Discovering the Right Things to Design with Artificial Intelligence

Nur Yıldırım

A dissertation submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy

CMU-HCII-24-104
July 2024

Thesis Committee

James McCann (Co-Chair), Robotics Institute
John Zimmerman (Co-Chair), HCI Institute
Jodi Forlizzi, HCI Institute
Kayur Patel, Meta

Human-Computer Interaction Institute
School of Computer Science
Carnegie Mellon University

This research was sponsored by the National Science Foundation (2007501, 2112633), the National Institutes of Health (R35HL144804), Accenture and Bloomberg. The author was also supported by the Center for Machine Learning and Health (CMLH) Translational Fellowships in Digital Health. Any opinions, findings, and conclusions or recommendations expressed in this material are solely those of the authors 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, Human Centered Artificial Intelligence, User Experience Design, Ideation, Brainstorming

Abstract

Advances in artificial intelligence (AI) enable impressive new technical capabilities: computers can diagnose diseases, translate between languages, and drive cars. Interestingly, today nearly 90% of AI initiatives fail; few projects survive until deployment. I argue that a lack of effective ideation leads teams to select suboptimal innovations to pursue. In addition, AI product teams fail to see low-hanging fruit, situations where simple predictive models can generate value for users and stakeholders. Currently, data science teams propose innovations customers do not want, while product teams ask for things AI cannot do. As AI capabilities become more pervasive and commoditized, discovering the right human problems to solve while mitigating potential harm remains a great challenge.

My research addresses this breakdown in early stage ideation and problem formulation. I studied practitioners and observed that teams better at ideating are more effective in developing AI solutions that generate value and minimize risk. Based on the industry best practices, I created new innovation processes and resources for helping cross-functional product teams effectively explore the AI solution space before selecting what to implement. I developed a taxonomy of AI capabilities and examples of these in product forms. These resources sensitize stakeholders to what AI can do and search for opportunities where these might be valuable. I developed a hybrid ideation method that blends technology-centered development and human-centered design. I conducted a preliminary assessment of these resources and processes through case studies with innovation teams working in critical care, radiology, insurance, and accounting. Overall, this dissertation provides a glimpse into the future of human-centered AI innovation, where human needs and concerns are given equal importance as technical advances in deciding what to build with artificial intelligence.

Acknowledgements

I would not be exaggerating to say that this is the most important section of this dissertation—after all, research is done with people, and I had some of the most brilliant and kind ones by my side.

First and foremost, I would like to thank my advisors John Zimmerman and Jim McCann. You took me on as someone who had never done any research and made me the researcher I am today. Having your support through every paper deadline, talk, and fellowship application felt like having superpowers. I don't recall a single conversation where I wasn't left inspired and energized to push forward. Thank you for being truly exceptional; none of this would have been possible without you.

I would also like to thank my committee members, Jodi Forlizzi and Kayur Patel, who provided invaluable feedback and excellent career advice. Thank you for asking the hard questions that I have yet to answer.

Beyond my advisors and committee members, many others at HCII supported my growth as a researcher: Jeff Bigham, Laura Dabbish, Sauvik Das, Motahhare Eslami, Ken Holstein, Scott Hudson, Geoff Kaufman, Niki Kittur, David Lindlbauer, Michael Madaio, Nik Martelaro, Dominik Moritz, Raelin Musuraca, Brad Myers, Adam Perer, Sarah Preum, Dan Saffer, Devansh Saxena, and Nesra Yannier, along with the incredible undergraduate and graduate students that have contributed to this work – thank you. Members of the HCII family, Queenie Kravitz, Marian D'Amico, Reenie Kirby, Leah Buffington, John Davis, Karen Harlan, Lindsay Olshenske, and Ryan Ries, thank you for making things feel so effortless.

The work in this dissertation would not have been possible without my collaborators across academia and industry. Jeremy Kahn, Leigh Bukowski and the Limeaid team at the University of Pittsburgh were incredibly generous with their time as I ventured into intensive care in the midst of a pandemic. Alex Kass, Teresa Tung, Connor Upton, and Medb Corcoran at Accenture generously opened up their teams and shared their unique expertise on AI innovation. I would also like to thank my collaborators in industry (who will remain anonymous) for trusting us to conduct AI Brainstorming workshops within their teams, as well as all the participants who contributed their time and insights.

Fernanda Viégas, Martin Wattenberg, and Mahima Pushkarna hosted me for my internship at Google's People + AI Research team and significantly widened my perspective on how AI research gets translated into products. Tolga Bölükbaşı, what a happy coincidence it was to run into you in Cambridge years after meeting at our alma mater. Thank you for the hallway conversations and bike rides. Gabe Clapper, the design extraordinaire, what you were capable of doing with AI inspired me on so many levels. Thank you for the rich discussions on AI sprints that sowed the seeds of the final chapter of this dissertation. I am also thankful to Clara Kliman-Silver and Tiffany Knearem for enabling many conversations with the talented designers of Material Design and AIUX at Google. Anja Thieme and Hannah Richardson at Microsoft Research became lifelong mentors, making my sojourn in Cambridge joyful and insightful, along with the rest of the Biomedical Imaging team. Cheers to our lovely conversations down at the pub.

I am indebted to many friends and peers at the HCII. To start with, my fellow designers and academic siblings, Qian Yang and Michal Luria, deserve the most credit. Qian paved the way for this dissertation through her groundbreaking work and patiently listened to my pointless worries about my research. Thank you for always looking out for me and for the afterwork cocktails at the Porch. Michal made me realize the importance of being critical when selecting problems, and that I can be productive without working on weekends. Thank you for telling me I'll be fine even if I pivot research directions. Lea Albaugh, thank you for suggesting that I could work with Jim, which might not have occurred to me otherwise, as well as for all the wisdom on how to PhD.

Frederic and Sümeyye Gmeiner, thank you for the dinner parties, late-night conversations, wine and laughter. The time we spent together brought me some of my most wholesome memories in Pittsburgh. Vikram Kamath Cannanure and Tomohiro Nagashima, thank you for adopting me as your officemate – running into you spontaneously always made my day. Textiles Labs folks, Jenny Lin, Vidya Narayanan, Evan Shimizu, Ticha Sethapakdi, thank you for the proofreading and camaraderie – it was wonderful to be exposed to the cool stuff you made with yarn.

So many people made Carnegie Mellon and Pittsburgh 'home' for the past six years: Lynn Kirabo, Julia Cambre, Samantha Reig, Stephanie Valencia Valencia, Jesus Silva Elizalde, Tiangying Chen, Jianzhe Gu, Amber Horvath, Hyeonsu Kang, Andrew Kuznetsov, Steven Moore, Huy Anh Nguyen, Sujeath Pareddy, Danny Weitekamp, Jason Wu, Ryan Kavanagh, Martin Michelini, Zahra Ebrahimi, Jaepil Lee, Joohyun Lee, Behnam Mohammadi, Elisabeth Estes, Mert Arslanoğlu, Yiğit Yakupoğlu, Umut Soysal, Mary Beth Kery, Michael Xieyang Liu, Michael Rivera, Siyan Zhao, Felicia Ng, Xu Wang, Anhong Guo, Zheng Yao, Kareem Bedri, David Gerritsen, Wesley Hanwen Deng, Luke Guerdan, Venkatesh Sivaraman, Tzu-Sheng Kuo, Katelyn Morrison, Kimi Wenzel, Hank Lee, Alicia Lee, Ji-Youn Jung, Minjung Park, Shixian Xie, and many others.

My former roommates at METU, Hande Güneş, Eti Ester Levi, Tuğba Pişkin, Deniz Seçilmiş, Duygu Tolunay, and Selen Yamak – having you in my life is one of my most prized accomplishments to date. Thank you for being just a video call away as we pursued PhDs in six different countries across four time zones. Our dear friends Mehtap Aydın, Ferhat Aydın, Clément Boullé, and Parama Gosh, thank you for hosting us from coast to coast, LA to NYC. Last but not least, fellow MSR UK interns Eloi Alonso, Adam Jelley, Pedro Sanchez, and Avijit Thawani, thank you for a fantastic summer of picnics, hikes, and Shakespeare plays.

I am immensely grateful to my parents, Gülser and Musa, and my sister Asu for their unconditional love and support. Thank you for encouraging me to carve my own path in life. I would also like to thank the Levy family, whose home became my second home.

Finally, I would like to thank my partner, Samuel Levy. *Tu me fais rire chaque jour.* We made it through a PhD, a global pandemic, and a two-body problem, and having you by my side since day one made all the difference. I cannot wait for the adventures ahead of us as we embark on a new chapter as faculty together. Thank you for being my person.

List of Publications

1. Nur Yildirim, Alex Kass, Teresa Tung, Connor Upton, Donnacha Costello, Robert Giusti, Sinem Lacin, Sara Lovic, James O'Neill, Rudi O'Reilly Meehan, Eoin Ó Loideáin, Azzurra Pini, Medb Corcoran, Jeremiah Hayes, Diarmuid Cahalane, Gaurav Shivhare, Luigi Castoro, Giovanni Caruso, Changhoon Oh, James McCann, Jodi Forlizzi, John Zimmerman. **How Experienced Designers of Enterprise Applications Engage AI as a Design Material.** In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* **(CHI'22)**. New Orleans, LA, USA.

2. Nur Yildirim, Mahima Pushkarna, Nitesh Goyal, Martin Wattenberg, Fernanda Viegas. **Investigating How Practitioners Use Human-AI Guidelines: A Case Study on the People + AI Guidebook.** In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* **(CHI'23)**. Hamburg, Germany. [Best Paper Honorable Mention]

3. Nur Yildirim, Changhoon Oh, Deniz Sayar, Kayla Brand, Supritha Challa, Violet Turri, Nina Crosby Walton, Anna Elise Wong, Jodi Forlizzi, James McCann, John Zimmerman. **Creating Design Resources to Scaffold the Ideation of AI Concepts.** In *Proceedings of the 2023 Designing Interactive Systems Conference* **(DIS'23)**. Pittsburgh, PA, USA. [Best Paper Honorable Mention]

4. Nur Yildirim, Susanna Zlotnikov, Aradhana Venkat, Gursimran Chawla, Jennifer Kim, Leigh Bukowski, Jeremy Kahn, James McCann, John Zimmerman. **Investigating Why Clinicians Deviate from Standards of Care: Liberating Patients from Mechanical Ventilation in the ICU.** In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems* **(CHI'24)**. Honolulu, HI, USA. [Best Paper Honorable Mention]

5. Nur Yildirim, Susanna Zlotnikov, Deniz Sayar, Jeremy Kahn, Leigh Bukowski, Sher Shah Amin, Kathryn Riman, Billie Davis, John Minturn, Andrew King, Dan Ricketts, Lu Tang, Venkatesh Sivaraman, Adam Perer, Sarah Preum, James McCann, John Zimmerman. **Sketching AI Concepts with Capabilities and Examples: AI Innovation in the Intensive Care Unit.** In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems* **(CHI'24)**. Honolulu, HI, USA.

6. Nur Yildirim, Hannah Richardson, Teo Wetscherek, Junaid Bajwa, Joseph Jacob, Mark Pinnock, Stephen Harris, Daniel Coelho de Castro, Shruthi Bannur, Stephanie Hyland, Pratik Ghosh, Mercy Ranjit, Kenza Bouzid, Anton Schwaighofer, Fernando Pérez-García, Harshita Sharma, Ozan Oktay, Matthew Lungren, Javier Alvarez-Valle, Aditya Nori, Anja Thieme. **Multimodal Healthcare AI: Identifying and Designing Clinically Relevant Vision-Language Applications**

for Radiology. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems* **(CHI'24)**. Honolulu, HI, USA.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Now that we can do anything, what will we do?
– Bill Buxton, Sketching User Experiences

1.1 Motivation

Advances in artificial intelligence (AI) enable a myriad of unprecedented technical capabilities. From mundane spam filters to predictive medicine and autonomous transportation, AI's technical advances are moving from research labs into the real world with a promise to improve the human experience. Clinical systems find anomalies in medical images to support clinicians in making life-and-death decisions. Cars assist people in driving and navigating while giving a real-time estimation of arrival time. Content platforms provide voice control, captions, and translation, making information accessible to everyone. The prevalence of AI makes it seem to be facile to innovate AI products and services.

In reality, there is an incredibly high failure rate with AI innovation. Despite the advances and investment, nearly 90% of AI projects fail to deploy, and only about half of the remaining projects make it from prototype to production. Researchers attribute this to a "hammer in search of a nail" approach to innovation, speculating that teams focus on problems that are too complex or do not address real human needs. Many "intelligent" features go unused, often because people do not find them useful.

Scholars studying AI product development highlight that most AI failures can be traced back to the early problem formulation phase, where teams generate and select project ideas. Data science teams often envision innovations that customers, frontline workers, or community stakeholders do not want. Meanwhile, product teams – product managers, designers, and domain experts – are often insufficiently involved in problem formulation. Even when they are included, they typically lack data and AI literacy and tend to think of innovations that cannot be built. Teams also tend to narrowly focus on critical, high-stakes problems requiring extremely robust models; they fail to consider low-hanging fruit: low-risk, medium-reward projects where moderate AI performance can generate value. Step counters, voicemail transcripts, media and shopping recommenders, and language translation of a friend's social media post all create value for users even when the model performance is below 90%. While it is currently quite challenging for developers to create systems with high model performance, systems with moderate model performance are often within easy reach.

There has been a growing interest in the human-computer interaction (HCI) community in designing human-centered AI systems that amplify human abilities and improve the human experience. A significant body of research investigated the challenges of designing human-AI interaction. Researchers delineated the unique characteristics of *AI as a design material*, such as the capability uncertainty or output complexity, that makes envisioning AI systems uniquely difficult compared to traditional software [158]. Practitioner-facing AI design resources and guidelines (e.g., [6, 111, 7] recently became available to address the challenges around usability, privacy, explainability, fairness, and algorithmic bias. While these resources provide teams with an initial set of tools, they mainly address the challenges of *making the thing right*, where a problem/solution is already formulated, and an AI system is already built. Resources and guidance for deliberating on *what is the right thing to make with AI* remain scarce.

Parallel to these efforts, a growing body of work raised concerns about participation and stakeholder engagement throughout the lifecycle of AI development, especially in high-stakes contexts where the input of impacted stakeholders is crucial. Researchers raised fundamental questions: *Who is involved in the design of AI systems, and who frames the discussion? What are the new design processes and patterns to make these systems successful for individuals, communities, and society?*[1] A relevant strand of research proposed creating resources for improving non-technical stakeholders' literacy in data science and AI concepts through hands-on experimentation with ML models and datasets. However, it is not clear what constitutes a *good enough* understanding to effectively engage in envisioning what AI can uniquely solve. As AI capabilities become more pervasive and commoditized, identifying innovation opportunities that both leverage AI's strengths and produce value for users and stakeholders while mitigating potential harms remains a great challenge.

