Profiling Artificial Intelligence as a Material for User Experience Design

Qian Yang

CMU-HCII-20-100
July 2020

Human-Computer Interaction Institute
School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213

Thesis Committee:
John Zimmerman (Chair)
Aaron Steinfeld (Co-Chair)
Carolyn Rosé
Saleema Amershi (Microsoft Research AI)

Submitted in partial fulfillment of the requirements

for the degree of Doctor of Philosophy.

This research was sponsored by the National Heart, Lung, and Blood Institute (NIH NHLBI 1R01HL122639-01A1), the National Institute on Disability, Independent Living, and Rehabilitation Research (NIDILRR 90REGE0007 and 90RE5011), and the National Science Foundation (SES-1734456). The author was also supported by the Center for Machine Learning and Health (CMLH) Fellowships in Digital Health and the 2019 Microsoft Research Dissertation Grant. The views and conclusions contained in this document are those of the author and should not be interpreted as representing the official policies, either expressed or implied, of any sponsoring institution, the U.S. government or any other entity.

Keywords: Human-AI Interaction, User Experience Design, Machine Learning, Clinical Decision Support Systems, Natural Language Generation

南大给我的爱

To my parents, with love

Abstract

From predictive medicine to autonomous driving, advances in Artificial Intelligence (AI) promise to improve people's lives and improve society. As systems that utilize these advances increasingly migrated from research labs into the real world, new challenges emerged. For example, when and how should predictive models fit into physicians' decision-making workflow such that the predictions impact them *appropriately*? These are challenges of translation: translating AI systems from systems that demonstrate remarkable technological achievements into real-world, socio-technical systems that serve human ends. My research focuses on this critical translation; on the user experience (UX) design of AI systems.

The prevalence of AI suggests that the UX design community has effective design methods and tools to excel in this translation. While this is true in many cases, some challenges persist. For example, designers struggle with accounting for AI systems' unpredictable errors, and these errors damage UX and even lead to undesirable societal impacts. UX designers routinely grapple with technologies' unanticipated technical or human failures, with a focus on mitigating technologies unintended consequences. What makes AI different from other interactive technologies? – A critical first step in systematically addressing the UX design challenges of AI systems is to articulate what makes these systems so difficult to design in the first place.

This dissertation delineates *whether*, *when*, and *how* UX of AI systems is uniquely difficult to design. I synthesize prior UX and AI research, my own experience designing human-AI interactions, my studies of experienced AI innovation teams in the industry, and my observations from teaching human-AI interaction. I trace the nebulous UX design challenges of AI back to just two root challenges: uncertainty around AI systems' capabilities and the complexity of what systems might output. I present a framework that unravels their effects on design processes; namely AI systems' "*design complexity framework*". Using the framework, I identify four levels of AI systems. On each level, designers are likely to encounter a different subset of design challenges: Current design methods are most effective in eliciting,

addressing, and evaluating the UX issues of Level 1 systems (probabilistic systems, systems with known capability with few possible outputs); Current methods are least effective for Level 4 systems (evolving, adaptive systems, systems that can learn from new data post-deployment and can produce complex outputs that resist abstraction or simulation). Level 2 and 3 are two intermediate levels.

I further demonstrate the usefulness of this framework for UX research and practice through two case studies. In both cases, I engaged stakeholders in their real-world contexts and addressed a critical challenge in fitting cutting-edge AI systems into people's everyday lives. The first is the design of a clinical decision-support system that can effectively collaborate with doctors in making life-and-death treatment decisions. It exemplifies Level 1 systems. The second project is an investigation of how Natural Language Generation systems might seamlessly serve the authors' communicative intent. This illustrates Level 4 systems. It reveals the limits of UX design methods and processes widely in use today. By teasing apart the challenges of routine UX design and those distinctively needed for AI systems, the framework helps UX researchers and design tool makers to address AI systems' design challenges in a targeted fashion.

Acknowledgments

This dissertation owes its development to so many people.

First, I would like to thank my advisors, John Zimmerman and Aaron Steinfeld. You were the first to bravely taking me on, and you have been my mentors, my confidants, my friends, and a never-ending source of emotional support since. I could never have done any of this, the research and writing that went into this dissertation, without the opportunities, lessons, and encouragement you gave me. For that, I am deeply grateful.

I would also like to thank the rest of my thesis committee for their support. Carolyn Rosé and Saleema Amershi have generously provided me with invaluable advice and comments on both my research projects and the many versions of this dissertation. Many ideas in this work, particularly in the first three chapters, had their origins in our conversations.

I have also been fortunate to have mentors and collaborators within and outside of Carnegie Mellon University. Jodi Forlizzi, Joep Frens, Karey Helms, Eunki Chung – thank you for giving me support and for the intellectual dialogues on design. Carolyn Rosé's group, particularly Michael Yoder, Qinlan Shen, Xu Wang, were particularly influential as I started adventuring language interaction design. At Microsoft Research, Justin Cranshaw, Shamsi Iqbal, Jaime Teevan, Michael Gamon, Edaena Salinas Jasso, Sujay Kumar Jauhar, Mark Encarnación, and so many others made every day of my research exciting, rich, and rewarding.

I have also been very lucky in that throughout my Ph.D. I have been able to concentrate mostly on my research. This is in part due to the gracious support of the Center for Machine Learning and Health (CMLH) and Microsoft Research. Parts of the work in this dissertation were also supported by grants from the National Heart, Lung, and Blood Institute (NIH NHLBI 1R01HL122639-01A1), the National Institute on Disability, Independent Living, and Rehabilitation Research (NIDILRR 90REGE0007 and 90RE5011), and from the National Science Foundation (SES-1734456).

Finally, I thank my family and friends who always believed in me, and always reminded me that I am more than my work. Thank you for being my rock.

Contents

List of Figures

List of Tables

Chapter 1

Motivation and Introduction

1.1 Understanding the Nature of AI's Design Challenges

1.1.1 A Motivational Case

Let me start with a story that exemplifies the central challenge this dissertation aims to address and the benefits of addressing it.

I began my Ph.D. research by studying the design of a decision support tool (DST) meant to aid clinicians in deciding whether and when to implant an artificial heart into an end-stage heart failure patient. The system extracts insights from previous implant recipients' medical records and then predicts the life expectancy of unseen patient cases. Over the past thirty years, a majority of such clinician-facing DSTs – from expert systems [6, 129] to regression risk models [29, 32, 131] – struggled when moving from the lab and into clinical practice. Despite compelling evidence of their effectiveness in research labs, in most cases clinicians rarely used these tools.

As a user experience (UX) researcher, I naturally took a user-centered design approach [24] to this challenge. I conducted a field study, observing how advanced heart failure teams made implant decisions in day-to-day practice. I observed various attitudinal and contextual barriers that are likely to prevent clinicians from slowing down or deviating from their work routine, to consult a computer [133].

These observations informed the design of a new form of DST, one that automatically generates the slides used in a clinician meeting [140]. It subtly embeds the machine predictions into clinicians' existing work routine, rather than pulling them away from it. It predicts post-implant complications to

inform clinicians' decision *discussion*, rather than making the decision for them. This new design draws inspirations from early HCI research, for example, Tolmie et al.'s classic notion of "Unremarkable Computing" [120], that technologies that augment users' existing routines can have significant importance for their lives yet remain unobtrusive. It also draws inspirations from an early HCI lesson in participatory design, that we need to make technical advances that skill workers instead of de-skilling them [31]. A simulation-based field evaluation demonstrated that clinicians were more likely to encounter and embrace such a DST in their practice. As such, this work offered one practical solution to the long-standing challenge of DSTs' real-world adoption.

Interestingly in the past few years, this work found renewed relevance as intense research interest rose in the intersection of user experience (UX) and artificial intelligence (AI). Researchers encountered similar barriers when moving state-of-the-art machine learning (ML) systems into clinical practice, barriers such as workflow integration and gaining clinicians' trust [4, 19, 59, 74, 130]. These are challenges that UX design communities have routinely grappled with, both in research and in practice. The heart-implant DST project, in this light, exemplifies how this rich body of UX research (e.g. designing skilling technologies, unremarkable computing) and practical methods (e.g. fieldwork, simulation-based rapid prototyping) can offer a valuable point of departure for improving human-AI interactions.

New questions also arise as researchers start to shift their attention from understanding the clinical contexts to crafting clinician-*AI* interaction. Consider a UX design issue of AI, for example, clinicians' trust AI suggestions. To what extent do the lessons learned in designing simple regression-based DSTs truly generalize to designing the deep-learning-based ones? Does "AI" bring *unique* challenges in ethics and fairness – ones that fundamentally differ from other data-driven, networked technologies – as public sentiment sometimes suggests? Consider UX design methods and tools – Are the existing UX design methods (e.g. field work, sketching, prototyping) that are sufficient for addressing UX issues for many DSTs, also sufficient for AI-powered ones? Current research does not always make these distinctions.

A critical and necessary first step in *systematically* moving clinical AI systems into practice – rather than every ethnographic and design work addressing one UX issue for a system that addressed one clinical decision – is to articulate what distinctive challenges make designing UX of AI so difficult in the first place.

1.1.2 Whether, Why, and How UX of AI Is Uniquely Difficult to Design?

One goal of this dissertation is to delineate *whether*, *why*, and *how* the user experience (UX) design of AI is distinctly different from the UX design of computational technologies in general. The preceding case study exemplifies the conditions that make now an opportune time for this inquiry:

1. *Bridging UX's and AI's ways of knowing and practices are becoming increasingly important*, as AI's technological advances increasingly migrate from research labs into the real world. From predictive medicine to autonomous driving, these technologies promise to improve people's lives and societies. Many have already enjoyed remarkable success; Others, however, faced new challenges. For example, how should predictive models integrate into physicians' decision-making processes, such that the predictions affect them appropriately? Modern Natural Language Generation (NLG) systems can provide phrase-or-sentence-level writing suggestions upon user request, but what kinds of suggestions do authors want? These are challenges of translation: translating AI as a technological advance in research labs into a real-world sociotechnical system effectively serving human ends.

   UX design research and practical methods have much to offer for addressing these challenges. Of particular relevance are bodies of research under the banners of human-centered machine learning, interactive machine learning, algorithm perception and interaction intelligibility, mixed-initiative systems, among others.

2. *Challenges persist in addressing the UX design challenges of AI,* especially among practitioners who do not specialize in both UX and AI. The prevalence of AI in today's society seems to suggest that the UX communities have already become experts at designing human-AI interactions. Indeed, AI-related HCI research has been in rapid growth, and many AI products and services have been remarkably successful. Interestingly, recent research also reveals something else; that even seasoned UX professionals can struggle with integrating AI into the practice [30, 41, 52, 60, 61]. Their reported challenges are rarely emphasized or formally studied.

3. *The core challenges AI brings to UX design are not yet well-understood and rarely formally studied.* There exists no agreed-upon set of root causes around which one can easily summarize the challenges that AI brings to UX design. Some researchers have speculated that AI systems' technical complexity

causes their UX problems, such as explainability and fairness. Others considered AI's unpredictable system behaviors as the cause [52]. Some argued that AI is just "*a new and difficult design material,*" suggesting that over time, known UX methods will likely address these challenges as UX professionals become more familiar with the technology [30]. Others argued that user-centered design needs to change in order to work for AI [35, 42]. These proposals rarely share key citations that indicate emerging agreements.

4. *Researchers have taken a remarkable heterogeneity of approaches to address AI's UX design challenges.* However, human-centered AI remains relatively disparate between its vast and soft goals (e.g., clinician-AI partnership) and valuable yet system-or-domain-specific solutions (e.g., improving physician-AI collaboration on a narrow clinical decision, with a particular kind of data-driven system). Researchers who study particular UX issues of AI struggle to tease out the unique challenges AI brings to the issue, or to avoid reinventing the wheel. Researchers who design particular systems struggle to assess or articulate the extent to which their solutions generalize to other systems or to other human conditions.

A critical first step in *systematically* addressing the UX design challenges of AI is to articulate the distinctive challenges that make it different or difficult to design in the first place. Researchers need to first understand how general HCI challenges and uniquely AI challenges are confounded, in order to tease them apart later and address them in a targeted fashion. It is in this context, that I worked to identify a useful structure to the currently fuzzy problem space of UX design of AI.

1.2 An Operational Bounding of AI for Understanding Its Design Complexities

A sophisticated understanding of AI's design challenges is hampered at the start by the difficulty of pinning down a precise definition of "AI". What is commonly referred to as AI encompasses many disconnected technologies (e.g., decision trees, Bayesian classifiers, computer vision, etc.). The technical boundary of AI, even in technical AI research communities, is disputed and continuously evolving. (More comprehensive reviews of AI definitions can be found elsewhere [63, 128]).

I would further argue that the question of "what is AI" is inherently intertwined with that of "what

about AI makes its UX so difficult to design". If a precise, agreed-upon definition of AI did exist, then UX researchers could just investigate the latter question by comparing the effectiveness of existing design methods on AI systems and non-AI ones. Teasing out AI's unique design challenges would not be a dissertation-worthy inquiry in the first place.

The concept of AI is so vast, nuanced, and controversial that it is worth deferring a precise definition for the moment. Instead, I will work to choose a loose, operational definition of "AI" as a starting place of my inquiry. I will then examine whether various systems that are considered as AI by this definition indeed require new HCI design methods.

Existing definitions of AI generally fall into two camps. One describes AI as computers that perform tasks typically associated with the human mind ("*AI is whatever machines haven't done yet*" [51]). The other defines AI in relation to computational mechanisms. I chose a widely-adopted definition from the latter camp, because the focus is this work is AI as a computational advance, rather than what people perceive as "intelligent".

> In this work, AI refers to computational systems that *interpret external data, learn from such data, and use those learnings to achieve specific goals and tasks through flexible adaptation.* [54]

Again, I do not intend to draw a technical boundary of what counts as AI here. I also do not consider this definition as valuable for UXers in working with AI. Instead, I will use this definition only as a starting place to examine AI's design complexities. For example, this definition describes AI as "learning" from data, yet does not specify what counts as "learning." (It remains an issue of debate in technical AI communities.) Therefore within this dissertation, I will consider the challenges designers reported in working with a full range of data-driven systems, including machine learning, classic expert systems, crowd-sourcing, etc. I will then examine whether the challenges are different across the spectrum from systems that most people would agree "learned" from data to those did not. This way, I can start to separate the design challenges that are unique to AI and those that UX design routinely copes with.

1.3 Research Methods Overview

With this initial bounding, I set out to investigate whether, why, and how human-AI interaction is uniquely difficult to design and innovate. I wanted to identify a coherent, useful framework that can give structure to the currently tangled problem space in the intersection between UX design and AI. Four threads of research have been leading me towards this goal.

First, I systematically analyzed research literature at the intersection of UX and AI. The analysis covers the research discourses under several different banners, such as human-centered machine learning, human-AI interaction design, AI/machine learning as design material, the design of intelligent systems, designing for/with data, and many more. I analyzed this body of literature by cataloging the many design challenges that literature has reported, as well as the solutions it has proposed (e.g. [131]). I also analyzed the literature with a practice focus. For example, we worked to identify clusters of prior HCI research where AI's technological advances (e.g. clinical machine learning) have frequently take similar interactive forms, indicating potential opportunities for design innovation [135].

I then started to design a variety of AI applications first-hand [133, 134, 139, 140, 146]. These applications span the domains of healthcare, mobile computing, online social networks, and more. I undertook empirical studies of actual stakeholders and use contexts as the basis of my design. Each design addressed a critical challenge in moving AI from research labs valuably into the real world, such as user acceptance, human-agent teamwork, accessibility, and human agency. These in-depth, hands-on Research through Design projects enabled me to develop a felt understanding of AI's design challenges, as well as the solutions that naturally emerged from the process.

Next, I investigated how the researchers and practitioners in the technology industry leveraged – or failed to leverage – AI in their respective domains of interest [109, 136]. This includes empirical studies with nearly 150 industry practitioners of the industry best practices as well as the intuitive approaches of HCI practitioners and AI engineers who are new to human-AI interaction. Similar to Research through Design, this empirical approach underscores that design knowledge arises from, and in response to, concrete problems and situations [46, 103]. This approach also complements my first-hand design experiences, since it covers a more diverse set of AI-related systems, many of which have been deployed for many years.

Synthesizing the observations and learnings from all these approaches, I worked to delineate whether, when, and how human-AI interaction is uniquely difficult to design with established HCI methods [141]. For over a year, I iteratively proposed and critiqued many candidate theoretical constructs. At the end of this process, I was able to trace the nebulous challenges of human-AI interaction design back to just two root challenges: 1) uncertainty surrounding AI's capabilities, and 2) AI's output complexity, spanning from simple to adaptive complex. I outlined a conceptual framework that explicates how the root challenges unfold in concrete UX design scenarios. This mapping also presents a prioritized, constructive agenda for future research opportunities.

One limitation of this work is that the case studies are mainly from my own research and design experiences. This is neither a representative sample nor a comprehensive one. The meta-analysis nature of this dissertation's research goal calls for an extensive collection of AI design projects, ideally covering *all* kinds of AI systems for *all* kinds of design contexts. This is beyond what one dissertation can achieve. The synthesis of my experience and the resulting framework is intended to serve as a moderate first step in this direction.

1.4 Thesis Overview

Like many other research efforts and knowledge creation processes, the process of formulating the conceptual framework is a lot messier than the above Methodology section implies. I therefore report my research findings in the order of general-specific-general, rather than chronologically.

This dissertation centers around the framework which outlines AI's design complexities, and proceeds in four stages:

- In Chapter 2, I set the stage by reviewing the routine challenges of UX design and the established methods and processes for addressing them (2.1). Against this backdrop, I catalog the many human-AI interaction design challenges that prior research has reported as well as solutions proposed (2.2). The contrast between the two (2.3) served as a springboard for rethinking how I can deepen the understandings of AI's design challenges.

  For example, traditional UX research highlights designers' ability to "have a reflective conversation" with their design material; that they can learn *what the technology can do* tacitly in the process of mak-

ing novel things with it (e.g., through fast, iterative prototyping). In contrast, prior HCI/AI research frequently cited AI systems' technical complexity as causing UX design challenges, therefore focused on teaching designers *how the technology works*. This contrasts puts the distinction between AI's capabilities and its inner workings in the spotlight. It foregrounds my investigations into AI's design challenges.

- In Chapter 3, I describe the *AI design complexity framework*, my answer to the question of whether, when, and how UX of *AI* is uniquely difficult to design. I identify two sources of AI's design complexities, and unravel their effects on design processes. I make the argument that UX design *expertise* remains valuable for AI. Established design methods and processes can readily address the UX issues of some AI systems, particularly those that produce simple outputs and do not continue to learn from new data post-deployment (e.g., most clinically-deployed DSTs). I refer to them as *Level 1 systems* for simplicity. Some other systems expose the limits of current UX design methods and practices. For example, some AI systems continue learn from new, unseen data post-deployment (e.g. Facebook news feed ranker) and can generate outputs that resist easy simulation (e.g. machine-generated utterances). Such systems problematize the conventional lab-based UX prototyping methods, most of which treat a system's capabilities and limits as bounded and interactions prescriptive.
- Chapters 4-6 describe case studies in detail, demonstrating the usefulness of the framework for future UX research. In Chapter 4, I describe the design process of a clinical decision support system for artificial heart implant patient selection (the project that was briefly described in section 1.1.1). This project exemplifies the design of *Level 1 systems*. Using the framework, I illustrate how existing UX design methods can address its many critical UX issues and encourage practitioners to embrace such AI systems in their daily practice.
- In Chapter 5, I describe the design process of a generative writing assistant, a *Level 4 system*. Modern Natural Language Generation (NLG) systems can provide phrase-or-sentence-level writing suggestions upon user request, but what kinds of suggestions do authors want? With the range of users' desired generative functionalities, which can the system *reliably* provide for users? Traditionally, UX designers explore these questions through rapid and iterative prototyping, probing user needs as well as technical capabilities and limits. However, machine-generated texts – especially their seemingly

unpredictable and bizarre errors – are difficult to simulate with traditional prototyping methods. In this case, "AI" reveals the limits of the existing UX design methods.

As part of this project, I created a set of tools for rapid prototyping human-AI collaborative writing in real-world contexts, in collaboration with many NLG researchers [139]. I created a new prototyping method that can simulate the interactions of various NLG systems, including their seemingly unpredictable and bizarre errors. This method enabled us to test the not-yet-built NLG systems in users' natural writing contexts. The observations from the user study became a vantage point for the later design and development of generative writing assistants. Using the framework, I discuss the advances these emergent design tools make for designing *Level 4 systems* more broadly and the challenges remain for future research.

- Chapter 6 continues the investigation into the design of *Level 4 systems*. I describe the design practices of some of the experienced UX designers who regularly create new products and services that use AI or machine learning to enhance UX; in order words, the current industry best practices of UX design of AI. I illustrate how their collective approach and reflections, though focused on varying systems and human contexts, echo many of the findings from the above-two projects. I discuss how the industry best practices reveal new insights around UX design education and insights on the kinds of design tools needed for supporting UX design innovation with AI.
- In Chapter 7, I summarize the takeaways and provide suggestions for future research.

This dissertation intends to make three contributions. First, it provides a synthesis of many human-AI interaction design challenges and emergent solutions in the literature. Second, it proposes a conceptual framework that gives structure to the currently fuzzy problem space of human-AI interaction design. The framework offers an alternative lens for understanding AI's UX design challenges. It draws attention to AI's design complexity rather than technical complexity; It draws attention to how AI hinders the interaction design process rather than the end product. Finally, this dissertation offers two detailed case studies, exemplifying how researchers can effectively broach into AI's UX design opportunities and challenges by distinguishing how they relate to and differ from UX design in general. Taken together, this dissertation provides a first step towards connecting islands of prior research and building a smoother pathway between AI's technological advances and valuable AI-enabled user experiences.

Chapter 2

Related Work

UX designers integrate known technologies into novel and valuable new applications and services [68]. To envision things that have never before existed, designers innovate by engaging in reflective conversations with the technology hand as their design materials [105]. Schön, and many design theorists that followed, have described how designers "*reflect in action*"; how they learn tacitly what the technology can do and conceive of what they want to make with it while in the act of designing and making [103].

In this chapter, I start by a brief review of building blocks of UX design expertise that enables and facilitates this creative process. This include the *cognitive skills and processes* that prior research identified as essential for designing good UX, the *hands-on design activities and techniques* through which these cognitive processes unfold, and the design *knowledge* accumulated through the process (Section 2.1). Against this backdrop, I review the challenges that prior research has reported in designing UX of AI as well as the proposed solutions (2.2). The juxtaposition of these two threads of work lays the ground for investigating the challenges of working with AI as a design material. It inserts new questions: Does AI require new design processes, new design activities, new methods, or some combination of the above? What makes UX of AI appear uniquely difficult to design? These questions drive the inquiry of this dissertation.

2.1 From Technological Advance to User Experience

Before discussing the UX design challenges of AI, it is useful to first consider what makes a UX design "good", what the common challenges are (in terms of design thinking, acting, and knowing), and how UX research has supported these aspects of UX design of previous new technologies.

2.1.1 Design Thinking: Cognitive Skills and Processes

Don Norman coined the term "*user experience*" as a counter-movement to the dominant, task-related "usability" paradigm at the time [85]. By his definition, UX encompasses "*all aspects of the end user's experience*" with a technology system. Engineering is the process of creating technologies that allow new technical capabilities. UX design is the process of creating everything that framed the experiences of known technologies [68].

Pinning down a universal matrix of "good" UX design is difficult, because of its all-encompassing nature [69]. While there exists widely-accepted rubrics for evaluating usability [83], an appropriate evaluation matrix for a UX design is often dependent on the technology and its design situation.

Facing this challenge, prior design and cognitive science research has considered the quality of a UX design through the cognitive sophistication involved in producing it [67]. In his seminal book *Sketching User Experiences: Getting the Design Right and the Right Design*, Buxton suggests that two considerations are necessary for a good UX design: *what* a system needs to accomplish for users ("*design the right thing*") and how it would accomplish it ("*get the design right*") [18].

Two broad kinds of cognitive skills are essential for achieving both goals, both identifying the right things to design and designing the thing right. One is divergent, creative thinking; The other is converging, judicious thinking [67, 106]. Through divergent thinking, designers acquire many possible frames of the users' true needs and consider a great number of alternative solutions in parallel. Through converging thinking, designers elaborate on these problem framings and the solution possibilities, anticipate the consequences of these choices, and consciously selecting the best ones. The iterative process of divergent-convergent thinking engenders "good" UX designs; designs that are both innovative and thoughtful.

This kind of cognitive process underpins some of the most widely adopted UX design processes

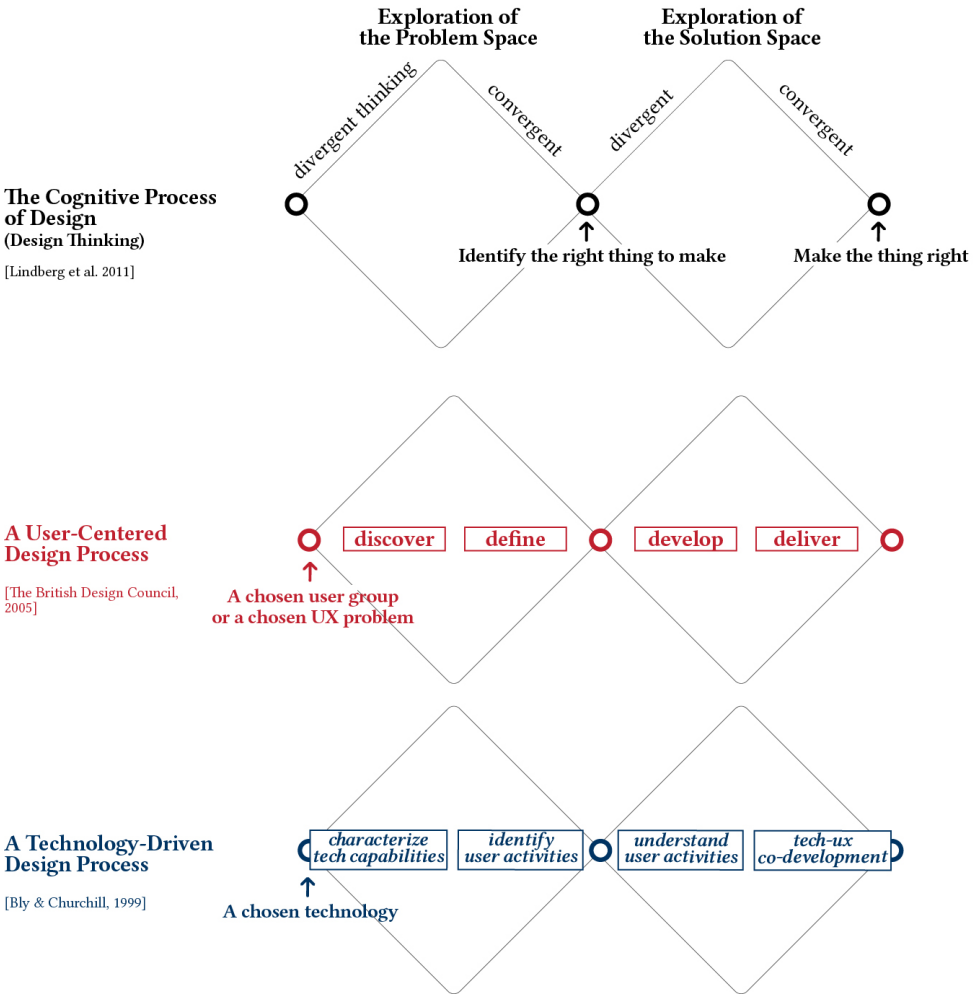

Figure 2.1: A technology-driven design innovation process [11, 97]

(Figure 2.1). The most well-known may be the "double-diamond" user-centered design process [24]. This workflow takes a target user group or a problematic situation as a starting place. It highlights the need to systematically investigate user needs (divergent thinking in the problem space) before developing any solutions; "*the last thing that you should do when sketching an interactive system is to write code*" [18].

