understanding (Chapter 3).

6.3 Conclusion and Future Work

In summary, my dissertation has contributed a qualitative formative study on the different challenges UX practitioners face when designing with AI, a design space of how practitioner tools support different forms of collaboration, and evaluation approaches that bridge HCI and AI. In future work, I hope to continue this line of research. An important limitation of this work is that I focused only on UX practitioners to describe collaborative and organizational challenges from their perspective. I also draw implications for designing collaborative tools but have not included the perspectives of AI practitioners. Doing so is important to uncover their perspectives and any barriers they may face for collaborating with product/UX functions but will be covered in future work. I would also like to expand my focus to study other collaborations

(e.g., domain experts, UX practitioners, AI practitioners / UX practitioners, AI practitioners, policy experts). Such an inquiry would provide more insight into how multiple stakeholders collaborate over AI design and development and help expand the design space.

In this work, I also contributed empirical insights into how UX practitioners interpret, understand, and apply insights from interactive visualizations of MLE results. I hope to design and evaluate other tools that can help UX practitioners adapt foundation models for specific downstream use cases in my future work. In future work, I plan to split my focus between UX designers and researchers. For example, UX researchers might be more interested in how Reinforcement Learning from Human Feedback (RLHF) is increasingly used as technique to ensure human-AI alignment by training the base model on responses selected by the human. Is RLHF the new UX for conversational AI experiences? How might we enable UX practitioners and other domain experts to participate in RLHF and inspect improvements in model behaviors and alignment? These are some questions I hope to tackle as part of future work.

References

1. Abubakar Abid, Ali Abdalla, Ali Abid, Dawood Khan, Abdulrahman Alfozan, and James Zou. 2019. Gradio: Hassle-Free Sharing and Testing of ML Models in the Wild. *arXiv.org*. Retrieved May 19, 2024 from https://arxiv.org/abs/1906.02569v1
2. Mark S. Ackerman, Juri Dachtera, Volkmar Pipek, and Volker Wulf. 2013. Sharing Knowledge and Expertise: The CSCW View of Knowledge Management. *Computer Supported Cooperative Work (CSCW)* 22, 4: 531–573. https://doi.org/10.1007/s10606-013-9192-8
3. Christopher Adolph. 2025. chrisadolph/tile-simcf. Retrieved July 19, 2025 from https://github.com/chrisadolph/tile-simcf
4. Jumana Almahmoud, Robert DeLine, and Steven M. Drucker. 2021. How Teams Communicate about the Quality of ML Models: A Case Study at an International Technology Company. *Proceedings of the ACM on Human-Computer Interaction* 5, GROUP: 222:1-222:24. https://doi.org/10.1145/3463934
5. Saleema Amershi, Andrew Begel, Christian Bird, Robert DeLine, Harald Gall, Ece Kamar, Nachiappan Nagappan, Besmira Nushi, and Thomas Zimmermann. 2019. Software Engineering for Machine Learning: A Case Study. In *2019 IEEE/ACM 41st International Conference on Software Engineering: Software Engineering in Practice (ICSE-SEIP)*, 291–300. https://doi.org/10.1109/ICSE-SEIP.2019.00042
6. 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. 2019. Guidelines for Human-AI Interaction. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (CHI '19), 1–13. https://doi.org/10.1145/3290605.3300233
7. Ian Arawjo. 2025. What AI engineers can learn from qualitative research methods in HCI. *Medium*. Retrieved July 31, 2025 from https://ianarawjo.medium.com/what-ai-engineers-can-learn-from-qualitative-research-methods-in-hci-5b29b9b7465a
8. Ian Arawjo, Chelse Swoopes, Priyan Vaithilingam, Martin Wattenberg, and Elena L. Glassman. 2024. ChainForge: A Visual Toolkit for Prompt Engineering and LLM Hypothesis Testing. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems* (CHI '24), 1–18. https://doi.org/10.1145/3613904.3642016
9. Alex Bäuerle, Ángel Alexander Cabrera, Fred Hohman, Megan Maher, David Koski, Xavier Suau, Titus Barik, and Dominik Moritz. 2022. Symphony: Composing Interactive Interfaces for Machine Learning. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (CHI '22), 1–14. https://doi.org/10.1145/3491102.3502102
10. Michel Beaudouin-Lafon. 2004. Designing interaction, not interfaces. In *Proceedings of the working conference on Advanced visual interfaces* (AVI '04), 15–22. https://doi.org/10.1145/989863.989865
11. Jesse Josua Benjamin, Arne Berger, Nick Merrill, and James Pierce. 2021. Machine Learning Uncertainty as a Design Material: A Post-Phenomenological Inquiry. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, 1–14. https://doi.org/10.1145/3411764.3445481
12. Avinash Bhat, Austin Coursey, Grace Hu, Sixian Li, Nadia Nahar, Shurui Zhou, Christian Kästner, and Jin L.C. Guo. 2023. Aspirations and Practice of ML Model Documentation: Moving the Needle with Nudging and Traceability. In *Proceedings of the 2023 CHI*

Conference on Human Factors in Computing Systems (CHI '23), 1–17. https://doi.org/10.1145/3544548.3581518