Within this context, I set out to explore this breakdown in the AI innovation process. I suspect that most of these failures and concerns stem from *a lack of effective ideation* – innovation teams do not seem to ideate when selecting real world, human problems to solve with AI. Throughout my Ph.D., I observed practitioners and teams working at the cutting edge of innovation of AI products and services. Teams that are better at ideation and problem formulation are more effective in developing AI solutions that provide value and minimize risk. Based on this observation, I claim :

> *We can improve AI innovation (reduce failure and reveal low-hanging fruit) by supporting cross-disciplinary teams in systematically and collectively exploring the problem-solution space before selecting what to implement.*

In this dissertation, I first provide evidence for the lack of ideation and human considerations in the problem formulation stage – one of the least investigated phases of AI product development. Acting on this insight, I take initial steps toward addressing this breakdown. I propose new innovation processes and resources to (1) support more effective collaboration and stakeholder engagement at the ideation and problem formulation stages of a project; (2) sensitize teams to AI capabilities and limitations; and (3) help teams more fully explore the problem-solution space. Using design research approaches, I

[1] From Stanford University's 2022 Fall Conference on Human-Centered Artificial Intelligence.

explore whether these processes and resources help teams identify strong matches between existing AI capabilities and actual needs, while simultaneously assessing each concept's risks and benefits.

1.2 Research Questions

This dissertation consists of three main threads of research: investigating challenges and best practices in the industry; developing resources for practitioners; and iteratively assessing and refining these resources with practitioners. In this section, I detail the research questions I explored and the research approach I followed.

I started by investigating the current challenges and best practices in the industry. Specifically, I sought to understand **(RQ1)** *What challenges do cross-disciplinary teams face in early phase AI product development? How do effective innovation teams tackle these challenges? What do these practices reveal about how to better innovate with AI?* I carried out empirical studies with nearly 50 practitioners across different disciplines (i.e., designers, data scientists, and product managers) and companies of varying sizes (e.g., large technology companies, startups, and nonprofit organizations). This formative fieldwork provided preliminary evidence of the lack of ideation and human considerations in the early phases as one of the significant root causes of AI failure. This empirical approach also helped me to uncover emergent best practices, such as design thinking and facilitation for collective ideation and problem formulation with diverse team members and stakeholders.

In parallel to fieldwork on industry best practices, I worked to develop resources that can support non-technical stakeholders in envisioning AI innovations. The practices of industry teams revealed that effective designers call upon internalized abstractions of AI capabilities when they ideate, and that they engage non-technical stakeholders in ideation using AI capability abstractions and product examples as resources. Building on these insights, I sought to explore **(RQ2)** *Can we create resources that document explicit AI capability abstractions? Can designers and AI innovators use these to improve their ideation, to help them think of things that can be built?* I set out to create a taxonomy of AI capabilities by systematically analyzing a corpus of 40 real-world, commercially successful AI applications. I conducted a pilot study to understand whether and how the taxonomy impacts the ideation of AI products and services.

Following resource development, I worked to understand how to integrate these resources into current practices. Data science teams struggle to engage product managers and domain experts in AI product development. There is no evidence that teams ideate. Design teams are often brought on after project selection, negating the skills they bring around ideation and broad exploration of the problem/solution space. Building on the industry best practices, I asked **(RQ3)** *Can designers scaffold cross-disciplinary teams in ideation?* To explore this question, I joined a cross-disciplinary medical innovation team composed of data scientists and critical care clinicians. I conducted multiple ideation workshops that operationalized the taxonomy to collectively brainstorm AI concepts for the intensive care unit. This hands-on Research through Design project served as a preliminary assessment of this new innovation approach: our team was able to broadly and systematically explore the problem-solution space.

Part of the challenge with ideation workshops and taxonomy piloting was assessing the quality of ideation. One key dimension that emerged was *AI model performance*: I observed that the selection of problems and applications seemingly ignore situations where model performance can be moderate and still provide value. Additionally, ideation often focused on more complex problems rather than searching for situations that would benefit from simpler predictions. Through additional case studies with innovation teams working in radiology, insurance, and accounting, I explored **(RQ4)** *Could priming teams with examples of simpler solutions and examples that produce value with moderate model performance improve the quality of ideation? Would this help teams consider this under-investigated search space?* These explorations provided preliminary evidence that this modified innovation process and resources may improve ideation and revealed future research directions.

1.3 Thesis Overview

This dissertation progresses through chapters detailing related work (Ch. 2), formative fieldwork (Ch. 3, Ch. 4), resource development (Ch. 5), case studies (Ch. 6, Ch. 7, Ch. 8), and conclusion (Ch. 9), organized as follows:

In Chapter 2, I draw on HCI and design literature to detail what sketching or ideating with AI means, and why it is more difficult compared to other technology materials. I then curate prior research that laid the groundwork for understanding the challenges in designing AI products and cross-disciplinary collaboration, especially in the early problem formulation phase. Finally, I briefly review innovation processes and methods as they relate to AI innovation.

In Chapter 3, I describe the current practices of cross-functional teams around designing and developing AI products and services. I investigated how designers and product managers use existing human-AI guidelines with an eye for discovering unmet needs and challenges. The findings from this interview study provided evidence of the lack of ideation and human considerations in the problem formulation phase. I also uncovered self-developed resources, such as AI capability frameworks, to span this gap. Reflecting on these, I discuss the opportunities for new design processes and resources to support the selection of the right problem to solve.

In Chapter 4, I describe how I gained insight into the best practices of innovation teams in the early phase AI innovation. This becomes a benchmark I want to bring less experienced teams up to. These experienced design and data science practitioners worked at an AI incubation center. I detail how they approached AI as a design material and effectively envisioned novel AI-driven products and services. I highlight a critical insight from this study, which later became a core idea in this dissertation: that human-centered approaches, particularly design thinking and facilitation, can effectively engage diverse stakeholders in collectively brainstorming, formulating, and assessing AI solutions. I conclude this chapter by discussing the kinds of design resources needed to enable cross-functional, collective ideation.

In Chapter 5, I describe a set of resources I developed to support systematic and collective ideation of AI systems. I created a taxonomy of AI capabilities and examples that sensitize non-technical stakeholders (e.g., product managers, designers, domain experts) to what AI can do and how it can generate value. I created communicative forms, such as web-based and print resources, to integrate these resources into current practice. I conducted a pilot study to understand if and how the capabilities would improve ideation. I discuss the learnings from the taxonomy development and assessment for advancing AI innovation.

In Chapter 6, I present a case study where I took on the challenge of facilitating AI ideation between data scientists and domain experts working in intensive care. I describe a design process where we broadly explored AI uses cases for supporting intensive care clinicians through a series of AI Ideation Workshops. I created and assessed a set of ideation tools as part of this project. I created workshop prompts using the taxonomy and examples, which sensitized domain experts to what AI can do. I created a new ideation worksheet for describing an AI system's behavior using non-technical terms for model, data, and interaction form. Together, these tools and approaches enabled our cross-disciplinary team to cover a broad range of ideas with high-impact and low-risk. I reflect on the challenges and lessons learned in facilitating ideation with data and AI.

In Chapters 7 and 8, I present additional case studies where I collaborated with innovation teams working in radiology, insurance, and accounting. These projects provided preliminary evidence that the resources and processes proposed in this work can support teams in discovering high-value, low-risk use cases. This line of work revealed open research questions around the formulation, assessment, and prioritization of AI concepts.

In Chapter 8, I reflect on my research journey and discuss how the contributions made move us toward a preferred future. Finally, I highlight promising future research directions and share my concluding thoughts.

1.4 Summary of Contributions

This dissertation makes three contributions.

First, it characterizes the breakdown in AI innovation as a gap between technology-centered and user-centered approaches, revealing that a lack of or ineffective ideation plays a prominent role in AI product failures. It positions HCI and design as core practices for facilitating early phase ideation and problem formulation. Second, it defines AI's problem-solution space by highlighting model performance and task expertise as key dimensions. It uncovers a richer, under-investigated solution space where moderate model performance and simpler inferences generate significant value for people. Third, it develops new resources that help non-data scientists better understand what AI can effectively do and proposes a new innovation process for systematically exploring matches between technological capabilities and human needs. Through case studies, it demonstrates the potential impact of these processes and resources for improved ideation and stakeholder engagement. Overall, this dissertation seeks to explore preferable futures for human-centered AI innovation, where human needs and concerns are equally important drivers as AI's technical advances, rather than being treated as an afterthought.

Chapter 2

Related Work

In this chapter, I briefly review HCI and design theory on ideation and its value for designing interactive systems. I then detail the challenges of designing AI products along with proposed solutions. Finally, I provide an overview of innovation processes for product development.

2.1 Ideation in Interaction Design

HCI research has characterized ideation (sketching) and prototyping as core activities for innovating interactive technologies. Interaction designers engage in ideation to discover the right thing to design, and in prototyping to iteratively design the right thing [21, 47]. Ideation involves *reflection in action*; generating many solutions that encode new problem framings for a problematic situation. Prototyping connects to *reflection on action*; stepping back to critique, assess and improve what was done [21, 123]. HCI research has shown that generating multiple ideas leads to better outcomes compared to iterative refinement of a single idea [44].

While HCI literature has a breadth of tools and methods for prototyping (e.g., paper prototyping, video sketches, and speculative enactments), tools for sketching and ideation remains relatively scarce [21]. As the HCI practice is ever expanding, new technology materials call for new ideation techniques and methods. In response, researchers have produced new methods that allow “tinkering with materials” to facilitate ideation with emergent technologies, such as bluetooth [138], internet of things [91], software [110], haptics [101], and soma experiences [139]. For technologies that are more difficult to “tinker with”, researchers facilitated ideation through abstractions of technology capabilities and experiential qualities [155]. I draw from this line of research by engaging data and AI as emergent design materials.

2.2 Designing AI Products and Services

In this section, I detail the current challenges and emergent solutions in AI product development processes. I describe these from the perspective of two distinct types of expertise: product teams and data science teams. Product teams involve practitioners who are experts in product development and business domains (e.g. product managers, business strategists, interaction and service designers, design researchers, software engineers, domain experts), but do not have a background in data science or AI. Data science teams involve practitioners who are experts in AI development (e.g. data scientists, AI engineers, research scientists), but might not be familiar with human-centered product development processes or business objectives.

2.2.1 Ideation and Problem Formulation

Challenges. Product teams struggle to understand AI capabilities – what AI can offer for designing new products and services [42]. They are not sensitized to think about data dependencies necessary to drive AI capabilities [160]. Without this ingrained understanding, they find it difficult to ideate many novel concepts, and often envision AI products that cannot be built [42]. Yang et al. [158] characterized AI's design space by the capability uncertainty and output complexity of an idea. They claim product teams tend to generate ideas and concepts that have high capability uncertainty and high output complexity (e.g. text generation). They think of things that cannot be built. They also overlook low hanging fruit – places where relatively simple, well established AI capabilities with few outputs – can still bring enough value for customers, users, and service providers.

Data science practitioners report that mapping business problems from a client's domain to tractable data science problems is one of the most difficult parts of their work [88, 104, 112]. Data science teams face challenges in eliciting user requirements from product managers and domain experts, and they tend to overlook how the AI system will generate value for users [88, 104, 113]. Without clear feedback from the product team, they tend to build AI products that customers and users do not want [155]. This gap between product and data science expertise creates challenges in early phase problem formulation where teams search for ideas that are both implementable and valuable for customers, users, and service providers [155, 146].

Proposed Solutions. A common practice among data science teams is to hold AI education sessions with product teams to overcome the knowledge gap. However, practitioners share that there is no effective way to teach AI concepts without getting into too much detail, resulting in frustration, and trial and error [113]. Researchers created resources to improve the data science and AI literacy of non-technical team members and stakeholders [62, 83, 45]. These resources describe how AI functions and detail learning mechanisms (e.g. supervised learning, neural networks, etc). However, it is not clear whether and how these approaches support product practitioners in early phase ideation.

The way experienced data scientists work resembles design practice; they midwife their stakeholders' desiderata by helping them articulate their real need as opposed to what they claim they need [88].

Research investigating how designers on product teams successfully envision AI concepts revealed that they only had a high level understanding of AI (e.g. what a model and label is) [157]. Instead, they worked with abstractions and examples that captured specific AI capabilities and value propositions (e.g. predicting user intent to surface relevant actions). Based on this insight, some researchers speculated that AI resources for product practitioners should detail what AI can do, instead of how AI works

[157, 154]. Currently, there is no consensus on what constitutes a good enough technical understanding for practitioners and stakeholders to effectively engage AI ideation and problem formulation [158].

2.2.2 Prototyping and Evaluation

Challenges. Designing AI products brings unique challenges due to their non-deterministic nature. Even simple AI applications (e.g. two-class classifiers) can make inference errors, leading to user experience (UX) breakdowns and harm [80]. A significant amount of research investigated the challenges of designing human-AI interaction, such as intelligibility [1], explainability [150, 90], user control [93, 128], feedback [136], error recovery [67], and setting user expectations [77, 85]. Despite these efforts, it remains difficult for product teams to anticipate potential errors a not-yet deployed AI system might cause [158].

AI-enabled experiences are difficult to rapidly prototype and assess. They are data dependent and contextual, which makes it incredibly difficult to create a functional prototype to simulate the experience. AI-enabled experiences are also dynamic; they adapt to different users and contexts over time and make bizarre, hard to anticipate errors. Current HCI methods, such as wizard-of-oz or rule-based simulators involve human commonsense; they do not allow product teams to experiment with the limitations of AI systems before the data is collected and a model is built [155].

Proposed Solutions. Technology companies proposed principles, guidelines, and design patterns for human-AI interaction to aid product practitioners in communicating the output of AI systems (e.g., Apple [7], Google [111], Microsoft [6], IBM [69]). These resources cover a comprehensive list of design considerations, such as explainability (e.g. "make clear why the system did what it did" [6]), control (e.g. "give people familiar, easy ways to make corrections" [7]), and feedback (e.g. "let users give feedback" [111]). In addition to general guidelines, some resources focus on specific interactions (e.g. voice interactions [5], chatbots [97]) or considerations (e.g. ethics [70]) (see a review in [9]). However, little is known about how practitioners use these guidelines and whether these are helpful for discovering problems where AI might be a good solution.

Another line of research developed prototyping tools with little or no code to allow first-hand experimentation with data and AI (e.g., Wekinator [48], Teachable Machine [27], ml5.js [99], Delft AI Toolkit [143]]. Building on the framing of *AI as a design material*, these tools allow interaction designers to work with pre-built AI models and libraries, such as image or sound classifiers, to gain a felt sense of what AI can do. While these tools make it easier to prototype AI-enabled interactions, it is not clear whether they help with envisioning. Additionally, due to their beginner-friendly nature, these prototyping toolkits often provide a limited set of AI capabilities and datasets which may not be applicable to complex, real world problems [158].

2.2.3 Anticipating Potential Harm

Challenges. Both data science and product teams face challenges in anticipating the potential harm of AI systems prior to deployment. Open-source ML fairness toolkits became available from large technology companies (e.g., IBM [12], Microsoft [14]) for detecting and mitigating algorithmic biases. However, these tools are often designed solely for the use of developers, and they typically support monitoring and debugging post deployment [65]. AI design guidelines intended for product teams also provide some guidance around fairness and bias, however, these tend to be high level and underdeveloped (e.g. "mitigate social biases" [6], "consider bias in the data collection and evaluation process" [111]).

Proposed Solutions. Recent work explored practitioners' experience addressing fairness issues, and their needs and desires for support [65, 37, 94]. Practitioners express challenges in engaging cross-disciplinary team members stakeholders, such as product managers and domain experts. Researchers note the need for tools and resources that support collectively anticipating and mitigating potential biases and harm.