Another example is the agile design/development process, a technology-driven innovation process [97]. In this case, designers chose a technology (e.g., a patent) as their starting place; as the *material* for their design. Designers then work to understand the capabilities and limits of the technology and the various design possibilities it might afford (divergent thinking in the problem space). They then systematically search for the users and activities that would benefit from the technology the most (convergent thinking, problem space). Next, they design a minimal viable product for the target users, refine

it iteratively, and pivot when necessary (divergent-convergent thinking in the solution space) [11]. This process leads designers to maximally explore both the problem and solution space of their design, even though they have committed to a technology as part of the solution from the beginning.

In this dissertation, I will consider both design conditions and examine whether and how AI destabilizes the cognitive processes of design. The case study in Chapter 4 starts with a pre-identified UX problem and explores a user-centered design process of AI. Chapter 5 starts with Natural Language Generation (NLG) as pre-identified part of the design solution.

2.1.2 Design Actions: Hands-on Activities and Methods

Designers carry out their cognitive processes through hands-on design activities; They reflect-in-action [103]. For example, they map out sticky notes when brain-storming; They draw sketches of the user interface; They build interactive prototypes. In undertaking these activities, designers are not merely externalizing the ideas that are already in their minds. Rather, new ideas emerge as they re-configure, amplify, or de-contextualize various factors of the diagrams, sketches, and prototypes.

Good design techniques and tools, such as sticky notes and paper prototypes, are valuable primarily as a tool for design thinking [18, 33, 45, 112]. Lim et al. have argued that good prototypes and prototyping techniques can serve as both manifestation of designers' emergent ideas as well as their filters [66].

When facing new or difficult technologies, research communities have created new techniques and tools to help. For example, new design methods such as service blueprint and customer journey map emerged from the newfound interest in services as a design material. Both provide new language for talking about the complex and amorphous social relations and technological ecologies surrounding the software and hardware offered to users [108]. This new language includes customer journey and touch points and value co-creation, to name a few [142]. Moussette's *Sketching Haptics* provides another example. He had a desire to work with haptics; however, he had no easy way of playing with haptics as a design material in order to develop tacit knowledge of what could be [77]. He therefore produced a set of haptic sketches; a set of physical prototypes that embodied his felt understanding of what experiential potentials haptics possess. These techniques and tools help designers with conceptualize the design space and externalize design ideas. Novel Wizard-of-Oz prototyping techniques have helped to expedite the feedback loop between the proposed design of complex systems and the resulting UX, thereby

facilitating the filtering of design ideas and provoking new ones [66, 96].

2.1.3 Design Knowing: Technologies as Design Material

Through the iterative process of design ideation and evaluation, designers also learn about technologies at hand. Specifically, designers investigate a technology's functional *capabilities and limits* as they manipulate it and use it in different designs. They develop a felt understanding of its *experiential affordances and qualities* [86] by observing how users interact with and experience it. These understandings then inform and expand the later design iterations. Prior research has refereed to this knowledge as a "*designerly understanding*" of technology, because it is tightly related to the specific purpose of design; of innovating new things out of existing technological capabilities [28, 123]. It differs from engineers' typical way of knowing, which focuses on creating new technical capabilities [68, 104].

This tacit, designerly way of knowing has tremendous implications for how UX research supports designers working with new or partially understood technologies in practice. The design methods and tools mentioned above – service blueprinting, haptic sketches, and various Wizard-of-Oz techniques – are good examples. Rather than educating designers about the technology directly, these methods aid designers to more easily "see" the new design possibilities it opens up as well as its manifestation on UX. In so doing, these methods allow designers to develop knowledge about the technology that influences their design.

UX research has also hosted workshops as a way of sensitising practitioners to emerging technological capabilities. For example, design researchers have made intriguing sensitizing concepts for interactive textiles [81]; they held workshops to expose haptics' design possibilities beyond a buzzing phone to design practitioners [77]. Similar to other forms of design knowledge transfer, the workshops transfer new technology to practitioners by facilitating them to grasp and feel its capabilities (e.g., the design possibilities interactive textiles offer), rather than to teaching them its inner workings (e.g., how interactive textiles work).

When working with complex technologies that are difficult to "play with" hands-on, designers often need to work with developers and collaboratively explore, evaluate, and understand the design opportunities the technology offers. HCI research has previously created boundary objects to facilitate such collaborations [89]. Boundary objects support dialog and consensus-building between people coming

from different perspectives (areas of expertise) [14].

The title of this dissertation "profiling AI as a design material" is a reference to this body of prior work. Through this dissertation, my goal is to provide an initial sketch of a designerly understanding of AI; an understanding of "AI" not in its role in technological advance, but in its characteristics relating to design thinking and acting. This understanding can benefit our field by establishing more solid ground upon which we can more purposefully innovate UX design methods for AI.

2.2 From AI's Technological Advances to User Experience

UX and AI researchers have produced a wealth of valuable, novel UX designs of AI in recent years (see [135] for a comprehensive review). Interestingly, they have also reported many challenges they encountered in the process [41, 52, 60, 61]. These challenges span across many application domains and various types of intelligent systems [15, 98, 125]. When discussing these challenges, researchers have chosen a number of different frames, including human-AI interaction design, AI/machine learning as a design material, the design of intelligent systems, designing for/with data, and many more [12, 34, 65, 88, 98, 102].

Below I catalog these challenges and emergent solutions identified in prior work. I map them to the double diamond user-centered design process (Figure 2.2) and to a diagram displaying a technology-driven innovation process (Figure 2.3). I will argue for a need to better unpack what is known and unknown about the UX design challenges of AI, particularly, whether AI calls for new design thinking, new design activities and methods, new design tools, or some combination of the above.

2.2.1 Reported Challenges

Across HCI and UX communities, researchers and practitioners have reported challenges in working with AI at almost every step of a user-centered design process. From left to right on Figure 2.2, they reported:

- *Challenges in understanding how AI works*: Recent research frequently cited AI systems' algorithmic complexity as the cause of many human-AI interaction problems. For example many deep learning systems' working mechanisms remain an active area of research within technical AI communities.

This raises UX concerns such as explainability and intelligibility, transparency, ethics [1, 2, 43]. Moreover, UX-practice-focused research has showed that some designers struggle with understanding less complex AI systems as well; They treated AI largely as black magic: "*inputs come in... some magic happens... and all your business needs are met!*" [30]

- *Challenges in understanding AI capabilities (first divergent thinking stage)*: Designers frequently report that it is difficult to grasp what AI can or cannot do. This hampers designers' brainstorming and sketching processes from the start [30, 52, 135, 138].

- *Challenges in envisioning many novel, implementable AI things for a given UX problem (in both divergent thinking stages)*: AI-powered interactions can adapt to different users and use contexts, and they can evolve over time. Even when designers understand how AI works, they often found it difficult to ideate *many* possible new interactions and novel experiences with much fluidity [30, 134].

- *Challenges in iterative prototyping and testing human-AI interaction (in both convergent thinking stages):* One core practice of HCI design and innovation is rapid prototyping, assessing the human consequences of a design and iteratively improving on it. HCI practitioners cannot meaningfully do this when working with many AI systems. As a result, AI's UX and societal consequences can seem impossible to fully anticipate. Its breakdowns can be especially harmful for under-served user populations, including people with disabilities [113].

  HCI researchers have tried two approaches to addressing this challenge. One approach is to create Wizard of Oz systems or rule-based simulators as an early-stage interactive AI prototype (e.g., as in [26, 58, 96, 111]). This approach enables HCI professionals to rapidly explore many design possibilities and probe user behaviors. However, this approach fails to address the UX issues that will come from unanticipated AI inference errors. The second approach is to create a functioning AI system, and deploy it among real users for a period of time [136]. This time-consuming, field-trial prototyping process enables designers to fully understand AI's intended and unintended consequences. However, it loses the value that comes from rapid and iterative prototyping. This approach does not protect teams from over-investing in ideas that will not work. It does not allow them to fail early and often.

- *Challenges in crafting thoughtful interactions (in the last convergent thinking stage)*: Designers struggled to set user expectations appropriately for AI's sometimes unpredictable outputs [2]. They also worried

about the envisioned designs' ethics, fairness, and other societal consequences [30, 52].

- *Challenges in collaborating with AI engineers (throughout the design process)*: For many UX design teams, AI technical experts can be a scarce resource [42, 136]. Some designers also found it challenging to effectively collaborate with AI engineers, because they lacked a shared workflow, boundary objects, or a common language for scaffolding the collaboration [42, 56, 139].

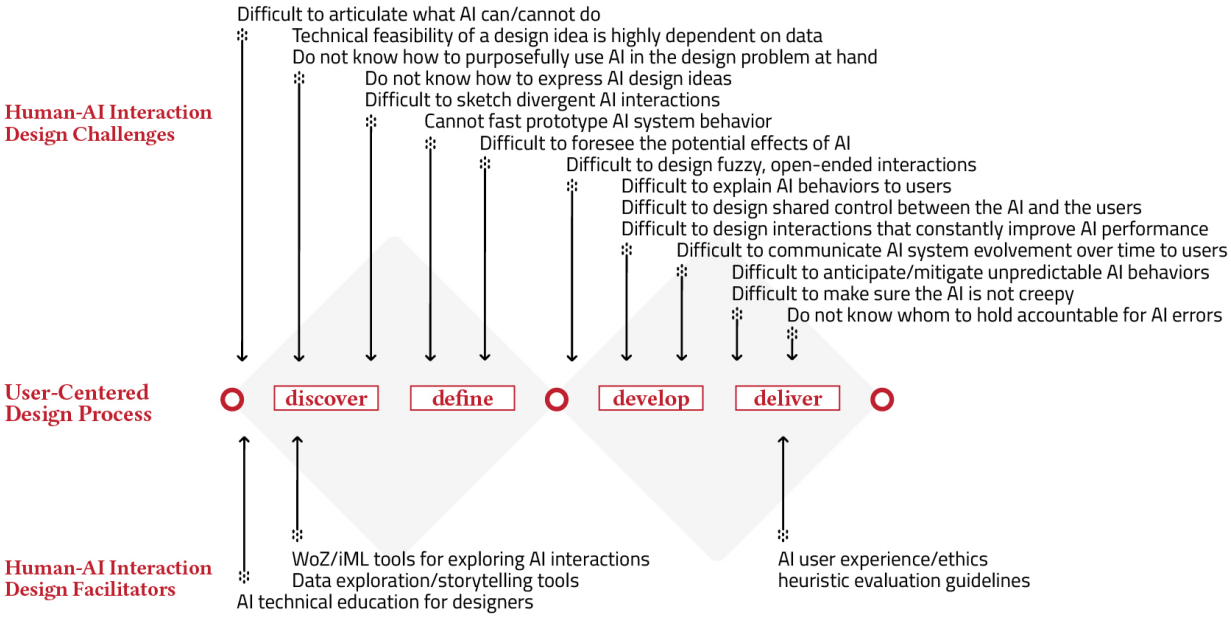

Figure 2.2: Mapping the human-AI interaction design challenges in the literature [30, 52, 134, 136] onto a user-centered design process (Double Diamond [24])

.

Propelled by these challenges, a few researchers speculated that, when working with AI, designers should start with an elaborate matching process that pairs existing datasets or AI systems with the users and situations that are most likely to benefit from the pairing [11, 135]. This approach deviates from more traditional user-centered design in that the target user or UX problem is less fixed. It is more similar to an agile, technology-driven innovation process that focuses on the creation and continual evaluation of a minimal viable product (MVP) [97]. In this light, I also mapped the human-AI interaction design challenges onto an MVP innovation process. However, it seems a similar set of design challenges that curbed user-centered design also thwarted technology-driven design innovations (Figure 2.3, from left to right). For example:

- Challenges in understanding AI capabilities;
- Challenges in mapping out the right user stories and user cases of a "minimum viable" AI system, or envisioning how it can be applied in less obvious ways [30];
- Challenges in collaborating with AI engineers.

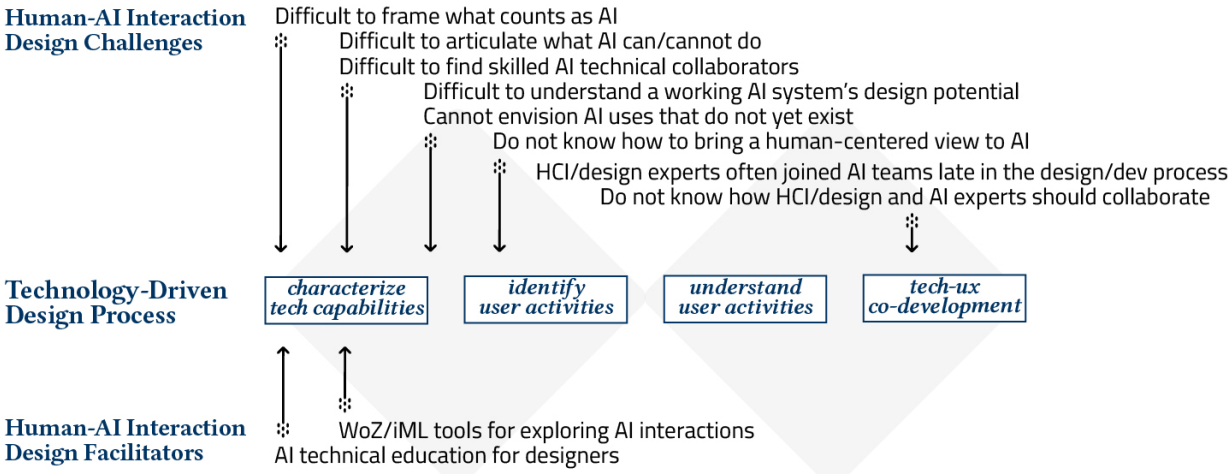

Figure 2.3: Mapping UX design challenges of AI in prior research on a technology-driven design innovation process [11, 97]

I found no agreed-upon set of root causes or themes around which one can easily summarize these challenges. Some researchers suggested that AI systems' technical complexity causes the interaction design problems [21]. Some considered the unpredictable system behaviors as the cause [52]. Some argued that AI appeared to be difficult to design because AI is just "*a new and difficult design material,*" suggesting that over time, known HCI methods will likely address these challenges [30]. Others argued that user-centered design needs to change in order to work for AI [42, 135]. These proposals rarely share key citations indicative of emerging agreements.

2.2.2 Proposed Facilitators

UX researchers have started to investigate how to make it easier to design UX design of AI. I identify five broad themes in this body of work; research under the banner of "human-AI interaction design" or "UX design of AI":

1. *Improving designers' technical literacy.* An increasing amount of research suggests that UX designers

need *some* technical understanding of AI to productively work with it. Designer-facing AI education materials have become available to help (e.g., [20, 44, 48, 49]). The book *Machine Learning for Designers*, for example, reviews concepts of supervised and unsupervised learning, as well as common analogies of how machine learning works [48]. However, substantial disagreement remains in what kinds of AI knowledge are relevant to UX design, and in how advanced a technical understanding is good enough for designers [21, 121, 137].

2. *Facilitating design-oriented data exploration.* This body of work encourages designers to investigate the lived-life of data and discover AI design opportunities [12, 13, 34]. For example, [88] investigated users' music app metadata as a material for designing contemplative music experience and [50] explored the design opportunities around intimate somatic data. Notably, this body of work often used terms like data-driven or smart systems; It was not always clear when the authors specifically aimed at AI.

3. *Enabling designers to more easily "play with" AI in support of design ideation, so as to gain a felt sense of what AI can do.* This work created interactive machine learning (iML) tools and rule-based simulators as AI prototyping tools, for example, Wekinator for gesture-based interactions [36] and the Delft AI Toolkit for tangible interactions [124]. This body of work can be seen as an extension of the classical ways of design knowledge transfer: by allowing designers to play with their design material and develop a tacit understanding of its capabilities and design affordances.

   Noteworthily, almost all iML tools are application-domain-specific. In order to make the systems accessible to designers and maximally automate data prepossessing and modeling, these systems had to limit the range of possible in/outputs, and therefore focused on particular application domains [91, 92].

4. *Aiding designers in evaluating AI outputs.* In recent years, technology companies have proposed more than a dozen human-AI interaction principles and guidelines (see a review by [114]). These guidelines covered a comprehensive list of design considerations such as "make clear how well the system can do, what it can do" [2] and "design graceful failure recovery" [47].

5. *Creating AI-specific design processes.* Some researchers have proposed that AI may require design processes less focused on one group of users, and instead on many user groups and stakeholders [38];

processes focused less on fast, iterative prototyping, and instead on existing datasets and functioning AI systems [135]; or processes focused less on one design as the final deliverable to engineers, and instead on closer, more frequent collaborations [42].

These themes demonstrated the remarkable heterogeneity of approaches researchers have taken to address the challenges around human-AI interaction design. Similar to most design methods published within HCI research, I found little to no empirical evaluations of the proposed design tools, guidelines, or workflows. It is difficult to control for and measure improvements in a design process to show that a method is producing better designs. Throwing AI into the mix only seems to increase this challenge.

2.3 Summary of Related Work

The preceding review on UX and AI revealed an remarkable set of insights and approaches to this complex problem space. In order to gain new insights, I have put this body of work against the backdrop of routine UX design challenges and facilitators. The juxtaposition of these two threads of work inserts new questions: Whether and when does AI require new design processes, new design activities, new methods, or some combination of the above? What exactly makes UX of AI appear uniquely difficult to design?

Prior research has not yet formally investigated these questions. AI brings challenges to almost all stages of a typical design process. However, the proposed AI design methods and tools have mostly focused on the two ends of this creative process (Figure 2.2 and 2.3); either helping designers to understand what AI is and can do generally, or enhancing the evaluation of the final design. The central activities of an interaction design process, (i.e. sketching and prototyping) and underlying cognitive design process (designerly conversation with AI as a design material), are under-explored. Research through Design (RtD) projects are rare when it comes to designing and innovating human-AI interaction [135]. In the following chapters, I take an RtD approach and search for a more incisive examination of AI's UX design challenges.

Chapter 3

A Framework of AI Systems' Design Complexities

In this Chapter, I will describe the conceptual framework, my answers to the question of whether, when, and how UX of *AI* is uniquely difficult to design (3.2). Specifically, I identify two sources of AI's design complexities, and unravel their effects on design processes. In 3.3, I demonstrate its usefulness to human-AI interaction designers, to researchers of AI's HCI issues, and to AI design method innovators and tool makers. The subsequent Chapters will present case studies that further unpack the uses and usefulness of the framework.

3.1 Research Process

Earlier in section 1.2, I identified an operational bounding of AI as a starting place of my inquiry: In this work, AI refers to computational systems that interpret external data, learn from such data, and use those learnings to achieve specific goals and tasks through flexible adaptation [54]. Within this bounding, I curated a set of AI design process from my own research, design, and teaching experience, in searching for a more incisive examination of AI's UX design challenges. Below is a brief overview of these projects. All projects described below except teaching have been published at DIS and CHI [132, 134, 136, 138, 139, 140].

3.1.1 Making AI Things via Research through Design

First, I draw on my own experience in designing a wide range of AI systems. These systems range from simple adaptive user interfaces [134], to large-scale crowdsourced transportation information systems [146]; from clinical decision supports [133, 140] to natural language productivity tools [139]. I undertook empirical studies of actual stakeholders and use contexts as the basis of each design. Each design addressed a critical challenge in moving AI from research labs valuably into the real world, such as user acceptance, human-agent teamwork, accessibility, and human agency.

For space considerations, this thesis will only detail one clinical decision support project (Chapter 4) and the natural language productivity tool project (Chapter 4). Both addressed the challenges of user acceptance, human-AI teamwork, and human agency, though among very different user populations and use scenarios.

3.1.2 Studying Practitioners

I have studied HCI/UX practitioners and their AI engineer collaborators in two projects. The first project focused on novice AI product designers [138]. I interviewed 14 product designers/managers and surveyed 98 more to understand how they incorporated, or failed to incorporate, AI in their products. I also interviewed the 10 professional AI engineers they hired to better understand where and how designers sought help. The second project focused on experienced UX practitioners [137]. I interviewed 13 designers who had designed AI applications for many, many years, in order to understand how they work with AI differently compared to working with other technologies. Synthesizing and contrasting the findings across these two studies, I was able to see how novice and expert designers approached designing AI differently.

3.1.2.1 Teaching UX Design of AI Applications

Another set of observations come from teaching. My collaborators and I hosted a series of *Designing AI* workshops. Each workshop lasted for a day, with one instructor working with 2-3 students. The instructor first gave a half-hour introduction to AI, and then provided students with a dataset and a demonstrational AI system. Students were asked to design new products/services with these materials

for an enterprise client. 26 HCI Master students from two universities attended the workshop. All had little to no technical AI background. Throughout the series, I experimented with different ways of introducing AI. I observed how students used the AI technical knowledge in their design, where and how they struggled, and which challenges they were able to resolve with known design methods.

We also taught a one semester design studio course: *Designing AI Products and Services*. Approximately 40 undergraduate and master students took the course. About half had a computer science or data science background. In comparison to the workshops, the course allowed us to observe students working with a more diverse set of AI systems and design tasks, e.g. designing crowd as a proxy for AI, designing simple UI adaptions, designing natural language interactions.

3.1.3 Synthesizing a Conceptual Framework

With this diverse set of design processes and observations, I synthesized a framework meant to give structure to the many challenges around human-AI interaction design. I started by proposing many themes that might summarize these challenges. I then analyzed the emergent themes via affinity diagramming, with a focus on the characteristics of AI that may scaffold a full range of design challenges. Specifically, I critiqued these frameworks based on three criteria:

- *Analytical leverage*: The framework should effectively scaffold a wide range of AI's design opportunities and challenges. It should help separate design challenges unique to AI from others;
- *Explanatory power*: The framework should help researchers articulate how a proposed design method/tool/workflow contributes to the challenges of human-AI interaction design, and the limits of its generalizability.
- *Constructive potential*: The framework should not only serve as a holder of AI's known challenges and solutions; It should also provide new insights for future research.

I proposed and discussed with mentors and collaborators more than 50 thematic constructs and frameworks. Two faculties within the Carnegie Mellon HCI institute, an external faculty, and an industry researcher participated in this process. All have spent at least 5 years researching AI and HCI. I also presented and discussed this work to two research groups. One included about 40 HCI researchers and the other included 12 machine learning researchers. They provided additional valuable critiques and

helped us refine the framework.

3.2 The AI Design Complexity Framework

My synthesis identified two attributes of AI that are central to the struggles of human-AI interaction design: *capability uncertainty* (uncertainties surrounding what the system can do and how well it performs) and *output complexity* (complexity of the outputs that the system might generate). Both dimensions function along a continuum. Together they map the problem space of human-AI interaction design.

3.2.1 Two Sources of AI Design Complexity

3.2.1.1 Capability Uncertainty

When speaking of the capabilities of AI, I broadly refer to the capability an AI system offers (e.g., detect spam emails, rank news feeds, find optimal driving routes), how well it performs, and the kinds of errors it produces. These characteristics are a result from available capabilities of learning algorithms, available datasets, and the interaction in-between. These characteristics determine the UX design possibilities it can afford.

The capabilities of AI are highly uncertain. I illustrate this by walking through the lifetime of an AI system, moving from an algorithmic emergent in AI research labs to situated user experience in the wild (Figure 3.1, left to right).

AI's capability uncertainty is at its peak in the early design/development stage, when designers work to understand what design possibilities an AI algorithm can offer generally before committing to a dataset. What might seem like a blue-sky AI design idea may suddenly become possible because of a newly available dataset. It may also become possible if designers can successfully harvest their own dataset from user interaction traces. This approach gives designers a relatively high degree of control over the data they will eventually work with. However, it is often very difficult to estimate how long it might take to collect enough high-quality data and to achieve the intended functionality. It can seem even more difficult to understand the gap between what the data appear to promise and what the AI system built from that data can concretely achieve. This great uncertainty in AI's capabilities makes it difficult for designers to evaluate the feasibility of their emergent ideas, thereby hindering their creative

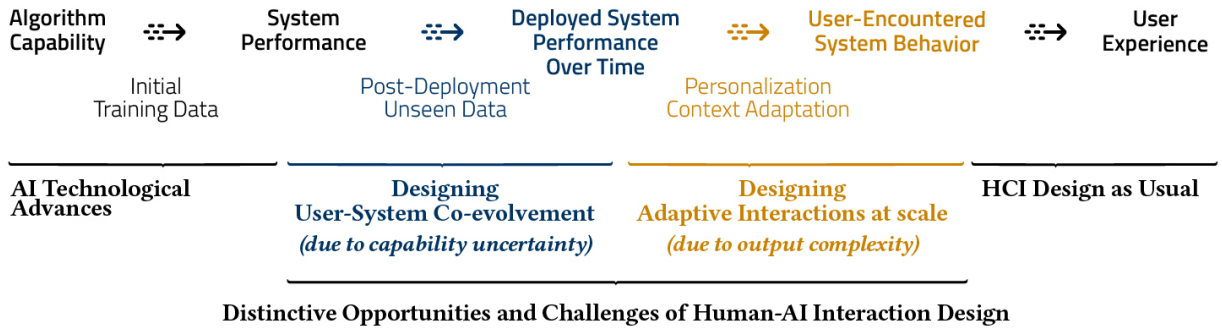

Figure 3.1: A framework of AI's UX design complexities. It illustrates The conceptual pathway translating between AI's capabilities and thoughtful UX designs. AI's capability uncertainty and output complexity add additional steps (the colored segments) to a typical HCI pathway, make some systems distinctly difficult to design. Designers encounter these challenges from left to right when taking a technology-driven innovation approach; right to left when following a user-centered design process.

processes.

What AI can do for a UX problem at hand becomes clearer once a functioning AI system is built. Designers can measure their performance and error modes, and then make design choices accordingly ("lab performance" in Figure 3.1).

Importantly, some AI systems continue to learn from new data after deployment (labeled as "deployed system performance over time"). In the ideal case, the system will "grow," integrating new insights from new data and adapting flexibly to more varieties of users and use contexts. Unfortunately, the new data might also drive system performance in the wrong direction. Tay, the Twitter bot, provides an extreme example [82]. More typically, the system's performance improves for users and use contexts that have produced rich data. It performs worse for less frequent users and less typical situations. That the system capability can constantly evolve, fluctuate, and diversify is another part of AI's capability uncertainty. For these "living systems" (systems that continue to learn from new data post-deployment), their lab performance should only be viewed as an initial estimate.

Finally, user profiles and use contexts could also impact an AI system's capability. Many context-aware and personalization systems fall into this category. Consider the social media news feed ranker, Amazon shopping recommendations, and ride-hailing app's driver-rider matching as examples. It is not difficult to conceptualize what these systems can do in general (e.g., ranking news, recommending items). However, it is no trivial task to envision, for a particular user in a particular use context, what

error the AI system might make, and how the user might perceive that error *in-situ*. Anticipating the situated, *user-encountered capability* of AI is difficult, yet it is fundamental to UX design.

3.2.1.2 Output Complexity

The second source of AI's UX design challenges concerns what an AI system produces as a possible output. While *capability uncertainty* is responsible for the HCI design challenges around understanding what AI can do, AI's *output complexity* affects how designers conceptualize the system's behaviors in order to choreograph its interactions.

Many valuable AI systems generate a small set of possible outputs. Designing interactions for these systems is similar to designing for non-AI systems that generate probabilistic outputs. A face detection tool, for example, outputs either "face" or "not face." To design its interactions, the designer considers four scenarios: when a face is correctly detected (true positive), when no face is detected (true negative), when there is no face and a face is mistakenly detected (false positive), and when the image contains a face but the system fails to detect it (false negative). Designers consider each condition and design accordingly.

When designing systems that produce many possible outputs, sketching and prototyping become more complex and cognitively demanding. Imagine designing the interactions of a driving route recommender. How many types of errors could the recommender possibly produce? How might a user encounter, experience, and interpret each kind of error, in various use contexts? How can interaction design help the user to recover from each error elegantly? Some simulation-based methods or iML tools can seem necessary for prototyping and accounting for the route recommender's virtually infinite variability of outputs. The route recommender exemplifies the many AI systems that produce open-ended, adaptive outputs. The traditional, manual sketching and prototyping methods struggle to fully capture the UX ramifications of such systems.