13. Jessica Y Bo, Sophia Wan, and Ashton Anderson. 2025. To Rely or Not to Rely? Evaluating Interventions for Appropriate Reliance on Large Language Models. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems* (CHI '25), 1–23. https://doi.org/10.1145/3706598.3714097
14. Tolga Bolukbasi, Kai-Wei Chang, James Zou, Venkatesh Saligrama, and Adam Kalai. 2016. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In *Proceedings of the 30th International Conference on Neural Information Processing Systems* (NIPS'16), 4356–4364.
15. Rishi Bommasani, Drew A. Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S. Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, Erik Brynjolfsson, Shyamal Buch, Dallas Card, Rodrigo Castellon, Niladri Chatterji, Annie Chen, Kathleen Creel, Jared Quincy Davis, Dora Demszky, Chris Donahue, Moussa Doumbouya, Esin Durmus, Stefano Ermon, John Etchemendy, Kawin Ethayarajh, Li Fei-Fei, Chelsea Finn, Trevor Gale, Lauren Gillespie, Karan Goel, Noah Goodman, Shelby Grossman, Neel Guha, Tatsunori Hashimoto, Peter Henderson, John Hewitt, Daniel E. Ho, Jenny Hong, Kyle Hsu, Jing Huang, Thomas Icard, Saahil Jain, Dan Jurafsky, Pratyusha Kalluri, Siddharth Karamcheti, Geoff Keeling, Fereshte Khani, Omar Khattab, Pang Wei Koh, Mark Krass, Ranjay Krishna, Rohith Kuditipudi, Ananya Kumar, Faisal Ladhak, Mina Lee, Tony Lee, Jure Leskovec, Isabelle Levent, Xiang Lisa Li, Xuechen Li, Tengyu Ma, Ali Malik, Christopher D. Manning, Suvir Mirchandani, Eric Mitchell, Zanele Munyikwa, Suraj Nair, Avanika Narayan, Deepak Narayanan, Ben Newman, Allen Nie, Juan Carlos Niebles, Hamed Nilforoshan, Julian Nyarko, Giray Ogut, Laurel Orr, Isabel Papadimitriou, Joon Sung Park, Chris Piech, Eva Portelance, Christopher Potts, Aditi Raghunathan, Rob Reich, Hongyu Ren, Frieda Rong, Yusuf Roohani, Camilo Ruiz, Jack Ryan, Christopher Ré, Dorsa Sadigh, Shiori Sagawa, Keshav Santhanam, Andy Shih, Krishnan Srinivasan, Alex Tamkin, Rohan Taori, Armin W. Thomas, Florian Tramèr, Rose E. Wang, William Wang, Bohan Wu, Jiajun Wu, Yuhuai Wu, Sang Michael Xie, Michihiro Yasunaga, Jiaxuan You, Matei Zaharia, Michael Zhang, Tianyi Zhang, Xikun Zhang, Yuhui Zhang, Lucia Zheng, Kaitlyn Zhou, and Percy Liang. 2022. On the Opportunities and Risks of Foundation Models. Retrieved May 20, 2024 from http://arxiv.org/abs/2108.07258
16. Karen Boyd. 2022. Designing Up with Value-Sensitive Design: Building a Field Guide for Ethical ML Development. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency* (FAccT '22), 2069–2082. https://doi.org/10.1145/3531146.3534626
17. Anders Bruun, Niels Van Berkel, Dimitrios Raptis, and Effie L-C Law. 2025. Coordination Mechanisms in AI Development: Practitioner Experiences on Integrating UX Activities. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, 1–14. https://doi.org/10.1145/3706598.3713200
18. Ángel Alexander Cabrera, Abraham J. Druck, Jason I. Hong, and Adam Perer. 2021. Discovering and Validating AI Errors With Crowdsourced Failure Reports. *Proc. ACM Hum.-Comput. Interact.* 5, CSCW2: 425:1-425:22. https://doi.org/10.1145/3479569
19. Ángel Alexander Cabrera, Erica Fu, Donald Bertucci, Kenneth Holstein, Ameet Talwalkar, Jason I. Hong, and Adam Perer. 2023. Zeno: An Interactive Framework for Behavioral Evaluation of Machine Learning. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (CHI '23), 1–14. https://doi.org/10.1145/3544548.3581268
20. Ángel Alexander Cabrera, Marco Tulio Ribeiro, Bongshin Lee, Robert Deline, Adam Perer, and Steven M. Drucker. 2023. What Did My AI Learn? How Data Scientists Make Sense of Model Behavior. *ACM Trans. Comput.-Hum. Interact.* 30, 1: 1:1-1:27. https://doi.org/10.1145/3542921

21. Shiye Cao, Anqi Liu, and Chien-Ming Huang. 2024. 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: 41:1-41:32. https://doi.org/10.1145/3637318
22. Tara Capel and Margot Brereton. 2023. What is Human-Centered about Human-Centered AI? A Map of the Research Landscape. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, 1–23. https://doi.org/10.1145/3544548.3580959
23. J. M. Carroll and W. A. Kellogg. 1989. Artifact as theory-nexus: hermeneutics meets theory-based design. In *Proceedings of the SIGCHI conference on Human factors in computing systems Wings for the mind - CHI '89*, 7–14. https://doi.org/10.1145/67449.67452
24. Ruijia Cheng, Alison Smith-Renner, Ke Zhang, Joel R. Tetreault, and Alejandro Jaimes. 2022. Mapping the Design Space of Human-AI Interaction in Text Summarization. https://doi.org/10.48550/arXiv.2206.14863
25. Elizabeth F. Churchill. 2020. HCI and UX as translational research. *Interactions* 27, 5: 22–23. https://doi.org/10.1145/3417108
26. Jacob Cohen. 2013. *Statistical power analysis for the behavioral sciences*. routledge. Retrieved July 20, 2025 from https://www.taylorfrancis.com/books/mono/10.4324/9780203771587/statistical-power-analysis-behavioral-sciences-jacob-cohen
27. Lucas Colusso, Cynthia L. Bennett, Gary Hsieh, and Sean A. Munson. 2017. Translational Resources: Reducing the Gap Between Academic Research and HCI Practice. In *Proceedings of the 2017 Conference on Designing Interactive Systems* (DIS '17), 957–968. https://doi.org/10.1145/3064663.3064667
28. Marios Constantinides, John Dowell, David Johnson, and Sylvain Malacria. 2015. Exploring mobile news reading interactions for news app personalisation. In *Proceedings of the 17th International Conference on Human-Computer Interaction with Mobile Devices and Services*, 457–462. https://doi.org/10.1145/2785830.2785860
29. Juliet M. Corbin and Anselm Strauss. 1990. Grounded theory research: Procedures, canons, and evaluative criteria. *Qualitative Sociology* 13, 1: 3–21. https://doi.org/10.1007/BF00988593
30. Anamaria Crisan, Margaret Drouhard, Jesse Vig, and Nazneen Rajani. 2022. Interactive Model Cards: A Human-Centered Approach to Model Documentation. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency* (FAccT '22), 427–439. https://doi.org/10.1145/3531146.3533108
31. Nigel Cross. 1982. Designerly ways of knowing. *DESIGN STUDIES* 3, 4.
32. Nigel Cross. 2001. Design cognition: Results from protocol and other empirical studies of design activity. *Design knowing and learning: Cognition in design education*: 79–103.
33. Peter Cummings. 2009. The Relative Merits of Risk Ratios and Odds Ratios. *Archives of Pediatrics & Adolescent Medicine* 163, 5: 438–445. https://doi.org/10.1001/archpediatrics.2009.31
34. Huw Talfryn Oakley Davies, Iain Kinloch Crombie, and Manouche Tavakoli. 1998. When can odds ratios mislead? *BMJ : British Medical Journal* 316, 7136: 989–991. https://doi.org/10.1136/bmj.316.7136.989
35. Wesley Hanwen Deng, Wang Claire, Howard Ziyu Han, Jason I. Hong, Kenneth Holstein, and Motahhare Eslami. 2025. WeAudit: Scaffolding User Auditors and AI Practitioners in Auditing Generative AI. https://doi.org/10.48550/arXiv.2501.01397
36. Wesley Hanwen Deng, Boyuan Guo, Alicia Devrio, Hong Shen, Motahhare Eslami, and Kenneth Holstein. 2023. Understanding Practices, Challenges, and Opportunities for User-Engaged Algorithm Auditing in Industry Practice. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (CHI '23), 1–18. https://doi.org/10.1145/3544548.3581026