In response, several researchers shared first-person accounts of engaging impacted stakeholders in addressing fairness issues. These include participatory approaches to elicit stakeholders' values in high stakes contexts (e.g., child mistreatment detection [79, 29], document review for civil litigation [35]); and user-driven algorithm auditing for identifying and reporting problematic algorithmic behaviors [38, 126]. While these approaches are helpful for evaluating existing AI systems, it is not clear how practitioners can engage stakeholders in anticipating the potential harm of AI systems that are not yet developed [120, 36].

2.2.4 Collaboration and Process

Challenges. Practitioners reported many barriers to cross-disciplinary collaboration due to a lack of shared workflow, shared language, and shared expertise [113, 104, 55, 81]. Most product teams do not have prior experience working with AI, and most data science teams have little experience with product development and human-centered design. Recent research indicates that most AI failures stem from miscommunications and misalignment between product and data science practitioners, especially in early phase problem formulation [46, 74, 146]. Product development and AI development life cycles have varying timelines which may not overlap; what is early phase for product teams may be late phase for data science teams. Some product roles, such as designers, typically join projects towards the end, they are rarely involved in early phase problem formulation [42, 157, 137]. Nevertheless, AI products require continued collaboration between product and data science teams due to potential changes in system behavior after product launch.

Proposed Solutions. Research investigating the industry best practices highlighted that successful teams form close, informal collaborations in early phases; and they share interim artifacts (e.g., data visualizations) to scaffold cross-disciplinary collaboration throughout the AI product development process [157, 137, 167, 88, 148, 113]. Close collaboration seems to sensitize practitioners to each others' expertise; where product teams learn about AI capabilities and data science teams learn about the problem domain [157, 88, 113].

2.3 Innovation Processes

In this section, I briefly review major innovation processes, methods and frameworks as they relate to AI development.

2.3.1 Technology-centered Innovation

Technology-centered innovation takes the capabilities and constraints of technology as a starting place for the development of new products and services. Notable frameworks include Lean Startup [118] and Agile development [64]. This innovation process often focuses on the development of a prototype or a minimum viable product (MVP) to test business hypotheses and validate product-market fit. Through an iterative process of building and testing, innovators rapidly identify what works and what does not, which allows them to *pivot* – changing direction based on the customer feedback (e.g., altering the target market, product features, or business model) [118]. A pitfall of this approach is starting with a single technology and user pair – it may take several cycles and pivots for innovators to find a technology-customer match [144, 16]. Matchmaking [16] offers a complementary approach, where innovators work from a technology towards the discovery of many technology-customer matches. However, following a Matchmaking process has its limitations in AI product development, as specific datasets might not translate to different users or domains.

2.3.2 User-centered Design

User-centered design (UCD) places the needs, preferences, and behaviors of users at the center of an innovation process [32]. This approach involves iterative cycles of user research, prototyping, testing, and refinement to ensure that the final product or service is both usable and valuable to its intended audience. Participatory Design (PD), which emerged in the 1970s, can be considered a precursor to UCD. PD aims to democratize the innovation process by ensuring that the voices of those impacted by technology are heard and considered [129, 122, 17]. Methods such as co-design workshops and focus groups are commonly used in UCD to facilitate stakeholder engagement and co-creation. UCD has been widely adopted in the industry through UX design practice. It is one of the most commonly used innovation processes because it effectively reduces the risk of developing technology that does not meet user needs or wants [21]. One limitation of this process is that it does not work well when the challenge is the discovery of a broad set of users who might benefit from a technology (such as artificial intelligence), as it begins with the selection of a target user group [144].

2.3.3 Service Design and Systems Thinking

Service Design (SD) focuses on the complex interactions between users, service providers, and service environments [51, 50]. While UX and SD both employ UCD, SD expands UCD's user- and product-centric model. By exploring the creation of value for both customers and service providers, SD results in new strategies. HCI research notes that Service Design and UX interact: first, Service Design develops a strategy, and then UX follows and creates the interactive artifacts [121]. Service Design tools such as service blueprints [15] and customer journey maps [61] are increasingly used in HCI [134, 51]. Systems Design builds on this approach for innovating complex systems (e.g., transportation networks, healthcare systems, and organizational infrastructures) where products, services, people and relationships are taken into account holistically [117, 125].

2.3.4 Design-led Innovation

With the popularization of Design Thinking [20], user-centered innovation approaches have been widely adopted across various industries ever the past decades. This process leverages design practitioners' expertise in ideation and diverse thinking. Using techniques such as brainstorming sessions and design sprints [84], designers facilitate ideation and collaboration across team members and stakeholders. A known pitfall in this process is *hill climbing* – user-centered design is suited for incremental innovation and rarely results in radical innovation [108]. A framework used at Google characterizes this continuum as Versioning (making incremental improvements to existing products and services), Visioning (envisioning new products, products features and services), and Venturing (moonshot projects and experiments with emergent technology). While design-led innovation approaches work well for mature technologies, it is unclear how to apply these when innovating with emerging technologies such as artificial intelligence.

Chapter 3

Investigating What Practitioners Need from AI Design Resources

As I outlined in the previous chapter, several resources and human-AI guidelines became available to aid practitioners in designing AI products. However, little is known about how these resources are *actually* used in practice and whether they help with the challenges in the early product development phase.

I collaborated with researchers at Google to investigate how practitioners working on AI product teams use the People + AI Guidebook [111] as an instance representative of existing AI resources. My focus was not on assessing whether guidelines are effective in improving the user experience of products or how industry AI guidelines compare to each other. Instead, I sought to understand the felt experience of practitioners in tackling the challenges of AI product development.

I interviewed 31 designers and product managers across 23 AI product teams from 14 organizations. I probed if and how guidelines help, as well as the remaining challenges and needs for additional support. The study revealed several interesting findings:

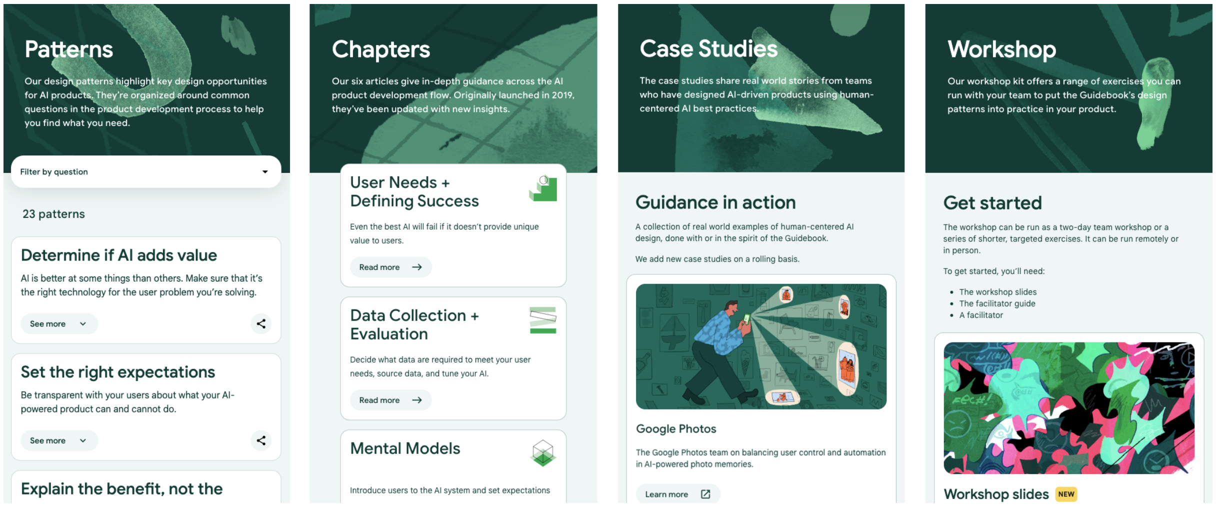
1. Practitioners not only use guidelines for addressing AI's design challenges but also for education, developing internal resources, cross-functional alignment, and buy-in within their teams and organizations.

2. Many practitioners had experienced AI product failures due to solving problems that do not address real needs. While guidelines are helpful for problem solving, practitioners desire better support for problem framing, ideation, and selecting the right problem.

3. I observed an emergent, hybrid design process reconciling user-centric and tech-centric approaches, where participants worked to match AI capabilities with human needs.

3.1 People + AI Guidebook

Building on the large body of HCI research on interaction with intelligent systems, technology companies have recently proposed principles, guidelines, and design patterns for human-AI interaction to



Patterns

Our design patterns highlight key design opportunities for AI products. They're organized around common questions in the product development process to help you find what you need.

23 patterns

Determine if AI adds value

AI is better at some things than others. Make sure that it's the right technology for the user problem you're solving.

Set the right expectations

Be transparent with your users about what your AI-powered product can and cannot do.

Explain the benefit, not the

Chapters

Our six articles give in-depth guidance across the AI product development flow. Originally launched in 2019, they've been updated with new insights.

User Needs + Defining Success

Even the best AI will fail if it doesn't provide unique value to users.

Data Collection + Evaluation

Decide what data are required to meet your user needs, source data, and tune your AI.

Mental Models

Introduce users to the AI system and set expectations

Case Studies

The case studies share real world stories from teams who have designed AI-driven products using human-centered AI best practices.

Guidance in action

A collection of real world examples of human-centered AI design, done with or in the spirit of the Guidebook.

We add new case studies on a rolling basis.

Google Photos

The Google Photos team on balancing user control and automation in AI-powered photo memories.

Workshop

Our workshop kit offers a range of exercises you can run with your team to put the Guidebook's design patterns into practice in your product.

Get started

The workshop can be run as a two-day team workshop or a series of shorter, targeted exercises. It can be run remotely or in person.

To get started, you'll need:

- The workshop slides
- The facilitator guide
- A facilitator

Workshop slides

FIGURE 3.1: The People + AI Guidebook [111] contains design patterns, chapters on AI's design considerations, case studies, and a workshop kit.

aid product professionals in designing AI products (e.g., [7, 111, 6, 69]). Prior research focused on validating the effectiveness of guidelines in improving the user experience [89], or studying guidelines comparatively to reveal the landscape of considerations [151, 72]. A study mapping existing guidelines and resources to current product development processes pointed out that AI design guidelines mainly help with late phase efforts, and that resources for the early phase are scarce [158]. Few researchers investigated how specific resources and guidelines are used by industry practitioners, such as fairness toolkits [37] and conversational UI guidelines [82]. In the same spirit, this work investigates if and how AI design guidelines in [111] are used, which I briefly introduce below.

Developed by Google, the People + AI Guidebook (the guidebook) is "a set of methods, best practices, and examples for designing with AI" [111]. Similar to AI design resources from Apple [7] and Microsoft [6], the guidebook contains principles, guidelines, and design patterns for human-AI interaction based on data and insights synthesized from Google products and academic research. The content is arranged around five sections (Figure 3.1): (1) chapters outlining design considerations (i.e. User Needs + Defining Success; Data Collection + Evaluation; Mental Models; Explainability + Trust; Feedback + Control; Errors + Graceful Failure), (2) design patterns with sensitizing examples demonstrating patterns and anti-patterns (e.g. "emphasize how the app will benefit users, avoid emphasizing the underlying technology"), (3) case studies of real-world AI products, (4) a workshop kit with a facilitator guide, (5) glossary of AI related terms. The guidebook employs three hypothetical app examples for design patterns – a running app, a plant classification app, and a learning app – to illustrate how principles might be applied in practice.

3.2 Method

I broadly defined "practitioners" as people who work on a team developing AI-enabled products and services. I aimed to recruit broadly across many roles (e.g., designers, product managers, data scientists, engineers, domain experts, etc.), hoping the target audience would emerge through the recruitment process. My inclusion criteria included having experience with the guidebook concerning AI/ML projects.

TABLE 3.1: Participants' teams and technology areas. Roles included product manager (PM), design manager (DM), user experience designer (UXD), and user experience researcher (UXR).

Team	Technology Area	Roles	ID	Size
1	Business analytics	UXD	P20	<100
2	Business analytics	DM, PM, UXD	P8, P10, P11	>10,000
3	Code completion	DM, UXD, UXR	P4, P9, P26	>10,000
4	Code completion	PM	P16	<100
5	Conversational AI	UXD	P21	<10,000
6	Financial forecasting	UXD, DM, UXD	P1, P2, P15	>10,000
7	Fraud detection	UXD, DM	P18, P23	>10,000
8	Image classification	DM	P7	>10,000
9	Medical diagnosis	UXD	P5	>10,000
10	Medical diagnosis	UXD	P30	>10,000
11	Personal healthcare	DM	P24	<100
12	Recommender system	UXR, UXD	P3, P12	>10,000
13	Recommender system	UXD	P14	100-1,000
14	Recommender system	DM	P28	<10,000
15	Recommender system	UXR	P31	<10,000
16	Resume screening	DM	P29	>10,000
17	Search & retrieval	UXD	P17	>10,000
18	Search & retrieval	UXD	P19	<10,000
19	Search & retrieval	PM	P22	>10,000
20	Speech recognition	UXR	P13	>10,000
21	Text prediction	UXD	P6	>10,000
22	Text prediction	UXD	P27	>10,000
23	Warranty processing	UXD	P25	<100

I paid attention to two aspects during recruitment. First, I recruited participants from a broad range of technology areas (e.g., natural language processing, information retrieval, recommender systems, etc.). Second, I tried to include organizations differing in size and service type (i.e., consumer or enterprise).

I conducted semi-structured interviews with 31 practitioners across 23 AI product teams from 14 companies (9 large companies, 4 startups, 1 nonprofit). While I did not limit participation to any role, the participants mainly included designers (e.g., design managers, user experience designers, and user experience researchers) and product managers. Although some participants shared that their engineering and data science colleagues used the guidebook, I could not recruit any participants from those roles. Participants worked on teams developing AI products across a wide range of technology areas, including recommender systems, medical diagnosis, text prediction, and more. Table 3.1 provides a summary of the participants' teams, roles, and technology areas.

The interview protocol had three main parts. First, I asked participants about their roles, practices, and workflows on AI projects to gain background information and context. Next, I asked them about the use of the guidebook. I asked them to walk me through a recent case where they incorporated the guidance in their work. I probed them about how they discovered the guidebook, and whether and how the guidebook helped with specific challenges. Finally, I asked them about the remaining challenges that were not covered by the guidebook, their needs for support, and any other areas for improvement. Whenever possible, I encouraged participants to share any artifacts created as part of using the guidebook or as aids for relevant challenges. These included worksheets, design patterns and examples, case studies, and workshop materials. All interviews were conducted remotely on a video conference platform.

I recorded and transcribed the interviews, and documented the artifacts participants shared during or after the interviews. Each interview lasted between 30 to 60 minutes, resulting in 18.5 hours of audio in total. I analyzed the transcripts using affinity diagramming [78, 96], pulling out insights and generating codes for participant utterances. Following a bottom up process, I iteratively reviewed and synthesized these into high-level themes related to current practices, challenges, and emergent approaches.

3.3 Findings

I present the findings around three themes: using the guidebook as a means for building a culture around human-centered AI, practices for putting AI design guidance into practice, and emergent practices for remaining challenges. Within each high-level theme, I describe sub-themes detailing specific challenges and the corresponding use of the guidebook for support. I share implications and design opportunities at the end of each high-level theme. The themes are not exclusive; some aspects spread across themes.