The system outputs that entail most design complexities are those that are difficult to simulate. Consider Siri as an example. Similar to route recommenders, Siri can generate many, many possible outputs. Yet unlike route recommenders, the relationship between Siri's in- and outputs follow complex patterns that cannot be concisely described. As a result, rule-based simulators cannot meaningfully simulate Siri's utterances; nor can a human wizard. I refer to such AI system outputs as "complex."

Notably, output *complexity* is not output *unpredictability*. While prior research often viewed AI systems' unpredictable errors as causing UX troubles, I argue that AI's output complexity is the root cause. Let us illustrate this by considering how designers might account for AI errors when designing two different conversational systems. One is Siri. The other is a system that always replies to user requests with a random word picked from a dictionary. While highly unpredictable, the interactions of the latter system can be easily simulated by a random word generator. Then following a traditional prototyping process, designers can start to identify and mitigate the AI' costly errors. In contrast, Siri's outputs are only quasi-random, therefore resist abstraction or simulation. To date, it remains unclear how to systematically prototype the UX of such systems, in order to account for its breakdowns.

3.2.2 Two Complexity Sources Taken Together

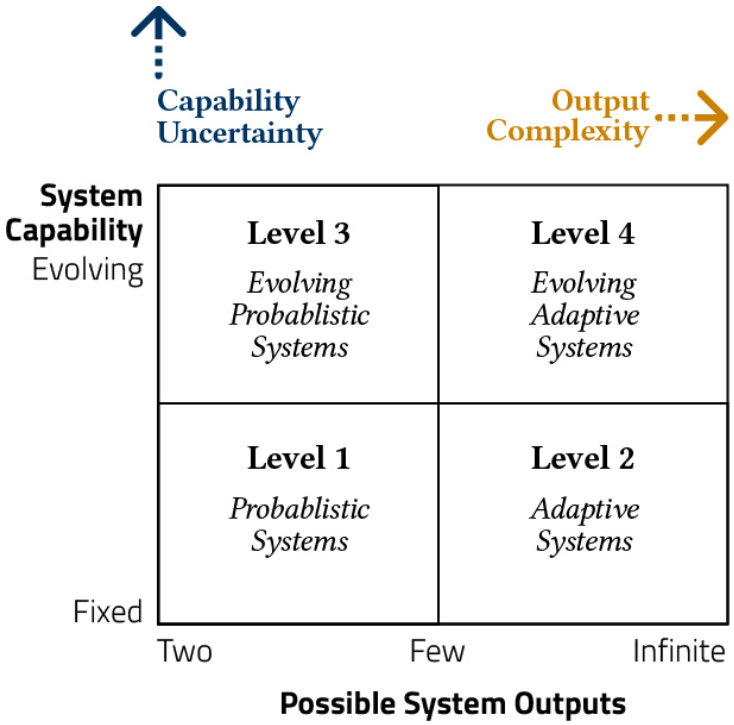

Figure 3.2: The AI design complexity map. Not all "AI" systems are equally difficult to sketch or prototype. This map shows what subset of the AI design challenges a system is likely to involve.

I argue that a wide range of UX design challenges stem from AI's *capability uncertainty* and *output complexity*. For instance, designers struggled to understand what AI can and cannot do even when they understood how AI works [30]; This is because the capabilities of an AI system can be inherently

uncertain and constantly evolving. Designers struggled to rapidly prototype human-AI interaction [139] because the interactions of two mutually adaptive agents resist easy abstraction or simulation. Designers struggled to follow a typical user-centered design workflow when designing human-AI interactions [42, 135]. This is because the central point of a double diamond process is to identify a preferred future, a defined design goal that existing technologies can achieve. However, AI systems have capabilities that do not fully take shape until after deployment, so the preferred future can seem like "*a funnel of what's possible*", rather than what is concretely achievable.

Figure 3.1 maps the challenges onto the translation process between technological capabilities and user experience. When taking a user-centered design approach, designers will encounter the challenges from the right to left. Taking a technology-driven design innovation approach, from left to right. This diagram explains why a similar set of design challenges appeared to have thwarted both technology-driven and user-centered AI design processes.

AI's evolving capabilities and adaptive behaviors have made it a particularly powerful material for HCI and UX design. The same qualities also bring distinctive design challenges. Human-AI interaction design and research, therefore, should not simplistically reject AI's capability uncertainty or output complexity/unpredictability. Rather, it is important to understand how to leverage these distinctive qualities of AI for desirable human ends, while minimizing their unintended consequences.

3.3 Effects on UX Design Processes and Activities

Below I demonstrate the usefulness of the framework. Specifically, I map how AI's design complexities unfold as designers undertake concrete UX design processes and activities (section 3.3.1) and address specific HCI issues (3.3.2). This mapping illuminates valuable research opportunities for AI design method innovators and tool makers (3.3.3).

3.3.1 Four Levels of AI Systems

The framework can help expose *whether* and *how* a given AI system is difficult to design with traditional HCI design processes and methods. The UX design challenges of Level 1 systems, systems with known capability with few possible outputs, overlap with and extend those that UX research has long been

investigating. Existing HCI sketching and prototyping methods, therefore, are better suited to cover Level 1 systems then others. New challenges emerge when designers work with systems that produce a broad set of possible outputs, and when the deployed system continues to learn from new user data. Therefore, for practitioners, the framework can help identify the low-hanging fruit in integrating AI into their practice. For HCI researchers, the framework can help identify the unique challenges of human-AI interaction design and make a targeted contribution.

To make the framework easier to use as an analytical tool, I summarized four levels of AI systems according to their design complexity (Figure 3.3). I demonstrate its usefulness using Levels 1 and 4 systems as examples since they represent the two extremes of AI's design complexity. The design challenges of Level 4 are also a superset of issues encountered in Levels 2 and 3.

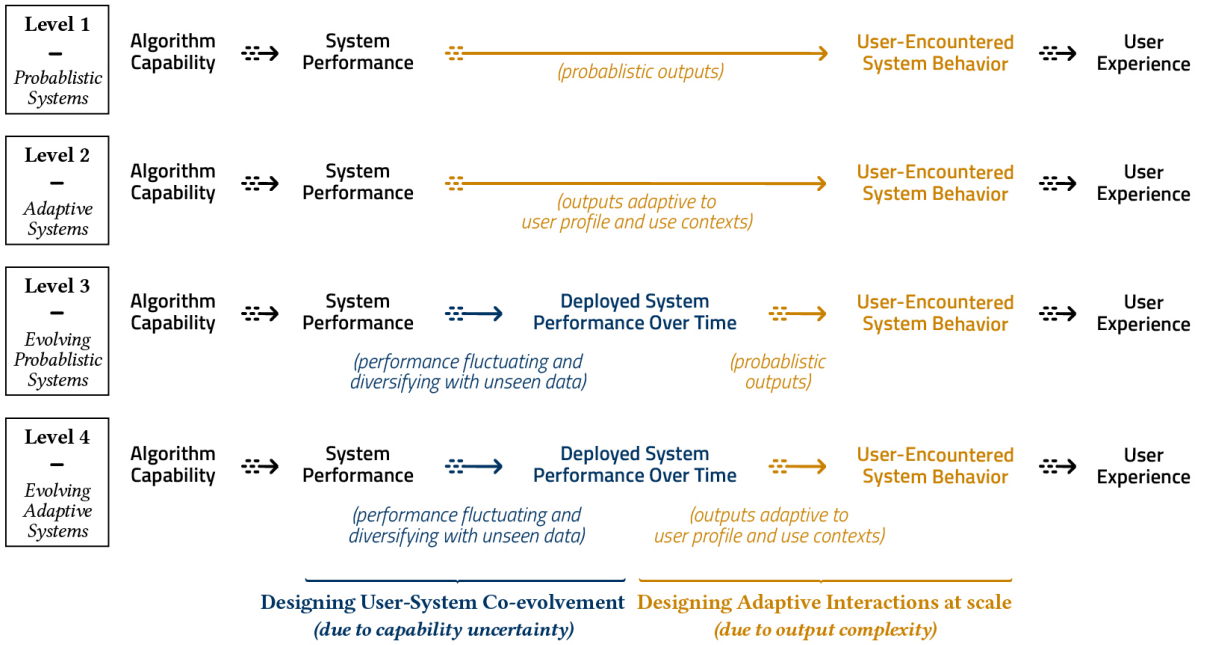

Figure 3.3: Four levels of AI systems, derived from AI's design complexity framework.

3.3.1.1 Level 1: Designing Probabilistic Systems

Level 1 systems learn from a self-contained dataset. They produce a small, fixed set of outputs. Examples of Level 1 systems include face detection in camera apps, adaptive menus that ranks which option the user is more likely to choose, text toxicity detectors that classify whether or not a sentence is profane,

and many more. The clinical machine learning system that predicts artificial heart implant outcomes, which I will study in detail in Chapter 4, also fits here.

I argue that designers can design the UX of these systems in relatively similar ways as designing non-AI, probabilistic systems. They are less likely to encounter the distinctive challenges of human-AI interaction design. Consider this design situation: a design team wants to help online community moderators to more easily promote civil discourse by using a text classifier that flags toxic comments.

- *Few challenges in understanding AI capabilities:* By "*playing with*" the system, the designers can develop a felt understanding of what the classifier can and cannot do: How well does it perform for different use contexts? What kinds of prediction errors (e.g., false positive and false negative errors) are likely? How do users (in this case, community moderators and participants) perceive and react to such errors? Because the system will not learn from new data, these understandings will remain valid post-deployment.

  Noteworthily, I am not arguing that understanding AI capabilities is easy. On the contrary, *conclusive* answers to these questions remain a topic of discussion among HCI/AI research communities. However, I argue that these challenges can be seen as an extension of the classic challenges of designerly conversation with their technological design materials (See 2.1) rather than brand new. Therefore, Level 1 systems can be seen as a place where the application of fairly traditional HCI knowledge and methods can help.

- *Few challenges in envisioning novel and technically feasible designs of the technology*: Designers can easily imagine many use scenarios in which the flagging-profane-text functionality can provide value.

- *Few challenges in iterative prototyping and testing:* Because the outputs of the system are limited (profane, not profane), designers can enumerate all the ways in which the interactions may unfold (false positive, false negative, etc.) and making interactive prototypes accordingly.

- *Few challenges in collaborating with engineers*: Once the designers understand the functionality and the likely performance and errors of the classifier, they can design as usual and provide wireframes as a deliverable to engineers at the end of their design process.

Language toxicity detection is a complex technical problem at the frontier of AI research. However, because the system's capabilities are bounded and the outputs are simple, existing HCI design methods

are sufficient in supporting designers in sketching, prototyping, and assessing its interactions. Language toxicity exemplifies Level 1 systems; They are valuable, low-hanging fruits for HCI practitioners to integrate into today's products and services.

3.3.1.2 Level 4: Designing Evolving, Adaptive Systems

Level 4 systems learn from new data even after deployment. They also produce adaptive, open-ended outputs that resist abstraction. Search engines, newsfeed rankers, automated email replies, a recommender system that suggests "items you might like," would all fit in this category. In Chapter 5, I will describe in detail my design process of a natural language generation system, a Level 4 system.

In designing such systems, designers can encounter a full range of human-AI interaction design challenges. Consider the face recognition system within a photos app. It learns from the photos the user uploaded, clusters similar faces across photos, and automatically tags the face with the name inferred from the user's previous manual tags.

- *Challenges in understanding AI capabilities*: The system's performance and error modes are likely to change as it learns from new images and tags. Therefore it is difficult to anticipate what the system can reliably do, when and how it is likely to fail. This, in turn, makes it difficult to design appropriate interactions for these scenarios.
- *Challenges in envisioning novel and technically feasible designs of the technology*: Re-imagining many new uses of a face-recognition-and-tagging tool – beyond tagging people on photos – can be difficult. This is because its capabilities are highly evolved and specialized for its intended functionality and interactions.
- *Challenges in iterative prototyping and testing:* The system's capabilities evolve over time as users contribute more images and manually tags, challenging the very idea of rapid prototyping.
- *Challenges in collaborating with engineers.* The system requires a closer and more complex HCI-AI collaboration than as in a traditional double-diamond process. Engineers and designers need to collaborate on understanding how the face-recognition performance will evolve *with* users' newly uploaded photos and tags, how to mitigate the AI's potential biases and errors, as well as how to detect AI errors from user interactions so as to improve system learning.

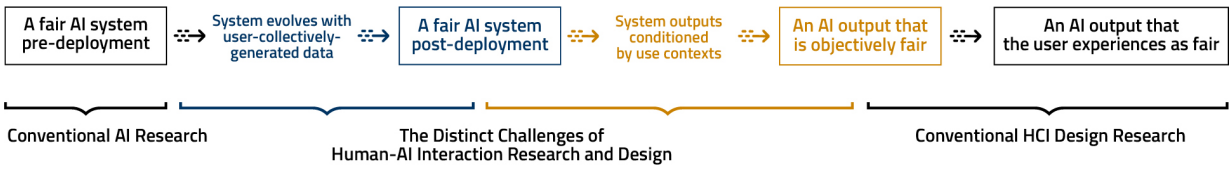

Figure 3.4: An example of the framework in use. Using the framework, researchers can easily outline the problem space of a human-AI interaction issue of their interest, for example, the issue of AI fairness.

Face recognition and tagging are a relatively mature technology that many people use every day. However, because its capabilities are constantly evolving and the outputs are diverse, systematically sketching, prototyping, and assessing the UX of face tagging remains challenging. This exemplifies Level 4 systems. These are opportune areas for HCI and RtD researchers to study human-AI interaction and design, without getting deeply involved in technological complexities.

3.3.2 The Anatomy of AI's HCI Issue

For researchers who study specific human-AI interaction design issues (e.g., fairness, intelligibility, users' sense of control, etc.), the proposed framework gives a preliminary structure to these vast issues. Take as an example the challenges surrounding *accounting for AI biases*, a challenge that many critical AI systems face across application domains such as healthcare and criminal justice. Building a "fair" AI application is widely considered as difficult, due to the complexity both in defining fairness goals, in detecting underlying biases, and in algorithmically achieving the defined goals. Prior research has been addressing these challenges by promoting interaction design guidelines [2, 72].

The framework provides a more holistic structure to the problem space of "AI fairness" (Figure 3.4). It illustrates that the current work has mostly focused on building "a fair AI system pre-deployment"; that algorithmic fairness is only part of the whole "AI fairness" problem space. There is a real need for HCI and AI research in collaboratively translating fairness as an optimization problem into a feature of AI the socio-technical system (Figure 3.4, blue segment), and into a situated, user experience of fairness (yellow segment). The framework suggests a tentative agenda for these important future research topics.

3.3.3 Implications for Design Methods and Tools

Finally, the proposed framework intends to allow for a more principled discussion on how research might support the UX design activities of AI (i.e., sketching and prototyping). It can help researchers to

articulate the contribution of their emergent AI design methods/tools/workflows as well as their scope of generalizability. Finally, it can provide new insights into how to address the remaining challenges. Consider UX prototyping methods of AI as an example.

1. Identifying root challenges. Current research typically attributes the difficulty of prototyping AI to AI's technical complexity or reliance on big data. However, HCI routinely grapples with complex, resource-intensive technologies using simple prototypes. What makes AI unique? The framework suggests that the root challenges are that AI's capabilities are adaptive and its outputs can autonomously diverge at a massive scale. Such systems problematize the conventional HCI prototyping methods that treat technology's affordance as bounded and interactions prescriptive. These methods can work when prototyping AI as an optimization system in the lab (Level 1). They could fail in fully addressing AI's ramifications over time as a real-world, sociotechnical system.

2. Articulating the contributions and limits of emergent design methods/tools/processes. To make prototyping human-AI interaction easier, researchers have created simple-rule-based simulators [13, 124]) as AI prototyping tools. Mapping the characteristics of rule-based interactions onto the AI design complexity map (Figure 3.3), it becomes evident that rule-based simulators are most effective in prototyping level 1-2 systems. They can be particularly valuable for systems that generate a broad set of outputs (level 2) where traditional, manual prototyping methods struggle. However, rule-based simulators cannot easily prototype systems that autonomously learn from user-generated data (level 3-4). These are living, sociotechnical systems; the rules that map their inputs to outputs evolve in complex ways over time.

3. Providing new insights for future research. Framing level 3 and 4 AI systems as living, sociotechnical systems reveal new insights into how we might more effectively prototype their interactions. For example, computer-supported cooperative work (CSCW) research has investigated how to prototype workplace knowledge sharing systems whose affordance co-evolves with its users' behaviors, the interactions among its users, and the organizational contexts at large [62]. These are living, sociotechnical systems with uncertain capabilities and complex outputs as well. This body of work, though not typically considered as related to AI, could offer a valuable starting place for considering how we might design prototype human-AI interactions in the wild, over time. In this light, the proposed conceptual framework offers actionable insights for addressing the challenges of prototyping AI methodologically.

Chapter 4

A Case Study of Designing Level One Systems (Probabilistic Systems)

In the previous chapter, I have identified two sources of AI's design complexities (namely, capability uncertainty and output complexity) and presented a framework that illustrates their effects on the design processes of four levels of AI systems. Level one systems learn from a self-contained dataset and produce a small, fixed set of out-puts. I have argued that fairly traditional UX design methods can help elicit, address, and evaluate the UX issues of these systems. More difficult are level four systems, systems that learn from new data even after deployment.

This Chapter describes the design process of a level one system (probabilistic system). It is a clinical decision support tool (DST) that aids artificial heart implant patient selection, a life-and-death decision. Currently, all clinically deployed AI and ML systems cannot learn from new training data without additional Food and Drug Administration (FDA) approval [37]. Therefore, clinically-deployed DSTs are by definition level one systems.

I first describe my user-centered design process, using a set of commonplace design methods. This process led to a near-future solution to the long-standing challenge of situating DSTs in doctors' critical decison making processes (section 4.3-4.4). I reflect on the virtue of the UX design processes and methods for designing level one systems, as well as opportunities for further research (section 4.5).

4.1 Designing a Decision Support Tool for Artificial Heart Implant

The idea of leveraging machine intelligence in healthcare in the form of decision support tools (DSTs) has fascinated healthcare and AI researchers for decades. With the adoption of electronic medical records and the explosive technical advances in machine learning (ML) in recent years, now seems a perfect time for DSTs to impact healthcare practice.

Interestingly, despite their success in labs, the vast majority of DSTs struggled when they moved to clinical practice [32, 53, 55]. Clinicians rarely use them [29, 32, 129]. Even when they do, the performance of the clinician-DST team rarely exceeds that of the clinicians alone [5, 6, 64]. In a review of deployed DSTs, healthcare researchers ranked the lack of HCI considerations as the most likely reason for failure [80, 126]. This includes a lack of consideration for clinicians' workflow and the collaborative nature of clinical work. However, little HCI research has studied the context of healthcare decision making with a focus on how to best integrate and situate a DST. Few studies that investigated DST in use are lab studies; instead, studies have often substituted undergraduate students for patients and medical students for clinicians. [107].

I collaborated with biomedical researchers on the design of a DST supporting the decision to implant an artificial heart. The artificial heart, VAD (ventricular assist device), is an implantable electromechanical device used to partially replace heart function. For many end-stage heart failure patients who are not eligible for or able to receive a heart transplant, VADs offer the only chance to extend their lives. Unfortunately, many patients who received VADs die shortly after the implant [8]. Modern DSTs can learn from previous implant outcomes (as in MIMIC II Databases [99]) and predict the likely trajectory a patient will take post-implant. Such a DST should help identify the patients who are mostly likely to benefit from the therapy.

Like almost all other prognostic DSTs, it currently takes a context-less, prototypical form: It takes in a list of patient condition measures and produces an individualized prediction of patient trajectory, such as likely post-implant life expectancy [7]. Given the known challenges of DST deployment and the wide gap between DST technology and clinical reality, I followed a user-centered design process to designing the heart-implant DST. This effort unfolded in three stages.

1. *Understanding clinical reality:* I first conducted a field study at three hospitals. I wanted to better

understand the clinical decision process around a VAD implant and the clinical reality in general, so as to identify relevant design requirements and the key touch points where I might situate a prognostic DST that clinicians would find useful in their practice.

2. *Designing a new DST for clinical practice:* I wanted to offer a concrete solution to the long-standing challenge of effectively situating DSTs in clinical practice.

3. *Field evaluation of the new design:* I wanted to probe clinicians' responses to the new DST design, when it is situated in their day-to-day workflow. I wanted to explore whether the insights and the design would likely to generalize to other clinical decisions beyond artificial heart implant patient selection.

4.2 Understanding Clinical Reality

4.2.1 Field Study Design

In the first stage of the project, I wanted to understand how the decision making process to implant a VAD unfolds in the clinical environment. I wanted to know who participates and where decision-making happens, and to probe on when clinicians think an intelligent system might offer support for their work. I wanted to identify contextual barriers that might prevent people from engaging with a DST and to identify the times and places it might add the most value.

To address these needs, I chose to conduct a qualitative field study consisting of observations and semi-structured interviews. I chose an ethnographic approach so as to capture the richness of context, and also because this has become a standard HCI approach when designing new software systems meant to improve work. We analyzed our data using affinity diagrams [73] and by creating a service blueprint [10] that documents the decision pathway for individual patients.

I carried out this research at three different implant hospitals all in the United States, all of which regularly perform VAD implantation. In two of the hospitals we performed interviews and observations. In the third, we only performed interviews, as we could not secure permission to make observations for legal and privacy reasons. The three facilities vary geographically and in scale. Their performance rankings range from top 5 to top 60 in the United States. Despite great inter-site differences we observed, I report findings that all three facilities share.

- Hospital 1: large-scale service performing over 60 heart transplants and over 100 VAD implants per year;
- Hospital 2: moderate-sized service performing over 20 heart transplants and over 30 VAD implants per year;
- Hospital 3: relatively small service performing about 20 heart transplants and 40-50 VAD implants per year;

I conducted observations in two Advanced Heart Failure services for 6 to 14 hours a day for 13 days. The observed VAD teams cared for approximately 75 patients who were formally or informally being considered for an implant. I followed attending cardiologists across all decision-related settings including morning rounds, clinician-patient consultations, clinician-to-clinician conversations, and weekly implant meetings. I observed out-patients from both General and Advanced Heart Failure clinics and in-patients from Advanced Heart Failure wards, Intensive Care Units, and Emergency Rooms.

I conducted IRB approved interviews with a total of 24 VAD clinical team members from 3 hospitals, covering many different roles and statuses that participate in decision-making. Interviewees were chosen according to their level of involvement in VAD decision-making. My research collaborators at each hospital recommended an initial set of interviewees. I then expanded this set by recruiting others we observed to play important roles in the decision-making. I confirmed our findings with a VAD cardiologist, a mid-level resident intern, and a VAD coordinator. Field notes were recorded using pen and paper. Interviews were audio-recorded and transcribed.

Below, I first give an overview of the decision process around a VAD implant, including the participants and their work practices. I then highlight the decision-makers' needs for decision support given the informational, social, and environmental contexts where the decisions get made. These needs are sometimes in tension with the stand-alone, walk-up-and-use form that existing DSTs typically take.

4.2.2 Overview of the Observed Decision Landscape

The clinical decision to implant a VAD involves many clinician roles and unfolds across many clinical contexts. Table 4.1 provides a high-level summary of the decision-makers and contexts.

The clinical environment is extremely hierarchical; however, it is also collaborative across status lev-

els. While many roles contribute to and execute on the implant decision, only a small and stable coalition has a final say. I refer to these ultimate decision-makers as *implant physicians*. These are mostly cardiologists, though at some sites surgeons and/or senior nurse practitioners also participate. The *midlevels* refer to other clinical members of the VAD team and also the non-clinical members who focus on insurance, social support, and VAD-related care coordination. The *consults* include other support services and physicians outside of the implant team. Implant physicians function at the top of the hierarchy, leading major decision-related activities. They decide who transitions from clinic to hospitalization and who gets classified as a difficult case and gets being discussed at an implant meeting.

At *clinics*, implant physicians monitor out-patients and hospitalize them for a formal VAD evaluation. When an out-patient gets hospitalized and becomes an in-patient, a group of clinicians visit the patient every morning during *rounds*: they visit each patient after a brief deliberation in the hallway outside the patient's room, where they establish a care plan for the day. The attending cardiologist of the week picks and presents the "difficult" cases during a *weekly implant meeting*, where all available clinicians can voice their opinions. The attending cardiologist and surgeon take away a collective decision for each presented case. If approved for implant, they pick a surgery date. They may stop the procedure if a patient's condition changes prior to surgery.

4.2.3 Potential Barriers of DST Adoption and Use

I observed many barriers that could negatively impact the use and perceived value of a prognostic DTS situated in VAD implant hospitals.

4.2.3.1 Attitudinal barrier

First and most importantly, VAD physicians expressed no desire for a prognostic decision support. They view the decision to implant a VAD as easy: As long as patients have no definitive exclusion conditions, they will all get a VAD after failing on an identical, escalating sequence of less aggressive treatments. Under this strategy, clinicians thoughtfully order tests to detect red flags, and then deliberately and iteratively adjust daily medications to resolve the red flags. They spend much more time on daily care decisions than on the implant decision itself.

In factoring patient condition to implant decisions, physicians' tried-and-true precedence works for

Participating Clinicians		Decision-Making Procedure		
		Clinic	*Ward Round*	*Weekly Meeting*
Implant Physicians	Cardiologists	■	■	■
	Surgeons	□	□	■
Medical Midlevels	Nurse Practitioners	□	□	■
	Fellow & Interns	□	■	□
	Physician Assistant	□		
	Registered Nurses		□	
Social Midlevels	VAD Coordinators			■
	Finance Coordinator			
	Social Workers			■
	Palliative Care			■
Consults	Pharmacists			
	Nutritionists		*On Demand*	
	Other physicians			

Table 4.1: Clinicians and activities of a VAD implant team. They unequally participate in routine decision-making activities. ■ marks the clinicians who lead or always attend the activity; □ marks those who attends occasionally or in a subset of hospital sites.

the majority of their cases. For the grey cases, implant physicians did not imagine that algorithmic predictions would help. While all physicians knew about the availability of VAD risk models, none used them in practice. Physicians' rationale for not using these models presented a number of barriers that a prognostic DST would likely face.

Clinicians know how to do their jobs. As trivial as it sounds, it is a missing perspective in DST literature that has instead focused on the clinicians as a source of errors, biases, overconfidence, and communication breakdowns. This assumption behind DST development and design, though not immediately evident in interfaces, perhaps seeds the attitudinal barrier I observed. Many of interview participants implied that makers of current prognostic systems want to replace their expertise with inhuman technology.

4.2.3.2 Informational Barrier

The commonly assumed function of a prognostic DST is to predict the likely post-implant life expectancy based on a list of quantitative measures. I observed a mismatch between clinicians' information needs and such a DST information flow.

At the input end, DSTs take in quantitative and explicit inputs, while challenging decisions are often characterized by unavailable or ambiguous medical and/or social evidence. Clinicians are unlikely to use a tool that only does the easiest part of their job; telling them a textbook case is, "textbook". Even if they approach the system when facing a difficult case, they might find it difficult to fill in some of the blanks, such as diagnosis for an emergency-room-path patient. They might find the information that most concerns them is not captured in the prediction, such as the patient's home life and social support, which are critical and difficult factors most often not captured in the medical history.

In terms of DST output, physicians need support for action taking. Consultation between cardiologist and surgeon best captures this: *Is this case too risky to operate on? No? Ok, then do it.* A probabilistic prediction can be obscure in telling whether to execute a therapy or not, to do it now or to "wait and see". DSTs only predict outcomes of "conducting a therapy now", with little sense of waiting and seeing.

4.2.3.3 Social Barrier

When faced with difficult cases, implant physicians turned to their colleagues. The consultative collaborations were frequent and clinicians generally found them efficient and effective. Implant physicians relied on teamwork. Within a shift cycle, one attending cardiologist cares for all in-patients: often more than 40 patients per week. Each patient gets assigned a primary nurse and resident intern who prepare information and monitor unfolding situations. The nurse and intern handle all reporting and documentation, and they prevent patients from falling though the cracks. Cardiologists also consult surgeons for surgical risks, and pharmacists for nuanced medication changes. For patients with other organ complications, they turn to physicians with corresponding expertise.