37. Wesley Hanwen Deng, Nur Yildirim, Monica Chang, Motahhare Eslami, Kenneth Holstein, and Michael Madaio. 2023. Investigating Practices and Opportunities for Cross-functional Collaboration around AI Fairness in Industry Practice. In *2023 ACM Conference on Fairness, Accountability, and Transparency*, 705–716. https://doi.org/10.1145/3593013.3594037
38. Gerardine DeSanctis and Marshall Scott Poole. 1994. Capturing the Complexity in Advanced Technology Use: Adaptive Structuration Theory. *Organization Science* 5, 2: 121–147.
39. Advait Deshpande and Helen Sharp. 2022. Responsible AI Systems: Who are the Stakeholders? In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society* (AIES '22), 227–236. https://doi.org/10.1145/3514094.3534187
40. Prasun Dewan. 1999. Architectures for collaborative applications. *Computer Supported Co-operative Work* 7: 169–193.
41. Mark Diaz, Isaac Johnson, Amanda Lazar, Anne Marie Piper, and Darren Gergle. 2018. Addressing Age-Related Bias in Sentiment Analysis. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (CHI '18), 1–14. https://doi.org/10.1145/3173574.3173986
42. Lucas Dixon, John Li, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. 2018. Measuring and Mitigating Unintended Bias in Text Classification. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society* (AIES '18), 67–73. https://doi.org/10.1145/3278721.3278729
43. Paul Dourish. 2003. The Appropriation of Interactive Technologies: Some Lessons from Placeless Documents. *Computer Supported Cooperative Work (CSCW)* 12, 4: 465–490. https://doi.org/10.1023/A:1026149119426
44. Graham Dove, Kim Halskov, Jodi Forlizzi, and John Zimmerman. 2017. UX Design Innovation: Challenges for Working with Machine Learning as a Design Material. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems* (CHI '17), 278–288. https://doi.org/10.1145/3025453.3025739
45. Graham Dove, Nicolai Brodersen Hansen, and Kim Halskov. 2016. An Argument For Design Space Reflection. In *Proceedings of the 9th Nordic Conference on Human-Computer Interaction* (NordiCHI '16), 1–10. https://doi.org/10.1145/2971485.2971528
46. P. Alex Dow, Jennifer Wortman Vaughan, Solon Barocas, Chad Atalla, Alexandra Chouldechova, and Hanna Wallach. 2024. Dimensions of Generative AI Evaluation Design. https://doi.org/10.48550/arXiv.2411.12709
47. Clarence A. Ellis, Simon J. Gibbs, and Gail Rein. 1991. Groupware: some issues and experiences. *Communications of the ACM* 34, 1: 39–58. https://doi.org/10.1145/99977.99987
48. Clarence Ellis and Jacques Wainer. 1994. A conceptual model of groupware. In *Proceedings of the 1994 ACM conference on Computer supported cooperative work* (CSCW '94), 79–88. https://doi.org/10.1145/192844.192878
49. Felix Anand Epp, Anton Poikolainen Rosén, Antti Salovaara, and Camilo Sanchez. 2024. Uncertainties as Generative Resources in Research through Design: Three Dynamics for Moving in a Design Space. *ACM Transactions on Computer-Human Interaction* 31, 6: 1–31. https://doi.org/10.1145/3689041
50. K. J. Kevin Feng, Q. Vera Liao, Ziang Xiao, Jennifer Wortman Vaughan, Amy X. Zhang, and David W. McDonald. 2025. Canvil: Designerly Adaptation for LLM-Powered User Experiences. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems* (CHI '25), 1–22. https://doi.org/10.1145/3706598.3713139
51. K. J. Kevin Feng and David W. Mcdonald. 2023. Addressing UX Practitioners' Challenges in Designing ML Applications: an Interactive Machine Learning Approach. In *Proceedings of the 28th International Conference on Intelligent User Interfaces* (IUI '23), 337–352. https://doi.org/10.1145/3581641.3584064