Related to participants' team roles, I observed a distinction between lead roles (i.e. design managers, product managers) and individual contributor roles (i.e. UX designers, UX researchers) regarding their challenges in developing AI products. While all participants operationalized the guidebook in some way, lead roles described it as more of a strategic resource, whereas individual contributor roles described it as a tactical resource. Although this was not unexpected, it became a relevant aspect to capture concerning different needs for support based on team role. In the detailed findings below, I note where this difference was evident.

3.3.1 Establishing a Culture Around Human-Centered AI

When I asked participants about their experiences with the guidebook, I mainly expected to hear how they used it to address AI's interaction design challenges. Interestingly, participants' answers revealed that they often used the guidebook as a means for building a culture around human-centered AI within their teams and organizations. In this section, I detail the use cases around (1) education; (2) developing internal resources; (3) cross-functional alignment; and (4) gaining credibility and buy-in. I discuss the implications for the design of future human-AI guidelines at the end of the section.

Participants used the guidebook for educating themselves, their teams, and organizations. Most participants reported that the literacy around data and AI is low among PMs and UX practitioners within their teams and organizations. Especially, participants in lead roles expressed ongoing educational efforts to address this key concern. For example, several participants (P8, P13, P14, P21, P22, P31) gave talks and presentations to large audiences using the guidebook content to create awareness on AI's design considerations: *"I gave an internal talk maybe to 60-80 designers and PMs with a lot of lessons from the people AI guidebook." (P21)* A few participants (P17, P28, P31) mentioned using the guidebook chapters as a skeleton to create an internal course for product teams –typically for PMs and designers– who are new to working with AI: *"We have recently created an ML for designers course internally as part of educating our peers … [We used] examples that we either have shipped or we just experimented with and haven't shipped to tell the story of how to work with AI and ML." (P28)*

I also observed that participants who did not have any prior experience working with AI (P3, P4, P12, P14, P15, P18, P25) used the guidebook for self-education: *"My first objective was: I need to get smarter about this topic. … [after digesting the guidebook content] Then my next step was, how do I communicate this to my PM, engineering and UX team?" (P3)* Several participants found the guidebook approachable as an entry point: *"I went from knowing nothing to being someone who has a substantive amount of knowledge in the UX of ML." (P4)* All participants shared that they became proponents of the guidebook, educating their teams through sharing specific chapters or patterns or through short talks and presentations. Similarly, a majority of participants reported learning about the guidebook through their colleagues or broader professional connections. A few participants mentioned becoming aware of it through industry conferences or discovering it organically when searching for resources.

The guidebook provided a benchmark for developing internal resources and guidelines. Many participants (11 teams out of 23) revealed that they have self-created AI design resources and guidelines for internal use, and that they referred to the guidebook as a benchmark when they were developing their own resources. Among these, some participants noted that they benchmarked multiple resources, including industry guidelines (e.g. [7, 6, 69, 114]) and academic papers (P21, P28, P29, P30). Few participants (P4, P23) also shared that they have plans to develop internal resources, and are *"using this as a source of truth in the meantime." (P23)* Additionally, within large technology companies, I observed multiple internal resources and guidelines that were customized for a product area (e.g. healthcare AI design guidelines).

Internal resources often took the form of playbooks, documentation hubs, and slide decks, and included domain-specific patterns and examples [detailed in section 3.2.2]. A design lead mentioned using the guidebook design patterns when developing a design system and component library for AI-based products: *"We referenced several of the patterns when creating design components specification." (P29)* For example, they created components for representing intelligent agents, such as chatbots, in a way that sets user expectations and mental models.

When probed about the need to create domain- or organization-specific resources, participants highlighted the importance of relevance to own organizational context and domain: *"There is a strong need to appropriate it to your own domain, to your own industry, and to your own organization as you probably have very different needs. … Regardless of how good the guidebook or these kinds of tools could be, we would have to do that appropriation step and pick and choose what is relevant for us." (P30)* Interestingly, several participants noted that even with the custom developed internal resources, creating awareness within their organizations remained a challenge.

Participants used the guidebook for easing the challenges of cross-functional collaboration. Participants brought up several challenges around cross-functional collaboration in AI-based projects, and they described how they used the guidebook to overcome such challenges. For example, several participants mentioned establishing a shared language as a major benefit: *"Establishing the vocabulary around the space is one of the biggest values that I see in an artifact like this. Because sometimes we use the same terms and don't refer to the same thing, or use different terms for the same thing." (P20)* Some participants noted that it empowered them in participating discussions and product decisions: *"Anytime we work from a principles based approach, it levels the playing field." (P4)*

Participants in lead roles spoke of using the guidebook to sensitize their engineering and data science teams to AI's design considerations, and to human-centered AI in general:

> *"[We did a workshop using the guidebook chapters] focusing on how do we make these features more human centered … The engineering and the data science team found it incredibly useful because it helps them to better understand everything you may have to think about when designing for ML. … That was very beneficial for building the human centered muscle within the team, regardless of the actual outputs of the exercises." (P8)*

Participants shared that these efforts helped to facilitate a shared understanding and alignment within their teams, and were well received by cross-functional partners.

The guidebook was used for gaining credibility and buy-in for design recommendations. Several participants spoke of using the guidebook as an artifact for negotiation and alignment on design recommendations (P3, P6, P10, P20, P25, P26, P31). As a PM working on business analytics put it, *"Without the principles that guide the conversation, it can be everyone's personal preference or opinion on how a feature should be or not … What's a stronger pitch or recommendation is, we should do this, because it's tied to this guideline or principle." (P10)* They shared that being able to reference such a resource brought credibility and helped to convince their teams, as the guidance showed tried and true solutions synthesized across many products and academic studies. Interestingly, all UX researchers (P3, P13, P26, P31) described using the guidebook as a literature review; they cited the relevant guidance when presenting research findings: *"[When we shared our interview study findings] data scientists' first* *reaction was, you make all these claims and recommendations, how generalizable are these? It was really nice to be able to say, we looked at the literature and the people AI guidebook, and brought these things together." (P26)*

Implications for AI Design Resources.

- **Support learning within AI design resources and guidelines.** While human-AI design guidelines are primarily designed to support practitioners in problem solving, the findings indicate that practitioners also use these resources for self and organizational education. Future AI design resources might be designed to better support bi-directional learning: human-centered AI concepts for data science and engineering teams, and AI concepts as it relates to product development for product teams. For example, resources can be presented in various forms, such as short videos or interactive modules for self-learning, or as slides for delivering talks and presentations. In contrast to the abundance of educational resources for technical AI/ML concepts (e.g. ML crash-courses [56, 106]), there is little educational content for human-centered AI ([45, 115] as rare examples). Future resources can be served in a course-like format to support both individual and organizational learning.

- **Support cross-functional communication and collaboration.** A large body of HCI research has suggested creating boundary objects to effectively scaffold collaboration between AI practitioners [22, 155]. Findings from this study offer preliminary evidence highlighting the potential of AI design resources to facilitate the collaboration between team members and stakeholders from different disciplines. I see a great opportunity for future resources to be explicitly designed for use by multiple teams and roles across product development. Future research should investigate the specific needs of different roles to better understand how to support team members with different knowledge and expertise.

- **Support adaptation and appropriation of resources for the development of domain- and organization-specific guidance.** Prior research discussed the inherent trade-off between generality and specialization for developing AI resources and guidelines, noting that specific applications might require specialized guidelines [6]. The results of this study confirm this; I found that teams and organizations have a strong desire to develop their own resources to better contextualize the guidance for their specific application domains. To this end, resources and guidelines from industry and academia served as a benchmark for practitioners. One clear implication is to design resources for adaptation and appropriation by providing explicit guidance and affordances. For instance, resources might contain guiding questions to help practitioners assess which design considerations and guidelines might be most relevant to their domain, service type or AI technology areas. Similarly, guidelines can include multiple case studies and examples demonstrating how a particular design consideration might apply to different contexts and interaction scenarios.

3.3.2 Putting Human-AI Guidelines Into Practice

In this section, I present how practitioners addressed some challenges of designing AI products by incorporating concrete guidance from the guidebook. I detail how practitioners operationalized the guidebook, and their needs for enhanced guidance.

Framing Human-AI Problems. Participants often commented that AI projects are heavily driven by technical considerations, and that a human-centered approach can be missing. A major challenge in AI-centric development was ensuring that the AI product was addressing a real user need. Several participants brought up the chapter on defining user needs (chapter 1) and stressed the importance of problem framing for setting up AI projects for success. A PM working on search and recommenders (P22) systems shared:

> *"I use the guidebook for problem framing to make sure that machine learning is solving an actual people problem. … [Before] it was being treated as a technical solution, there was no thought into what kind of user problem is being solved. And because of that, a lot of the things that people were shipping didn't work."*

P22 gave an example where the data science team was assigned a "time to last result" metric to predict when all the relevant results have been loaded to stop the search. When the team conducted user research, they realized that it was a nonexistent problem: *"[Users] didn't even care if the results are still loading, they only cared about the first three results. [The data science team] were optimizing for a metric that wasn't even relevant." (P22)*

This was a shared sentiment regardless of whether participants worked at large software companies or startups. Several participants (P3, P13, P18, P19, P22, P27) similarly recalled projects that failed or got canceled, and reflected that they should have spent more time framing and defining the problem: *"We did some concepts and tested [the feature] with users, no one really wanted it. … In retrospect, I would have pushed back [on the idea early on]." (P19)* In some cases, participants were able to reframe the problem to iterate and pivot to a solution grounded in user needs: *"Is there a different experience we can create? Maybe it's a useful model. Working with data science, you can often pivot and do something else." (P19)*

Participants described using the workshop kit and worksheets in the guidebook for conducting problem framing workshops where they collectively defined user needs and success metrics with their cross-functional team members. A UX designer working on medical diagnosis shared an activity where they asked their team to separately fill out a worksheet detailing the intended user, the intended use, and the value for the user: *"There was actually divergence … Then it becomes a discussion to align on these." (P5)* Some participants shared that the guidance helped them assess whether AI was the right fit for a problem or whether they should use heuristic or rule based approaches instead (P3, P12, P17, P29, P31). However, some participants felt limited in questioning and reframing problems as their job was centered around providing AI solutions: *"I'm definitely not a person who believes machine learning should be used for everything, but a lot of the projects that I work on are pre-baked with machine learning. So I don't often get to ask 'should we actually use ML systems?'" (P21)*

Addressing AI's specific design considerations. Participants often used the guidebook to identify and address AI's design challenges around setting expectations and mental models, providing explanations, designing feedback and control mechanisms, and mitigating errors (chapters 3-6). Some participants noted that they had no prior knowledge around these challenges: *"A huge part of it was figuring out, what are the design considerations that we need to have when using machine learning?" (P19)* Most participants recalled cases where they used the guidance to influence product decisions: *"[Based on] the chapters in the guidebook, we provided a lot of feedback and explanations, and a lot of control to users." (P29)* Specifically, they mentioned examining user journeys and service blueprints of their products to identify where critical issues, such as trust issues or errors, might arise. A majority of participants shared conducting focused workshops and design sprints to collectively envision solutions that operationalize specific guidance (e.g. explainability).

Most participants became aware of problematic issues and UX breakdowns during testing or after launch. Indicators of breakdown were typically lack of user adoption, lack of trust or satisfaction, and error reports from users. Similar to reports in AI fairness literature [65], several participants commented that addressing design challenges *"tend to be reactive despite our best efforts" (P6),* and expressed a desire to anticipate these challenges early on: *"the majority of our learnings happened in internal beta testing when the feature is live, where it's time expensive and resource heavy to make any changes. … [Designing AI products] has been a learning curve for our team at every stage of the development cycle, from preparing for launch to post-launch to internationalizing." (P13)*

Other participants reported that while they were aware of AI's design considerations, it takes time and several iterations to craft and implement thoughtful solutions: *"We launched some of our initial [image classification] features without really good feedback loops … We've had to see the real user impact of some of the features in order to really put these principles into practice. So the guidebook was helpful in getting from nowhere to somewhere. We're seeing even now how much more we need to do, how we need to build controls that are more flexible, more granular or have better clarity to help people truly avoid harmful experiences." (P7)*

Explainability and trust was the most referred chapter; data collection and evaluation was the least referred chapter. While participants referred to all the aforementioned chapters, explainability and trust was the challenge they spoke of the most. Interestingly, I observed that participants used "trust" as a broad term to describe many problematic situations. Some participants spoke of intelligibility, where users did not understand what the AI system could do and would not use the suggested system actions (P1, P3, P9, P18). Others described usability and user acceptance issues, where users understood what the system does, but would not trust it to do what it claimed it was capable of doing (P13, P14, P24, P25). Few participants mentioned cases where the system produced biased or incorrect outputs when describing lack of user trust (P6, P7, P29). Several participants reflected on the complexity of trust as a concept: *"[part of the challenge] is thinking of the different dimensions of trust, as it's an open-ended and complex concept. It's almost like saying, what is love? It means different things to different people in different situations." (P6)*

I also noticed that only a few participants referred to the chapter around data collection and evaluation (chapter 2). These participants spoke of issues around bias in machine translation (P6), automatic speech recognition (P21), resume screening (P29). When explicitly probed, most participants shared that they did not use the guidance around data as they worked with pretrained models. Among the participants who were individual contributors (20 out of 31), only three were involved in the collection of initial training data. Instead, participants mostly worked on collecting user feedback data after launch (chapter 5). Only one participant, a design lead on business intelligence applications, mentioned using the guidance on data collection for designing systems for data labelers who annotate training data (i.e. *'design for labelers and labeling'*) (P29).

Patterns and examples work well, but there is a need for domain-specific examples. Nearly all participants referenced the design patterns and examples, whether for developing internal resources or for addressing specific design challenges as individual contributors. Several participants commented that the patterns made the overall content and guidance actionable and practical. The major challenge with design patterns was the relevance of examples for participants' own domain. Participants' criticisms included the lack of breadth; narrow focus on consumer-based examples; and limited use of real-life examples.

A majority of participants shared that they collected their own domain-specific design patterns and examples. When I asked participants how they curated these patterns and examples, few dimensions seemed important. First, participants wanted patterns that are hyper relevant: *"... it's not just medical, but specifically examples around medical imaging." (P5)* Similarly, a UX designer pointed out that best practices and patterns might differ within a specific technology area, such as language interactions: *"Confidence in voice is very different than confidence in chat. In voice, confirming what the user said would be the best practice; for chat it's surfacing multiple options." (P21)* Second, participants searched horizontally and looked at adjacent domains to see a breadth of examples. For example, P13 working on mobile communication referred to examples from email products; and P24 working on personal healthcare devices looked at personalized home energy products for patterns around automation. Third, relevance to own service type – consumer vs enterprise applications – stood out as a salient aspect in collected examples.

In addition to collecting existing examples, several participants mentioned creating hypothetical concepts and UI mocks for contextualizing how a pattern might apply to their product. They often referred to the guidebook's pattern representation illustrating patterns and anti-patterns as *industry standard*, noting that they adopted this structure when creating their internal patterns. A UX lead (P7) shared an internal example where they introduced even more granularity in the pattern and anti-pattern spectrum. For example, they sketched four UI mocks for collecting user feedback: one without any feedback mechanism, one with feedback mechanism but without any explanations, one with a generic explanation, and one with a better use of explanations. Internal pattern collections became a part of internal resources [detailed in section 3.2.1] and were typically documented using slide decks or visual workspaces (e.g. Figma [49]). However, several participants desired more effective mechanisms for sharing and reuse of this knowledge with other teams.