Attending cardiologists fluently integrate inputs from colleagues through various routine and ad-hoc activities. During rounds, for example, they request midlevel follow-ups right after visiting a patient; they call other cardiologists whenever a problem emerges; they always consult pharmacists right after rounds and before ordering medications. Unlike EMR use, these collaborations happen when and where decisions get made. The implant physicians trust this social decision support process; they often immediately act on their colleagues' input.

Such a hierarchical but collaborative clinical culture poses a two-fold challenge for DST use. First, decision makers (physicians) and computer users (the midlevels) rarely overlap at the point of decision-

making. Second, physicians have great trust in their social network of other physicians, who help them make more difficult decisions. It seems unlikely they will move towards computational support and away from social support when things are difficult.

4.2.3.4 Environmental Barrier

Interaction with computers presents many challenges in a ward environment where most in-patient clinical decisions happen. During the 4-to-6-hour rounds, clinicians visit more than 30 patient rooms. They are constantly moving and conversing, logging in and out of the EMR. Everything they have with them must fit into their pockets because before and after visiting each patient's room they must wash their hands, and sometimes put on and take off disposal gowns and gloves as well [71].

These barriers naturally stratified across decision makers and computer users. For example, cardiologists give oral orders during meetings with patients, and a midlevel will take notes and enter them into EMR at a later time. A few midlevels would carry a computer with them when rounding. They often skipped the in-room patient conversations because of the hassle hand washing presented. As a result, *almost no decision-making ever takes place in front of a computer.*

4.3 Designing an "Unremarkable" AI

The observations above forced me to reflect on the traditional forms most prognostic DSTs take. Most require clinicians to recognize when computational advice would be useful and then make an explicit effort to access a DST [80]. In addition, most imagine a single decision maker participating in making the decision at a single time and place [116].

With the field observation findings in mind, I set out to design a new form of DST for implant patient selection and, more broadly, explore how to overcome its real-world adoption barriers that many prognostic DSTs face. I had two design goals:

1 - *Embedding DST in current workflow*: Clinicians, especially cardiologists and surgeons, need to naturally encounter the DST within their current decision-making workflow, because they are unlikely to recognize when they might need help and then walk up to a computer for help;

2 - *Slowing down decision-making only when necessary*: The DST outputs need to be easily ignored

in most patient cases that are textbook. However, it should also be present enough to slow the decision-making down when there is a meaningful disagreement between the clinicians' view and the DSTs view of the situation;

These views are very different from the convention of DST design in which decision supports are always available, waiting for clinicians to walk up and use at any point across the decision-making process. Instead, I wanted to tailor the DST for particular moments in the process, such that clinicians do not have to take pause and invent sequences of action anew. I wanted the DST to naturally augment the actions of decision making, rather than pulling the user away from doing their routine work.

4.3.1 Making Clinical DST Unremarkable

Tolmie et al. [120] introduced the notion of unremarkable computing when discussing how ubiquitous computing should arrive and create its place in people's homes. They argued that technology can augment people's actions in ways that have a wealth of significance but seem unremarkable, because its interactions are "so highly situated, so fitting, so natural". They argued that home technology should not only be more intelligent, it should also be more subservient to people's daily routines. In doing so, the technology becomes part of the routines, part of the very glue of their everyday life.

I draw connections between this ambition and my aforementioned design goals. I also draw connections between this notion of routine and VAD decision making. While these are daunting life-and-death decisions, the implant decisions are part of a work routine for clinicians. To fit into their practice, the DST needs to be subservient to the day-to-day decision-making workflow they engage in.

I wanted to operationalize this idea of unremarkable technology in the context of critical, clinical decision making. This is a difficult goal because it requires a right level of "unremarkableness" such that the DST does not constrain clinicians' decision making flow *except when it needs to.*

4.3.2 Design Process

To situate a DST into the current VAD decision-making routine, I first needed to identify a time and place where clinicians should naturally and impactfully encounter it. I chose the multidisciplinary patient evaluation meetings, for a number of reasons. First, the meeting is a rare social touch point where most clinicians involved in the decision are present, and they are actively forming a collective decision about

patient treatment. Second, it is one of the few decision points where a computer is present and being used. Third, decision meetings are common across hospital sites. VAD centers in the US are legally required to take a multidisciplinary approach to patient care, therefore regularly scheduled meetings are common. Globally, these meetings are also recommended [110]. Fourth, multidisciplinary meetings have become an increasingly common best practice in organ transplantation [94]. Designing DST for decision meetings therefore could potentially generalize beyond VADs to include a number of other clinical decisions.

Next, I considered how to fit the DST comfortably within the meetings. Drawing lessons from prior work [90, 133], I wanted to embed the DST into Electronic Medical Records (EMR) to minimize the effort needed from clinicians to type in patient information. I also wanted to augment clinicians' paperwork to provide them additional motivation for adoption. I therefore integrated the DST output into a meeting slide generator, a system that automatically extracts patient information from EMR and populates slides for the decision meeting, which could be projected or printed.

I sketched what the DST predictions output might look like. I iterated on the design based on feedback of two collaborating clinicians (an attending cardiologist and a nurse practitioner). The final design was a small line chart that showed a patient's predicted chance of survival (Figure 4.1). It also showed the most likely causes of death, such as right ventricular failure or renal failure. These predictions *inform* clinicians' discussion about the implant decision, rather than indicating the decision to them. Taking a lesson from early HCI work in participatory design, I need to make technical advances that skill workers instead of de-skilling them [31].

I placed this chart in the top-right corner of the slide summarizing an individual patient's current state. The subtlety was a deliberate choice toward achieving the right level of unremarkableness. In the most common case, when the DST agreed with the clinicians' assessment, the visual display of the agreement could help clinicians gain trust in the system without slowing them down. In the rare case that the DST prediction conflicted with the clinicians' assessment, the DST could slow the decision down. Everyone attending the meeting would see the disagreement. I speculated this would apply social pressure on the senior physicians to rationalize and articulate their decision making. I speculated it could also encourage the medical students, residents and other mid-level clinicians to participate in the discussion when they disagreed with the senior clinician's decision. It could allow them to disagree by

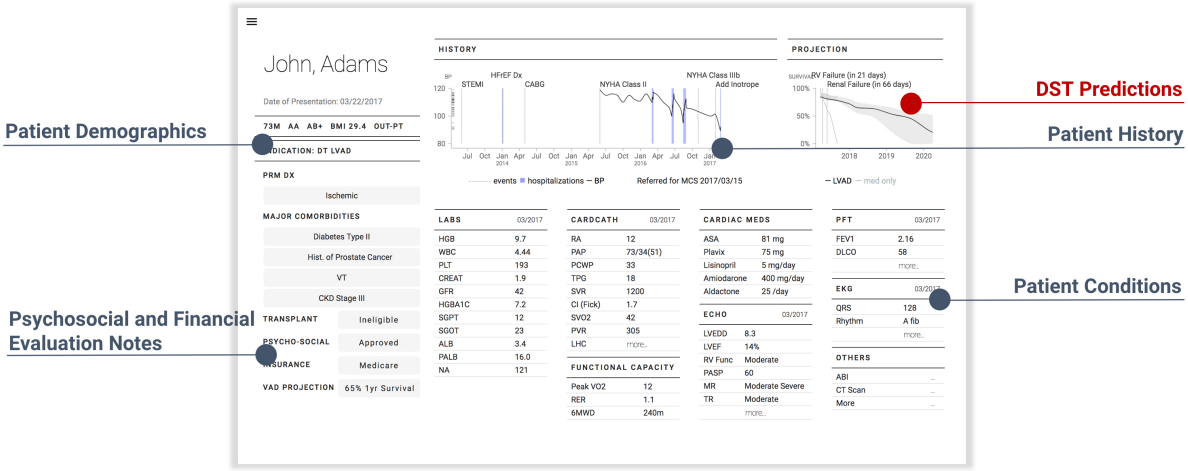

Figure 4.1: The decision meeting slide design. We designed a DST that automatically generates decision-meeting slides for clinicians with subtly embedded machine prognostics at the top right corner.

pointing to the conflict with the DST and not claiming that they personally knew more than the senior physician.

I worked out the detailed contents of the slide with the two collaborating clinicians. I also referenced the meeting printouts and workup checklists currently in use.

I wanted to finalize the design by populating with real patient data. However, a variety of policies and legal regulations would not allow this. As a work-around, I asked our clinical collaborators to help us populate the slides with synthetic patient cases. Interestingly, they found it very challenging to generate a prototypical patient case including dozens of vital signs and test results. They instead selected elements across several of their former patient cases, removing identifiable demographic information and molding parts of the medical condition to disguise the identity.

In my final design (Figure 4.1), the DST outputs are in the top right corner of the slide, next to a summarized patient history visualization. Patient test results are categorized and put in the center. The patient demographics and links to social and financial evaluations are on the left.

4.4 Experience Prototyping and Evaluation

Next, I conducted a field evaluation of the new DST design. I had several questions I wanted to answer with the assessment, including: (1) Would clinicians naturally encounter the DST within their

current workflow? (2) Would clinicians accept computational decision support in the public context of the meeting? (3) Does placing the prediction in the corner present the right amount of unremarkability? Specifically, does the DST get ignored when its predictions align with the clinicians' judgment, and would it slow decisions down when its output conflicts with clinicians?

4.4.1 Methods

4.4.1.1 Assessment in VAD Implant Centers

I gained access to three U.S. hospitals that regularly perform VAD implantation. Two were sites from our formative field study and one was new. The facilities varied geographically and in scale. The smallest we studied performs about 40 VAD implants a year; the largest performs over 100.

I wanted to assess our design within the context of an actual implant decision meeting in order to observe whether it impacted discussion. Unfortunately, this proved to be impractical. None of the sites would allow me to present slides showing information for the patients they were currently implanting. All felt this could impact the life and death decision. The clinicians doing the VAD implants were quite busy. They would only agree to interact with a single design. They did not have the time for me to make revisions and then revisit. Finally, one of the sites had a specific policy preventing us from observing the decision meeting. They would only participate in one-on-one interviews.

In reaction to these restrictions, I re-designed the evaluation process with the goal of making the most use of the participant pool within one round of assessment. I carried out all following procedures in hospital C. In hospital B, I carried out all except (3) presenting at a decision meeting. In hospital A, I carried out all procedures except (4) interviewing all physicians and surgeons.

(1) At each site, I first interviewed the mid-levels to understand their practice around the decision meeting, and to probe the DST design's fit in their respective hospitals. When necessary, I adjusted the designs to fit specific hospital's routine practice;

(2) My research collaborator at each site recommended one attending physician to be our confederate. I conducted interviews with them, discussing the DST design and confirming there was no glaring mismatch between the design and the practice at their respective sites;

(3) The confederate physician presented the patient case with the DST on display in the decision meeting. I observed clinicians' responses and discussions;

(4) Finally, I interviewed the rest of the VAD team to further individually discuss the DST design.

In total, I interviewed nine attending cardiologists or surgeons and eight mid-level clinicians. Each interview lasted for at least one hour. The DST design was presented in two hospitals' multidisciplinary decision meetings. Field notes were recorded using pen and paper. Interviews were audio-recorded and transcribed. I analyzed the data using affinity diagrams [73] and by performing thematic analysis.

4.4.1.2 Assessing Generalizability of the DST Design

I chose to situate the DST within slides used for decision meetings partially because these meetings are best practices in other critical medical domains as well. To gain some insights as to if this design might generalize, I chose to probe a small set of clinicians from other medical domains who participate in these meetings.

To recruit these participants, I asked participants from the VAD study to help me identify other clinical domains and decisions that have interdisciplinary decision meetings. I then interviewed 6 physicians from these domains. Their practices include decisions meetings for pediatric surgery, pediatric critical care, adult cardio-thoracic surgery, internal medicine emergency care, orthopedic surgery, and obstetrics/gynecology. I audio-recorded, transcribed and analyzed these interviews using the same methods as we used for our VAD participants.

4.4.2 Findings

4.4.2.1 Validating the New Design Goal of "Unremarkable" AI

My observations suggested that most clinicians involved in the VAD implant decision would likely encounter the DST output if it was included as part of an individual patient's information presented at the decision meeting. All three facilities hosted a weekly implant decision meeting. Clinicians of all ranks and roles attended, ranging from seasoned surgeons to residents, to nurse practitioners to social workers to palliative care coordinators. Although the weight that the meetings carried for influencing an implant decision appeared to vary across the three sites, the occurrence of the meetings was one of the few events that happened everywhere.

These meetings offered one of the extremely few situations where senior clinicians actively discussed decisions in proximity of a computer. Meetings in all three hospitals had a shared computer projecting

patient information. Two hospitals projected dedicated meeting materials. The other projected patient profiles from the EMR. Clinicians described the other key decision points as "*just talk on the fly*" with no EMR access or paper records in hand. The other decision points most often only included attending physicians and surgeons. "*Everything is happening live.*" Mid-level clinicians, who spend more time with each individual patient, did not participate in the decisions made outside of the meeting.

None of my interview participants expressed any resistance to the including DST output within the context of the decision meeting. One site (Hospital C) had already made the effort to manually include DST data into their meeting but had abandoned this practice due to their loss of confidence in its quality. Seasoned physicians and surgeons voiced their appreciation for what a prognostic DST might bring, stating that it would "*give its perspective*" and offer a chance for an "*occasional recalibration.*" Clinicians also shared that making an objective decision could sometimes be hard. The decision to not implant was usually a death sentence for a patient. "*When I really like this patient, really want to help him or her, it sometimes helps to get a more factual view.*" (Cardiologist, B5)

Seasoned physicians shared that their dream DST should play a role similar to mid-level clinicians. They should provide additional context for the seasoned physicians' decision. The DST could provide additional context and a different perspective to the senior physicians. They recognized the value a DST might bring from its statistical consideration across many cases. "*The value is you are looking at thousands of cases, I'm looking at 100 and overweighting the last three I saw.*" They also shared that input from mid-levels was not always "*taken really into account*".

Mid-levels agreed they only inform and support the discussions. They did not make decisions.

> *My role in selecting patients for VAD... hmm. I don't select patients. But I do talk about it... We are there to help discuss patients.* (Nurse practitioner, B2)

> *A lot of what I do in that meeting is to give people perspective and context.* (VAD Coordinator, B1 and A3)

Mid-level clinicians enthusiastically welcomed the idea of a decision meeting slide generator. They envisioned a number of possible benefits. They shared that the slide generator would automate work that is not currently billable. At hospital A and B, meeting slides were prepared by staff who had little to no medical training. Physicians could get frustrated with the result, characterizing the unfiltered mate-

rials as being prepared by "amateurs." These staff members could not personalize patient presentations because they could not risk skipping information that might prove to be critical. Mid-levels felt they could benefit from the automation and seasoned physicians felt they would benefit by the removal of the copious, irrelevant data being pulled out of the EMR.

Mid-level clinicians viewed the slides as a potentially important vehicle for communicating their opinions to physicians. In all three hospitals, senior physicians set the agenda for decision meetings. They decided which patients to present, and during the meeting, they called out the information that they felt was important enough to discuss. This hierarchical culture was well captured by the design of a custom patient review tool at hospital C. Two VAD coordinators customized a patient review dashboard within EMR in order to help themselves better track medical tests and share results within the team. Although cardiologists and surgeons rarely used the tool, they controlled which pieces of information could be placed on the dashboard and which elements would not be included when the patient case was classified as urgent.

Mid-levels often doubted that their voice was heard or that their expertise was considered. They were hesitant to directly disagree with a physician. They described the situation as more complicated than just the power dynamics. They shared that the cardiologists were incentivized to implant more patients and to implant sicker patients. They found themselves often advocating for patient mortality (let the patient die). Mid-levels felt their opinions focused on post-implant quality of life. Unlike the physicians, mid-levels worked intimately "*with all the problems that can come from a patient that maybe shouldn't have been implanted.*" They noted there was no right or wrong answer between length of life and quality of life. They shared it was often hard to argue with great confidence that letting patients die was better than offering them a small chance to live. In such situations, mid-levels frequently cited "*you never know what will happen*" as a reason to not to pursue further discussion with attending physicians. Some shared that over time, they had slowly removed themselves from the decision making.

> *There is risk stratification for each patient, but I don't know... It's like, we talk about it, but I don't know if it's really taken really into account.* (Nurse practitioner, B2)

Mid-levels consider the ability to organize the contents of meeting slides as one way to increase their influence. Meeting slides provide additional, visual presence they could use in support of the facts they felt were important. This would make it less like they were only sharing an opinion with the physicians.

The meeting slides could be facts in a space where only the seasoned physicians' opinions carried any weight. They felt the formality the meeting slides carried was unparalleled to any other artifact they had access to. A prognostic DST that indicates post-surgery quality of life could potentially amplify their voices.

> *There is not a way to present (my reasoning) formally. It's just me saying: 'This, this and this'. [...] I think it's good to have something visual for anybody to see. It's like, OK. LOOK. Let's slow down a bit here.* (Nurse practitioner)

4.4.2.2 Intricacies of Designing a Right Level of Remarkableness

Both seasoned physicians and mid-levels expressed appreciation for DSTs that could slow them down "only when necessary". They liked this aspect of our design. Furthermore, clinicians' discussions and questions depicted many unexpected intricacies in this notion of the "right" level of unremarkableness. These discussions offer valuable insights for further refinement of the new DST design.

- *Is the Model Validated by Clinical Trials?* Clinicians commonly expressed a need to know more about the model's source and credibility. When they learned that the model presented has not been rigorously validated through clinical trials and published in prestigious clinical journals, they suggested I was wasting their time. Physicians also desired a model that had been validated with data from their own hospital. "*It's better to be home-grown.*"
- *Are the Predictions Based on Clinicians' Best Efforts?* Physicians highlighted that the predictive models, regardless of how well they measure medical uncertainties, would never replace human, clinical decision-making. They viewed their own decision making as focused on managing and reducing uncertainties. "*If we think that we will be able to tell everybody what to do based on a model, we ignore the fact that we also have tools and mechanisms for dealing with the uncertainty that is inherent when putting VADs in patients.*" (Cardiologist)
- *Does DST Prediction Mean Causality or Correlation?* There was a sense that if the DST predictions were not based on causal factors, then the predictions should not be presented at all. Clinicians described differentiating correlation (predication) versus causality as a central part of their clinical decision making.

- *Are Data-Driven Prognostics Facts OR Predictions?* Clinicians frequently asked us to clarify whether DST prognostics are predictions that carry agency and subjectivity, or if predictions are facts rooted in historic data. I sensed they wanted to limit discussions to facts, including how heart failure has played out for the patient they were treating and the statistics from previous, similar cases.
- *Are the Predictions Individual Medicine OR Population Medicine?* Most clinicians seemed to find the notion of personalized predictions difficult to grasp. Some voiced strong concerns that using DST was the same as applying "*populational statistics*" to individual patient decision making.
- *What Does "Now" Mean in DST Predictions?* The DST visualized the patient outcome predictions. For example, it shows that the patient's post-implant life expectancy is 21 days if a VAD was implanted now, under the condition shown on the slides. Clinicians were confused by this notion of "now" because it was extremely unlikely that they would implant a patient on the same day as the decision meeting. "*Is that 21 days from today? If we are gonna lose the patient in 21 days [21 days following after implant], can we just wait?*"
- *DSTs Do Not Account For the X Factors.* Clinicians said that the DST would only ever be one factor in their decision because of "*X factors*"; the many factors beyond a patient's condition that impacts the implant decision. One X factors they spoke of was O/E ratio (observed-to-expected mortality ratio). The O/E ratio is a rating that measures the surgeon and care teams' performance. Surgeons cared about keeping a high rating. They described the implant decision for high-risk patients as "*taking on new O/E ratio debts.*" This seemed to strongly influence whether they take on another high-risk patient. It seemed to depend strongly on how many patients had recently had poor outcomes.

4.4.2.3 Generalizability Beyond Aritificial Heart Implant

The interviews with clinicians outside of VAD centers showed that multidisciplinary decision meetings take place across many clinical domains for some of their most aggressive interventions. They are also referred to as internal medicine panel meetings, tumor boards, or floor meetings (referring to meetings between critical and general care physicians). These meetings happen widely because for patients are very sick and are being considered for their last-option surgical intervention, their illness usually have involved multiple organs. Treating them requires physicians from multiple clinical domains. Multidis-

ciplinary meetings therefore occurred naturally.

> *Esophageal cancer, COPD, diabetes, cystic fibrosis, LITERALLY everything in psychiatry, gastric bypass, end stage renal disease, hernia repair, syndromes like Down and Turner, any disease that requires management with meds with nasty side effects, and even emergency room situations to expedite processes. Any of the above diseases the approach has to be multidisciplinary almost by definition because they affect multiple systems and usually but not always the last option is a surgical intervention.* (Pediatric surgeon)

To summarize, the findings of the field evaluation suggested that an "unremarkable" DST may more effectively fit into clinical practice, as it can naturally augment clinicians' current routine of decision making, rather than pulling them away from it. Taking lessons from prior HCI work, we should not only make AI more intelligent, but make them highly situated in people's routines. In doing so, AI can become part of the decision-making routines, part of the very glue of clinicians' everyday work. The DST as a meeting slide generator offers an initial design examplar in this new direction.

4.5 Reflecting on UX Design Expertise and Methods for AI

Next, I want to take a step back and examine this case study from a UX design method and process perspective. The design process I followed (user study, design, and then user testing) is not new. Nor are the design methods used (fieldwork, simulation-based prototype and testing). Yet, this classic approach provided a solution to the long-standing challenge of enabling clinician-AI collaborative decision making. What has previous research missed? What can future research in designing AI draw from this case study?

4.5.1 Designing the User *Experience*, rather than the Usability, of AI

DSTs, despite compelling evidence of their effectiveness in lab studies, have often failed in clinical practice, in improving patient outcomes [5, 19]. Prior research investigated this challenge mainly through lab studies, in part because of the restricted access to the clinical environment [25, 115, 119, 131]. Most lab studies took *what will a DST predict* as a given, and focused on other critical issues including better information presentation and visualization, the accuracy of risk communication, trustworthiness, ease

of use for medical information, etc. Borrowing language from Buxton [18], most prior works focused on designing the predictions right, rather than considering what are the right predictions to make for clinicians in the first place.

This project focused mostly on identifying the right AI thing to offer clinicians. What predictions, if any, do clinicians desire and perceive as valuable? Among these desired predictions, which can existing datasets and algorithms reliably offer? What role should the DST place functionally and socially in the clinical decision-making process? Relatedly, when and where should it intervene? Only limited prior work has asked these questions.

My exploration in this previously under-explored problem space revealed new design opportunities. On the AI side (Figure 4.2, top row), end-of-life healthcare decision-making is a very human decision, while data-driven DSTs, even if they offer 100% accurate predictions, embody evidence-based part of the decision making. This realization pushed me to design a DST that informs decision *discussion* rather than the decision itself. It pushed me to think beyond binary life-or-death predictions, and instead try to illustrate risks of post-implant complications and quality-of-life losses, as much as available EMR datasets and algorithmic can offer. On the human side (Figure 4.2, bottom row), I leveraged the rich social context and made it a part of the clinician-DST interaction design. Situated in the decision meetings, a socially aggregated decision point, the DST could leverage mid-level clinicians to advocate for its information and value to the decision-makers. These observations and design ideas are unlikely to come out of usability-focused lab studies or out-of-context clinician interviews.

This shift of focus from usability to UX in DST design is not trivial. I analogize this shift to that from designing desktop computers to designing ubiquitous computing [127]. Both break away from the convention of designing a more user-friendly AI system/desktop computer, and instead try to reshape AI/computation to better fit in people's day-to-day lives. Both mark an HCI paradigm shift that calls for:

1. *New ethnographic and design work* that re-imagines when, where, and how people may perceive AI/computation as valuable;
2. *New technological and interaction design innovations* that together enable AI/computing devices to deliver value when, where, and how it is needed;
3. A renewed understanding of *what "AI"/"computer" is and means* for people and for societies at

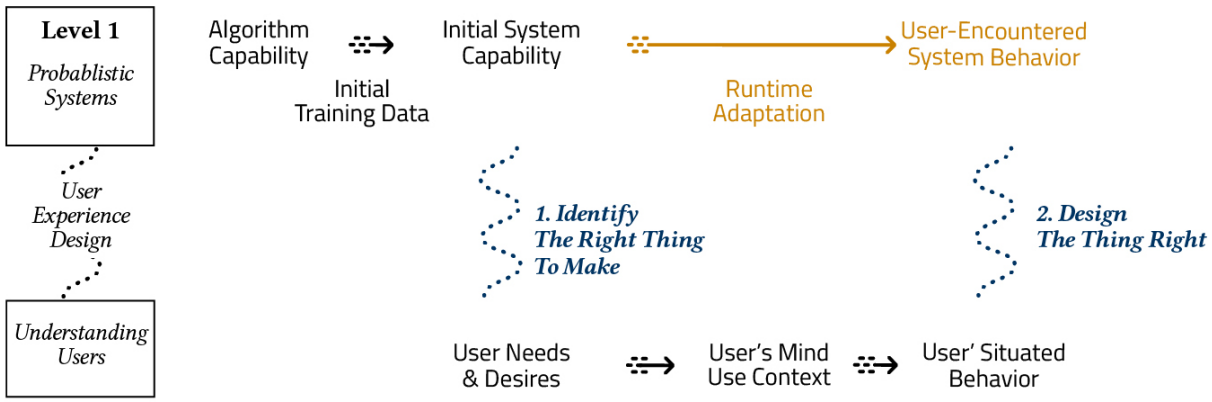

Figure 4.2: case study of level one system design process, against the backdrop of the AI design complexity framework.

large;

Future research shall advance this work by systemically searching for new opportunities in situating AI into people's everyday work and lives meaningfully yet unobtrusively (even when that everyday work is making life-and-death decisions!). The software development community has learned over many years that HCI and UX should be considered early in the development process and not added as an afterthought at the end. Similarly, this works promotes the idea that users' needs, desires, and work contexts should also be considered at the early stages of the AI design and development process.

4.5.2 Experience Prototyping Clinician-AI Interaction

Previous research frequently suggests that human-AI interaction is extremely difficult to "*prototype*", citing AI systems' technical complexity and unanticipated errors as the reason [30, 52]. As a result, simulation-based evaluation of AI systems is uncommon in the research literature. (An exception is prototyping natural language interactions [96], which the next chapter will discuss in detail.)

This case study provides an alternative perspective on this common perception. It demonstrates that many UX design methods (fieldwork, simulation-based prototyping and testing) remained valuable and sufficient for designing level one AI systems – systems that produce simple outputs and do not continue to learn post-deployment. The primary goal of prototyping is to allow designers, users, and other stakeholders to gain a first-hand appreciation of the future conditions the design creates, through

active engagement with the prototype [17]. The simulation-based prototype used in this study achieved this goal because it was able to recreate the critical experiential aspects of the DST use (the social and physical contexts of clinical decision-making) without making a fully working prototype. These early lessons from research on experience prototyping and remain valuable to UX design of AI [17, 66].

Chapter 5

A Case Study of Designing Level 4 Systems (Evolving, Adaptive Systems)

Let us now shift our attention to Level four systems. Level four systems learn from new data even after deployment. They also produce adaptive, open-ended outputs that resist abstraction. I have previously (3.3) argued that level four systems reveal the limits of the UX design methods and processes widely in use today. This chapter demonstrates this. More importantly, I demonstrate that a clearer mapping of AI's design challenges can illuminate opportunities for future research on new design methods and tools.

5.1 Project: Designing an Intelligent Text Editor

This project is a collaboration with a group of NLP researchers on integrating Natural Language Generation (NLG) Systems into a Word document editor, with the goal of improving the authors' writing experience. Prior HCI research has utilized NLG for providing writing assistance in a number of ways, most typically, suggesting next sentences as inspiration [23, 100]. But what functionalities do authors want, and what interaction design can allow the machine generated text seamlessly serve authors' communicative intent?

With these questions in mind, I attempted to rapidly experiment with many tentative NLP design ideas and broadly explore how NLP might improve the authoring experience. I wanted to use story-

boards, UI wireframes, paper prototypes, and other simple, tangible instruments to sketch out early design ideas and probe users' reactions [18, 33]. I soon encountered many unexpected challenges. Common sketching tools and techniques deal with tangible interactions and are ineffective at abstracting the experience of language or a conversation. A number of technical aspects of language intelligence further complicate its UX design. For example, data-driven interactions vary across users, adapt to different contexts, and evolve over time, and it can be difficult for designers to envision such divergent courses of interaction or to visualize using traditional wireframes and prototypes [3, 30, 48, 134].