52. Jules Françoise, Baptiste Caramiaux, and Téo Sanchez. 2021. Marcelle: Composing Interactive Machine Learning Workflows and Interfaces. In *The 34th Annual ACM Symposium on User Interface Software and Technology* (UIST '21), 39–53. https://doi.org/10.1145/3472749.3474734
53. Susan N. Friel, Frances R. Curcio, and George W. Bright. 2001. Making Sense of Graphs: Critical Factors Influencing Comprehension and Instructional Implications. *Journal for Research in Mathematics Education* 32, 2: 124–158. https://doi.org/10.2307/749671
54. William Gaver. 2011. Making spaces: how design workbooks work. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (CHI '11), 1551–1560. https://doi.org/10.1145/1978942.1979169
55. William W. Gaver. 1991. Technology affordances. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (CHI '91), 79–84. https://doi.org/10.1145/108844.108856
56. Fabien Girardin and Neal Lathia. 2017. When user experience designers partner with data scientists. In *2017 AAAI Spring Symposium Series*.
57. Nitesh Goyal, Ian D. Kivlichan, Rachel Rosen, and Lucy Vasserman. 2022. Is Your Toxicity My Toxicity? Exploring the Impact of Rater Identity on Toxicity Annotation. *Proc. ACM Hum.-Comput. Interact.* 6, CSCW2: 363:1-363:28. https://doi.org/10.1145/3555088
58. Colin M. Gray. 2016. "It's More of a Mindset Than a Method": UX Practitioners' Conception of Design Methods. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* (CHI '16), 4044–4055. https://doi.org/10.1145/2858036.2858410
59. Colin M. Gray, Erik Stolterman, and Martin A. Siegel. 2014. Reprioritizing the relationship between HCI research and practice: bubble-up and trickle-down effects. In *Proceedings of the 2014 conference on Designing interactive systems*, 725–734. https://doi.org/10.1145/2598510.2598595
60. Saul Greenberg. 1991. Personalizable groupware: Accommodating individual roles and group differences. In *Proceedings of the Second European Conference on Computer-Supported Cooperative Work ECSCW '91*, Liam Bannon, Mike Robinson and Kjeld Schmidt (eds.). Springer Netherlands, Dordrecht, 17–31. https://doi.org/10.1007/978-94-011-3506-1_2
61. Jonathan Grudin. 1988. Why CSCW applications fail: problems in the design and evaluationof organizational interfaces. In *Proceedings of the 1988 ACM conference on Computer-supported cooperative work - CSCW '88*, 85–93. https://doi.org/10.1145/62266.62273
62. Jonathan Grudin. 1994. Groupware and social dynamics: eight challenges for developers. *Communications of the ACM* 37, 1: 92–105. https://doi.org/10.1145/175222.175230
63. Laura J. Gurak and Nancy L. Bayer. 1994. Making gender visible: Extending feminist critiques of technology to technical communication. *Technical Communication Quarterly* 3, 3: 257. https://doi.org/10.1080/10572259409364571
64. Kim Halskov and Caroline Lundqvist. 2021. Filtering and Informing the Design Space: Towards Design-Space Thinking. *ACM Transactions on Computer-Human Interaction* 28, 1: 1–28. https://doi.org/10.1145/3434462
65. Emma Harvey, Emily Sheng, Su Lin Blodgett, Alexandra Chouldechova, Jean Garcia-Gathright, Alexandra Olteanu, and Hanna Wallach. 2024. Gaps Between Research and Practice When Measuring Representational Harms Caused by LLM-Based Systems. https://doi.org/10.48550/arXiv.2411.15662
66. Amy K. Heger, Liz B. Marquis, Mihaela Vorvoreanu, Hanna Wallach, and Jennifer Wortman Vaughan. 2022. Understanding Machine Learning Practitioners' Data Documentation Perceptions, Needs, Challenges, and Desiderata. *Proceedings of the ACM on Human-Computer Interaction* 6, CSCW2: 340:1-340:29. https://doi.org/10.1145/3555760

67. Erin R. Hoffman, David W. McDonald, and Mark Zachry. 2017. Evaluating a computational approach to labeling politeness: Challenges for the application of machine classification to social computing data. *Proceedings of the ACM on Human-computer Interaction* 1, CSCW: 1–14.
68. Lars Erik Holmquist. 2017. Intelligence on tap: artificial intelligence as a new design material. *Interactions* 24, 4: 28–33. https://doi.org/10.1145/3085571
69. Matthew K. Hong, Adam Fourney, Derek DeBellis, and Saleema Amershi. 2021. Planning for Natural Language Failures with the AI Playbook. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (CHI '21), 1–11. https://doi.org/10.1145/3411764.3445735
70. Sungsoo Ray Hong, Jessica Hullman, and Enrico Bertini. 2020. Human Factors in Model Interpretability: Industry Practices, Challenges, and Needs. *Proceedings of the ACM on Human-Computer Interaction* 4, CSCW1: 1–26. https://doi.org/10.1145/3392878
71. Jakob Hummes and Bernard Merialdo. 2000. Design of Extensible Component-Based Groupware. *Computer Supported Cooperative Work (CSCW)* 9, 1: 53–74. https://doi.org/10.1023/A:1008761709799
72. Lilly C. Irani and M. Six Silberman. 2013. Turkopticon: interrupting worker invisibility in amazon mechanical turk. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (CHI '13), 611–620. https://doi.org/10.1145/2470654.2470742
73. Susan Jamieson. 2004. Likert scales: how to (ab)use them. *Medical Education* 38, 12: 1217–1218. https://doi.org/10.1111/j.1365-2929.2004.02012.x
74. Anniek Jansen and Sara Colombo. 2023. Mix & Match Machine Learning: An Ideation Toolkit to Design Machine Learning-Enabled Solutions. In *Proceedings of the Seventeenth International Conference on Tangible, Embedded, and Embodied Interaction* (TEI '23), 1–18. https://doi.org/10.1145/3569009.3572739
75. Akshita Jha and Radhika Mamidi. 2017. When does a compliment become sexist? Analysis and classification of ambivalent sexism using twitter data. In *Proceedings of the Second Workshop on NLP and Computational Social Science*, 7–16. https://doi.org/10.18653/v1/W17-2902
76. Ellen Jiang, Kristen Olson, Edwin Toh, Alejandra Molina, Aaron Donsbach, Michael Terry, and Carrie J Cai. 2022. PromptMaker: Prompt-based Prototyping with Large Language Models. In *Extended Abstracts of the 2022 CHI Conference on Human Factors in Computing Systems* (CHI EA '22), 1–8. https://doi.org/10.1145/3491101.3503564
77. Robert Johansen. 1988. Current user approaches to groupware. *Groupware: Computer support for business teams*: 12–44.
78. Eunice Jun, Maureen Daum, Jared Roesch, Sarah Chasins, Emery Berger, Rene Just, and Katharina Reinecke. 2019. Tea: A High-level Language and Runtime System for Automating Statistical Analysis. In *Proceedings of the 32nd Annual ACM Symposium on User Interface Software and Technology*, 591–603. https://doi.org/10.1145/3332165.3347940
79. Alex Kale, Matthew Kay, and Jessica Hullman. 2020. Visual Reasoning Strategies for Effect Size Judgments and Decisions. https://doi.org/10.48550/arXiv.2007.14516
80. Maurits Kaptein and Judy Robertson. 2012. Rethinking statistical analysis methods for CHI. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (CHI '12), 1105–1114. https://doi.org/10.1145/2207676.2208557
81. Claire Kayacik, Sherol Chen, Signe Noerly, Jess Holbrook, Adam Roberts, and Douglas Eck. 2019. Identifying the Intersections: User Experience + Research Scientist Collaboration in a Generative Machine Learning Interface. In *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems*, 1–8. https://doi.org/10.1145/3290607.3299059
82. Sunnie S. Y. Kim, Jennifer Wortman Vaughan, Q. Vera Liao, Tania Lombrozo, and Olga Russakovsky. 2025. Fostering Appropriate Reliance on Large Language Models: The Role of