Participants observed a positive impact on products, but shared many challenges around measuring and evaluating impact. A large body of work in HCI has demonstrated the effectiveness of the guidelines [89, 166]. Similarly, participants often observed a positive impact on products, and reported overall positive perceptions around the usefulness and effectiveness of the guidebook. However, several participants expressed challenges around measuring and evaluating the impact of guidelines on their products (P2, P6, P11, P13, P25, P29). Participants mostly spoke of impact through metrics such as increased adoption rate, satisfaction, engagement, task completion rate, and resolution time, measured through qualitative or quantitative studies. However, they desired better support for measuring the impact of incorporating AI design guidance in their products, including metrics and methods: *"It would be great to hear what others have done to understand how to measure success or understand what is 'good enough' for launch." (P13)*

Related to measuring impact, several participants emphasized that it was critical to communicate the product impact and the business case to the larger product team for the guidelines to be prioritized. A UX researcher (P3) spoke of constraints of the product development process: *"The only push back I got from my team was, what priority is this? Should this be part of the MVP (minimum viable product)?"* They shared that the design team was able to make a case for incorporating explanations by creating a few design concepts and validating the concepts through a user evaluation study.

Implications for AI Design Resources.

- **Provide context-specific design patterns and guidance.** Prior research has emphasized the need for contextualization in AI resources and toolkits for fairness [37, 149]. The findings of this study echo this; participants worked to contextualize the design considerations and patterns to their products and applications. Future resources and guidelines should provide patterns and examples that (i) are **domain-specific** (e.g., best practices for providing explanations in healthcare); (ii) include **a breadth of service types** (e.g. consumer vs enterprise applications); and (iii) display **a breadth of patterns** for specific AI design considerations (e.g. different ways to collect explicit feedback). Future work is needed to explore ways to index and organize such a pattern library where practitioners can contribute their work. What would be a good level of abstraction for contributing unpublished or hypothetical examples remains an open question.

- **Provide both patterns and anti-patterns with varying degrees of granularity.** Traditionally, HCI research has followed well-defined formats for design patterns, such as Alexandrian and Tidwell forms [140, 31, 4]. Interestingly, a majority of participants in this study seemed to have adopted the pattern and anti-pattern structure that is commonly used in industry design resources [59, 57] when developing their internal patterns. Studying how practitioners create and appropriate AI design patterns marks a clear space for HCI and design research. Future work can investigate these forms through artifact analysis to identify which dimensions are critical for practitioners to effectively understand and apply AI design patterns "in the wild".

- **Provide guidance on measuring and evaluating the impact of Human-AI guidelines.** HCI literature has provided several methods for evaluating design principles' impact on usability and user experience, such as comparative usability testing and heuristic evaluation [107, 26]. However, applying these methods for measuring the impact of AI design guidelines can be costly and complex [89]. The findings confirm this challenge; many participants expressed difficulties in measuring the impact of guidelines and principles in the context of their specific applications. Future resources should provide explicit guidance on measuring and evaluating the impact, including relevant metrics and methods for pre-launch (e.g. lab studies) and after launch (e.g. real-world data). For instance, recent research suggested using factorial surveys for evaluating AI systems that do not yet exist [89]. Guidance on evaluating the impact may help practitioners in getting organizational buy-in and championing human-centered AI practices forward.

3.3.3 Emergent Practices for Broader Challenges

In this section, I present additional challenges that were not supported by the guidebook, and detail the emergent practices I observed in response to these challenges.

Practitioners need more support for early phase AI ideation and problem formulation. A major critique from participants was that while the guidelines were helpful for solving existing problems, little help was provided for finding the right problems: *"The guidebook really focuses on optimizing a design you already have … like diagnosing why something might not be working versus the envisioning step." (P13*) Several participants highlighted needs for support in the early stages of the product development for ideation and problem formulation with cross-functional partners. As a PM put it (P22), *"There is even a step before [defining user needs]: what is a problem that you could apply machine learning for? … [Product managers] don't even know how to ask a data scientist whether we could solve this through machine learning."* They emphasized the reciprocal relationship between product development and model development, as they worked to find use cases where machine intelligence can improve products and user experience:

> *"Given a current solution or an envisioned future solution, how do we ideate about what AI could be doing or what the data requirements would be? … I feel like that probably happens before you even get to such a guidebook." (P30)*

Interestingly, there were few participants who created their own resources and artifacts for scaffolding ideation and problem formulation with AI. Similar to reports in prior literature [157, 161], these were AI capability abstractions that delineated a subset of relevant capabilities and value propositions. For example, a UX designer working on business analytics (P1) shared an "AI pillars" framework they created with their data science team prior to conducting a visioning workshop. The framework involved the capabilities *discovery, recommendation, prediction,* and *automation,* where each capability was detailed with the types of inferences and value (e.g. *discover seasonality and industry trends to help businesses understand market and customers*). A UX researcher (P13) shared a similar framework outlining AI capabilities (e.g. *identify, inform, anticipate,* etc) and value propositions (e.g. *save time, save* *effort, provide assistance, reduce distractions,* etc). Reflecting on how this capability framework helped with brainstorming, P13 noted:

> *"The framework offered [the designers and PMs] a way to think about what ML could do, whereas beforehand, it was just the ML team who understood what it could do. There wasn't this shared understanding of what [ML] could offer. … Before, people were throwing out ideas [sporadically]. Now [our] team has a hyper focused mindset of 'what are all the things we could do?' … [to] brainstorm [new product] features within that capability." (P13)*

Practitioners described a hybrid design process to reconcile user-centric and AI-centric approaches. When talking about the broader challenges of designing AI products, several participants expressed tensions between user-centered design (UCD) and AI product development. Some participants reflected on the fact that having an already built AI model limits the solution space: *"We were already coming at it from, where can we add intelligence to add value? Not necessarily from a blank slate like, what pain point should we solve?" (P3)* Other participants recalled cases where the UCD process alone was not effective in identifying AI opportunities. For example, a UX researcher tried engaging end users in AI product development, but simply asking users where they need intelligence did not work: *"How do we develop a method that isn't just asking people where they want machine intelligence? Instead we can learn what's painful about their process … [with an eye] out for where we can fit [intelligence] in." (P26)*

Additionally, there were challenges in brainstorming without thinking about limitations and constraints of data and AI: *"[In visioning workshops] teams tend to think of ideas that are too far ahead, then you have challenges in grounding [ML ideas]." (P17)* Some design leads and PMs simply recognized this as a shortcoming of UCD: *"[For] the ML or AI side, we don't really have a process baked in at all. How do we discover a new solution for customers or a feature or product that might be a good fit for an AI application?" (P23)* Interestingly, several participants described a hybrid process where they started both with user needs *and* AI models or data, and sought to explore good matches in this problem-solution space. As a PM put it (P16):

> *"[We do a lot of sketching] who are these personas, why [is this] useful? What are the range of interactions that could happen? [Then thinking about] all the different inputs for training the model that may have downstream value … [not only to ensure relevance but also] to minimize the downstream risk."*

Implications for AI Design Resources.

- **Broaden the scope of human-AI resources and guidelines to provide support for early phase ideation and problem formulation.** Recent literature on human-centered AI and fairness in AI have highlighted the importance of early phase problem formulation, reframing, and ideation for solving the right problem [112, 158, 161]. The findings confirm prior speculations on the lack of guidance and support for early phase AI product development [158]. In response, some researchers explored curating AI capability abstractions and value propositions to sensitize product teams to what AI can do [154]. The findings of this study echo this approach; several participants expressed a need to understand what AI can do for ideating and formulating problems; and a few participants shared similar AI capability-value frameworks they had created as aids. Future resources could include taxonomies and frameworks delineating what AI can do, and what types of problems are best suited for AI to better support practitioners in ideation and problem formulation.

- **Support practitioners in incorporating resources into their existing processes and workflows.** Similar to findings from prior literature [157, 137, 42, 148], the findings point out that practitioners start AI product development in different stages of AI model development; and that the entry point might not be apparent for every team or product. Future resources can better support practitioners in incorporating guidelines and design considerations into their existing processes by explicitly connecting them to different stages in AI product development. For instance, resources around data collection can be used in upstream product development for concept generation; and in downstream product development for assessing current data collection practices. Similarly, providing actionable forms and methods, such as design workshops and patterns, may help practitioners in operationalizing design guidance. In particular, workshop resources that can be adapted to collaborative work (e.g. visual workspaces such as Miro [98] and Figma [49]); and to time constraints (e.g. one hour vs one day long workshop kits with facilitator resources) can lower the barrier to entry for practitioners.

3.4 Discussion

3.4.1 Selecting the Right Human-AI Problem

HCI researchers have cleverly distinguished problem setting from problem solving [123, 21]. Our community has specialized methods for sketching – selecting the right thing to build; and prototyping – building the thing right. This study revealed that while AI resources and guidelines support prototyping, they offer little support for sketching. AI products and experiences are difficult to ideate: once a dataset is collected and a model is built, the design space becomes limited to certain problem framings [158]. Participants in the study recognized this challenge, and expressed needs for support in early phase problem formulation and ideation. I hope that the emergent AI capability-value frameworks presented in this work will inspire future resources and guidance to support practitioners in selecting the right human-AI problem.

On a higher level, these findings point to a more consequential problem: AI products fail when human centered perspectives are lacking, especially in the early problem formulation phase. Several participants shared hard learned lessons through project failures, highlighting how problem framing is essential for success yet overlooked. Despite recent evidence [146], AI product failure is rarely discussed in the HCI literature. Most HCI research around AI is focused on usability; yet our study shows that practitioners need more support for assessing usefulness. While the lack of methods for usefulness is not specific to working with data and AI [58], identifying the right thing to design becomes more critical for AI products as it can be extremely difficult and costly to make changes once an AI system is built. I see a real opportunity for HCI practitioners to play a critical role in ensuring that AI products are solving real problems. Future research should provide methods and resources for iteratively framing, reframing, and pivoting to ensure a match between AI solutions and human needs.

3.4.2 Reconciling User-centric and AI-centric Approaches

Previous work surfaced the tensions between user-centered design process and the AI development process, and speculated that AI product development may require design processes beyond UCD [50]. The findings echo this: I observed emergent processes that seemed to be a hybrid between user-centric and AI-centric processes. Several teams shared that during problem formulation and ideation, they limited their solution space to existing AI models and data sets. Similarly, teams spoke of their process as customer or feature discovery, where they conducted needfinding studies with an eye for particular AI solutions. As prior research noted [157], their process resembled technology-centric product development (e.g. lean startup [116], matchmaking [16]); yet it differed in that participants were not considering all possible customers and users. Instead, they were looking at a small set of users related to their products and services.

Future research should further investigate these new, hybrid design processes for reconciling user-centric and AI-centric approaches. Similar to [112], ethnographic studies exploring different modes of AI product development across different teams, organizations, and product settings might reveal insights into how product ideas are selected, defined, and prioritized during early problem formulation. New knowledge into AI product innovation processes may open up new opportunities for embedding human-centered design approaches throughout the product development lifecycle. I encourage HCI and design researchers to explore, speculate, and formalize the processes of practitioners by studying these in the wild.

3.4.3 AI Design Guidance around Fairness

This study revealed parallels between AI's design challenges and AI's fairness challenges. Similar to prior literature [65, 37], practitioners in this study expressed difficulties in mitigating potential bias (P6, P7, P16, P20, P21, P29, P31). While industry AI design guidelines provide some guidance around fairness and bias, these tend to be high level and underdeveloped [151] (e.g. "mitigate social biases" [6], "consider bias in the data collection and evaluation process" [111]). The findings show that practitioners currently lack a vocabulary to describe fairness issues. They need more granular guidance detailing risks around trust and fairness. How might fairness and bias issues manifest themselves in specific domains and AI applications? Madaio et al.'s work detailing practitioners' fairness desiderata provides a great starting point [94]. Future research should explore how to categorize and communicate fairness considerations in a way that is easy to incorporate into current practices (e.g. design patterns, case studies).

From a broader perspective, this study highlights an overlapping scope between AI design guidelines and AI fairness toolkits. Currently, AI fairness toolkits are typically targeted towards engineering and data science roles who drive the upstream AI development processes (e.g. data collection, model building). However, this study and recent literature indicate that many other product roles (e.g. product managers, UX designers, software engineers, domain experts) inform data collection practices and contribute to improving fairness efforts [65, 94]. I suspect that the emergent practices around ideation and problem formulation provide a great opportunity for identifying potential fairness risks early in the development process collaboratively. Future research should investigate how these seemingly separate guidelines overlap, and how to scaffold fairness efforts across multiple team roles and stakeholders.

3.5 Limitations

This study has several limitations. First, the study focused on a single resource, the People + AI Guidebook, as an instance of AI design guidelines from large technology companies. Future research should explore the broader landscape of AI design guidance, including resources from industry, academia, and more. Second, by recruiting participants who already have used the guidebook, I may have sampled practitioners who find AI design guidelines useful and are unusually motivated to use such resources. Third, the participants mainly included designers and product managers, there are many other roles involved in AI product development (e.g. software developers, researchers, data scientists, domain experts, policy makers) whose perspectives were not covered in this work. Future work should take a broader recruitment approach to recruit participants from a wider range of roles, teams, and organizations, including people outside of academia or industry (e.g. government agencies, civil society organizations). Finally, these findings are based on retrospective interviews that were conducted over a limited period of time. Future research should include longitudinal ethnographic studies to fully understand the uses and impact of AI design guidelines and resources.

Chapter 4

Industry Best Practices for AI Innovation

In this chapter, I investigate the industry best practices for cross-disciplinary AI innovation with a focus on early stage ideation and problem formulation.

I collaborated with researchers at Accenture, a company that develops AI-powered enterprise software for many industrial clients, to explore how their cross-functional teams navigate the challenges of AI innovation. These were design and data analytics teams at an incubation hub, working on early stage, first of its kind AI products and services. For example, they had recently designed and developed a human-in-the-loop logistics platform where workers could customize the optimized route plans to create detailed delivery schedules [2].

Partnering with these researchers and practitioners, I led a series of design workshops and discussions. The workshops produced several interesting findings:

1. Designers bring more impact at early stage AI innovation through design thinking, systems thinking, and service design, as opposed to searching for opportunities at the user interface level of a project.

2. Designers had developed internal AI educational resources delineating AI capabilities with example applications. These resources scaffolded cross-disciplinary, collective ideation with stakeholders that did not have AI or data science background.

3. Close collaboration between design and data science offered a successful path for cross-disciplinary AI innovation. Designers served as facilitators in early stage ideation and problem formulation.

4.1 Method

We formed an team composed of 22 participants, including academic and industry researchers, interaction and service designers, data scientists, research scientists, and AI engineers. We conducted three design workshops with two teams, resulting in a total of six workshops. Each workshop session lasted between 1-2 hours (10 hours in total). All participants had worked in professional practice for more

TABLE 4.1: Our team comprised of researchers (R) and practitioners (P) with experience across AI, HCI, design and data science.