These challenges led me to explore how to design NLP-powered user experiences, and what sketching and prototyping actually mean in the context of intelligent language interactions. I took a Research through Design approach [145] to our project at hand. My goal was twofold:

1. Provide a rare, first-person account of the sketching and prototyping process of a NLP-based product, as well as an articulation of the challenges we encountered.
2. Explore new design methods and tools for designing intelligent language interactions. My reflection-in-action during this project provides some solutions to these challenges.

In what follows, I offer a first-person account of the intelligent text editor project. I identify five challenges that are central to designing language interactions in practice. I also describe a set of instruments that became effective for my design process: a new form of wireframes that illustrated abstract language-interaction design ideas and became an effective boundary object; a set of NLP technical properties that are closely relevant to UX design; and a new prototyping method that enabled us to rapidly simulate various kinds of NLP errors. Finally, I will discuss how these findings reveal under-explored research questions and new insights in supporting UX design of NLP and AI more broadly.

5.2 Related work

When I speak of NLP technologies, I broadly refer to any computer manipulation of natural language, ranging from simply counting word frequencies, to giving meaningful responses to human utterances [9]. Some examples of modern canonical NLP problems are information retrieval, machine translation, dialogue systems, and question answering.

HCI research on NLP systems does not discuss the design *process*. While offering creative, user-

centered systems, researchers in this area typically describe one design solution, followed by its implementation and subsequent user study evaluation (e.g., [23, 75, 100]).

An exception is the work on prototyping conversational AI. Researchers simulated system behaviors with wizard or rule-based simulators so as to rapidly explore many interaction possibilities [22, 27, 57, 111]. For example, prototyping tools for speech interfaces [58] enabled designers to quickly test their conversation scripts in Wizard-of-Oz (WoZ) experiments. Unlike the aforementioned, common sketching methods, WoZ does not facilitate designers to experiment a technology's capabilities and limits. Instead, it frees designers from the technical complexities and limitations of NLP and facilitate experimentation on interactions. In this light, recent HCI work started to call for demystifying NLP [78], arguing that UX designers need to possess some technical understanding of NLP to be able to design with it [75, 101].

5.3 Method

5.3.1 Research Through Design

I wanted to identify challenges of sketching NLP within the context of one specific project and share my learning.

I chose a Research through Design (RtD) approach because, in alignment with my goals, RtD underscores that design knowledge arises from, and in response to, concrete problems and situations [46, 103]. I first immersed myself in the concrete design problems of the project (designing a writing assistant), and then offered an intentional accounting of the project to allow for objective reflections on procedural, pragmatic, and conceptual insights [40]. To achieve the methodological transparency needed for capturing my own design activities, I followed Bayazit's three-stage process [93]:

(1) *Knowledge elicitation in an unstructured and unanalyzed form.* Throughout the project, I wrote project dairies and weekly summaries, documenting my design activities. I documented all regular project meeting and impromptu conversations with my collaborators (n=24), and how the conversations affected my later design activities. The regular meetings took place among all HCI and NLP researchers in the project three times a week. Additionally, in the final weeks of the project, I conducted 14 formal interviews with 9 external NLP researchers. I recorded audio of these meetings, each lasting approx-

imately one hour. This resulted in more than 36 pages of description of my own design thinking and activities, as well as 9 hours of interview recordings, documenting major conversations between design and NLP expertise.

(2) *Data analysis and interpretation.* After the project ended I performed a thematic analysis on the data collected to identify key instances of challenges and reflection-in-action that happened during the design process. I transcribed the meeting recordings, and reflected on whether and how they shaped my later design trajectory. Finally, I sought agreement on interpretations across project members.

(3) *Finding validation.* I presented the findings respectively to all project members as well as to external NLP researchers. They validated my interpretations of our design journey and understandings of NLP's technical capabilities.

5.3.2 Collaboration with AI Researchers

The project team included a number of HCI researchers as well as 4 NLP researchers. One specialized in computational linguistics; the other three in language modeling and deep learning. Later in the project, I started to design with techniques that are not typically used in writing assistance. I interviewed other NLP researchers in the organization. Their expertise ranged from conversational agents, search, machine translation and more.

5.4 Findings

Below I will first provide an overview of the design process of the intelligent text editor. I will then detail five challenges we encountered when sketching, as well as the solutions emergent in my reflection in action.

5.4.1 Overview of the Challenges Encountered

I began by following a traditional user-centered design process [24]. First, I conducted a contextual inquiry study of 18 participants to understand their needs and wants in writing. I invited them to record their screen for 40 minutes as they were writing one of their own documents. I then conducted an 1-hour interview. Participants walked us through their thought process in writing during the time of

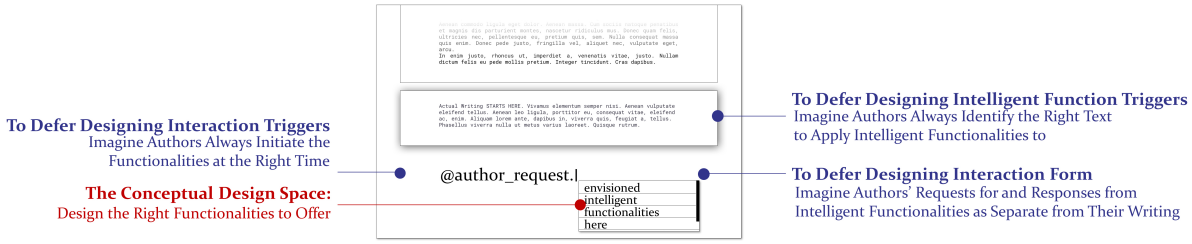

Figure 5.1: "The notebooks", an emergent form of wireframe for sketching abstract, language interactions. It bounded my design thinking to focus on envisioning "the right thing to design" and deferred detailed interaction design tasks.

screen recording and discussed their unmet needs and wants toward writing assistance.

Next, I envisioned many intelligent functionalities that users would be likely to find valuable. When doing so, I encountered several challenges: (1) How can I sketch language interactions abstractly? (2) How can I design with data scientists without data at hand? (3) How can I better understand and stretch NLP's technical limits? (4) Within these limits, how can I envision novel, less obvious applications of NLP?

After addressing these challenges, I proceeded with a small set of design ideas. For example, I envisioned an ask-your-reader function that compares a user's writing with examples from their target venues, helping them to account for readers' likely expectations. I created prototypes of these early ideas and tested them in a second user study. In this process, I encountered another challenge: (5) How can I prototype an intelligently flawed UX?

5.4.2 How to Abstractly Sketch Language Interactions?

Early in the project, I wanted to focus on "designing the right thing" rather than making detailed interaction design choices. Surprisingly, untangling the two turned out to be a challenge.

Traditionally designers address this challenge by drawing storyboards. Storyboards capture the contexts and the holistic experiences of a macroscopic design idea, while dismissing its interface and interaction details. This doesn't work for language interactions; language as a form of interaction carries both the interface and the utility it manifests. I did not know how to sketch one without the other, nor did I know how to sketch language interactions abstractly.

Adding to this challenge is the transient nature of authors' need for assistance. This entails two complex design tasks: Designing the trigger of the interaction so that the authors only interact with

the NLP function when they want to; and designing the trigger of the NLP function such that it applies to the right part of the author's writing. While these two are interaction details, I found them difficult to ignore because they are significant mediators of the perceived value of our designs. As a result, the looming question "*Am I designing another Clippy?*" frequently derailed the team discussion of a design idea from its utility to its interaction details.

How to stay abstract when sketching NLP? My solution to this challenge was a new format of wireframe, namely *the notebook* (Figure 5.1). It is an abstract representation of the moment when a user requests intelligent assistance: against the backdrop of what has been written in the document, the writer selects a part of the text and requests an intelligent assistance function from a drop-down menu. The user may additionally specify whether the assistance function should overwrite their writing, or display the response elsewhere as a reference. The user may also specify other information as additional inputs into the intelligent function.

Notably, the notebook does *not* depict a design idea. It is highly unlikely that the final design will require users to type their requests via such computer-program-like commands. Rather, it is an instrument that facilitates my design thinking and sketching.

The notebook bounded my design problem at hand, that is, *assume that users have made the right judgments on what intelligent function to apply to which text, and they are willing to make great efforts to make the function happen, what functions can AI offer?* The notebook prompted us to freely imagine valuable functionality offerings while deferring other detailed design choices.

Before I created the notebook, I had drawn many different representations of language interactions; some were literal and visually resembled a text editor, others abstract and conceptual. Only the notebook caught on and was later organically adopted by the whole team. Upon later reflection, the notebook caught on because it embodied the team's initial stances in designing intelligent writing assistance. These stances are 1) we wanted machine intelligence to support authors' writing as a process, not a resulting product [79]; and 2) we refused to assume that authors need or want help in writing, hence the intelligent assistance is passive by default and only became proactive upon user request.

Embodying these stances, the notebook became "*a very effective problem framing*" (designer diary, week 3) for me. For the rest of the design process, the notebook representation evolved every time when I reframed the design problem. For instance, after I had discovered in the user study that most participants

Axes of NLP Capabilities	What It Takes to Extend the Capabilities
Text Length. Words are easier to computationally process than phrases, than sentences, than paragraphs and finally a document. Knowledge beyond the written texts (e.g., common sense) is the most difficult to process.	Escalating an intelligent functionality, for example, from word level to sentence level, requires building **new models** and "*there is no guarantee how well it will work.*"
Classification - Comprehension - Generation Assessing or classifying a piece of text is easier than comprehending it (i.e. pinpointing the problem in this text), than text generation.	Escalating intelligent functionality along this axes requires building a **new model** and collecting additional or new labeled **datasets** for building it.
[classification only] Efforts Needed to Label Training data. If it is *easy* and *fast* for humans to make an *agreed-upon* judgment of its class, curating a labeled dataset for this intelligent classifier is likely to be practical.	Time, effort and often financial costs. Medicated by the amount of labeled data needed.
Likelihood to find training data that resemble the envisioned input/output pairs. We cannot presume a model that performs well on a benchmark research corpus would naturally perform in other texts.	Transferring or generalizing an existing model to a different corpus requires building **new models** and "*there is no guarantee how well it will work.*"

Table 5.1: NLP Capabilities, Limits and What It Takes to Extend the Capabilities

outlined in the same document what they want to say before they worked to improve on how to say it, I included outlines as part of the notebook framework. These outlines externalize users' communicative goals and can serve as a valuable source for more situated and personalized interactions and functionality. Including the outline in the notebook suggests that: 1) I can imagine new intelligent design possibilities with outlines as a resource; and 2) motivating authors to externalize their communicative goals will be one of my later interaction design goals.

5.4.3 How to Design with Data Scientists Without Data?

With the notebook framework, I started to ideate many writing assistance utilities that, based on our user study, users are likely to find useful. The first round of ideation generated 19 design ideas. To my great surprise, according to the NLP researchers, *none* of the ideas were promising from a technical feasibility perspective. These ideas "*need ten more years to make it happen,*" they said, only half-joking.

Eager for more insights, I asked the NLP researchers: why are these designs technically unfeasible? Why does this functionality work in this research publication, but doesn't work for this design? What *is* feasible then? However, technical researchers were unable to answer these questions. They could not articulate NLP's technical limits with our abstract design ideas. "*It's difficult to say; It depends on data.*"

"*The function you described is too abstract; I need to look at the data.*" For example, when I asked *Would these two models you are building work for our users?* I received:

> Scientist 2: *Both models are sort of data agnostic. So as long as the data [users' writing] is somewhat analogous to what we have now, it should, theoretically, translate very easily.*
>
> Scientist 5: *That's true for any model, right? So really, we don't know. We need to look at the data.*

In order to rapidly explore the design space, I could not afford to collect data before sketching. Data collection, prepossessing and exploration take up more than 80% of the total ML effort [143]. The question became: how can I partner with NLP scientists without a text corpus at hand?

After experimenting with numerous ways to explain my design ideas, I arrived at one boundary object [14] that effectively scaffolded my conversation with NLP researchers – a more developed version of the notebook. This version of the notebook was projected on a Text Editor wireframe, resembling a user interface (Figure 5.2a). When I embedded my design ideas within the notebook framework, NLP researchers no longer asked for data and became able to engage in feasibility discussions about abstract design ideas.

Interpreted as a design problem framing for me, the notebook represents a language model for NLP researchers: when a user triggers an intelligent function to be applied to a selected snippet of text, the content and paratext of the Word document at that moment constitutes potential training data. The selected text is the model's runtime input. The envisioned function outputs are modeling goals.

The notebook became a shared representation and a means of translation between the two worlds of UX and NLP. It scaffolded my discussions with NLP researchers for the rest of the project. I described my design ideas by describing what they would look like on the notebook. NLP researchers then gave feedback on whether these are sufficient sources of data for building intelligence. They also proposed additional kinds of data that could boost model performance. Drawing on user study findings, I considered what additional data might be present or attainable through user interactions, and iterated on our designs. This process iterated smoothly and required no data collection or cleaning efforts.

5.4.4 How to Understand and Stretch Technical Limits?

You are really good at designing things we cannot build. We are good at making things that users don't use. (NLP researchers 2 & 9, weeks 3 & 5)

My first-round sketching produced design ideas that are uniformly beyond the limits of existing technical capabilities or existing datasets. It is worth noting that, when I envisioned these designs, I did not imagine NLP as a crystal ball. I drew the ideas from NLP literature; I intended to innovate writing assistance by amplifying or re-contextualizing these existing techniques. Below are two examples of my initial design ideas:

- Rephrasing the selected text in a more positive tone (seems possible based on existing work on style transfer between different sentiments);
- Identifying whether the selected text is logically coherent with its context (seems possible based on textual entailment analysis techniques [39]);

How can I understand NLP's technical capabilities and limits from an UX perspective? Realistically, to what extent can I push these limits to enable novel designs?

A set of technical boundaries became clear to me after many discussions with NLP researchers, through negotiating with them and iterating on my design ideas. I describe these boundaries via four measures of NLP's technical difficulty: text length, text classification-comprehension-generation, effort needed for labeling (for classification problems only), and likelihood to find training data that resemble the envisioned input/output pairs (Table 5.1). This set of measures enabled me to eventually find the intersection design space between what is valuable to users and what is technically feasible.

Taken together, the four axes depict an algorithmic approach to language processing that is quite different from typical authors'. Authors start writing with a big idea in mind, then scaffold its constituent supporting ideas, structure paragraphs, sentences, and so on. In contrast, language models first parse a sub-word, then a word, then a phrase, a sentence, and so on. Comprehending the big idea underlying the written texts is considered "*the holy grail of NLP research*"; It is so challenging that "*when we figure this out, the whole field of NLP would have become a solved problem*".

Because of this difference, my early design ambition to support the *experience* of writing – the ongoing process of translating the big idea to texts – has unknowingly led us toward technically challenging

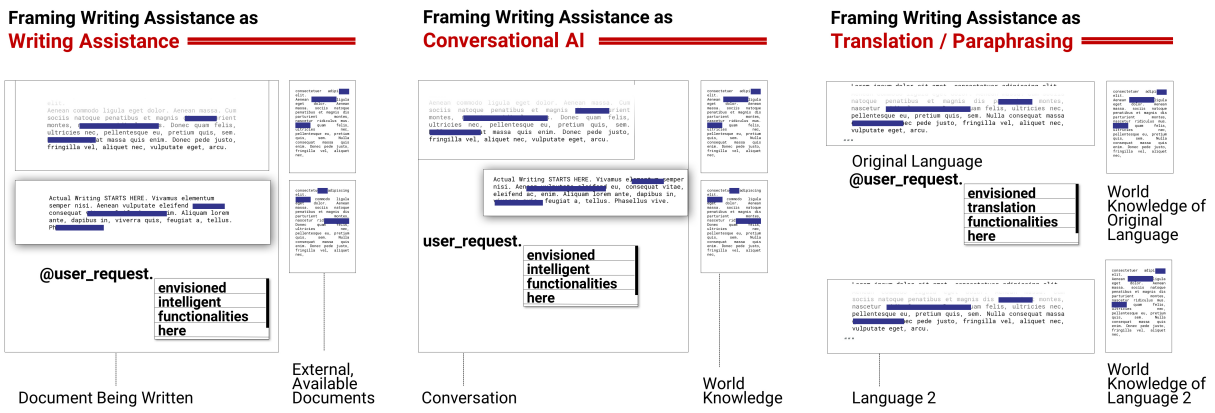

Figure 5.2: Left: A developed version of the Notebook (Figure5.1), used as boundary object between HCI and NLP researchers. "Contexts" that can help inform intelligent function outputs are marked blue. Middle and Right: Reframing the problem of designing writing assistance as other canonical NLP technical problems. This expands our design space to the intersection between what authors want and what existing NLP capabilities can do.

designs. In order to generate any implementable design ideas, I cannot naively project authors' goals onto intelligent functions. I need to identify technically achievable intermediate steps toward their goals.

By quickly assessing the four measures I was able to "gut-check" the feasibility of new design ideas and weigh their promise of UX gain against the technical effort they required. For example, the "efforts required for labeling" stands as a reminder that I cannot presume many seemingly trivial skills that authors master can be easily automated. This is because "*humans generally don't have a sense of classification or labelling. It's unclear when they used common sense, biases, or their world knowledge. But algorithms require labelling (to learn these)*" (Designer note, week 7).

Take the aforementioned design idea "paraphrasing in a more positive tone" as an another example. It is technically out of reach because of its low likelihood to find training data that resemble the envisioned input/output pairs (axis 4), as an NLP researcher explained:

> *We don't have the training data at the scale that we need. [If] Say: here is a mangled version of the sentence, here is a cleaned up, positive version of it, times 50 million pairs. If we had that, it (building the intelligence) will be trivial. But it's very unlikely to have this kind of data.*

5.4.5 How to Envision Less Obvious NLP Applications?

With some understanding of the technical limits, I sketched new design ideas that NLP researchers consider as implementable, or "*at least have a clear direction to work from*". However, I found my own design ideas rather unsatisfying. State-of-art NLP can assess or classify writings, but cannot easily pinpoint causes of the problems or generate suggestions on how to improve. This seemed a textbook recipe of a frustrating user experience. As a result, most of my design ideas are word or phrase-level alternations, "*many variations of auto-complete and auto-correct basically*".

How can I envision technically feasible NLP designs that have not been imagined before? How can I expand this narrow intersection between what is value to users and what can be built?

I addressed these questions with a classic designerly approach: taking designing writing assistance as a wicked problem [16] and seeking reframings. My initial problem framing underlying the original notebook wireframe is that the writing assistance functionality comprehends and generates texts as the author writes. As described in the last section, this framing is prone to technically challenging design solutions. Authors are inherently better than algorithms at comprehending their unfinished writing and at predicting their unformed ideas, which is a wicked problem that cannot be accurately modeled.

I reframed the relationship between authors and writing assistants as other canonical NLP problems, specifically human-AI conversations, information retrieval/search, and question answering. (Figure 5.2 illustrates how I abstracted users' writing into many text components, and then mapped them onto other NLP problems.) Each of these alternative framings exposed me to a new set of technical capabilities in a different NLP sub-domain. I demonstrate how these new framings broadened the design space through the two design ideas they spurred.

5.4.5.1 A Context-aware, Rhetorical Search Function

Search is a relatively matured NLP sub-domain. Conceptualizing AI-assisted writing as a search experience introduced many near-future design possibilities into my design space. Instead of algorithmically generating responses to authors, a search function can simply retrieve relevant writings based on author requests. It does not require large datasets or the collection of labels.

I started to ideate intelligent search tools that users are likely to find useful. I observed and in-

terviewed participants in our user studies about how they sought information during writing. I noticed that, prior to writing, most participants outlined and organized their thoughts in the forms of bullet lists, tables, and even drawings. Yet many struggled with translating the organization of thought into a linear, natural flow. Authors therefore searched online for rhetorical structures that they could borrow, for example, one participant, P7, Google'd "*[quotation mark][comma] in comparison to [quotation mark]*" to search for examples of connecting ideas of contrastive relationship. However, this carefully constructed search query does not actually work as he expects. Modern search engines expand and rewrite search queries based on similar searches, user search history and so on, optimizing for finding content that is relevant to the query topically rather than rhetorically. Participants like P7 could not find the writing examples he sought.

To support this unmet need I sketched a rhetorical search function. It searches the web for text that is similar in language *structure* and *composition* to the author's query. It takes into considerations the topic and style of the authors' current document to optimize the relevance of the search results. When an author selects a part of their bullet-list outline (e.g., "Issue A: good/bad examples") the writing assistance tool then searches for contents online that contain contrastive examples relevant to Issue A and sorts by different ways of transitioning between them. Rather than optimizing for topic relevance, this search functionality helps users find better ways to organize and connect their thoughts. It can be implemented with readily available search techniques.

These search functions can also serve as a stepping stone to future, more ambitious intelligent designs. The search results that authors adopted as training data for future generative language models.

5.4.5.2 An Asking-Your-Reader Function

Another useful reframing is conversational AI, that is, reframing the role of writing assistance as a conversation partner of the author. This reframing asserted new design questions: Whom would authors like to talk to during writing and for what purpose? What information can conversational assistance offer? These design questions naturally expanded our design space beyond "helping writers verbalize what they have in mind", and prompted us to imagine utilities NLP can provide as an outsider to the author's world.

With these questions in mind, I asked participants whom and how they asked for feedback while

writing. I found they often picked those who are close to their target readers as their "beta-readers". Participants worked to translate their often egocentric writing into a style that meets the expectations and needs of their target readers. Many read other documents from their target venue to infer the expected length, lexical complexity, or level of detail that they should write in.

I see this as an excellent opportunity for NLP technologies to help authors, as algorithms are good at rapidly summarizing or characterizing a sizable collection of documents. I therefore designed an "ask your reader" function. It mines documents from an author-identified venue. The author can request insights about these documents or make comparisons between their own writing against it. For instance, "Am I writing too formally?" "How long is a typical introduction section in [venue]?" In this design writing assistance does not assist authors in writing *per se*, but supports their communication with their target readers.

Through these two design exemplars, I demonstrated that design problem reframing helped the team envision novel forms and functions of existing NLP techniques, expanding the design space of technically feasible writing assistance.

5.4.6 How to Prototype an Intelligently Flawed UX?

I generated a prioritized set of intelligent functions offering ideas informed by the our initial user study and bounded by existing NLP capabilities. I then turned to building a low-fidelity prototype to rapidly experiment on these ideas with users. I wanted to test the ideal behavior of the envisioned intelligent functionalities with users to see if I was pursing the right design directions. I also wanted to probe users' reactions to a more realistic range of NLP-powered behaviors and errors to account for these reactions and expectations when improving on our design.

But how can I realistically simulate NLP's errors without spending months building fully-functioning systems? I experimented with a series of prototyping methods in collaboration with 9 NLP researchers.

5.4.6.1 Failed Attempts

Wizard-of-Oz is a common way to prototype NLP. However, I learned early in the project that algorithms make errors that are unlike human errors. For example, even state-of-art NLP can fail in text comprehension or generation because of a lack of common sense knowledge. To enable wizards to simulate NLP

behaviors we need to prevent them from accessing their common sense, which is extremely difficult.

Beyond unrestricted Wizard-of-Oz experiments I also considered using a rule-based simulator to prototype intelligent input/outputs. I encoded some rules (e.g., manually specified decision trees) into the prototype. However, NLP researchers pointed out that a rule-based simulator at best could behave as well as a rudimentary ML system. Their capabilities are far behind state-of-art technology.

I attempted to use publicly available, pre-built NLP models to power the prototype, yet failed for similar reasons. These application-agnostic toolkits only include the most matured kinds of NLP technologies. Their level of sophistication is not close to state-of-art NLP technologies either.

I then experimented building simple ML/NLP models to simulate modern NLP's behaviors, using publicly available datasets and off-the-shelf toolkits such as AllenNLP [118]. This failed for a number of reasons. First, preparing the datasets is itself a daunting task. Next, integrating NLP toolkits that were built upon different platforms, in different programming languages into one prototype further complicates the prototype building. Finally, I built a simple model yet its performance was just not good enough for a user study. When an algorithm-generated sentence makes sense but reads awkwardly, the awkwardness washed out all other user "experiences". The sentence reads simply, awkward.

5.4.6.2 Successful Attempts

I prototyped my design ideas with an alternative WoZ method. For each NLP-powered interaction, I designed a different hybrid of WoZ and off-the-shelf toolkits to best simulate the likely errors. The design of each hybrid mimics the likely architecture of its underlying NLP system. This method highlights that different intelligent features produce different kinds of errors, each of which can have different UX consequences. In order to better capture these consequences, I needed to better orchestrate WoZ behaviors to simulate NLP behaviors. Below are a few examples:

(1) Simulating context-awareness with machine translators: Most of my designs took authors' writing as an input and provide context-aware, personalized writing suggestions. Our prototype read the authors' writing in English, translated to a foreign language using existing machine translation services, and translated back to English. The output of this process simulated the noise in "context" detection.

Language technologies could fail at extracting relevant contexts from authors' writing. There instead of taking their writing into full account, my prototype removed parts of it that algorithms could not

easily comprehend through the two rounds of translations. I used online machine translation services to build this prototype. To simulate context awareness of lower quality, I selected the second, third, and fourth ranked translations that the translator provided, so that more context and meaning were lost in the translation.

(2) Intelligent functions that assess or categorize author's writing: I simulated the results based on what kinds of errors were more likely to happen (precision, recall, etc.) and which assessments/categorizations were more error-prone.

(3) Simulating generative writing assistance with a multi-wizard simulator: When a user study participant requested a piece of machine-generated text, multiple wizards and a meta classifier worked in the background. Each produced a response that excelled at one aspect of the text generation. For example, one wizard produced a topically relevant response, the second wizard took charge of the response fluency, the third focused on the coherence between the generated text and the writers, the forth added domain knowledge to the response, the fifth generated random words, and so on. The meta classifier assembled all of the wizards' responses into the final response returned to the user.

I designed these wizards' roles based on common models of generative neural networks. I simulated different kinds/degrees of generative errors by tuning the weights that each wizard carried. As such, I was able probe user study participants on their preference among various designs of a generative writing assistance as well as their error tolerance. Below I briefly describe some of the user study results and discuss how they informed my later design iterations and refinement.

5.4.6.3 Effectiveness of the Prototypes

In the second user study, I invited the 18 participants to use the prototypes as they were writing one of their own documents. To my knowledge, this is different from almost all previous HCI work on writing assistance systems, in which researchers typically invited participants to write on pre-determined topics for a particular time duration (e.g., as in [23, 100]). My prototypes triggered unexpected reactions, ones that are distinctively different from either previous HCI studies or what the participants verbally described as desirable from an intelligent text editor.