Explanations, Sources, and Inconsistencies. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems* (CHI '25), 1–19. https://doi.org/10.1145/3706598.3714020

83. Tzu-Sheng Kuo, Aaron Lee Halfaker, Zirui Cheng, Jiwoo Kim, Meng-Hsin Wu, Tongshuang Wu, Kenneth Holstein, and Haiyi Zhu. 2024. Wikibench: Community-Driven Data Curation for AI Evaluation on Wikipedia. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems* (CHI '24), 1–24. https://doi.org/10.1145/3613904.3642278
84. Vivian Lai, Samuel Carton, Rajat Bhatnagar, Q. Vera Liao, Yunfeng Zhang, and Chenhao Tan. 2022. Human-AI Collaboration via Conditional Delegation: A Case Study of Content Moderation. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (CHI '22), 1–18. https://doi.org/10.1145/3491102.3501999
85. Vivian Lai, Chacha Chen, Alison Smith-Renner, Q. Vera Liao, and Chenhao Tan. 2023. Towards a Science of Human-AI Decision Making: An Overview of Design Space in Empirical Human-Subject Studies. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency* (FAccT '23), 1369–1385. https://doi.org/10.1145/3593013.3594087
86. Michelle S. Lam, Mitchell L. Gordon, Danaë Metaxa, Jeffrey T. Hancock, James A. Landay, and Michael S. Bernstein. 2022. End-User Audits: A System Empowering Communities to Lead Large-Scale Investigations of Harmful Algorithmic Behavior. *Proc. ACM Hum.-Comput. Interact.* 6, CSCW2: 512:1-512:34. https://doi.org/10.1145/3555625
87. Philip A. Laplante. 2007. *What Every Engineer Should Know about Software Engineering*. CRC Press.
88. Charlotte P. Lee. 2007. Boundary negotiating artifacts: Unbinding the routine of boundary objects and embracing chaos in collaborative work. *Computer Supported Cooperative Work (CSCW)* 16: 307–339.
89. Charlotte P. Lee and Drew Paine. 2015. From The Matrix to a Model of Coordinated Action (MoCA): A Conceptual Framework of and for CSCW. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing* (CSCW '15), 179–194. https://doi.org/10.1145/2675133.2675161
90. Hao-Ping (Hank) Lee, Advait Sarkar, Lev Tankelevitch, Ian Drosos, Sean Rintel, Richard Banks, and Nicholas Wilson. 2025. The Impact of Generative AI on Critical Thinking: Self-Reported Reductions in Cognitive Effort and Confidence Effects From a Survey of Knowledge Workers. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems* (CHI '25), 1–22. https://doi.org/10.1145/3706598.3713778
91. Michelle Seng Ah Lee and Jat Singh. 2021. The Landscape and Gaps in Open Source Fairness Toolkits. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (CHI '21), 1–13. https://doi.org/10.1145/3411764.3445261
92. Mina Lee, Katy Ilonka Gero, John Joon Young Chung, Simon Buckingham Shum, Vipul Raheja, Hua Shen, Subhashini Venugopalan, Thiemo Wambsganss, David Zhou, Emad A. Alghamdi, Tal August, Avinash Bhat, Madiha Zahrah Choksi, Senjuti Dutta, Jin L.C. Guo, Md Naimul Hoque, Yewon Kim, Simon Knight, Seyed Parsa Neshaei, Antonette Shibani, Disha Shrivastava, Lila Shroff, Agnia Sergeyuk, Jessi Stark, Sarah Sterman, Sitong Wang, Antoine Bosselut, Daniel Buschek, Joseph Chee Chang, Sherol Chen, Max Kreminski, Joonsuk Park, Roy Pea, Eugenia Ha Rim Rho, Zejiang Shen, and Pao Siangliulue. 2024. A Design Space for Intelligent and Interactive Writing Assistants. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems* (CHI '24), 1–35. https://doi.org/10.1145/3613904.3642697
93. Q. Vera Liao, Daniel Gruen, and Sarah Miller. 2020. Questioning the AI: Informing Design Practices for Explainable AI User Experiences. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (CHI '20), 1–15. https://doi.org/10.1145/3313831.3376590

94. Q. Vera Liao, Hariharan Subramonyam, Jennifer Wang, and Jennifer Wortman Vaughan. 2023. Designerly Understanding: Information Needs for Model Transparency to Support Design Ideation for AI-Powered User Experience. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, 1–21. https://doi.org/10.1145/3544548.3580652
95. Q. Vera Liao, Mihaela Vorvoreanu, Hari Subramonyam, and Lauren Wilcox. 2024. UX Matters: The Critical Role of UX in Responsible AI. *Interactions* 31, 4: 22–27. https://doi.org/10.1145/3665504
96. Q. Vera Liao and Ziang Xiao. 2025. Rethinking Model Evaluation as Narrowing the Socio-Technical Gap. https://doi.org/10.48550/arXiv.2306.03100
97. Hongjin Lin, Naveena Karusala, Chinasa T. Okolo, Catherine D'Ignazio, and Krzysztof Z. Gajos. 2024. "Come to us first": Centering Community Organizations in Artificial Intelligence for Social Good Partnerships. *Proc. ACM Hum.-Comput. Interact.* 8, CSCW2: 470:1-470:28. https://doi.org/10.1145/3687009
98. Fang Liu, Junyan Lv, Shenglan Cui, Zhilong Luan, Kui Wu, and Tongqing Zhou. 2024. Smart "Error"! Exploring Imperfect AI to Support Creative Ideation. *Proc. ACM Hum.-Comput. Interact.* 8, CSCW1: 121:1-121:28. https://doi.org/10.1145/3637398
99. Allan MacLean, Richard Young, Victoria Bellotti, and Thomas Moran. 1991. Questions, Options, and Criteria: Elements of Design Space Analysis. *Human-Computer Interaction* 6: 201–250. https://doi.org/10.1080/07370024.1991.9667168
100. Michael Madaio, Shivani Kapania, Rida Qadri, Ding Wang, Andrew Zaldivar, Remi Denton, and Lauren Wilcox. 2024. Learning about Responsible AI On-The-Job: Learning Pathways, Orientations, and Aspirations. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency* (FAccT '24), 1544–1558. https://doi.org/10.1145/3630106.3658988
101. Tim Miller. 2019. Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence* 267: 1–38. https://doi.org/10.1016/j.artint.2018.07.007
102. Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. 2019. Model Cards for Model Reporting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency* (FAT* '19), 220–229. https://doi.org/10.1145/3287560.3287596
103. Naomi Miyake and Donald A. Norman. 1979. To ask a question, one must know enough to know what is not known. *Journal of Verbal Learning and Verbal Behavior* 18, 3: 357–364. https://doi.org/10.1016/S0022-5371(79)90200-7
104. Steven Moore, Q. Vera Liao, and Hariharan Subramonyam. 2023. fAIlureNotes: Supporting Designers in Understanding the Limits of AI Models for Computer Vision Tasks. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (CHI '23), 1–19. https://doi.org/10.1145/3544548.3581242
105. Meredith Ringel Morris, Carrie J. Cai, Jess Holbrook, Chinmay Kulkarni, and Michael Terry. 2023. The Design Space of Generative Models. https://doi.org/10.48550/arXiv.2304.10547
106. Meena Devii Muralikumar, Yun Shan Yang, and David W. McDonald. 2023. A Human-centered Evaluation of a Toxicity Detection API: Testing Transferability and Unpacking Latent Attributes. *ACM Transactions on Social Computing* 6, 1–2: 4:1-4:38. https://doi.org/10.1145/3582568
107. Nadia Nahar, Haoran Zhang, Grace Lewis, Shurui Zhou, and Christian Kästner. 2023. A Meta-Summary of Challenges in Building Products with ML Components -- Collecting Experiences from 4758+ Practitioners. https://doi.org/10.48550/arXiv.2304.00078
108. Nadia Nahar, Shurui Zhou, Grace Lewis, and Christian Kästner. 2022. Collaboration challenges in building ML-enabled systems: communication, documentation, engineering,