ID	Session	Professional Role	Exp.	Org.
R1	All	Principal Director/Fellow	10+yrs	Industry
R2	1-5	Chief Technologist	10+yrs	Industry
R3	All	HCI Researcher/Designer	10+yrs	Academia
R4	All	HCI Researcher/Designer	10+yrs	Academia
R5	4-6	HCI Researcher/Designer	10+yrs	Academia
R6	All	HCI Researcher/Designer	5-7 yrs	Academia
R7	1-3	HCI Researcher/Designer	5-7 yrs	Academia
P1	1-3	UX Designer	7-9 yrs	Enterprise
P2	1-3	UX Designer	7-9 yrs	Enterprise
P3	4-6	Design Research Lead	10+yrs	Enterprise
P4	4-6	Service Design Lead	5-7 yrs	Enterprise
P5	4-6	Service Design Lead	5-7 yrs	Enterprise
P6	4-6	Service/UX Designer	3-5 yrs	Enterprise
P7	4-6	Data Designer	3-5 yrs	Enterprise
P8	4-6	Data Designer	10+yrs	Enterprise
P9	4-6	Data Designer	10+yrs	Enterprise
P10	4-6	Design Lead	10+yrs	Enterprise
P11	4-6	Group Design Director	10+yrs	Enterprise
P12	6	AI R&D Managing Director	10+yrs	Enterprise
P13	6	AI Research Engineer	10+yrs	Enterprise
P14	6	AI Research Principal	7-9 yrs	Enterprise
P15	6	Data Architect	3-5 yrs	Enterprise

than 3 years. Table 4.1 provides a summary of our research teams' composition, relevant experience, and the participants' involvement in workshop sessions.

Each workshop session focused on a different stage in the innovation process, roughly to correspond to the early phase (Discover, Define), mid-phase (Define, Develop), and late phase (Develop, Deliver) [32]. We asked practitioners to complete a prework activity prior to workshop sessions, which scaffolded the workshop activities described below:

Workshop 1. Designers created customer journey maps of their design process to detail their current workflow, tools, and stakeholders as prework. Researchers shared a short presentation of their projects on design-driven AI innovation, including a case study and design patterns for adaptive mobile experiences [160, 170, 172]. Designers reflected on their recent enterprise design projects where they improved UX with AI. Researchers probed whether, when, and how they envisioned AI-driven interactions. We discussed whether interface-level AI innovations provide an opportune space for design-driven AI innovation.

Workshop 2. Designers shared projects showcasing an AI experience they designed. We discussed practices for ideating and prototyping AI-enabled products and services. We reflected on scenario and wireframe generation with data and AI to understand how practitioners represented data dependency and user labelling of data in interfaces.

Workshop 3. Designers reflected on the tools, methods, and processes they used to capture, document, and transfer their AI-enabled designs. Researchers introduced the concept of boundary objects for collaboration [133, 30]. Designers and data scientists collectively discussed their innovation process with a focus on cross-disciplinary communication and collaboration, and the role of design in facilitating AI innovation. Participants were asked to articulate challenges, pain points and best practices for AI innovation. Following this session, we conducted an additional meeting with participants for debrief.

Workshops were held over video conference, due to the Covid-19 pandemic. I recorded and transcribed the workshop sessions, and documented the artifacts generated during the sessions or shared prior or afterward. I analyzed the transcripts using thematic analysis [19], and discussed emerging themes with our research team. I iteratively reviewed and synthesized the insights to identify thematic patterns.

4.2 Findings

4.2.1 AI's Design Innovation Space

In response to questions about how they envision AI-driven product and service innovations, designers shared several examples where they recognized opportunities for AI to improve a situation. Prior literature reported on designers innovating on the interface-level product features and interactions [157]. Interestingly, participants shared that interface-level opportunities only partly cover how they innovate with AI. In explaining how they worked, designers sketched a diagram to help illustrate their view of the AI innovation space, the space where design thinking had the most impact (Figure 4.1).

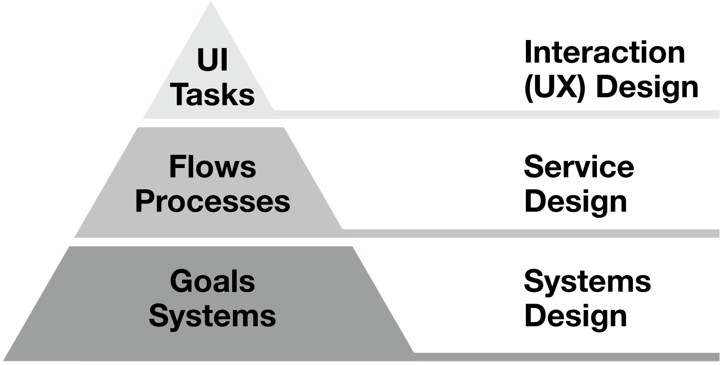


Figure 4.1: Three levels where designers can recognize or discover ways for AI to improve work. The width of each level indicates the value and impact design brings.

The diagram illustrates design activities at three levels: interaction (UX) design, service design, and systems design. At the top, AI optimized repetitive tasks that happen in an interface. At the bottom, AI helped to improve workers' performance, often by offering new insights or by augmenting their capabilities. Participants reflected on the tensions between automation and augmentation, and referred to the bottom level as a richer space for design: *"Often we hear the narrative "we'll automate the low value tasks for people to move to better jobs", but no one designs what those jobs are. [Designers can create value in cases where] you're not going to reach a 100% automation as the data itself is changing over time, and there is a role for the human in the loop to deal with the hard cases, but also to directly train the algorithm ... [Design can have a positive impact] for the users as they can employ their skills; the future of their job takes on some ownership for that model." (P11)*

Designers spoke about these three levels as a continuum, and they shared examples of recognizing opportunities at each level:

Interaction Design (User Interface, Tasks): Designers had worked on a tool to classify financial transactions. This application divided operational work between the AI system and the human workers. The AI system automatically classified the most common and easily recognized transactions, and the workers classified the infrequent, uncommon transactions that required their expertise and common sense. As the interface design took shape, designers realized that the AI system could also classify many of the uncommon transactions correctly. This switch accelerated the pace of the work. Based on the new interaction design, workers only had to confirm if the classification was correct, or they would repair the misclassified transaction. *"The labels were there already. So we could use that as a placeholder to say 'is this right' rather than asking [people] to fill it from scratch." (P5)*

Service Design (Flows, Processes): When designing an AI decision support tool for the pharmaceutical industry, designers conducted design research to understand scientists' mental models and workflows. They found out that scientists search various web sites for data regarding clinical trials. Realizing that this was crucial to their workflow, the team worked to ingest some of the data into the AI system: *"We circumvented linking off to a website that doesn't necessarily fit with their flow and pulled that data in and reorganized that in a way that suits them better." (P5)*

Systems Design (Goals, Systems): Designers had worked on an AI system to discover relevant relationships between medical diagnoses and treatments [147]. They created a tool where clinical experts (typically nurses) reviewed the discovered relationships in order to validate that this might be relevant. When a relationship was approved, it was forwarded to data scientists who used it to update the knowledge graph. One human-AI challenge was to motivate high quality work from the clinicians. The team reframed the role of these experts, shifting away from thinking of them as "coders" and explicitly referring to them as "AI curators". This shift served to *"upskill them, allowing them to be AI producers without becoming data scientists" (P12).* The interaction design used clinicians' expertise to build the AI's knowledge graph; clinicians directly trained and maintained their AI system. The new design simultaneously enhanced job satisfaction and improved AI learning and knowledge discovery.

Designers shared that the complexity of the AI system changed the level where design thinking could impact innovation. P11 shared that for simpler AI systems, *"...designers can focus on the UX, we can use traditional [UX] methods and the target of analysis can be the human user."* For more complex AI systems, *"...design initially brings more value by mapping the functional system, its goals, and its causal relationships. The methods move towards systems design, and the target of analysis is the socio-technical system."* The challenge of envisioning how humans and AI systems collaborate and the division of work between these two different types of intelligence created a richer space for designers to draw on their creative skills.

Recognizing an AI opportunity required three things. First, designers needed an internalized understanding of AI's capabilities, and they needed to notice the availability of data required for a specific capability. Second, they needed to conceptualize how the idea would lead to a co-creation of value between the user and the service. Third, their ideas needed to be viable, meaning that any AI feature requiring an additional cost would be assessed in terms of its value generation against the cost of development: *"It's always come from this joint realization of, this is doable within the budget constraints and access to data we have, but also this feature that we're asking the user to do is really boring and repetitive without it." (P5).*

When asked which actions helped them recognize AI opportunities, designers spoke about observing users, creating scenarios, holding co-design workshops, and wireframing. While conducting research, they would notice user behavior patterns and repetitive tasks. P9 shared that they spotted opportunities *"... whenever I start looking at different user types ... repetitive tasks, processes, and routines."* They mentioned recognizing opportunities during ideation, when they sketched and generated scenarios. This happened both when working alone as well as in more structured group activities such as co-design workshops. They shared that wireframing offered one of the best moments for discovery: *"...when you put the input area on an interface... you could say there's an autocomplete here, or a suggestion box. That's the exact moment for me." (P4)* Below I detail the practices and approaches of participants. While some of these are specific to working with AI, some are applicable to more general design and data work (e.g. data-driven design).

AI Capability, Data, and the Data Pipeline

When starting a new project, designers invested significant time to understand the AI system described in the design brief. They worked closely with data scientists, software engineers, and AI engineers to understand how the proposed AI system would work and the data it would require. P5 described this as, *"trying to understand the technicalness of what's going on with an AI solution ... what data is there that we can use as a material. Almost like treating data as you would a dropdown or other design material. What does the system know that we can then leverage for the UX?" (P5)* They established a shared understanding of the AI system.

Through a process of gaining an understanding of the intended AI system across several projects, designers developed a deeper understanding of AI's overall capabilities. This helped them learn to recognize new opportunities: *"The guise of us [designers] coming up to speed on what's going on in the project – as a consequence of that, we start to see those opportunities, saying, we know this already, so we can use that to drive some other feature." (P5)* Designers used several techniques to engage their technical collaborators in understanding a system including diagramming, mapping logic flows, and data visualizations. These functioned as boundary objects between the design and data science expertise.

> *A lot of our job as a design team on this project was sitting with data scientists and making them draw on a white board how it worked over and over again. ... I need to draw out a flow box diagram, or logic flows, and sit down and explain it with people until we get a shared understanding or mental model of what [the AI system] does. (P5)*
>
> *We generated loads and loads of R plots to see what the outputs might look like, how you might classify those outputs. And then we could go back to the data scientists with these plots on the wall, and start asking them to annotate, label, and explain to us. (P8)*

Designers largely worked on the first versions of a new system, making access to user logs impossible. Data was frequently unavailable or difficult to access. As a workaround, they sometimes worked with a scheme describing the structure of data that might be available or they worked with training data being used to prototype the AI system. Even without ready access to data, designers invested great effort in exploring data and worker benefit as part of the system concept. *"When I'm doing data exploration, I look for hooks and what those hooks can mean in terms of functionality or interactions that you have in your user interface (UI). . . . for example if there's longitude or latitude, it suggests a map or some mapping functionality." (P8)* They frequently asked, *"What is the action you want to take from that dataset or what is the insight it's telling us?"* They frequently pushed clients and other stakeholders to gain access to data resources. They noted that the design and analytics teams often worked together to figure out if more data is needed: *"Say, there are six things we need to stitch together to find an answer. We know the three of those. We're still trying to get the other three attributes of the data." (P4)*

In addition to the data and its structure, designers stressed the importance of understanding the data pipeline: *"It's not just understanding what data we have available, but it's understanding which systems the data sits on, and whether data can be transferred across systems to be used together. So really the piping, can we pipe this data out of this system into this other system in order to achieve a particular goal?" (P10)* They often made system maps to learn the overall data flow between the front end and back end. Diagramming and mapping helped reveal design opportunities beyond the interface (P5, P9, P10). While designers shared that combining datasets across sources may reveal new design possibilities, they raised several concerns around privacy and ethics (P1, P3, P9, P10, P11, P12). Participants were skeptical of the use of data and AI for adaptation or personalization in enterprise applications. They noted that features requiring user models may enable employers to infer worker's productivity, and could be instead designed through customization without the use of data or AI. While our focus was on envisionment, our discussions surfaced tensions around system boundaries – how much a system knows about its users versus how much it *actually needs* to know.

Tools, Methods, and Resources

Designers talked about their use of design tools and methods: *"Going from UX to service, it's a lot of the same [design] tools but you're expanding your reach." (P10)"* They used a combination of UX methods (e.g. interaction flows, personas, scenarios, user journeys), and service and systems design methods (e.g., service blueprints, systems mapping, causal loop diagrams). To address the challenges of working

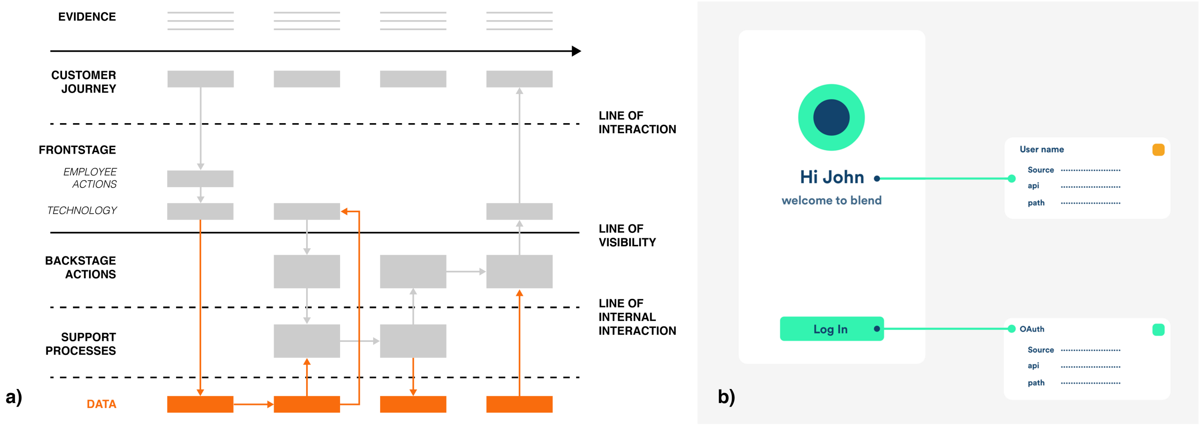


FIGURE 4.2: Augmented tools, such as (a) service blueprints with a data swim lane, and (b) annotated wireframes supported designers in understanding and communicating the role of data within their design.

with AI and AI's need for data, they augmented service blueprints, adding data as a distinct swim lane (Figure 4.2a). They spoke about this as a way of visualizing the data pipeline. They annotated the data swim lane to describe the role data played, and they annotated wireframes to indicate the data source for specific UI features (Figure 4.2b).

In addition to augmenting service blueprints, they also tweaked the service model canvas, creating what they referred to as a "data-driven service design canvas" (Figure 4.3a). They frequently used this tool to support ideation and team alignment around data needs. Based on the canvas, they created a set of logic statements using the structure, *"if this, then that."* These statements aided ideation, exploration, and scenario construction: *"We give people post-its where they put [if, and, then] clauses together with actions, so 'if nothing was rejected on the last delivery, then repeat shopping list'." (P9)* The canvas explicitly prompted designers to think about the AI's value proposition and the data requirements through questions such as *"how will this service help to make people's lives better?"*, *"when is the service triggered?"* and *"what data is needed at each point?"*. This exercise helped them build sophisticated data-driven services.