Is Adopting Machine Generated Writing Plagiarism? One of the functionalities offered in the prototype is generating sentence-or-phrase-level writing suggestions. For example, when participants

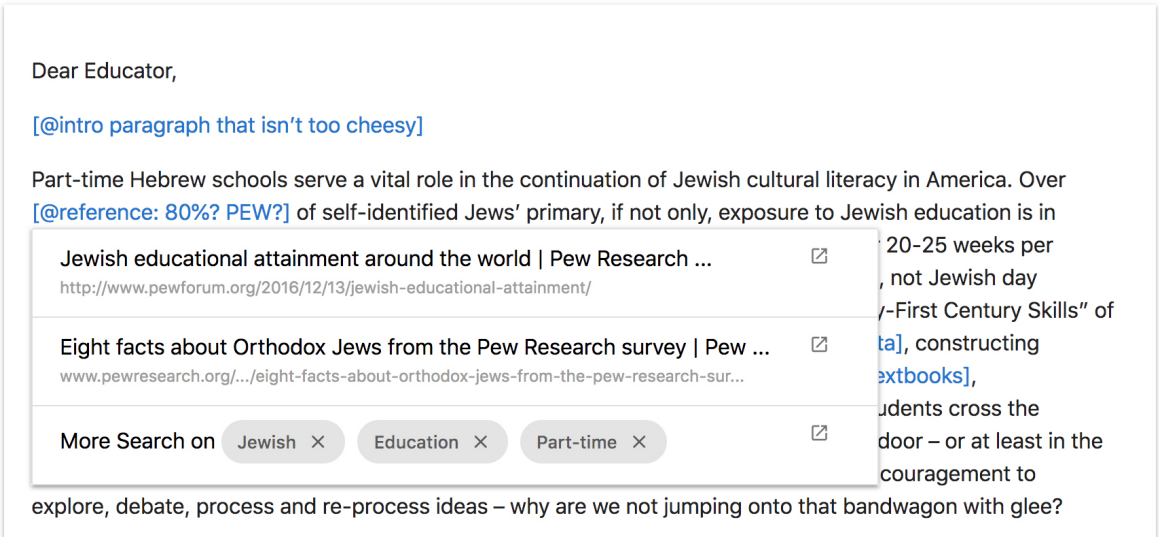

Figure 5.3: This prototype interface is a simple text editor. At any time of their writing, users type @ to signal the start of an intelligent function request and Enter to end. When they click on a request, intelligent assistance pops out. This prototype probes users' needs and wants for writing assistance, and their reactions to the simulated intelligent responses.

type "@ add an opening with a quote" (@ signals a request for intelligent assistance) and click on the request, the prototype surfaces a list of opening sentence suggestions. Participants in the initial user study expressed a desire for such functionalities for they can save writers' efforts to search for relevant quotes, examples, or references online and to integrate into their own writings.

> "*Even if I just liked this scaffolding (in the machine-generated suggestion), I wouldn't take these exact words. It's like... In my school, five or six consecutive words from any other piece of media that isn't referenced as a quote are considered plagiarism. People get expelled from school [...] It's a societal judgemental thing.*" (P9)

Interestingly, when participants saw the machine-generated suggestions to their own writing, they instantly became more resistant and expressed a much stronger sense of ownership or their writing than they had initially expressed. "*Isn't this plagiarism?*" Few participants described adopting a machine suggestion directly as "*stealing a sentence from another article*" and "*it just feels wrong!*"

Instead of accepting machine-suggested sentences, most participants browsed many, many suggestions, and from different suggestions picked parts of the sentences, semantics, word choices or references that they liked, and integrated them into their own writing. Some participants clicked the "refresh" key more than 20 times – corresponding to more than 60 different writing suggestions, "*just to get a vibe.*"

Even machine-generated sentences have ramifications beyond themselves.

"*Some of the things, the inspirational stuff (quote), I need to know that (machine-generated) content better as my name is attached to this document, and I need to refer to it and talk to it.*" (P9)

"*I need to scout out the (machine-generated) sentence. That kind of sentences has been written a million times. It is not really the point that they are trying to get. That's just a way of seeding the context.*" (P5)

Almost all participants firmly believed that the machine-generated sentence suggestions have larger contexts and ramifications. The true intent of sentences, the philosophical stance of the author, reside in these larger contexts, "*at least two paragraphs later*". Participants shared that they needed to "*make sure this article (source article of the suggested sentence) is going where I thought was going*", in order to assess whether the suggestion aligns with their writing. However, our prototype, as well as almost all sentence-level generative algorithms, do not produce such contexts, therefore simply is unable to respond to such user requests.

5.4.7 Summary of Emergent Solutions

Earlier in this chapter, I detailed the five challenges we encountered. In the process, three instruments became useful to the design process.

- The notebooks. The notebooks are a set of wireframes that illustrate abstract language interaction design ideas. The notebooks ended up playing three important roles in the design process: They enabled me to externalize and communicate early-stage, abstract, language-based design ideas (challenges 1 & 4). They also served as a boundary objects between designers and data scientists (challenge 2) and between designers and users (challenge 5), which enabled conversations among UX, design and NLP expertise;
- A set of NLP properties that are closely relevant to UX design, including "axes of NLP capabilities" and "what it takes to extend them". Understanding these properties helped me frame the design space within current technical limits (challenge 2);
- An alternative WoZ prototyping method. For each NLP-powered interaction, I designed a different hybrid of WoZ and off-the-shelf toolkits to best simulate the likely errors (challenge 5). This method

shares the goals of traditional WoZ in enabling fast prototyping of NLP. In addition, my method highlights that different intelligent features produce different kinds of errors; each kind can have very different UX consequences.

5.5 Reflecting on UX Design Expertise and Methods for AI

This project offers a point of reference for discussing and addressing the challenges of designing generative language interactions. Do the challenges I encountered in this project generalize to other AI systems? What are the root causes of these different challenges? Relatedly, to what extent do the emergent solutions generalize to other design situations? Answers to these questions have the potential to improve the UX design and innovation of many natural language generation and AI systems at large.

To jump start this discussion, below I discuss the challenges I encountered and solutions emerged in this case study again the AI design complexity framework as backdrop. My goal is to extend the value of my situated findings from this case study to designing level four AI systems more generally, in a structured and rigid manner.

This project highlights the value of UX design expertise for making AI's technical advances valuable for people in real-world scenarios. The team identified a priorities list of users' desired NLP functionalities before spending months and even years in collecting data and training a generative model, thereby minimizing the risks of making the wrong thing. It was particularly important in the early design stages – the process of exploring many broad ideas before drafting concrete UIs or dialogues. For example, what kind of writing assistance do people even want? Identifying "the right thing to make" and prototyping early on in the design/development process can minimize the risks of spending months and even years in collecting the wrong data or training an undesired generative model.

This study also surfaced a number immediate research opportunities in supporting UX practice: 1) developing new sketching and prototyping techniques for language interaction design, 2) understanding NLP's design affordance and limits, and 3) designing and evaluating the UX of evolving AI systems.

5.5.1 Sketching and Prototyping Techniques for NLP

Sketching and rapid prototyping are cornerstones of HCI's creative activities [18]. Designers carry out their design thinking through these hands-on activities [103]. However, it was not intuitive how to conceptualize, design, or evaluate language interactions abstractly.

NLP systems are difficult to sketch or prototype because language interactions are difficult to abstract for both. On one hand, most sketching techniques and tools in designers' tool-belts, such as storyboards and wireframes, have evolved over the last two decades under the dominance of the graphical user interface, which is not directly applicable to language interactions. On the other hand, machine-generated language can make errors incomprehensible to humans and are difficult-to-anticipate. It could seem that only building a working NLP system can reveal its likely behaviors. Abstraction is essential to any early design ideation, yet a missing perspective in NLP HCI literature. Most assume that designers start designing by "*writing linear dialog examples*" [58]. WoZ studies often simulated NLP interactions with rule-based systems or crowd intelligence; Deliberations are lacking on whether these are effective abstractions of NLP system outputs.

Upon reflection, all five challenges I encountered involve some aspect of abstracting NLP interactions. For example, challenge 1 dealt with abstracting language interaction design ideas into different problem framings effective for design deliberation. Challenges 2, 4, and 5 struggled to elicit relevant experiential qualities of language interactions, in order to create meaningful boundary objects or low-fidelity prototypes.

In this light, I argue that supporting sketching and prototyping language interactions *abstractly* is an important yet under-engaged issue for HCI/design research. To summarize, this case study revealed three aspects of abstraction, offering a starting place for this line of research.

1. Abstracting language *interactions* as ways of framing its design problems
2. Abstracting NLP *capabilities* to frame its design space realistically
3. Abstracting NLP's *experiential qualities* to enable rapid UX prototyping

Figure 5.4 maps these three aspects onto the problem space of human-AI interaction design. The first two aspects concern how to "identify the right thing to design" with NLP; the other "design the thing right". In addressing these challenges, future work may take inspirations from previous work on

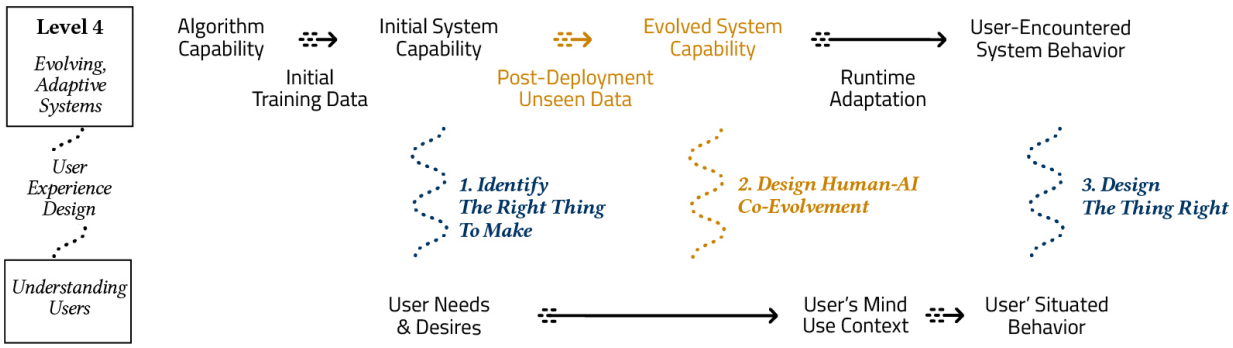

Figure 5.4: Case study of level one system design process, against the backdrop of the AI design complexity framework.

designerly abstractions of difficult technology materials, such as visualizations, taxonomic vocabulary, and sensitizing concepts [76, 87, 95]. In doing so, the UX community can develop a robust family of methods for designing language interaction abstractly, thereby enabling its UX design innovation.

5.5.2 Understanding NLP's Design Affordance and Limits

Let me expand on the issue of "abstracting NLP capabilities".

Previous research reported that some UX designers could fail to understand NLP/machine learning "specifically," even when they understand how the systems generally work [30]. My experience in this project echos this observation. My early sketches of writing assistance revealed a significant gap between how I wanted to support users ideally and what NLP can build realistically. This gap is hazily assumed yet rarely discussed in HCI research. Much work – exemplified by the many unrestricted WoZ studies – instead has focused on the design *possibilities* NLP inspired. This orientation leads to some NLP researchers' claim that "*HCI people design useful things that we cannot build; we make things that nobody uses.*"

In parallel to works that freely imagine possible futures, there should also be research on creating products that wisely attend to state-of-art NLP's capabilities and limits. Towards this goal, more work needs to investigate respective advantages and disadvantages of human and artificial language intelligence in order to choreograph harmonious interactions in-between.

In this case study, three aspects of NLP were relevant to the writing assistant design, as they emerge naturally in my design activities: (1) High-level understandings of NLP's capabilities and limits, which

oriented our design ideation; (2) NLP's capabilities given the available data and development resources, which informed our design deliberation and the trade-off between UX gains and technical investigations; and (3) Each design's likely errors and other experiential qualities, which enabled rapid prototyping and helped us account for unexpected system behaviors. Table 5.1 provides a glimpse into what a designerly understanding of NLP capabilities might look like.

Future research should evaluate and improve this set of NLP design properties. Moreover, enabling practitioners to develop their own tacit understanding of NLP opens up new research opportunities and promises real impact on UX practice. For example, what new boundary objects that help designers more effectively collaborate with NLP scientists and understand the technical capabilities and limits applicable to their respective design problems?

5.5.3 Designing and Evaluating the UX of Evolving AI Systems

One limitation of this case study is that it does not include an evaluation of the envisioned NLP systems. I do not know whether my understanding of NLP's capabilities and limits has indeed led to novel and technically feasible designs, as my NLP researcher collaborators believed. Beyond the early prototyping study, I do not have evidence that the users will indeed appreciate and enjoy using the writing assistants. This evaluation is difficult. Deep-learning-powered NLP systems require an unwieldy amount of data, and collecting the data from users' organic interaction traces can take years. This is beyond what the timeline of this project allowed.

But more pertinently, even if there was enough time, it is unclear how to evaluate the UX design of an evolving AI system; a system whose technical performance and error modes are constantly fluctuating and evolving with its users' interaction traces. Consider Tay the Microsoft Twitter bot as an extreme example. Positive results of a UX evaluation today do not indicate that the system will continue to deliver good UX tomorrow.

In this sense, level four AI systems (systems that continue to learn from new data post-deployment) revealed a non-trivial limitation of current UX prototyping and evaluation methods: They cannot manifest or evaluate the evolving UX of an evolving AI system over time (Figure 5.4 yellow parts). It is very difficult to fully anticipate how users' behaviors will evolve over time and how the interactive NLP systems' behaviors will evolve accordingly. In this case study, the team worked to anticipate and mitigate

this human-AI co-evolvement based on the designer's and the NLP researchers' intuition. Are there better, more rigid ways to design the interactions of two mutually adaptive agents? Future research should critique and improve the design process described in this case study, seeking better, more systematic methods of designing and prototyping living AI systems.

Chapter 6

Beyond Case Studies: Investigating Industry Best Practice of Designing AI

Previous chapters have presented Research through Design projects as case studies. They demonstrated the benefits and limits of UX design methods in designing AI systems, and also raised new questions for UX design research. For example, how can we design or evaluate the interactions between two mutually-adaptive agents, an evolving AI system and its users?

In this chapter, I investigate the design practices of some of the few UX designers who regularly create new products and services that use ML to enhance UX. I hoped that their rich experience and practices could reveal new insights into the questions around UX design of AI.

I interviewed 13 designers who all had at least four years of experience designing AI/machine learning-enhanced UX. The interviews produced several interesting findings: 1) Designers shared that they knew very little about how machine learning (ML) works, and this was not a priority for them. They instead used designerly abstractions and popular exemplars to explain what ML is and to communicate design ideas with each other. 2) ML projects are longer in preparation and scope than other design projects. During the preparation stage, designers evolved their ideas in close collaboration with data scientists; They did not deliver fully formed designs to a technical team. 3) New design activities and techniques have emerged as designers try to embrace the data-driven culture. They "play with" quantitative data – in different ways and for different purposes – throughout all phases of a design project.

Education	Professional Role	Example Design Project	Org Size	Exp. (years)
HCI	Design Manager	Intelligent features in multiple messaging or conversation apps	>10,000	10+
Design	Design Manager	Intelligent features in a media consumption app to better match content providers and users	>10,000	10+
Design	UX Designer	Decision support app for physicians	1,000 - 10,000	10+
Design	Designer-turned PM	Language learning app; patient coaching app.	100 - 1,000	10+
Design & HCI	Designer	Intelligent tutoring feature for education apps	<100	10+
CS and HCI	Designer-turned entrepreneur	New messaging and efficiency apps	<100	10+
Design & HCI	Designer	Recommender in a shared service app	>10,000	7-10
CogPsy	UX researcher	Intelligent reminder in a social media app	>10,000	7-10
EE & Design	UX designer	New wearable health product	>10,000	5-7
HCI	Designer	Recommender in a text processing software	>10,000	
Design	Service designer	New health coaching app	>10,000	5-7
Psychology	UX researcher	Intelligent feature in a social media app	>10,000	4
HCI	UX researcher	Intelligent content recommendation for a media consumption product	>10,000	4

Table 6.1: Interview Participants. I interviewed UX designers who had more than four years of experience in designing ML-enhanced products. Many had more than 10 years of related experience.

6.1 Method

I conducted retrospective interviews with UX design practitioners who have played an active role in creating widely adopted ML-driven products. I interviewed 13 participants. All had designed products and services that enhance UX with ML for more than 4 years. Most had worked on these types of products and services for more than 10 years. Nine participants designed products used by more than one billion users. Two worked on successful special-purpose ML applications: a clinical decision support systems and a wearable health coaching system. The others worked to extend existing products with new ML features. Table 6.1 provides a summary of the participants' background and the type of projects they described in our study.

I asked all participants to complete a pre-interview survey where they described their education and professional background, as well as their familiarity with concepts from UX design, statistics, machine

learning, and data storage. I then interviewed participants, asking them to walk us through a recent design case where they used ML to improve the UX. Throughout the interviews, I probed them to get details on their process and to surface the triggers that drove specific decisions for what to do and how to work.

At the end of the interview, I asked participants to reflect on what they viewed as the major differences between how they design when working with ML and when working on products and services that don't use ML. I asked them to share, "*the things you wish you had known about before you started your career in designing ML systems*".

I recorded and transcribed the interviews. I then reviewed the transcripts, pulled out important insights, and used affinity diagrams to synthesize across the interviews in order to identify thematic patterns. I created and consolidated process models detailing how the different projects unfolded and how the designers collaborated with data scientists.

6.2 Findings

I organized my findings around three themes: work participants did to understand ML capabilities, changes to the design process that seem directly related to working with ML, and new, AI-specific design activities participants have undertaken.

6.2.1 Designerly Understanding of AI

Participants characterized their ML literacy as "*understanding at a very high level... [at the level of] knowing what a classifier is and what a label is.*" Interestingly, participants did not feel their lack of technical knowledge hindered their ability to design or to collaborate with data scientists. They shared that they worked on the design issues, not the technical issues, and that working with ML required "*... more design savvy*". Several participants claimed designing the user interaction was "*the actual challenge.*"

> *I don't think of ML as affecting my interaction designs or not. I think about it more like impacting certain algorithms that are inputs into users' experience.* (P14)

Most claimed that they learned to work with ML similarly to how they learned to work with other interactive technologies. "*ML is just like JavaScript*", several participants claimed. They did not seem to

view themselves as technology experts, but instead as UX design experts who had great comfort working with a variety of technologies, and ML was one of these.

The way designers spoke about ML has little overlap with the way education materials meant to teach ML for engineers. Participants rarely spoke of ML in technical terms. For example, they never talked about supervised or unsupervised learning (common starting point for teaching ML). Instead, they appeared to think with, and work with, abstractions; simple insights about an ML capability had had implicitly linked with generating value for users. These were much more abstract than design patterns. More similar to design patterns, they often used exemplars to communicate these abstractions.

> Q: What does machine learning do?
>
> P7: *Some try to recognize intent, a bit like auto-correct. Some are intent prediction like Clippy. Anyone who is working on assistive technology, is working on some class of that problem.*

In the excerpt above, P7 describes ML using abstractions of its capabilities: recognize intents and predict intents. He then ground these capabilities through the use of exemplars: auto-correct and Clippy. I found this manner of describing ML across most participants.

Participants most often described the capabilities of ML as it related to the user's utility. Their abstractions narrowly oriented towards both the users and scenarios related to their designs. They never spoke of general taxonomies of ML functionality or specific algorithms. Some examples:

> *We use machine learning so that we can build something that can personalize for a lot of people.* (P3)
>
> *In consumer tech, we try to raise the level of abstraction [of user commands] rather than doing everything manually.* (P7)
>
> *ML gets users directly into the task they really need to do.* (P1)
>
> *We are doing a recommendation system of sorts. As a product designer [not a technologist], I think about that as how can we show an evolving relationship between a user and our service... [I want users] foreseeing our relationship improve, where the relationship is the recommendations we are giving them.* (P14)

The abstractions almost always appeared with exemplars. The abstractions served as a general insight about an ML capability and provided an understanding of how it worked. The design exemplars

provided specific interaction possibilities and a glimpse of a possible felt experience. In our interviews, participants frequently referenced widely-known exemplars including Clippy, autocorrect, email spam filtering, and Tay the Twitter bot. They used these to help describe the capabilities of ML; the form, function, and user experience; and the potential breakdowns that might occur.

The number and variety of exemplars participants used varied wildly, and those with the largest working sets seemed to be the most successful and comfortable at using ML to enhance UX. Participants working at AI-focused organizations had a significant advantage in building their working set of exemplars and abstractions. These organizations had data scientists frequently giving demos as one way to sensitize their design teams to emerging ML capabilities. "*So many people demo for me. I don't even know whom to call if I have an idea and want to consult a data scientist.*" Participants at smaller or less AI-focused organizations had much more limited access to data scientists. They appeared to consider and propose fewer design alternatives. They were also more likely to use only the most familiar interaction forms, such as recommender or reminder, when describing their process.

6.2.2 Design Process and Collaboration

Participants shared that working with ML took much longer than when designing other UX products and services. I wanted to synthesize the actual time span ML projects required. However, I found that none of our participants had worked on an ML product from its initial ideation until its final release. I could see that some ML projects had lasted for more than four years.

> *ML has a different time-frame for design iteration. Longer initial development, but then ongoing iteration. It felt like building a feature versus building a framework. When you ship it, it's not the end of it. They cycle, and data drives the next step.* (P3)

A consolidation of participants' design process narratives revealed an ML design process. It starts with a long preparatory stage (stage 1), during which designers and data scientists identify a design goal that is both technically viable and that appears to have a high likelihood of improving the user's experience. Once a team had settled on the design goal, scientists and developers implemented the system and designers crafted its interaction design (stage 2). In stage 3, they together invested in frequent iterative releases and assessments in order to improve user adoption.

All participants described stage 2 in their ML projects. Stage 2 is a central stage that all projects that

participants described went through. Stage 1 happened only at the large, AI-focused organizations. Few projects had made it to stage 3. Below I describe three distinct design stages.

6.2.2.1 Stage 1: Concept Development

New ML design seemed to only happen within the AI-focused organizations. It started with a long, preparatory stage involving two activities. First, participants collected log data form current services their organization provided. Participants examined the log with sensitivity to specific patterns. They often seemed to search for patterns they thought might be there more than they would data mine to discover unexpected patterns. They imagined what user behavior might be worth learning and what learned interactions might be valuable for users. One designer described this process as "*just me being really excited about the product*" (P14).

Second, participants would share these inchoate ideas with data scientists. This happened informally, outside of any specific development projects, and they had iterative interactions, where they might return to the same data scientists with new or refined ideas. They noted that the data scientists had a very different view of scenario-based design. They used scenarios to validate what an idea might be instead of using scenarios to generate and refine new ideas. "*Data scientists use scenarios to validate designers' assumptions about how the product should work, [and then] toss back ideas*".

The conversation between designers and data scientists focused on identifying a design goal worth pursuing. Often this would lead the discussion away from working with ML. The discussions focused on coming up with a "good enough" idea, and it did not address either its exact technical feasibility or experiential quality. Participants particularly pushed the data scientists to understand what might be technically possible.

> *(I) framed the questions not as do you know what would work, but in your gut, do you think this would be possible. Possible on a scale of 1-10.* (P8)
>
> *The level of detail I'll need to discuss with [the data scientists] is understanding the capability of what could be possible. I didn't need to get into specific detail about it.* (P3)
>
> *Those [design ideas] are a series of aspirations. Rather than saying what data do we have [...] Could we challenge the data team to figure out how to get close to that?* (P9)

The collaboration at this stage focused on co-evolving a shared vision between the two areas of expertise, an ideal user experience that was worth pursing and ML could potentially help achieve. This shared vision often took form of a unique abstraction of ML capabilities that emerged out of both UX and ML. Instead of saying "machine learning", designers and data scientists might select an expression better suited to the context of their product. For example, some used "affinity" to describe the match between a user and a piece of content. Others used "personalization" to describe the intended, evolution of the relationship between the product and the user. Participants characterized these discussions as a chance for both sides to learn from the others' expertise.

> *There is no such a framework or something, but I think later there is a kind of an acknowledgment when we talk about "personalization"... An acknowledgment that a more personalized experience is a better experience that one is less likely to walk away from.* (P14)

> *I gained a better understanding of the capabilities of the algorithms. The data scientist gained a better understanding of what was worth pursuing.* (P3)

6.2.2.2 Stage 2: Interaction Design and Assessment

Stage 1 produced preliminary understanding of the data, and it established a shared vision between the designers and the data scientists that met the company's goals. Stage 2 focused on refining this vision. In many organizations, this was the stage when UX designers were invited to join a project. During stage 2, participants described developing "*a funnel of visions, a funnel of what exists and what is possible in the company*" (P8). This would advance towards defining a single, valuable ML feature that was both experientially valuable and technically feasible.

> *It's definitely an ongoing process. We had a lot of basic things to get working. [...] Once we get that improved, then we will probably be adding more ML stack onto the system.* (P14)

> *Design is always about a new product and a complete, larger vision. But that doesn't make it easy to build. We chunk. We talk about stages of development.* (P8)

Once a product idea was clearly formulated and agreed upon, participants shared that they would next put the required resources in place. Participants would start to design the interactions in parallel with the technical development. Participants explored many possibilities. They also spoke about

assessing their interaction designs based on following criteria:

- *Could users produce effective labels needed to train the ML system?*
- *Could users make sense of the ML inference [or adaptation] and did they view it as valuable?*
- *Could users easily recover from ML [inference] errors?*

Participants used a combination of traditional UX methods such as user studies, sketching, and usability testing. They also engaged in continued negotiations with technologists. Collectively, these activities allowed them to craft the user experience. This iterative process continued until the interaction design had matured to a point it could be handed over to a front-end development team.

Today, most new features for online products and services only see broad release after passing a series of A/B tests, and the same process was used for ML features. These tests are meant to reduce the risk of user abandonment and/or reduction in new user conversion rates. Almost all participants spoke of A/B testing as a critical part of improving their design, and this related to both improving the algorithms' performance and improving the interaction. In most cases, the participants shared that they offered users the option to turn a new ML feature off. This appeared to help them view passing of the A/B tests as less challenging.

6.2.2.3 Stage 3 - Release and Continued Refinement

Few of the projects described by participants made it to stage 3. One notable exception was P7's worked on a message classifier. At the beginning of stage 3, they worked to improve the poor adoption rate. This became "*a really key moment in thinking about design of systems like this*". Stage 3 had to do with refinement that could make a project successful.

> *We had a classifier that predicts whether or not an incoming message is important … and the performance is scary good. … However, a vast majority of users didn't turn the feature on or soon opted out. When you dug into why that was, people would say "I don't even know what's important to me, how am I gonna trust the ML system to know it's important to me?"*
>
> *The core insight my research team had was that people are trying to figure out whether they are gonna trust that system, and the way they figure it out to think of it as a person. If I hand a stack of messages to a person, someone I don't know, is it possible that person could figure out*

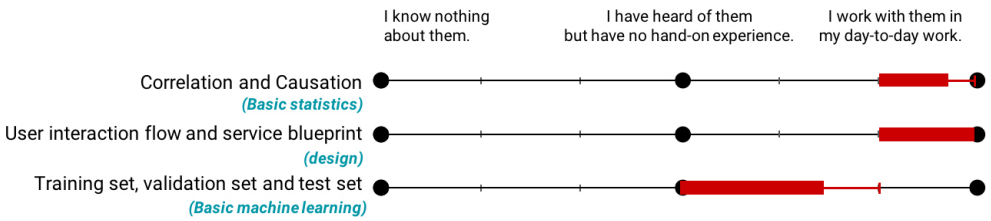

Figure 6.1: Participants' familiarity with concepts from statistics, UX design and machine learning.

> *with some accuracy if that message is important to me. Most people's take was "no, I barely know it myself".*

Stage 3 sent the UX designers into the field to investigate how people used their system in order to improve the design. The team improved the design by moving away from a binary classification (important or not important) to a small set of categories. The new design made it clear why something might not be important by classifying it as "promotional materials," something the team felt any human would be able to do. The new design used the same ML technology and proved to be a huge success.

The few participants who had projects reach stage 3 stated that such a re-design was inevitable because some mental model issues are difficult to catch before products launch. They felt they "*have to do this level of work on the design side*" (P2, 7, 13) so that users cannot only recover from the errors, but can understand the errors, and at the same time preserve some level of trust in the system.

> P1 describing why design problems for a chatbot auto-reply feature were unforeseeable: *We help you [users] say that you already gonna say. We do it a bit faster. We are actually influencing what you are saying, but not predicting what you are saying.*

6.2.3 New Design Activities To Embrace a Data-Driven Culture

I probed participants on what they needed to be effective and what UX students should know in order to effectively envision ML innovations. Their responses collectively revealed an acknowledgement that they all worked in data centric environments, and that designers needed to embrace this data centricity in order to have impact. They spoke of the importance of learning to speak the language of quantitative data and data science (e.g., telemetry, analytics, A/B testing, covariance, correlation). In the pre-interview surveys, participants responded that they worked with these concepts constantly (Figure 6.1). "*This is how engineers measure and businesses do things. This can influence how you design.*"

To help embrace a more data-centric culture, participants shared that they engaged in new design activities. They used a combination of qualitative and quantitative methods to advocate for their design. Specifically, they developed new skills around collecting telemetry data (data remotely collected from current products and services), and they generated data visualizations as one way of making sense of their data. In addition to working with data, participants shared that their collaborating data scientists share their role as a user advocate and as facilitator of incorporating domain experts' insights; something that did not happen with other technical development efforts.