and process. In *Proceedings of the 44th International Conference on Software Engineering* (ICSE '22), 413–425. https://doi.org/10.1145/3510003.3510209

109. Wanda J. Orlikowski. 1992. The Duality of Technology: Rethinking the Concept of Technology in Organizations. *Organization Science* 3, 3: 398–427.
110. Wanda J. Orlikowski and Debra C. Gash. 1994. Technological frames: making sense of information technology in organizations. *ACM Transactions on Information Systems* 12, 2: 174–207. https://doi.org/10.1145/196734.196745
111. Anna-Marie Ortloff, Julia Angelika Grohs, Simon Lenau, and Matthew Smith. 2025. A Qualitative Study on How Usable Security and HCI Researchers Judge the Size and Importance of Odds Ratio and Cohen's d Effect Sizes. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems* (CHI '25), 1–16. https://doi.org/10.1145/3706598.3714022
112. Anna-Marie Ortloff, Florin Martius, Mischa Meier, Theo Raimbault, Lisa Geierhaas, and Matthew Smith. 2025. Small, Medium, Large? A Meta-Study of Effect Sizes at CHI to Aid Interpretation of Effect Sizes and Power Calculation. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems* (CHI '25), 1–28. https://doi.org/10.1145/3706598.3713671
113. Antti Oulasvirta and Kasper Hornbæk. 2022. Counterfactual Thinking: What Theories *Do* in Design. *International Journal of Human–Computer Interaction* 38, 1: 78–92. https://doi.org/10.1080/10447318.2021.1925436
114. Lawrence A. Palinkas, Sarah M. Horwitz, Carla A. Green, Jennifer P. Wisdom, Naihua Duan, and Kimberly Hoagwood. 2015. Purposeful sampling for qualitative data collection and analysis in mixed method implementation research. *Administration and policy in mental health* 42, 5: 533–544. https://doi.org/10.1007/s10488-013-0528-y
115. Soya Park, April Yi Wang, Ban Kawas, Q. Vera Liao, David Piorkowski, and Marina Danilevsky. 2021. Facilitating Knowledge Sharing from Domain Experts to Data Scientists for Building NLP Models. In *26th International Conference on Intelligent User Interfaces*, 585–596. https://doi.org/10.1145/3397481.3450637
116. Savvas Petridis, Michael Terry, and Carrie J Cai. 2024. PromptInfuser: How Tightly Coupling AI and UI Design Impacts Designers' Workflows. In *Proceedings of the 2024 ACM Designing Interactive Systems Conference* (DIS '24), 743–756. https://doi.org/10.1145/3643834.3661613
117. David Piorkowski, Inge Vejsbjerg, Owen Cornec, Elizabeth M. Daly, and Öznur Alkan. 2023. AIMEE: An Exploratory Study of How Rules Support AI Developers to Explain and Edit Models. *Proceedings of the ACM on Human-Computer Interaction* 7, CSCW2: 1–25. https://doi.org/10.1145/3610046
118. Snehal Prabhudesai, Leyao Yang, Sumit Asthana, Xun Huan, Q. Vera Liao, and Nikola Banovic. 2023. Understanding Uncertainty: How Lay Decision-makers Perceive and Interpret Uncertainty in Human-AI Decision Making. In *Proceedings of the 28th International Conference on Intelligent User Interfaces* (IUI '23), 379–396. https://doi.org/10.1145/3581641.3584033
119. Inioluwa Deborah Raji and Joy Buolamwini. 2019. Actionable Auditing: Investigating the Impact of Publicly Naming Biased Performance Results of Commercial AI Products. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society* (AIES '19), 429–435. https://doi.org/10.1145/3306618.3314244
120. Inioluwa Deborah Raji, Peggy Xu, Colleen Honigsberg, and Daniel Ho. 2022. Outsider Oversight: Designing a Third Party Audit Ecosystem for AI Governance. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society* (AIES '22), 557–571. https://doi.org/10.1145/3514094.3534181
121. Judy Robertson. 2012. Likert-type scales, statistical methods, and effect sizes. *Communications of the ACM* 55, 5: 6–7. https://doi.org/10.1145/2160718.2160721