Drawing from data science's use of the terms insight, action, and outcome, designers created an exercise meant to capture the connections between these elements [92]. Data science team members were asked to complete cards that displayed the prompts: *"I want to know [Insight] so that I can [Action] to enable [Outcome]"* (Figure 4.3c). This exercise enabled the team to identify useful features and functions of the AI system: *"There were many different features that [the technical team] could begin to engineer but we were trying to figure out what was important to see, what was important to interact with. ... [This exercise] allowed us to spill out bite sized, finite pieces of ideas to then together formulate something coherent." (P8)*

To help improve their collective understanding of AI capabilities, designers created new resources [41, 40], such as an AI Capability Matrix (Figure 4.3b). This was meant to translate well known AI mechanisms such as natural language processing and computer vision into AI capabilities designers

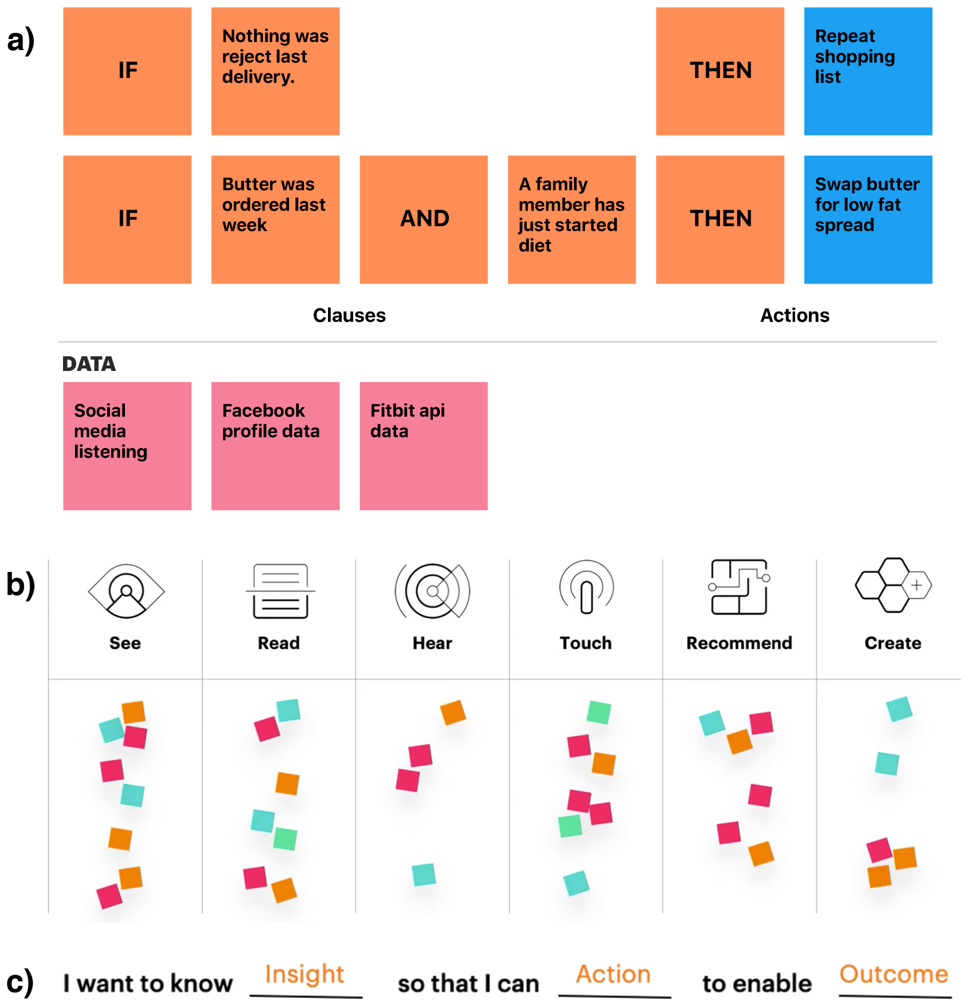


FIGURE 4.3: (a) Data-driven service design and systems design methods scaffolded designers' thinking around the AI system and its dependency on labelled data, (b) the AI Creativity Matrix was used to learn about and ideate AI capabilities, (c) Insights-Actions-Outcomes cards helped breaking down AI features into pieces for formulating ideas.

could envision from. The matrix used action verbs such as *see, read,* and *hear,* which made the capabilities explicit and put them in approachable terms. They explained that the capabilities within the matrix also needed AI exemplars to become actionable to designers: *"One of the key things around our design and AI educational materials was examples. For a lot of people you actually have to show, you need to give them those inspirational examples." (P11)* For instance, when talking about *"seeing"* as a capability, they described a system that used computer vision and natural language processing. This system could *"see"* text on packaging. It then *"read"* the text it found, extracting the ingredients in order to monitor for a conflict with a known set of food allergies. They thought of these capabilities as functions that could be combined, such as *"seeing"* text and then *"reading"* any found text. Using action verbs instead of technical AI terms and mechanisms made ideation workshops more accessible for designers, product managers, clients, and other stakeholders that did not have AI or data science training.

4.2.2 Barriers to AI's Design Innovation

Designers spoke frequently about value delivered by AI innovations using service design language. For example, they talked about the co-creation of value between the company (service provider) and workers (users). The terms *"accelerators"* and *"enhancers"* were used to talk about two major types of value propositions: accelerators speed up the pace or reduce the effort for current work, often fully automating input tasks; enhancers improve the quality of output and the experience of work.

Proposals for a novel use for AI or a new form of human-AI interaction had to be proven as a business case to justify the investment. The value co-creation embodied in designers' concepts needed to easily outweigh the development and operational costs for building and deployment. Value for an accelerator was easier to quantify, estimate, and justify. Enhancers seemed more challenging to justify. Some value propositions designers used included increasing job satisfaction, enhancing decision making, improving the quality of data collection, and capturing organizational knowledge that might have potential future use. Designers shared that experiential value was often impossible to estimate without building experience prototypes: *"It's not just user acceptance testing. If you're actually measuring the [AI system's] impact, you have to simulate the thing that you want to measure." (P11).* They gave an example of an interactive decision support system where they had to build a simulator with a new set of metrics and key performance indicators (KPIs) to assess the value that the AI system might generate [142].

Innovations at the interface level most frequently took the form of accelerators that speeded work. Interestingly, while the value in terms of saved worker time was easier to estimate, designers described these innovations as much harder to pitch as a convincing business case. *"If you start [innovating] from the UI, you need to convince the next levels as the business case provides the constraints." (P11)* Business constraints, including project timelines and budget, played a critical role in determining what would and would not be included in a design. In most cases, the easily estimated value for an interface-level innovation was considered too low compared to possible development costs or risk to tight project timelines.

Designers used a simple heuristic to think about the cost of an AI feature: *(1) The idea is doable with the existing AI model and the dataset, (2) It requires collecting new data, (3) It requires building a new AI model and collecting new data.* Consequently, the AI opportunities that designers searched for would often repurpose existing data or include only small extensions to the currently collected data. If an idea required AI development with additional cost, such as building a new learning model, it would likely move out of the current plan and get added to the future product roadmap. In some cases, the value for an innovation was not considered large enough, while in other cases, the value proved difficult to estimate. When deciding whether to pitch an idea, designers faced a challenge in that they did not have a good sense of the development effort for their idea in terms of cost or time.

Designers all agreed that estimating software costs was not their job. The opacity of this estimate seemed to discourage them from suggesting AI proposals. They tended to refer to any AI applications that required additional investment as "costly" or "expensive". We discussed several AI applications in existing products and services to delineate what is cheaper, and what is more expensive. For example, building a single model for a system would be cheaper than building a separate model for each user. Similarly, building static models or models that are updated infrequently would be cheaper than building models that required constant data collection and frequent model building. Our conversations surfaced these cost-related properties as key aspects for designers to get a rough estimate of "how expensive" a proposed AI innovation might be. However, little confidence was expressed in designers' ability to make an accurate estimate, with the exception of design leads. The "expensiveness" of different AI capabilities and data collection remained more elusive.

In discussing an innovation's value and the challenge of making a business case, designers frequently described their work as defining an MVP – the minimum viable product. They described the culture of their work as very fast paced with strong budget and time constraints. This demanded a focus on articulating and refining a core set of features that co-created value. They used tools such as the impact-effort matrix [60] to rapidly qualify and assess ideas: *"If we recognize that the feature has a very high value for the user and requires a very low [design and implementation] effort from the team, that's a quick win. If it is something that has some uncertain value to offer the user and it costs a lot, usually you tend to park that under the 'nice to have' or 'let's consider this in phase 2'." (P1)* Only high impact-low effort AI innovation ideas moved to the development phase.

4.2.3 Cross-functional AI Innovation

We asked participants about the best practices to overcome the gaps in collaboration between design and data science. Teams emphasized co-located, informal collaboration in successfully spanning the cross-disciplinary gap. Below, I describe the role of designers in AI innovation teams, and then discuss artifacts and boundary objects that facilitated collaboration.

Role of Designers in AI Innovation

Design practitioners supported AI teams in three ways: 1) designing the human-AI interaction, 2) facilitating alignment, and 3) facilitating collective ideation. The first two activities were part of all projects, while the third seemed to happen less frequently on select projects.

Designing human-AI interaction. A principal task for designers was to design human-AI interaction. Designers typically joined ongoing projects, after the data science team performed analytics and developed a proof of concept algorithm for a particular use case. Joining a team meant that at this stage, designers worked on an AI solution predefined by the clients and shaped by the team: *"There may be a very defined business area where the client already has data and they know what the value is." (P11)*

The design process for designing human-AI interaction did not differ radically from traditional design processes. However, designers spent more time in the early stages of product development to be able to *"frame what it is [the technical teams and the client] want to do." (P9)* Designing human-AI interaction required special attention to usability issues and user acceptance. An essential part of designers' work was situating AI in users' workflow: *"If we're designing an AI solution that's going to take away part of [users'] work or responsibility, we really need to understand how they think about that so that they can trust the outcome, but also it fits better to their workflow." (P5).* Participants highlighted the importance of design research to understand user needs around trust, explainability, interpretability, transparency and acceptance of AI systems.

Designers spoke of "design as communication": *"We usually work with the outputs [of the model]. We innovate by improving flows, interaction concepts or improving the adoption of UI components." (P1)* This often involved presenting the complex output of AI systems to end users in ways that supported their mental models. For example, when designing a contract risk analysis tool, designers worked on representing the level of risk for a given contract. Design, data science, and business teams worked to define confidence score thresholds to rank and surface riskier contracts for users to immediately act on. Designers often thought about error recovery and potential errors users can introduce when generating labels. They frequently asked data scientists *"What would happen if the user did this, is that going to ruin the algorithm?"* Some projects required making speed and accuracy trade-offs, and an understanding of users' willingness to wait and tolerate delays. In such cases, interaction design had a direct impact on algorithm selection (P1, P11).

Facilitating alignment. Similar to research describing designers' role as facilitators [95], participants often spoke of themselves as facilitators on AI teams. They played a key role in alignment between disciplines: *"The designer is someone who has to go between data scientists and software engineering because design speaks visual language and everyone can look at it and point at things." (P5)* In initial phases, designers worked to facilitate stakeholder alignment in terms of setting project goals, requirements, and success metrics. They used knowledge elicitation exercises and boundary objects [detailed in next section] to help technical teams and clients articulate their objectives around target inferences. In later development stages, they held workshops with stakeholders to discuss and prioritize features through scenarios and wireframes: *"That's where we get together all ideas and make the plan for what the product or prototype should be and set those functional requirements." (P4)*

Facilitating collective ideation. Depending on the project, designers occasionally engaged in problem formulation – envisioning and reframing to align on the right design. As described in early stage AI strategy projects: *"We have clients who come in for an open innovation session where they have some key strategic areas they want to go. Part of the goal is helping them ideate about potential applications of AI using design thinking methods. If you get [clients] at that stage, you don't know if they have the data [for an AI application]." (P11)* In these cases, designers facilitated early phase collective ideation and problem formulation. They held co-creation workshops with cross-functional teams and stakeholders, and engaged their clients in idea generation using self-developed AI ideation tools, such as the AI Capability Matrix. Together, the team, the client, and stakeholders mapped out the problem space using concept mapping, and worked to identify models and data types that can drive particular design goals. This approach to envisioning AI strategy seemed like an approach in between matchmaking [16] and traditional user-centered design.

There were also cases where designers were able to broaden the project framing using design thinking and design research to get to the root causes of problems. One example was a project in the public safety domain: *"The primary objectives given by the client would only solve a subset of issues. So how do you then also solve the periphery issues that are associated with making this more effective?" (P3)* They used systems design methods including systems mapping and causal loop diagramming to explore the relationships between technical, cultural and organizational challenges, leading to discovery of emergent user and business value (P3, P4, P11).

Collaboration and Boundary Objects

Co-located work was an intentional organizational decision that made design and data science collaboration easier. Data scientists stated that by working with designers, they became more conscious of their assumptions about end users' understanding of AI system outputs: *"As a data scientist, I bring a lot of assumptions, 'of course [the users] are going to understand this' ... [I realized] that we need to take more time thinking about the outputs with the designers." (P12, P13).* It was noted that collaborating with designers early in the process leads to *"a happier marriage" (P13).* One data scientist shared that in addition to increasing business value, they became aware of the "experiential value" – value of improved user experience: *"[The design] allowed [users] to share their expertise. This was an experience that they really value now in their new role. ... I wouldn't have been sensitized to the potential value on that experience side if it weren't for the designers on the team." (P14).* The data science team also benefited from working with designers in eliciting requirements from domain experts, as design practice has well-established methods and tools for knowledge elicitation from end users (P10, P11, P12).

Designers shared that working with data scientists made them more data aware. They also seemed to be running their ideas by data scientists, using their expertise as a proxy [157] for understanding AI's design possibility space:

> *As a designer, I'm not often sure of what's required to run a particular idea I have. That's sort of the nature of the conversation, which is I draw something. I'll go to someone on the team and say, can I talk to you about this? I'm thinking it's going to be useful, I think it's possible, but I actually don't know. So I need your input. Hopefully they come with a constructive attitude to build on the idea. (P5)*

Designers stated that working with data scientists helped them to identify the root cause of pain points, especially in the research phase. They shared including all team members in design research as a best practice, especially with data scientists: *"A data scientist will want to know something very particular, like 'do [users] use this kind of data?' They get an opportunity to reprobe certain areas that we might not have probed as we don't have the experience or knowledge." (P10)* They noted that this collective approach to design research yields more value (P2, P5, P8, P10).

Close collaboration was not without challenges. Confirming prior literature [113], teams expressed that the lack of shared language was a challenge: *"Even though [the data scientists] are speaking English, it's like they have a different language. What does it mean to have outputs from a knowledge graph in* *practical terms?" (P9)* We discussed artifacts that facilitated communication and collaboration across roles throughout the AI development process. Participants readily identified several boundary objects, including whiteboard sketches and visualizations, annotated wireframes, and service blueprints with data annotations:

> *When that data layer was added [to service blueprints], it made a huge difference in terms of the data scientists being able to talk through the process with the designers. Because it's really important for us where the data feeds in, so we know when we can use it for our analytics and AI. (P12)*
>
> *[Annotated wireframes] was designed to make our conversations with the development team a lot easier, because you're drawing this box, but where does this box pull its data from? (P5)*

These responses provided evidence of the use of boundary objects for scaffolding conversations between design and data science. Boundary objects were used in sketching to facilitate a shared understanding, and in prototyping to detail data dependencies.