6.2.3.1 Designing Telemetry Data

Participants frequently asked ML specialists about how people use the product, how they feel about it, and what problems they had that could be rectified by design. To help answer these questions, participants shared that they designed telemetry systems so their products can better track users' interactions. They used the data to capture meaningful and accurate snapshots of user behaviors.

Designers began to define their design goals in relation to the user behavioral data collected telemetrically, in the form of a "*matrix for a good experience*" (P2, 8, 12). This matrix would later be used to measure the success of their designs through a suite of A/B tests. Designers use these matrices to "*influence the engineers to think a different way, not in terms of single A/B test, but a suite of tests.*" The telemetry data thus became a space of exchange between designers and data scientists. The data helped to expand the narrow scope of A/B tests and to more holistically address user experience over a larger course of interactions.

6.2.3.2 Designing Data Visualizations

Participants worked to interpret user behavior data. Most utilized customized dashboards that translated mundane log data snapshots into user stories and insights. Several taught themselves to create data visualization tools and visualizations for the data scientists, engineers, and fellow designers in their organizations. They designed dashboards and visualizations to combine an immersive and an analytical way of understanding, so that the quantitative analysis of user behavior "*do not privilege data scientists*".

> *Data scientists have their methods, and I have my bag of tricks. They have kernels and clusters, and we are good at telling rich, compelling user stories. How can we look at hundreds and*

thousands of attempts, and also reach out to inquire more about what happened, what the breakdown was? (P14)

A few designers further utilized these data visualizations as vehicles for conducting user studies. One participant recalled that the first thing he did when joining the project was to list "*a series of questions about user behavior that data analytics can answer*" (P8). They then bookmarked the corresponding data matrix on the telemetry analytic tool the team used. Every morning they checked the new incoming log data against this matrix. When they noticed an intriguing user behavior, they emailed the user to inquire the back story, details, and sometimes set up follow-up interviews. They labeled this method as "*qualitative study in a quantitative scale*."

6.3 Reflecting on UX Design Expertise and Methods for AI

6.3.1 Towards Designerly Understandings of AI

The case studies in the previous chapter demonstrated the complexities around understanding AI systems' design affordance and limits. This study echoes this challenges. The designers comprehend ML in notably different ways than its textbook definitions; they understand ML largely through abstractions and examplars.

The abstractions served as a frame through which designers reflected on the design challenges at hand and made new assertions about how ML might provide value for users. They freed designers from grappling with technical limitations when sketching, empowering them to envision ML applications that moved beyond current archetypical forms. They served as boundary objects, allowing designers to discuss what users value with data scientists and to address issues of context. They also fostered new design ideas, serving as bridges between technical capabilities and design possibilities.

I suspect that many of the abstractions participants shared would generalize well beyond the specific applications they were working on. For example, participants stated that ML enables "*an experience personalized for everyone*" (P3), "*an evolving relationship with the users*" (P14), and "*handling more abstract user instructions*" (P7), and these typically matched with other UX values that HCI research has raised over time [42, 135]. Extending, evaluating, and documenting these abstractions offers a clear space for design research.

A robust set of these abstractions would help to evolve the understanding of ML as a design material. It could function as a kind of taxonomy that is likely to be radically different from ones used by data scientists; a taxonomy focused on the match of contextual capability and user value.

6.3.2 Boundary Objects for Bridging UX and AI Technical Expertise

The practitioner interviews captured an intimate, constant, cross-disciplinary collaboration when creating ML products. This is somewhat different from the design and technical collaborations found in a traditional UX design [24], where designers typically deliver a fully formed design to a technical team to implement. I see opportunities for new collaboration tools that help designers better work with data scientists. Previous work, including the case study in the previous chapter, has proposed the use of boundary objects that scaffold the conversation between UX and ML expertise in creating AI-mediated interactions [134]. Similar work could be potentially valuable for many other application domains of ML.

The interviews showed that designers who lacked effective access to data scientists explored fewer design ideas and more often quickly resorted to familiar designs of ML. Previous design research has focused on enabling and improving designer and data scientist collaboration, assuming that capable data scientists are readily available [42, 134]. However, this study indicates that this assumption in not always true. Many designers in our study lacked access to dedicated or even proficient data scientists, especially the ones working at startups, small technology companies, and non-IT-focused companies. Despite efforts to make ML available to everyone [117, 122], the fact that proficient data scientists were scarce might be a reality that most designers will face.

There is a real need for design tools and methodologies that support designers who lack constant access to capable data scientists. For example, ML tools for designers could simulate the role of the data scientists, enabling designers to quickly evaluate the feasibility of their ideas when sketching. I also see opportunities for constructive design research to demonstrate creative designs that use off-the-shelf, ML plug-ins; designs that do not need intensive ML development effort to implement.

6.3.3 Designing and Evaluating the UX of Evolving AI Systems

Previously in the "sketching NLP" case study (5.5.3), I raised the question of how to evaluate the UX design of evolving AI systems. These systems continue to learn post-deployment, and therefore their technical performance and error modes are constantly fluctuating and evolving with user interaction traces. The practices of experienced UX designers offer one possible solution – a divide-and-conquer approach. They first identified the right problem for the system to solve, validated its technical feasibility, and then iterated on the diverging-converging ideation process to craft the right manifestation of the intelligent functionality. In this regard, I found their overall approach to be similar to mine in the sketching NLP project.

What is unique about the experienced practitioners' approach, however, is the new design activities they have undertaken in order to anticipate and monitor the users' interaction traces. They taught themselves to capture "rich and compelling user stories" from telemetry data that were different from the data scientists' insights. By "playing with" data, designers frequently ask: How do users use the system? How will their interactions affect the systems' behaviors? What is the matrix for a good user experience? Answers to these questions would help designers anticipate and design the co-evolvement of user and system behaviors (Figure 6.2 yellow parts).

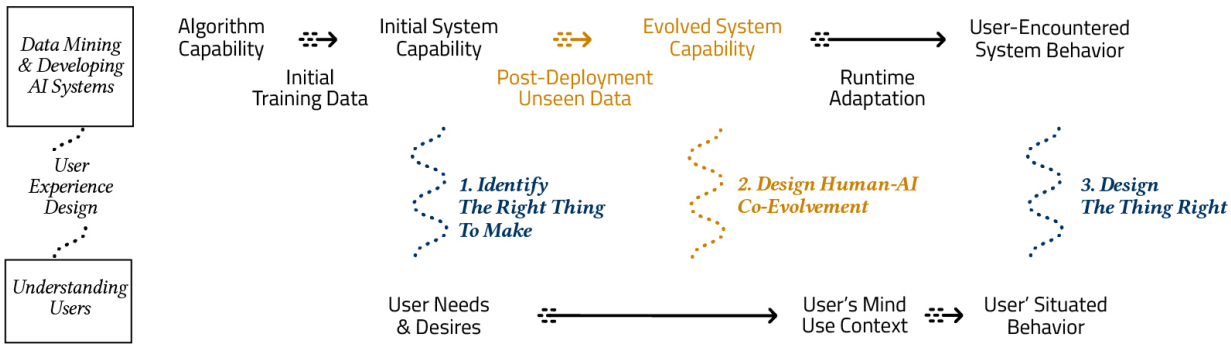

Figure 6.2: The current industry best practice in designing and developing UX of AI, against the backdrop of the AI design complexity framework.

Future research should examine and improve this design process. One advantage of this process is that it did not require designers to acquire extensive ML skills, use many new tools, or radically change their familiar design activities. After a working ML application was launched, designers would

go through a second design and evaluation process, fixing the design problems that the initial iteration failed to capture. However, a complete ML design process seemed to take way more time than a conventional Double Diamond UX process [24]. Can the function and form of human-AI interaction design be addressed at once? Are there new ways to prototype and evaluate the technical viability and design quality before product launch? Answers to these questions have the potential to radically lower the cost of developing ML systems, and to enable many more resource-sensitive organizations to introduce ML to their products and services.

At a higher level, the procedural knowledge of designing AI marks a clear space for HCI design research. Existing work has offered valuable declarative knowledge and conceptual understandings of ML from a design perspective (i.e., [65, 70]). Embedding this growing body of new knowledge into organizational and procedural contexts opens up new research opportunities and promises real impact on UX practice. The AI design complexity framework has offered an initial outline of the problem space. Now it is an opportune time for the HCI design community to have a reflective discussion on how designers might most productively traverse this problem space, in order to create novel and meaningful user experiences.

Chapter 7

Summary and Future Work

From predictive medicine to autonomous driving, AI promises to improve people's lives and societies. On the path to realize these promises, new challenges have emerged as these systems migrate from research labs into the real world. How should predictive models integrate into physicians' decision-making processes, such that the predictions affect them appropriately? How could natural language generation systems provide personalized, valuable writing suggestions without being perceived as plagiarism? These are challenges of translation: translating AI's remarkable technological advance in research labs into real-world sociotechnical systems valuable to human efforts. These are challenges where UX design expertise has much to offer and challenges that design practitioners and academics will continue to face in the future.

The case studies in the dissertation work have illustrated these challenges in addition to the value UX design expertise brings to them. I consider this work as a continuation of the many prior HCI research efforts of the human-centered design tradition. When computers started evolving from something that only trained operators use to one that everyone can use, user-centered design methods and processes emerged in response [84]. When the tech industry began moving from hardware products to hardware platforms (e.g., from mobile phone as a hardware product to smartphones as one component of a software-hardware-service system), service design techniques such as service blueprinting emerged in response [144]. In these previous waves of technological advances, new design methods and processes followed. They respond to the design complexities of the new technology and supports designers in bringing their human-centered sensitivities to bear on these challenges.

Prior research in the intersection of UX and AI suggests (possibly unknowingly) that we are at the cusp of a similar transformation. Designers have reported various challenges in integrating AI into their current practice. They have speculated that familiar user-centered design methods and processes need to change for AI [38, 42]. However, the core characteristics of AI that create a misfit to existing UX design methods have rarely been formally studied. As a result, there exists no principled discussion on to what extent and how UX design needs to change for AI.

It is in this context that I undertake this investigation into whether, when, and how AI is uniquely difficult to design. I have arrived on an initial AI design complexity framework that can explicate the nebulous challenges of human-AI interaction design and trace them back to just two root challenges: evolving capabilities whose limits are difficult for designers to grasp and complex, adaptive interactions that resist simulation. I argued that systems that share these characteristics – whether they technically are "AI" or not – problematize the conventional HCI prototyping methods that treat technology's affordance as static or bounded and interactions as prescriptive. These systems call for new UX design methods and tools.

This framework is not fixed nor final. It is limited in that it draws mainly from my own research, design, and teaching experiences. The case studies presented in this work, though illustrative, are neither a representative sample nor a comprehensive one. The meta-analysis nature of my research goal calls for an extensive collection of AI design projects, ideally covering *all* kinds of AI systems for *all* kinds of design contexts. This is beyond what the course of a Ph.D. can achieve. The synthesis of my research experience and the resulting framework is intended to serve as a moderate first step in the direction towards fostering an accurate and insightful understanding of AI as a design material. I hope more researchers will join me, reflecting on their respective design and research experiences, critiquing and improving this framework.

This is not to undermine the advance this dissertation makes. I argue that its primary contribution is **framing AI as a *material* for UX design**. It takes as a starting point a move from design as an afterthought to machine learning and system building, to design as a way of thinking and acting about what AI thing to make in the first place. As trivial as it might sound to those from a human-centered tradition, this is a largely missing perspective in the current human-AI interaction research. Most have instead solely focused on UX design of AI as issues of better communicating predictions in order for

users to adopt, appreciate, or follow AI's right suggestions while rejecting its errors. The case studies presented in this work have provided vivid examples: Despite a large body of research in AI, healthcare, and HCI that has focused on the adoption challenges decision support systems face, few have studied how clinicians work and how AI might fit into their day-to-day work. Despite much research on natural language generation (NLG) and its human evaluation, few have asked: What do authors want and don't want from the technology? Also illustrated in the case studies is the impact asking these questions could bring to AI: Clinical machine learning systems can more effectively impact clinicians' decision making, not only in research labs, but in clinical practices. NLG systems can mitigate and even prevent user adoption challenges when they are created with human needs and agencies in mind from the start.

At a higher level, I analogize the implications of taking AI as a design material to those of Weiser's vision of *Ubiquitous Computing* [127]. Both break away from the convention of designing a more user-friendly AI system/desktop computer, and instead systematically reshape AI/computation to better fit in people's day-to-day lives. Both mark a paradigm shift that calls for:

1. *New ethnographic and design work* that re-imagines when, where, and how people may perceive AI/computation as valuable;
2. A new wave of *technological and interaction design innovations* that together enable AI/computing devices to deliver value when, where, and how it is needed;
3. A renewed understanding of *what "AI"/"computer" is and means* for people and for societies at large.
4. New *practitioner-facing processes and tools* that then integrate and transfer the above innovations and insights to practicing communities, thereby materializing AI/computation's impact in everyday technology products and services.

This dissertation offers an initial step along these lines, towards this ambitious paradigm shift.

Bibliography

[1] Ashraf Abdul, Jo Vermeulen, Danding Wang, Brian Y. Lim, and Mohan Kankanhalli. Trends and trajectories for explainable, accountable and intelligible systems: An hci research agenda. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, CHI '18, pages 582:1–582:18, New York, NY, USA, 2018. ACM. ISBN 978-1-4503-5620-6. doi: 10.1145/3173574.3174156. URL http://doi.acm.org/10.1145/3173574.3174156. 2.2.1

[2] Saleema Amershi, Dan Weld, Mihaela Vorvoreanu, Adam Fourney, Besmira Nushi, Penny Collisson, Jina Suh, Shamsi Iqbal, Paul N. Bennett, Kori Inkpen, Jaime Teevan, Ruth Kikin-Gil, and Eric Horvitz. Guidelines for human-ai interaction. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, CHI '19, page 1–13, New York, NY, USA, 2019. Association for Computing Machinery. ISBN 9781450359702. doi: 10.1145/3290605.3300233. URL https://doi.org/10.1145/3290605.3300233. 2.2.1, 4, 3.3.2

[3] Jonathan Bean and Daniela Rosner. Big data, diminished design? *interactions*, 21(3):18–19, 2014. 5.1

[4] Emma Beede, Elizabeth Baylor, Fred Hersch, Anna Iurchenko, Lauren Wilcox, Paisan Ruamviboonsuk, and Laura M. Vardoulakis. A human-centered evaluation of a deep learning system deployed in clinics for the detection of diabetic retinopathy. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, CHI '20, page 1–12, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450367080. doi: 10.1145/3313831.3376718. URL https://doi.org/10.1145/3313831.3376718. 1.1.1

[5] Emma Beede, Elizabeth Baylor, Fred Hersch, Anna Iurchenko, Lauren Wilcox, Paisan Ruamviboonsuk, and Laura M. Vardoulakis. A human-centered evaluation of a deep learning system deployed in clinics for the detection of diabetic retinopathy. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, CHI '20, page 1–12, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450367080. doi: 10.1145/3313831.3376718. URL https://doi.org/10.1145/3313831.3376718. 4.1, 4.5.1

[6] Michael Z. Bell. Why expert systems fail. *The Journal of the Operational Research Society*, 36(7):613–619, 1985. ISSN 01605682, 14769360. URL http://www.jstor.org/stable/2582480. 1.1.1, 4.1

[7] Riccardo Bellazzi and Blaz Zupan. Predictive data mining in clinical medicine: current issues and guidelines. *International journal of medical informatics*, 77(2):81–97, 2008. 4.1

[8] Raymond L Benza, Dave P Miller, Robyn J Barst, David B Badesch, Adaani E Frost, and Michael D McGoon. An evaluation of long-term survival from time of diagnosis in pulmonary arterial hypertension from the reveal registry. *CHEST Journal*, 142(2):448–456, 2012. 4.1

[9] Steven Bird, Ewan Klein, and Edward Loper. *Natural language processing with Python: analyzing*

text with the natural language toolkit. " O'Reilly Media, Inc.", 2009. 5.2

[10] Mary Jo Bitner, Amy L. Ostrom, and Felicia N. Morgan. Service blueprinting: A practical technique for service innovation. *California Management Review*, 50(3):66–94, 2008. doi: 10.2307/41166446. URL https://doi.org/10.2307/41166446. 4.2.1

[11] Sara Bly and Elizabeth F Churchill. Design through matchmaking: technology in search of users. *interactions*, 6(2):23–31, 1999. (document), 2.1, 2.1.1, 2.2.1, 2.3

[12] Kirsten Boehner and Carl DiSalvo. Data, design and civics: An exploratory study of civic tech. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, CHI '16, pages 2970–2981, New York, NY, USA, 2016. ACM. ISBN 978-1-4503-3362-7. doi: 10.1145/2858036.2858326. URL http://doi.acm.org/10.1145/2858036.2858326. 2.2, 2

[13] Sander Bogers, Joep Frens, Janne van Kollenburg, Eva Deckers, and Caroline Hummels. Connected baby bottle: A design case study towards a framework for data-enabled design. In *Proceedings of the 2016 ACM Conference on Designing Interactive Systems*, DIS '16, pages 301–311, New York, NY, USA, 2016. ACM. ISBN 978-1-4503-4031-1. doi: 10.1145/2901790.2901855. URL http://doi.acm.org/10.1145/2901790.2901855. 2, 3.3.3

[14] Geoffrey C Bowker and Susan Leigh Star. *Sorting things out: Classification and its consequences*. MIT press, 2000. 2.1.3, 5.4.3

[15] Tone Bratteteig and Guri Verne. Does ai make pd obsolete?: Exploring challenges from artificial intelligence to participatory design. In *Proceedings of the 15th Participatory Design Conference: Short Papers, Situated Actions, Workshops and Tutorial - Volume 2*, PDC '18, pages 8:1–8:5, New York, NY, USA, 2018. ACM. ISBN 978-1-4503-5574-2. doi: 10.1145/3210604.3210646. URL http://doi.acm.org/10.1145/3210604.3210646. 2.2

[16] Richard Buchanan. Wicked problems in design thinking. *Design issues*, 8(2):5–21, 1992. 5.4.5

[17] Marion Buchenau and Jane Fulton Suri. Experience prototyping. In *Proceedings of the 3rd Conference on Designing Interactive Systems: Processes, Practices, Methods, and Techniques*, DIS '00, page 424–433, New York, NY, USA, 2000. Association for Computing Machinery. ISBN 1581132190. doi: 10.1145/347642.347802. URL https://doi.org/10.1145/347642.347802. 4.5.2

[18] Bill Buxton. *Sketching user experiences: getting the design right and the right design*. Morgan Kaufmann, 2010. 2.1.1, 2.1.1, 2.1.2, 4.5.1, 5.1, 5.5.1

[19] Carrie J. Cai, Emily Reif, Narayan Hegde, Jason Hipp, Been Kim, Daniel Smilkov, Martin Wattenberg, Fernanda Viegas, Greg S. Corrado, Martin C. Stumpe, and Michael Terry. Human-centered tools for coping with imperfect algorithms during medical decision-making. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, CHI '19, page 1–14, New York, NY, USA, 2019. Association for Computing Machinery. ISBN 9781450359702. doi: 10.1145/3290605.3300234. URL https://doi.org/10.1145/3290605.3300234. 1.1.1, 4.5.1

[20] Shan Carter and Michael Nielsen. Using artificial intelligence to augment human intelligence. *Distill*, 2017. doi: 10.23915/distill.00009. https://distill.pub/2017/aia. 1

[21] Amber Cartwright. Invisible Design: Co-Designing with Machines, 2016. URL http://airbnb.design/invisible-design/. 2.2.1, 1

[22] Ana Paula Chaves and Marco Aurelio Gerosa. Single or multiple conversational agents?: An interactional coherence comparison. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, CHI '18, pages 191:1–191:13, New York, NY, USA, 2018. ACM. ISBN

978-1-4503-5620-6. doi: 10.1145/3173574.3173765. URL http://doi.acm.org/10.1145/3173574.3173765. 5.2

[23] Elizabeth Clark, Anne Spencer Ross, Chenhao Tan, Yangfeng Ji, and Noah A. Smith. Creative writing with a machine in the loop: Case studies on slogans and stories. In *23rd International Conference on Intelligent User Interfaces*, IUI '18, pages 329–340, New York, NY, USA, 2018. ACM. ISBN 978-1-4503-4945-1. doi: 10.1145/3172944.3172983. URL http://doi.acm.org/10.1145/3172944.3172983. 5.1, 5.2, 5.4.6.3

[24] Design Council. The 'double diamond' design process model. *Design Council*, 2005. (document), 1.1.1, 2.1.1, 2.2, 5.4.1, 6.3.2, 6.3.3

[25] David Coyle and Gavin Doherty. Clinical evaluations and collaborative design: Developing new technologies for mental healthcare interventions. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '09, pages 2051–2060, New York, NY, USA, 2009. ACM. ISBN 978-1-60558-246-7. doi: 10.1145/1518701.1519013. URL http://doi.acm.org/10.1145/1518701.1519013. 4.5.1

[26] Justin Cranshaw, Emad Elwany, Todd Newman, Rafal Kocielnik, Bowen Yu, Sandeep Soni, Jaime Teevan, and Andrés Monroy-Hernández. Calendar. help: Designing a workflow-based scheduling agent with humans in the loop. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, pages 2382–2393. ACM, 2017. 2.2.1

[27] Justin Cranshaw, Emad Elwany, Todd Newman, Rafal Kocielnik, Bowen Yu, Sandeep Soni, Jaime Teevan, and Andrés Monroy-Hernández. Calendar. help: Designing a workflow-based scheduling agent with humans in the loop. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, pages 2382–2393. ACM, 2017. 5.2

[28] Nigel Cross. Designerly ways of knowing: Design discipline versus design science. *Design Issues*, 17(3):49–55, 2001. ISSN 07479360, 15314790. URL http://www.jstor.org/stable/1511801. 2.1.3

[29] Srikant Devaraj, Sushil K Sharma, Dyan J Fausto, Sara Viernes, and Hadi Kharrazi. Barriers and facilitators to clinical decision support systems adoption: A systematic review. *Journal of Business Administration Research*, 3(2):p36, 2014. 1.1.1, 4.1

[30] Graham Dove, Kim Halskov, Jodi Forlizzi, and John Zimmerman. 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*, pages 278–288, New York, New York, USA, 2017. ACM Press. ISBN 9781450346559. doi: 10.1145/3025453.3025739. URL http://dl.acm.org/citation.cfm?doid=3025453.3025739. (document), 2, 3, 2.2.1, 2.2, 2.2.1, 2.2.1, 3.2.2, 4.5.2, 5.1, 5.5.2

[31] Pelle Ehn. Scandinavian design: On participation and skill. *Participatory design: Principles and practices*, pages 41–77, 1993. 1.1.1, 4.3.2

[32] Glyn Elwyn, Isabelle Scholl, Caroline Tietbohl, Mala Mann, Adrian GK Edwards, Catharine Clay, France Légaré, Trudy van der Weijden, Carmen L Lewis, Richard M Wexler, et al. "many miles to go...": a systematic review of the implementation of patient decision support interventions into routine clinical practice. *BMC medical informatics and decision making*, 13(Suppl 2):S14, 2013. 1.1.1, 4.1

[33] Daniel Fallman. Design-oriented human-computer interaction. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 225–232. ACM, 2003. 2.1.2, 5.1

[34] Melanie Feinberg. A design perspective on data. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, CHI '17, pages 2952–2963, New York, NY, USA, 2017. ACM. ISBN 978-1-4503-4655-9. doi: 10.1145/3025453.3025837. URL http://doi.acm.org/10.1145/3025453.3025837. 2.2, 2

[35] Melanie Feinberg. A design perspective on data. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, pages 2952–2963. ACM, 2017. 3

[36] Rebecca Fiebrink and Perry R Cook. The wekinator: a system for real-time, interactive machine learning in music. In *Proceedings of The Eleventh International Society for Music Information Retrieval Conference (ISMIR 2010)(Utrecht)*, 2010. 3

[37] US Food and Drug Administration (FDA). Proposed regulatory framework for modifications to artificial intelligence/ machine learning (ai/ml)-based software as a medical device (samd)—discussion paper and request for feedback. Technical report, US Food and Drug Administration (FDA), 2019. 4

[38] Jodi Forlizzi. Moving beyond user-centered design. *Interactions*, 25(5):22–23, August 2018. ISSN 1072-5520. doi: 10.1145/3239558. URL http://doi.acm.org/10.1145/3239558. 5, 7

[39] Matt Gardner, Joel Grus, Mark Neumann, Oyvind Tafjord, Pradeep Dasigi, Nelson F. Liu, Matthew Peters, Michael Schmitz, and Luke Zettlemoyer. AllenNLP: A deep semantic natural language processing platform. In *Proceedings of Workshop for NLP Open Source Software (NLP-OSS)*, pages 1–6, Melbourne, Australia, July 2018. Association for Computational Linguistics. doi: 10.18653/v1/W18-2501. URL https://www.aclweb.org/anthology/W18-2501. 5.4.4

[40] William Gaver. What should we expect from research through design? In *Proceedings of the SIGCHI conference on human factors in computing systems*, pages 937–946. ACM, 2012. 5.3.1

[41] Marco Gillies, Bongshin Lee, Nicolas D'Alessandro, Joëlle Tilmanne, Todd Kulesza, Baptiste Caramiaux, Rebecca Fiebrink, Atau Tanaka, Jérémie Garcia, Frédéric Bevilacqua, Alexis Heloir, Fabrizio Nunnari, Wendy Mackay, and Saleema Amershi. Human-Centred Machine Learning. In *Proceedings of the 2016 CHI Conference Extended Abstracts on Human Factors in Computing Systems - CHI EA '16*, pages 3558–3565, New York, New York, USA, 2016. ACM Press. ISBN 9781450340823. doi: 10.1145/2851581.2856492. URL http://dl.acm.org/citation.cfm?doid=2851581.2856492. 2, 2.2

[42] Fabien Girardin and Neal Lathia. When user experience designers partner with data scientists. In *The AAAI Spring Symposium Series Technical Report: Designing the User Experience of Machine Learning Systems*. The AAAI Press, Palo Alto, California, 2017. ISBN 978-1-57735-754-4. URL https://www.aaai.org/ocs/index.php/SSS/SSS17/paper/view/15364. 3, 2.2.1, 2.2.1, 5, 3.2.2, 6.3.1, 6.3.2, 7

[43] Randy Goebel, Ajay Chander, Katharina Holzinger, Freddy Lecue, Zeynep Akata, Simone Stumpf, Peter Kieseberg, and Andreas Holzinger. Explainable ai: The new 42? In Andreas Holzinger, Peter Kieseberg, A Min Tjoa, and Edgar Weippl, editors, *Machine Learning and Knowledge Extraction*, pages 295–303, Cham, 2018. Springer International Publishing. ISBN 978-3-319-99740-7. 2.2.1

[44] Mayank Goel, Nils Hammerla, Thomas Ploetz, and Anind K. Dey. Bridging the Gap: Machine Learning for Ubicomp - Tutorial @UbiComp 2015. https://openlab.ncl.ac.uk/bridging-the-gap/, 2015. 1

[45] Gabriela Goldschmidt. The dialectics of sketching. *Creativity research journal*, 4(2):123–143, 1991. 2.1.2

[46] Elizabeth Goodman, Erik Stolterman, and Ron Wakkary. Understanding interaction design practices. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '11, pages 1061–1070, New York, NY, USA, 2011. ACM. ISBN 978-1-4503-0228-9. doi: 10.1145/1978942.1979100. URL http://doi.acm.org/10.1145/1978942.1979100. 1.3, 5.3.1

[47] Google. People + ai guidebook: Designing human-centered ai products. pair.withgoogle.com/, 2019. 4

[48] Patrick Hebron. *Machine learning for designers*. O'Reilly Media, 2016. 1, 5.1

[49] Patrick Hebron. New York University Tisch School of the Arts Course: Learning Machines, 2016. URL http://www.patrickhebron.com/learning-machines/. 1