122. Samantha Robertson, Zijie J. Wang, Dominik Moritz, Mary Beth Kery, and Fred Hohman. 2023. Angler: Helping Machine Translation Practitioners Prioritize Model Improvements. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, 1–20. https://doi.org/10.1145/3544548.3580790
123. Mattia Samory, Indira Sen, Julian Kohne, Fabian Flöck, and Claudia Wagner. 2021. "Call me sexist, but..." : Revisiting Sexism Detection Using Psychological Scales and Adversarial Samples. *Proceedings of the International AAAI Conference on Web and Social Media* 15: 573–584. https://doi.org/10.1609/icwsm.v15i1.18085
124. Schlegel Benjamin E. 2024. glm.predict: Predicted Values and Discrete Changes for Regression Models. Retrieved from https://cran.r-project.org/package=glm.predict
125. Kjeld Schmidt and Carla Simonee. 1996. Coordination mechanisms: Towards a conceptual foundation of CSCW systems design. *Computer Supported Cooperative Work (CSCW)* 5, 2: 155–200. https://doi.org/10.1007/BF00133655
126. Donald A. Schön. 1987. *Educating the reflective practitioner: Toward a new design for teaching and learning in the professions*. Jossey-Bass, San Francisco, CA, US.
127. Roy Schwartz, Jesse Dodge, Noah A. Smith, and Oren Etzioni. 2020. Green AI. *Communications of the ACM* 63, 12: 54–63. https://doi.org/10.1145/3381831
128. Nikhil Sharma and George Furnas. 2009. Artifact usefulness and usage in sensemaking handoffs. *Proceedings of the American Society for Information Science and Technology* 46, 1: 1–19. https://doi.org/10.1002/meet.2009.1450460219
129. Mary Shaw. 2012. The Role of Design Spaces. *IEEE Software* 29, 1: 46–50. https://doi.org/10.1109/MS.2011.121
130. Herbert A. Simon. 2019. *The Sciences of the Artificial, reissue of the third edition with a new introduction by John Laird*. MIT Press.
131. Petr Slovák, Christopher Frauenberger, and Geraldine Fitzpatrick. 2017. Reflective Practicum: A Framework of Sensitising Concepts to Design for Transformative Reflection. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, 2696–2707. https://doi.org/10.1145/3025453.3025516
132. Susan Leigh Star and James R. Griesemer. 1989. 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.
133. Erik Stolterman. 2008. The Nature of Design Practice and Implications for Interaction Design Research.
134. Erik Stolterman and James Pierce. 2012. Design tools in practice: studying the designer-tool relationship in interaction design. In *Proceedings of the Designing Interactive Systems Conference*, 25–28. https://doi.org/10.1145/2317956.2317961
135. Hariharan Subramonyam, Jane Im, Colleen Seifert, and Eytan Adar. 2022. Solving Separation-of-Concerns Problems in Collaborative Design of Human-AI Systems through Leaky Abstractions. In *CHI Conference on Human Factors in Computing Systems*, 1–21. https://doi.org/10.1145/3491102.3517537
136. Hariharan Subramonyam, Colleen Seifert, and Eytan Adar. 2021. Towards a process model for co-creating AI experiences. In *Designing Interactive Systems Conference 2021*, 1529–1543.
137. Hariharan Subramonyam, Colleen Seifert, and Eytan Adar. 2021. ProtoAI: Model-Informed Prototyping for AI-Powered Interfaces. In *26th International Conference on Intelligent User Interfaces* (IUI '21), 48–58. https://doi.org/10.1145/3397481.3450640
138. Hanna Wallach, Meera Desai, A. Feder Cooper, Angelina Wang, Chad Atalla, Solon Barocas, Su Lin Blodgett, Alexandra Chouldechova, Emily Corvi, P. Alex Dow, Jean Garcia-Gathright, Alexandra Olteanu, Nicholas Pangakis, Stefanie Reed, Emily Sheng, Dan Vann, Jennifer Wortman Vaughan, Matthew Vogel, Hannah Washington, and Abigail Z. Jacobs.

2025. Position: Evaluating Generative AI Systems Is a Social Science Measurement Challenge. https://doi.org/10.48550/arXiv.2502.00561
139. Michael D. Ward and John S. Ahlquist. 2018. *Maximum Likelihood for Social Science: Strategies for Analysis*. Cambridge University Press.
140. Zeerak Waseem and Dirk Hovy. 2016. Hateful Symbols or Hateful People? Predictive Features for Hate Speech Detection on Twitter. In *Proceedings of the NAACL Student Research Workshop*, 88–93. https://doi.org/10.18653/v1/N16-2013
141. Laura Weidinger, Inioluwa Deborah Raji, Hanna Wallach, Margaret Mitchell, Angelina Wang, Olawale Salaudeen, Rishi Bommasani, Deep Ganguli, Sanmi Koyejo, and William Isaac. 2025. Toward an Evaluation Science for Generative AI Systems. https://doi.org/10.48550/arXiv.2503.05336
142. Laura Weidinger, Maribeth Rauh, Nahema Marchal, Arianna Manzini, Lisa Anne Hendricks, Juan Mateos-Garcia, Stevie Bergman, Jackie Kay, Conor Griffin, Ben Bariach, Iason Gabriel, Verena Rieser, and William Isaac. 2023. Sociotechnical Safety Evaluation of Generative AI Systems. https://doi.org/10.48550/arXiv.2310.11986
143. Etienne Wenger. 1998. Communities of practice: Learning as a social system. *Systems thinker* 9, 5: 2–3.
144. Maximiliane Windl, Sebastian S. Feger, Lara Zijlstra, Albrecht Schmidt, and Pawel W. Wozniak. 2022. 'It Is Not Always Discovery Time': Four Pragmatic Approaches in Designing AI Systems. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (CHI '22), 1–12. https://doi.org/10.1145/3491102.3501943
145. Richmond Y. Wong, Michael A. Madaio, and Nick Merrill. 2023. Seeing Like a Toolkit: How Toolkits Envision the Work of AI Ethics. *Proceedings of the ACM on Human-Computer Interaction* 7, CSCW1: 1–27. https://doi.org/10.1145/3579621
146. Ellery Wulczyn, Nithum Thain, and Lucas Dixon. 2017. Ex Machina: Personal Attacks Seen at Scale. In *Proceedings of the 26th International Conference on World Wide Web* (WWW '17), 1391–1399. https://doi.org/10.1145/3038912.3052591
147. Qian Yang, Alex Scuito, John Zimmerman, Jodi Forlizzi, and Aaron Steinfeld. 2018. Investigating How Experienced UX Designers Effectively Work with Machine Learning. In *Proceedings of the 2018 Designing Interactive Systems Conference* (DIS '18), 585–596. https://doi.org/10.1145/3196709.3196730
148. Qian Yang, Aaron Steinfeld, Carolyn Rosé, and John Zimmerman. 2020. Re-examining whether, why, and how human-AI interaction is uniquely difficult to design. In *Proceedings of the 2020 chi conference on human factors in computing systems*, 1–13.
149. Qian Yang, Aaron Steinfeld, Carolyn Rosé, and John Zimmerman. 2020. Re-examining Whether, Why, and How Human-AI Interaction Is Uniquely Difficult to Design. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, 1–13. https://doi.org/10.1145/3313831.3376301
150. Runlong Ye, Matthew Varona, Oliver Huang, Patrick Yung Kang Lee, Michael Liut, and Carolina Nobre. 2025. The Design Space of Recent AI-assisted Research Tools for Ideation, Sensemaking, and Scientific Creativity. https://doi.org/10.48550/arXiv.2502.16291
151. Zining Ye, Xinran Yuan, Shaurya Gaur, Aaron Halfaker, Jodi Forlizzi, and Haiyi Zhu. 2021. Wikipedia ORES Explorer: Visualizing Trade-offs For Designing Applications With Machine Learning API. In *Designing Interactive Systems Conference 2021*, 1554–1565.
152. Nur Yildirim, Alex Kass, Teresa Tung, Connor Upton, Donnacha Costello, Robert Giusti, Sinem Lacin, Sara Lovic, James M O'Neill, Rudi O'Reilly Meehan, Eoin Ó Loideáin, Azzurra Pini, Medb Corcoran, Jeremiah Hayes, Diarmuid J Cahalane, Gaurav Shivhare, Luigi Castoro, Giovanni Caruso, Changhoon Oh, James McCann, Jodi Forlizzi, and John Zimmerman. 2022. How Experienced Designers of Enterprise Applications Engage AI as a Design Material. In *CHI Conference on Human Factors in Computing Systems*, 1–13. https://doi.org/10.1145/3491102.3517491