[50] Karey Helms. Do you have to pee?: A design space for intimate and somatic data. In *Proceedings of the 2019 on Designing Interactive Systems Conference*, DIS '19, pages 1209–1222, New York, NY, USA, 2019. ACM. ISBN 978-1-4503-5850-7. doi: 10.1145/3322276.3322290. URL http://doi.acm.org/10.1145/3322276.3322290. 2

[51] Douglas R Hofstadter et al. *Gödel, Escher, Bach: an eternal golden braid*, volume 20. Basic books New York, 1979. 1.2

[52] Lars Erik Holmquist. Intelligence on tap: artificial intelligence as a new design material. *interactions*, 24(4):28–33, 2017. (document), 2, 3, 2.2, 2.2.1, 2.2, 2.2.1, 4.5.2

[53] Monique WM Jaspers, Marian Smeulers, Hester Vermeulen, and Linda W Peute. Effects of clinical decision-support systems on practitioner performance and patient outcomes: a synthesis of high-quality systematic review findings. *Journal of the American Medical Informatics Association*, 18(3): 327–334, 2011. 4.1

[54] Andreas Kaplan and Michael Haenlein. Siri, siri, in my hand: Who's the fairest in the land? on the interpretations, illustrations, and implications of artificial intelligence. *Business Horizons*, 62 (1):15–25, 2019. 1.2, 3.1

[55] Kensaku Kawamoto, Caitlin A Houlihan, E Andrew Balas, and David F Lobach. Improving clinical practice using clinical decision support systems: a systematic review of trials to identify features critical to success. *Bmj*, 330(7494):765, 2005. 4.1

[56] Claire Kayacik, Sherol Chen, Signe Noerly, Jess Holbrook, Adam Roberts, and Douglas Eck. 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*, CHI EA '19, pages CS09:1–CS09:8, New York, NY, USA, 2019. ACM. ISBN 978-1-4503-5971-9. doi: 10.1145/3290607.3299059. URL http://doi.acm.org/10.1145/3290607.3299059. 2.2.1

[57] Bogyeong Kim, Jaehoon Pyun, and Woohun Lee. Enhancing storytelling experience with story-aware interactive puppet. In *Extended Abstracts of the 2018 CHI Conference on Human Factors in Computing Systems*, CHI EA '18, pages LBW076:1–LBW076:6, New York, NY, USA, 2018. ACM. ISBN 978-1-4503-5621-3. doi: 10.1145/3170427.3188515. URL http://doi.acm.org/10.1145/3170427.3188515. 5.2

[58] Scott R. Klemmer, Anoop K. Sinha, Jack Chen, James A. Landay, Nadeem Aboobaker, and Annie Wang. Suede: A wizard of oz prototyping tool for speech user interfaces. In *Proceedings of the 13th Annual ACM Symposium on User Interface Software and Technology*, UIST '00, pages 1–10, New York, NY, USA, 2000. ACM. ISBN 1-58113-212-3. doi: 10.1145/354401.354406. URL http://doi.acm.org/10.1145/354401.354406. 2.2.1, 5.2, 5.5.1

[59] Ajay Kohli and Saurabh Jha. Why cad failed in mammography. *Journal of the American College of Radiology*, 15(3):535–537, 2018. 1.1.1

[60] Mike Kuniavsky, Elizabeth Churchill, and Molly Wright Steenson. Designing the user experience of machine learning systems, papers from the 2017 AAAI spring symposium, technical report ss-17-04, palo alto, california, usa, march 27-29, 2017, 2017. URL https://mikek-parc.github.io/AAAI-UX-ML/. 2, 2.2

[61] Mike Kuniavsky, Elizabeth Churchill, Molly Wright Steenson, and Phil Van Allen. The design of the user experience for artificial intelligence (the ux of ai), papers from the 2018 AAAI spring symposium, palo alto, california, usa, march 26-28, 2018, 2018. URL https://mikek-parc.github.io/AAAI-UX-AI/. 2, 2.2

[62] Esko Kurvinen, Ilpo Koskinen, and Katja Battarbee. Prototyping social interaction. *Design Issues*, 24(3):46–57, 2008. 3.3.3

[63] Shane Legg and Marcus Hutter. A collection of definitions of intelligence. In *Proceedings of the 2007 Conference on Advances in Artificial General Intelligence: Concepts, Architectures and Algorithms: Proceedings of the AGI Workshop 2006*, page 17–24, NLD, 2007. IOS Press. ISBN 9781586037581. 1.2

[64] Constance D Lehman, Robert D Wellman, Diana SM Buist, Karla Kerlikowske, Anna NA Tosteson, and Diana L Miglioretti. Diagnostic accuracy of digital screening mammography with and without computer-aided detection. *JAMA internal medicine*, 175(11):1828–1837, 2015. 4.1

[65] Brian Y Lim, Anind K Dey, and Daniel Avrahami. Why and why not explanations improve the intelligibility of context-aware intelligent systems. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 2119–2128. ACM, 2009. 2.2, 6.3.3

[66] Youn-Kyung Lim, Erik Stolterman, and Josh Tenenberg. The anatomy of prototypes: Prototypes as filters, prototypes as manifestations of design ideas. *ACM Transactions on Computer-Human Interaction (TOCHI)*, 15(2):7, 2008. 2.1.2, 4.5.2

[67] Tilmann Lindberg, Christoph Meinel, and Ralf Wagner. Design thinking: A fruitful concept for it development? In *Design thinking: Understand – Improve – Apply*, pages 3–18. Springer, 2011. 2.1.1

[68] Panagiotis Louridas. Design as bricolage: anthropology meets design thinking. *Design Studies*, 20 (6):517–535, 1999. 2, 2.1.1, 2.1.3

[69] Hassenzahl Marc and Tractinsky Noam. User experience - a research agenda. *Behaviour & Information Technology*, 25(2):91–97, 2006. doi: 10.1080/01449290500330331. URL https://doi.org/10.1080/01449290500330331. 2.1.1

[70] Betti Marenko and Philip Van Allen. Animistic design: how to reimagine digital interaction between the human and the nonhuman. *Digital Creativity*, 27(1):52–70, 2016. 6.3.3

[71] Yatin Mehta, Abhinav Gupta, Subhash Todi, SN Myatra, DP Samaddar, Vijaya Patil, Pradip Kumar Bhattacharya, and Suresh Ramasubban. Guidelines for prevention of hospital acquired infections. *Indian journal of critical care medicine: peer-reviewed, official publication of Indian Society of Critical Care Medicine*, 18(3):149, 2014. 4.2.3.4

[72] Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. Model cards for model reporting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, FAT* '19, page 220–229, New York, NY, USA, 2019. Association for Computing Machinery. ISBN 9781450361255. doi: 10.1145/3287560.3287596. URL https://doi.org/10.1145/3287560.3287596. 3.3.2

[73] Bill Moggridge and Bill Atkinson. *Designing interactions*, volume 17. MIT press Cambridge, MA, 2007. 4.2.1, 4.4.1.1

[74] Enid Montague, John D Lee, and P Carayon. Trust in health technologies. *Handbook of human factors and ergonomics in health care and patient safety*, 2:281–291, 2012. 1.1.1

[75] Robert J. Moore, Raphael Arar, Guang-Jie Ren, and Margaret H. Szymanski. Conversational ux design. In *Proceedings of the 2017 CHI Conference Extended Abstracts on Human Factors in Computing Systems*, CHI EA '17, pages 492–497, New York, NY, USA, 2017. ACM. ISBN 978-1-4503-4656-6. doi: 10.1145/3027063.3027077. URL http://doi.acm.org/10.1145/3027063.3027077. 5.2

[76] Camille Moussette. *Simple haptics: Sketching perspectives for the design of haptic interactions*. PhD thesis, Umeå Universitet, 2012. 5.5.1

[77] Camille Moussette. *Simple haptics: Sketching perspectives for the design of haptic interactions*. PhD thesis, Umeå Universitet, 2012. 2.1.2, 2.1.3

[78] Cosmin Munteanu and Gerald Penn. Speech-based interaction: Myths, challenges, and opportunities. In *Proceedings of the 16th International Conference on Human-computer Interaction with Mobile Devices & Services*, MobileHCI '14, pages 567–568, New York, NY, USA, 2014. ACM. ISBN 978-1-4503-3004-6. doi: 10.1145/2628363.2645671. URL http://doi.acm.org/10.1145/2628363.2645671. 5.2

[79] Donald Murray. Teach writing as a process not product. *The Leaflet*, 71(3):11–14, 1972. 5.4.2

[80] Mark A Musen, Blackford Middleton, and Robert A Greenes. Clinical decision-support systems. In *Biomedical informatics*, pages 643–674. Springer, 2014. 4.1, 4.3

[81] Sara Nabil, David S. Kirk, Thomas Plötz, Julie Trueman, David Chatting, Dmitry Dereshev, and Patrick Olivier. Interioractive: Smart materials in the hands of designers and architects for designing interactive interiors. In *Proceedings of the 2017 Conference on Designing Interactive Systems*, DIS '17, page 379–390, New York, NY, USA, 2017. Association for Computing Machinery. ISBN 9781450349222. doi: 10.1145/3064663.3064745. URL https://doi.org/10.1145/3064663.3064745. 2.1.3

[82] GINA NEFF and PETER NAGY. Talking to bots: Symbiotic agency and the case of tay. *International Journal of Communication (19328036)*, 10, 2016. 3.2.1.1

[83] Jakob Nielsen and Rolf Molich. Heuristic evaluation of user interfaces. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '90, page 249–256, New York, NY, USA, 1990. Association for Computing Machinery. ISBN 0201509326. doi: 10.1145/97243.97281. URL https://doi.org/10.1145/97243.97281. 2.1.1

[84] Don Norman. *The design of everyday things: Revised and expanded edition*. Basic books, 2013. 7

[85] Don Norman and Jakob Nielsen. The definition of user experience (ux). *Nielsen Norman Group Publication*, 1, 2016. 2.1.1

[86] Marianna Obrist, Sue Ann Seah, and Sriram Subramanian. Talking about tactile experiences. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '13, pages 1659–1668, New York, NY, USA, 2013. ACM. ISBN 978-1-4503-1899-0. doi: 10.1145/2470654.2466220. URL http://doi.acm.org/10.1145/2470654.2466220. 2.1.3

[87] Marianna Obrist, Rob Comber, Sriram Subramanian, Betina Piqueras-Fiszman, Carlos Velasco, and Charles Spence. Temporal, affective, and embodied characteristics of taste experiences: A

framework for design. In *Proceedings of the 32Nd Annual ACM Conference on Human Factors in Computing Systems*, CHI '14, pages 2853–2862, New York, NY, USA, 2014. ACM. ISBN 978-1-4503-2473-1. doi: 10.1145/2556288.2557007. URL http://doi.acm.org/10.1145/2556288.2557007. 5.5.1

[88] William Odom and Tijs Duel. On the design of olo radio: Investigating metadata as a design material. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, CHI '18, pages 104:1–104:9, New York, NY, USA, 2018. ACM. ISBN 978-1-4503-5620-6. doi: 10.1145/3173574.3173678. URL http://doi.acm.org/10.1145/3173574.3173678. 2.2, 2

[89] Fatih Kursat Ozenc, Miso Kim, John Zimmerman, Stephen Oney, and Brad Myers. How to support designers in getting hold of the immaterial material of software. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 2513–2522. ACM, 2010. 2.1.3

[90] Annette M O'Connor, John E Wennberg, France Legare, Hilary A Llewellyn-Thomas, Benjamin W Moulton, Karen R Sepucha, Andrea G Sodano, and Jaime S King. Toward the 'tipping point': decision aids and informed patient choice. *Health Affairs*, 26(3):716–725, 2007. 4.3.2

[91] Kayur Patel, James Fogarty, James A. Landay, and Beverly Harrison. Examining difficulties software developers encounter in the adoption of statistical machine learning. In *23rd AAAI Conference on Artificial Intelligence and the 20th Innovative Applications of Artificial Intelligence Conference*, pages 1563–1566, Chicago, IL, United States, 2008. 3

[92] Kayur Dushyant Patel. *Lowering the Barrier to Applying Machine Learning*. PhD thesis, University of Washington, 2012. 3

[93] Owain Pedgley. Capturing and analysing own design activity. *Design studies*, 28(5):463–483, 2007. 5.3.1

[94] Brindha Pillay, Addie C Wootten, Helen Crowe, Niall Corcoran, Ben Tran, Patrick Bowden, Jane Crowe, and Anthony J Costello. The impact of multidisciplinary team meetings on patient assessment, management and outcomes in oncology settings: a systematic review of the literature. *Cancer treatment reviews*, 42:56–72, 2016. 4.3.2

[95] Johan Redström, Maria Redström, and Ramia Mazé. *IT+Textiles*. IT Press, 2005. ISBN 9518267936. URL http://eprints.sics.se/3632/. 5.5.1

[96] Laurel D. Riek. Wizard of oz studies in hri: A systematic review and new reporting guidelines. *J. Hum.-Robot Interact.*, 1(1):119–136, July 2012. ISSN 2163-0364. doi: 10.5898/JHRI.1.1.Riek. URL https://doi.org/10.5898/JHRI.1.1.Riek. 2.1.2, 2.2.1, 4.5.2

[97] Eric Ries. *The lean startup: How today's entrepreneurs use continuous innovation to create radically successful businesses*. Crown Books, 2011. (document), 2.1, 2.1.1, 2.2.1, 2.3

[98] Antonio Rizzo, Francesco Montefoschi, Maurizio Caporali, Antonio Gisondi, Giovanni Burresi, and Roberto Giorgi. Rapid prototyping iot solutions based on machine learning. In *Proceedings of the European Conference on Cognitive Ergonomics 2017*, ECCE 2017, pages 184–187, New York, NY, USA, 2017. ACM. ISBN 978-1-4503-5256-7. doi: 10.1145/3121283.3121291. URL http://doi.acm.org/10.1145/3121283.3121291. 2.2

[99] Mohammed Saeed, Mauricio Villarroel, Andrew T Reisner, Gari Clifford, Li-Wei Lehman, George Moody, Thomas Heldt, Tin H Kyaw, Benjamin Moody, and Roger G Mark. Multiparameter intelligent monitoring in intensive care ii (mimic-ii): a public-access intensive care unit database. *Critical care medicine*, 39(5):952, 2011. 4.1

[100] Rushit Sanghrajka, Daniel Hidalgo, Patrick P Chen, and Mubbasir Kapadia. Lisa: Lexically in-

telligent story assistant. In *Proceedings of the 13th Artificial Intelligence and Interactive Digital Entertainment Conference*, 2017. 5.1, 5.2, 5.4.6.3

[101] Ari Schlesinger, Kenton P. O'Hara, and Alex S. Taylor. Let's talk about race: Identity, chatbots, and ai. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, CHI '18, pages 315:1–315:14, New York, NY, USA, 2018. ACM. ISBN 978-1-4503-5620-6. doi: 10.1145/3173574.3173889. URL http://doi.acm.org/10.1145/3173574.3173889. 5.2

[102] Albrecht Schmidt. Implicit human computer interaction through context. *Personal technologies*, 4 (2-3):191–199, 2000. 2.2

[103] Donald Schön and John Bennett. Reflective conversation with materials. In *Bringing design to software*, pages 171–189. ACM, 1996. 1.3, 2, 2.1.2, 5.3.1, 5.5.1

[104] Donald A Schon. *The reflective practitioner: How professionals think in action*, volume 5126. Basic books, 1984. 2.1.3

[105] Donald A Schön. *The reflective practitioner: How professionals think in action*, volume 5126. Basic books, 1984. 2

[106] Marilyn Schwartz. *Guidelines for Bias-Free Writing*. Indiana University Press, 1995. 2.1.1

[107] Victoria A Shaffer, C Adam Probst, Edgar C Merkle, Hal R Arkes, and Mitchell A Medow. Why do patients derogate physicians who use a computer-based diagnostic support system? *Medical Decision Making*, 33(1):108–118, 2013. 4.1

[108] G Lynn Shostack. How to design a service. *European journal of Marketing*, 16(1):49–63, 1993. 2.1.2

[109] Patrice Simard, Saleema Amershi, Max Chickering, Alicia Edelman Pelton, Soroush Ghorashi, Chris Meek, Gonzalo Ramos, Jina Suh, Johan Verwey, Mo Wang, and John Wernsing. Machine Teaching: A New Paradigm for Building Machine Learning Systems. Technical report, Microsoft Research, jul 2017. URL https://arxiv.org/pdf/1707.06742. 1.3

[110] Mark S. Slaughter, Francis D. Pagani, Joseph G. Rogers, Leslie W. Miller, Benjamin Sun, Stuart D. Russell, Randall C. Starling, Leway Chen, Andrew J. Boyle, Suzanne Chillcott, Robert M. Adamson, Margaret S. Blood, Margarita T. Camacho, Katherine A. Idrissi, Michael Petty, Michael Sobieski, Susan Wright, Timothy J. Myers, and David J. Farrar. Clinical management of continuous-flow left ventricular assist devices in advanced heart failure. *The Journal of Heart and Lung Transplantation*, 29(4, Supplement):S1 – S39, 2010. ISSN 1053-2498. doi: https://doi.org/10.1016/j.healun.2010.01.011. URL http://www.sciencedirect.com/science/article/pii/S1053249810000434. Clinical Management of Continuous-flow Left Ventricular Assist Devices in Advanced Heart Failure. 4.3.2

[111] Lisa Stifelman, Adam Elman, and Anne Sullivan. Designing natural speech interactions for the living room. In *CHI '13 Extended Abstracts on Human Factors in Computing Systems*, CHI EA '13, pages 1215–1220, New York, NY, USA, 2013. ACM. ISBN 978-1-4503-1952-2. doi: 10.1145/2468356.2468574. URL http://doi.acm.org/10.1145/2468356.2468574. 2.2.1, 5.2

[112] Erik Stolterman. The nature of design practice and implications for interaction design research. *International Journal of Design*, 2(1), 2008. 2.1.2

[113] Maria Stone, Frank Bentley, Brooke White, and Mike Shebanek. Embedding user understanding in the corporate culture: Ux research and accessibility at yahoo. In *Proceedings of the 2016 CHI Conference Extended Abstracts on Human Factors in Computing Systems*, pages 823–832. ACM, 2016. 2.2.1

[114] Jennifer Sukis. Ai design & practices guidelines (a review). https://medium.com/design-ibm/ai-design-guidelines-e06f7e92d864, 2019. 4

[115] Alan R Tait, Terri Voepel-Lewis, Brian J Zikmund-Fisher, and Angela Fagerlin. The effect of format on parents' understanding of the risks and benefits of clinical research: a comparison between text, tables, and graphics. *Journal of health communication*, 15(5):487–501, 2010. 4.5.1

[116] Jonathan M Teich, Pankaj R Merchia, Jennifer L Schmiz, Gilad J Kuperman, Cynthia D Spurr, and David W Bates. Effects of computerized physician order entry on prescribing practices. *Archives of internal medicine*, 160(18):2741–2747, 2000. 4.3

[117] TensorFlow: Smarter machine learning, for everyone, 2016. https://www.google.com/intl/en/about/main/tensorflow/. 6.3.2

[118] The AllenNLP toolkit demo: Text Entailment. http://demo.allennlp.org/textual-entailment, 2017. 5.4.6.1

[119] Danielle Timmermans, Bert Molewijk, Anne Stiggelbout, and Job Kievit. Different formats for communicating surgical risks to patients and the effect on choice of treatment. *Patient education and counseling*, 54(3):255–263, 2004. 4.5.1

[120] Peter Tolmie, James Pycock, Tim Diggins, Allan MacLean, and Alain Karsenty. Unremarkable computing. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '02, pages 399–406, New York, NY, USA, 2002. ACM. ISBN 1-58113-453-3. doi: 10.1145/503376.503448. URL http://doi.acm.org/10.1145/503376.503448. 1.1.1, 4.3.1

[121] Mary Treseler. *Designing with Data: Improving the User Experience with A/B Testing*, chapter Designers as data scientists. O'Reilly Media, 2017. URL http://radar.oreilly.com/2015/05/designers-as-data-scientists.html. 1

[122] Udemy Online Course: Applied machine learning for Everyone, 2017. https://www.udemy.com/applied-machine-learning-for-everyone/. 6.3.2

[123] Anna Vallgårda and Johan Redström. Computational composites. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '07, pages 513–522, New York, NY, USA, 2007. ACM. ISBN 978-1-59593-593-9. doi: 10.1145/1240624.1240706. URL http://doi.acm.org/10.1145/1240624.1240706. 2.1.3

[124] Philip van Allen. Prototyping ways of prototyping ai. *Interactions*, 25(6):46–51, October 2018. ISSN 1072-5520. doi: 10.1145/3274566. URL http://doi.acm.org/10.1145/3274566. 3, 3.3.3

[125] Philip van Allen and Ben Hooker. ArtCenter College of Design Course: Internet of Enlightened Things, 2017. URL https://canvas.instructure.com/courses/1111888. 2.2

[126] Robert L Wears and Marc Berg. Computer technology and clinical work: still waiting for godot. *Jama*, 293(10):1261–1263, 2005. 4.1

[127] Mark Weiser. The computer for the 21 st century. *Scientific american*, 265(3):94–105, 1991. 4.5.1, 7

[128] Wikipedia contributors. Artificial intelligence — Wikipedia, the free encyclopedia. https://en.wikipedia.org/wiki/Artificial_intelligence, 2019. 1.2

[129] Jeremy C Wyatt and Douglas G Altman. Commentary: Prognostic models: clinically useful or quickly forgotten? *Bmj*, 311(7019):1539–1541, 1995. 1.1.1, 4.1

[130] Yao Xie, Ge Gao, and Xiang 'Anthony' Chen. Outlining the design space of explainable intelligent systems for medical diagnosis. *CoRR*, abs/1902.06019, 2019. URL http://arxiv.org/abs/

1902.06019. 1.1.1

[131] Qian Yang, John Zimmerman, and Aaron Steinfeld. Review of Medical Decision Support Tools : Emerging Opportunity for Interaction Design. In *Proceedings of the 6th IASDR (The International Association of Societies of Design Research Congress*, IASDR '15, pages 2366–2382, Australia, 2015. IASDR (The International Association of Societies of Design Research). 1.1.1, 1.3, 4.5.1

[132] Qian Yang, John Zimmerman, Aaron Steinfeld, Lisa Carey, and James F Antaki. Investigating the heart pump implant decision process: Opportunities for decision support tools to help. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, pages 4477–4488. ACM, 2016. 3.1

[133] Qian Yang, John Zimmerman, Aaron Steinfeld, Lisa Carey, and James F. Antaki. Investigating the heart pump implant decision process: Opportunities for decision support tools to help. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, CHI '16, pages 4477–4488, New York, NY, USA, 2016. ACM. ISBN 978-1-4503-3362-7. doi: 10.1145/2858036.2858373. URL http://doi.acm.org/10.1145/2858036.2858373. 1.1.1, 1.3, 3.1.1, 4.3.2

[134] Qian Yang, John Zimmerman, Aaron Steinfeld, and Anthony Tomasic. Planning Adaptive Mobile Experiences When Wireframing. In *Proceedings of the 2016 ACM Conference on Designing Interactive Systems - DIS '16*, pages 565–576, Brisbane, QLD, Australia, jun 2016. ACM Press. ISBN 9781450340311. doi: 10.1145/2901790.2901858. URL http://dl.acm.org/citation.cfm?id=2901790.2901858. (document), 1.3, 2.2.1, 2.2, 3.1, 3.1.1, 5.1, 6.3.2

[135] Qian Yang, Nikola Banovic, and John Zimmerman. Mapping Machine Learning Advances from HCI Research to Reveal Starting Places for Design Research. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems - CHI '18*, CHI '18. ACM, 2018. ISBN 9781450356206. 1.3, 2.2, 2.2.1, 2.2.1, 2.2.1, 5, 2.3, 3.2.2, 6.3.1

[136] Qian Yang, Alex Scuito, John Zimmerman, Jodi Forlizzi, and Aaron Steinfeld. Investigating how experienced ux designers effectively work with machine learning. In *Proceedings of the 2018 Designing Interactive Systems Conference*, DIS '18, pages 585–596, New York, NY, USA, 2018. ACM. ISBN 978-1-4503-5198-0. doi: 10.1145/3196709.3196730. URL http://doi.acm.org/10.1145/3196709.3196730. (document), 1.3, 2.2.1, 2.2, 3.1

[137] Qian Yang, Alex Scuito, John Zimmerman, Jodi Forlizzi, and Aaron Steinfeld. Investigating how experienced ux designers effectively work with machine learning. In *Proceedings of the 2018 Designing Interactive Systems Conference*, DIS '18, pages 585–596, New York, NY, USA, 2018. ACM. ISBN 978-1-4503-5198-0. doi: 10.1145/3196709.3196730. URL http://doi.acm.org/10.1145/3196709.3196730. 1, 3.1.2

[138] Qian Yang, Jina Suh, Nan-Chen Chen, and Gonzalo Ramos. Grounding interactive machine learning tool design in how non-experts actually build models. In *Proceedings of the 2018 Designing Interactive Systems Conference*, DIS '18, pages 573–584, New York, NY, USA, 2018. ACM. ISBN 978-1-4503-5198-0. doi: 10.1145/3196709.3196729. URL http://doi.acm.org/10.1145/3196709.3196729. 2.2.1, 3.1, 3.1.2

[139] Qian Yang, Justin Cranshaw, Saleema Amershi, Shamsi T. Iqbal, and Jaime Teevan. Sketching nlp: A case study of exploring the right things to design with language intelligence. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, CHI '19, pages 185:1–185:12, New York, NY, USA, 2019. ACM. ISBN 978-1-4503-5970-2. doi: 10.1145/3290605.3300415. URL http://doi.acm.org/10.1145/3290605.3300415. 1.3, 1.4, 2.2.1, 3.1, 3.1.1, 3.2.2

[140] Qian Yang, Aaron Steinfeld, and John Zimmerman. Unremarkable ai: Fitting intelligent decision

support into critical, clinical decision-making processes. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, CHI '19, pages 238:1–238:11, New York, NY, USA, 2019. ACM. ISBN 978-1-4503-5970-2. doi: 10.1145/3290605.3300468. URL http://doi.acm.org/10.1145/3290605.3300468. 1.1.1, 1.3, 3.1, 3.1.1

[141] Qian Yang, Aaron Steinfeld, Carolyn Rosé, and John Zimmerman. 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*, CHI '20, page 1–13, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450367080. doi: 10.1145/3313831.3376301. URL https://doi.org/10.1145/3313831.3376301. 1.3

[142] Daisy Yoo, Anya Ernest, Sofia Serholt, Eva Eriksson, and Peter Dalsgaard. Service design in hci research: The extended value co-creation model. In *Proceedings of the Halfway to the Future Symposium 2019*, HTTF 2019, New York, NY, USA, 2019. Association for Computing Machinery. ISBN 9781450372039. doi: 10.1145/3363384.3363401. URL https://doi.org/10.1145/3363384.3363401. 2.1.2

[143] Shichao Zhang, Chengqi Zhang, and Qiang Yang. Data preparation for data mining. *Applied Artificial Intelligence*, 17(5-6):375–381, 2003. doi: 10.1080/713827180. URL https://doi.org/10.1080/713827180. 5.4.3

[144] John Zimmerman and Jodi Forlizzi. Service design. *The Encyclopedia of Human-Computer Interaction, 2nd Ed.*, 2020. URL https://www.interaction-design.org/literature/book/the-encyclopedia-of-human-computer-interaction-2nd-ed/service-design. 7

[145] John Zimmerman, Jodi Forlizzi, and Shelley Evenson. Research through design as a method for interaction design research in hci. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 493–502. ACM, 2007. 5.1

[146] John Zimmerman, Anthony Tomasic, Charles Garrod, Daisy Yoo, Chaya Hiruncharoenvate, Rafae Aziz, Nikhil Ravi Thiruvengadam, Yun Huang, and Aaron Steinfeld. Field trial of tiramisu: Crowdsourcing bus arrival times to spur co-design. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '11, page 1677–1686, New York, NY, USA, 2011. Association for Computing Machinery. ISBN 9781450302289. doi: 10.1145/1978942.1979187. URL https://doi.org/10.1145/1978942.1979187. 1.3, 3.1.1