153. Nur Yildirim, Changhoon Oh, Deniz Sayar, Kayla Brand, Supritha Challa, Violet Turri, Nina Crosby Walton, Anna Elise Wong, Jodi Forlizzi, James McCann, and John Zimmerman. 2023. Creating Design Resources to Scaffold the Ideation of AI Concepts. In *Proceedings of the 2023 ACM Designing Interactive Systems Conference* (DIS '23), 2326–2346. https://doi.org/10.1145/3563657.3596058
154. Bowen Yu, Ye Yuan, Loren Terveen, Zhiwei Steven Wu, Jodi Forlizzi, and Haiyi Zhu. 2020. Keeping Designers in the Loop: Communicating Inherent Algorithmic Trade-offs Across Multiple Objectives. In *Proceedings of the 2020 ACM Designing Interactive Systems Conference* (DIS '20), 1245–1257. https://doi.org/10.1145/3357236.3395528
155. Sabah Zdanowska and Alex S Taylor. 2022. A study of UX practitioners roles in designing real-world, enterprise ML systems. In *CHI Conference on Human Factors in Computing Systems*, 1–15. https://doi.org/10.1145/3491102.3517607
156. Amy X. Zhang, Michael S. Bernstein, David R. Karger, and Mark S. Ackerman. 2024. Form-From: A Design Space of Social Media Systems. *Proc. ACM Hum.-Comput. Interact.* 8, CSCW1: 167:1-167:47. https://doi.org/10.1145/3641006
157. Tianyi Zhang, Faisal Ladhak, Esin Durmus, Percy Liang, Kathleen McKeown, and Tatsunori B. Hashimoto. 2024. Benchmarking Large Language Models for News Summarization. *Transactions of the Association for Computational Linguistics* 12: 39–57. https://doi.org/10.1162/tacl_a_00632
158. 2024. NIST Launches ARIA, a New Program to Advance Sociotechnical Testing and Evaluation for AI. *NIST*. Retrieved July 31, 2025 from https://www.nist.gov/news-events/news/2024/05/nist-launches-aria-new-program-advance-sociotechnical-testing-and
159. Machine Bias — ProPublica. Retrieved May 8, 2024 from https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing
160. People + AI Guidebook. Retrieved June 29, 2023 from https://pair.withgoogle.com/guidebook
161. Responsible AI: Ethical policies and practices | Microsoft AI. Retrieved July 29, 2025 from https://www.microsoft.com/en-us/ai/responsible-ai
162. Google AI - AI Principles. Retrieved July 29, 2025 from https://ai.googleprinciples/
163. ACM FAccT - 2018 Information for Press. Retrieved July 29, 2025 from https://facctconference.org/2018/press_release
164. ACM Global Technology Policy Council Releases Joint Statement on Principles for Responsible Algorithmic Systems by US and Europe Policy Committees. Retrieved April 15, 2024 from https://www.acm.org/articles/bulletins/2022/november/tpc-statement-responsible-algorithmic-systems
165. Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. Retrieved July 29, 2025 from https://proceedings.mlr.press/v81/buolamwini18a.html?mod=article_inline&ref=akusion-ci-shi-dai-bizinesumedeia
166. The Reflective Practitioner | How Professionals Think in Action | Dona. Retrieved January 24, 2024 from https://www.taylorfrancis.com/books/mono/10.4324/9781315237473/reflective-practitioner-donald-sch%C3%B6n
167. Grasping AI: experiential exercises for designers | AI & SOCIETY. Retrieved August 1, 2025 from https://link.springer.com/article/10.1007/s00146-023-01794-y
168. ModelLens: An Interactive System to Support the Model Improvement Practices of Data Science Teams | Companion Publication of the 2019 Conference on Computer Supported Cooperative Work and Social Computing. Retrieved July 26, 2025 from https://dl.acm.org/doi/10.1145/3311957.3359512

169. A.I. Has a Measurement Problem - The New York Times. Retrieved July 26, 2025 from https://www.nytimes.com/2024/04/15/technology/ai-models-measurement.html
170. Perspective | Developers. Retrieved July 19, 2025 from https://developers.perspectiveapi.com/s/?language=en_US
171. Relative risk ratios and odds ratios. *free range statistics*. Retrieved July 20, 2025 from https://freerangestats.info/blog/2018/08/17/risk-ratios.html
172. Parting Crowds | Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing. Retrieved September 16, 2024 from https://dl.acm.org/doi/abs/10.1145/2818048.2820016
173. Truth Is a Lie: Crowd Truth and the Seven Myths of Human Annotation | AI Magazine. Retrieved July 21, 2025 from https://ojs.aaai.org/aimagazine/index.php/aimagazine/article/view/2564
174. Counterfactuals | Wiley. *Wiley.com*. Retrieved July 20, 2025 from https://www.wiley.com/en-us/Counterfactuals-p-9780631224259