4.3 Discussion

Prior research detailed AI's design innovation with existing AI products and services [157]. This work revealed that design expertise can play a more critical role in early phase AI product development by facilitating ideation and problem formulation across a broad set of stakeholders from different disciplines. Below, I reflect on these emergent best practices and discuss how these inform future research directions.

4.3.1 Design Facilitation for Cross-functional AI Innovation

Prior work reports a gap in AI innovation: data science teams envision AI applications that users do not want; designers and domain experts propose AI applications that cannot be built [155]. The design and data science teams in this study were able to span this gap through co-located, informal collaboration. Data scientists helped designers assess the technical feasibility of their ideas. Designers supported data scientists in working with domain experts for knowledge elicitation. Although in select projects, design teams do engage in problem setting and reframing *the right AI thing to design*, activities that are considered designers' forte [169, 47]. Role of designers expanded beyond designing human-AI interactions to ideate on the value AI *could* and *should* bring.

This finding about design facilitation in early phase ideation and problem formulation marks a clear space for further research. What roles can and should design expertise play in cross-functional AI innovation? What types of value and impact could designers bring when they are involved in ideating and selecting the right AI thing to build? Recent research has highlighted the need for broadening stakeholder participation in AI through design to account for inclusiveness or fairness goals [36, 131]. This study showed that design thinking and facilitation may provide a path forward.

4.3.2 Resources for Facilitating Ideation and Problem Formulation

Previous research investigating AI as a design material has shown that designers who worked with a large set of AI capability abstractions and examples were more successful and comfortable at working with AI [157]. Practices of design teams in this study confirmed this observation. Participants had an implicit understanding of AI's capabilities and they frequently envisioned and recognized opportunities where AI could add value. In addition to a general understanding of AI capabilities, designers worked to gain an in-situ understanding of the AI system – what the AI system knows and what it is doing in that particular context with that particular data. Teams who wish to innovate with AI will need to prepare themselves for noticing the availability of a dataset and exploring it; envisioning the data pipeline; and effectively communicating how value co-creation is likely to generate sufficient, measurable impact.

This study provided details on the types of tools, methods, and exercises that helped to envision novel forms and functions of AI. Innovation teams thought of AI capabilities as action verbs (e.g. *read, see, listen*) as opposed to thinking about the technical mechanism (e.g., *neural networks, collaborative filtering*). I suspect that the identified AI capabilities would generalize to many contexts and applications beyond the enterprise. The self-made tools participants developed for internal use, such as the AI Capability Matrix (Figure 4.3b), suggest a need for new tools to help with envisioning. Taxonomies and resources can be created that explicitly document AI capabilities with exemplars to help innovation teams operationalize AI concepts. Notably, the study showed that stakeholders without data science expertise benefit from these resources in participating collective AI ideation.

Innovation teams would also benefit from new knowledge elicitation tools and exercises that support team alignment on AI projects. Participants used diagrams, data visualizations, and service and systems design tools to elicit knowledge from data scientists and domain experts (e.g., subject matter experts, clients) about the AI system, the dataset, and the data pipeline. These tools served as boundary objects between different types of expertise; they helped in discussing the data dependencies, in identifying root causes, and in formulating novel and coherent AI system behaviors.

Chapter 5

Developing A Taxonomy of AI Capabilities

The previous chapter detailed the best practices of cross-functional teams around AI innovation. I highlighted the role design thinking can play in facilitating collective ideation and problem formulation. I also presented practitioner-created resources to help teams identify AI opportunities.

This chapter describes the development process of a taxonomy of AI capabilities, a resource to scaffold early phase ideation.

I took a Research through Design approach, engaging in an expansive design process that spanned four years. I conducted two design experiments where the outcome of each experiment reframed the research goals and questions:

Design Experiment 1 focused on creating a resource that captures AI capabilities and examples. I curated a collection of 40 AI examples used in many commercial products and services across a wide range of domains *(e.g., spam filter, language translation, product review analytics).* Using a bottom-up process, I iteratively analyzed each example and delineated *what AI can do* from *how AI works.* This experiment resulted in a taxonomy of 8 high-level AI capabilities (*detect*, *identify*, *estimate*, *forecast*, *compare*, *discover*, *generate*, and *act*), a collection of 40 AI examples with granular capabilities, and a grammar for describing and extending the taxonomy with new capabilities and examples. The experiment led me to further probe the usefulness and the impact of this resource on ideation.

Design Experiment 2 focused on understanding the usefulness of this resource. *Do capability abstractions and examples improve the ideation of AI concepts? How and when should these be considered in the design process? What should more successful ideation and related outcomes look like?* To explore these questions, I created a set of slides documenting high-level AI capabilities and examples. I asked designers to ideate AI-enabled product features before and after reviewing the slides. The resource helped them to consider more AI capabilities, but designers still generated ideas that would be difficult to build. They mostly focused on situations that require near-perfect model performance. This experiment revealed *AI model performance* as a key consideration, and resulted in a Task Expertise-AI Performance matrix. The mapping of AI examples on the matrix suggested that innovators should search for places where moderate model performance creates value. I also observed that a user-centered approach (identifying pain points prior to considering what AI can do) unintentionally limited the ideation of buildable concepts and the exploration of the problem-opportunity space.

Below, I unpack each experiment, detailing the research goals, the design process, and my reflections on the insights gained.

5.1 Design Experiment 1: Collecting AI Examples

I had three requirements for the AI capability taxonomy:

1. **Capabilities over mechanisms.** I wanted to capture capabilities; *what* AI can do. This is in contrast to most AI literature that focuses on describing mechanisms; *how* AI makes an inference (e.g., deep neural networks, etc).

2. **Useful.** I wanted a *useful collection* that could guide designers and non-data scientists away from envisioning things that cannot be built. I cared less about capturing *everything* AI might possibly do. I wanted this resource to capture what designers could reasonably ask AI to do.

3. **Extensible.** I wanted the resource to be extensible. AI keeps growing and changing, and I wanted to make it easy to add new examples and capabilities to keep up with the advances in technology.

5.1.1 Design Process

This iterative process involved three main activities: collecting examples, drawing out and abstracting capabilities, and critiquing this emerging resource. This process took four years and involved several complete restarts. I kept working on this until achieving what felt like a stable collection of AI examples, detailed low-level capabilities, and links showing how these lead to eight high-level capability abstractions. As part of the process, I met with AI experts to discuss and critique the collection in order to discover gaps and missing capabilities.

One continuous challenge was defining what counts as AI, a point the experts repeatedly raised. Prior research noted an absence of discussion on *"what AI means as it relates to HCI or UX design"* [158]. Unfortunately, there is no agreed upon definition of AI, even within the AI research community. While the term is broadly used, it is also disputed and its meaning remains in flux [135]. In my search for AI examples and capabilities, I chose not to employ any specific definition of AI. Instead, I used "artificial intelligence" as a search term and accepted the search results I got back. I view this as an *operational definition of AI* [158] that collectively comes from people writing about and discussing AI.

Collecting AI Examples

I first generated a small set of AI examples by drawing on my personal experience designing and studying human-AI interaction. The initial set included 15 products and product features across various tech companies *(e.g. Amazon Alexa, Google Translate, etc).* The critique of this initial set raised several criteria that I used for the remainder of the project to improve the selection of examples: granularity, generality, and breadth.

Table 5.1: The collection of 40 AI Examples across 14 domains.

Domain	AI Example
Education	Automated Essay Scoring
	Personalized Lesson Plans
Energy & Infrastructure	Home Energy Optimization
	Predictive Maintenance
Finance	Robotic Invoice Processing
	Stock Trading Recommendations
Governance & Policy	Child Welfare Risk Assessment
	Infectious Disease Forecasting
Healthcare	Drug Discovery
	Medical Imaging Analysis
	Synthetic Health Data Generation
	Smartwatch Workout Detection
Hospitality	Review Analytics
	Smart Pricing
Human Resources	Resume Screening
	HR Chatbot
	Workforce Scheduling
Leisure, Content & Media	Smart Speaker Question Answering
	Media Feed
	Game Player
	Image Style Transfer
	Mobile App Face Filter
	Deepfakes
Manufacturing & Agriculture	Crop Monitoring
	Defect Detection
	Robotic Pick and Place
Marketing & Sales	AR Item Viewer
	Personalized Advertisements
	Web Usage Analytics
Office Productivity & Business Workflow	Text Generation
	Spam Filter
	Language Translation
	Meeting Summarization
Risk Mitigation & Security	Biometric Security
	Fraudulent Transaction Detection
Science	Aerial Wildlife Monitoring
	Weather Prediction
Transportation	Lane Departure Prediction
	Navigation Route Planner
	Autonomous Parking

Granularity. The initial set of examples mixed whole products, like Amazon's Alexa, and product features, like a spam filter found in most email clients. Products proved to be way too complicated. They often involved many unrelated AI capabilities as well as lots of non-AI technology. I refocused on AI-enabled features within products and services *(e.g. email spam filter, smart speaker question answering, fraudulent transaction detection).* For the remainder of the project, when we critiqued the examples, we focused on if the feature felt self-contained and if it matched the level of granularity of the other examples.

Generality. The initial set of examples included AI products specific to a single company and AI features found across companies. For example, Alexa was specific to Amazon while spam filters could be found in many email applications. Given my focus on supporting ideation, I decided to focus on AI features that were not bound to any specific company. The fact that a feature repeatedly showed up across companies and products offered a soft guarantee that it could be financially viable and technically achievable. I felt this quality could increase the ideation of buildable AI concepts.

Breadth. I noticed our initial set of examples almost exclusively contained consumer-facing products and services. The examples were mostly from mobile apps and online services. The collection did not have things like *Defect Detection* used in manufacturing nor any business-to-business services. I realized I needed to broaden the search beyond my personal AI experience to better capture more of the ways AI could co-create value for different stakeholders.

I shifted my search strategy, first focusing on identifying a set of industrial domains. I conducted online searches for industries most impacted by AI and synthesized various lists. These lists came from industry-focused news and media, research articles, and white papers. The synthesis resulted in a list of 14 domains (Table 5.1). Next, I searched for the most common AI applications and features for each domain. From these, I selected two to four examples for each domain. I then searched across all of the examples and eliminated ones that had a large overlap. This process was impacted by critiques to examine granularity and generality, and by the meetings with AI experts to find gaps. The final list included 40 examples related to the 14 domains. For each example, I created a short definition, described how value was co-created between the service and the customer, and I classified the example as being either business-to-business or business-to-consumer.

Extracting Capabilities from AI examples

I conducted a bottom up analysis of the examples, identifying the specific capabilities each required. I searched for explanations, triangulating across various sources including research papers, business and news articles, marketing product descriptions, and API documentation. In deciding what should count as a capability, I made distinctions between the *inference*, and the *reaction* an application has following the inference. For example, email applications classify emails as spam or not spam and then sort them into the inbox and spam folder. I considered the classification of the email as an AI capability. I did not include automatically sorting classified documents, viewing this as disconnected from the AI capability. Similarly, I worked to separate the user interface presentation of AI output (its form) from the capability. For example, I captured that a retail service's recommender compares and ranks all items for sale as an AI capability. However, I viewed the choice to present these in the form of a ranked list of recommendations as a design choice and not as an AI capability.

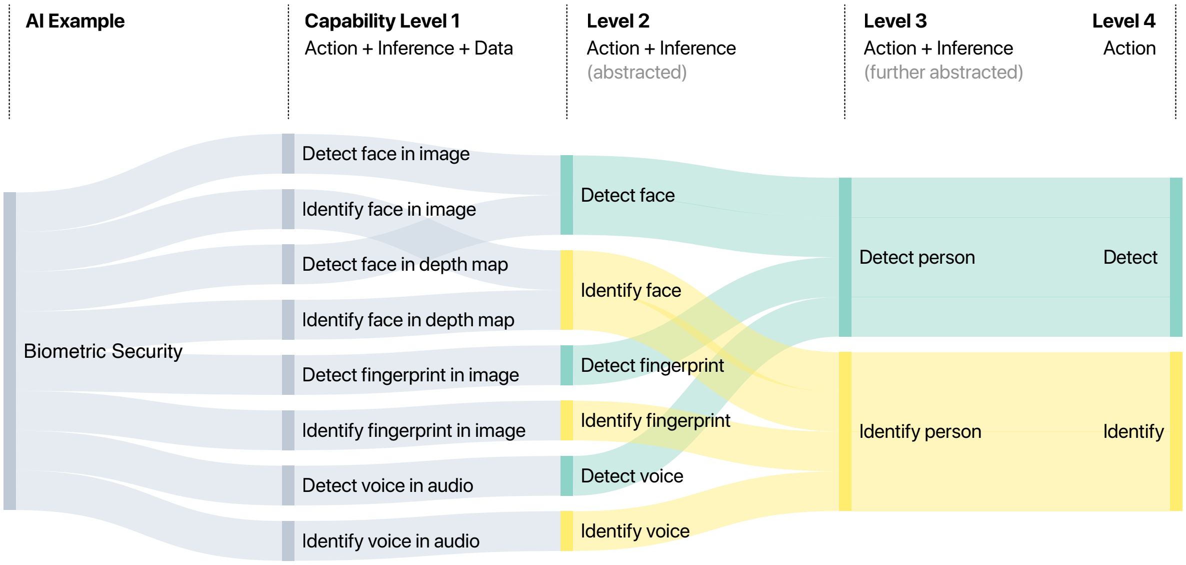


FIGURE 5.1: AI example *Biometric Security* has eight unique capabilities converging into high-level capabilities *Detect* and *Identify*.

I searched for an appropriate form to capture capabilities by writing terse descriptions. As I worked across examples, a simple grammar emerged: *Action + Inference + Data/Entity/Metric*. Each example had several capabilities captured in this terse structure. For example, *Biometric Security* lets users unlock things with their face. The example has the capability to Detect *(action)* + a face *(inference)* + in an image *(data)*. *Detecting* things (e.g., is there a person or an object in this image?) is different from *Identifying* things (e.g., is this Jane's face?). Each individual capability captured a distinct inference or data type (e.g. *face in image, face in depth map, voice in audio*) (Figure 5.1).

I worked on two additional tasks in parallel to the effort to capture a precise set of capabilities: 1) I developed consistent terms for everything labeled as an *Action*, *Inference*, or *Data/Entity/Metric*; 2) I worked to move up to higher levels of abstraction from the terse, detailed description of the capabilities. I tried many different verbs to describe the actions, many different terms to describe the inference, and many terms to describe the data, entity (the subject of an inference), and the metric. For example, the capability *Estimate size of tumor* has an entity (tumor) that is the subject of the inference (size). Through an iterative process, I consolidated these into a non-overlapping set. This resulted in 8 high-level Actions (Level 4) and 17 inference clusters (Level 3).

Inspired by scientific work on taxonomies, I wanted to capture a similar hierarchy for the AI capabilities hoping that this would make them more understandable and useful. I tried various ways of visualizing the connections between the AI examples, the first level of the AI capabilities, and the higher level capabilities, eventually settling on a Sankey diagram. This made it easy to see clusters forming at different levels. For instance, *Identifying a face* (Level 2) is ultimately about *identifying a person* (Level 3). A person can be identified by their face in an image, or by their name in text, or by their voice in audio. All these low-level inferences would abstract to “person” (Figure 5.1).

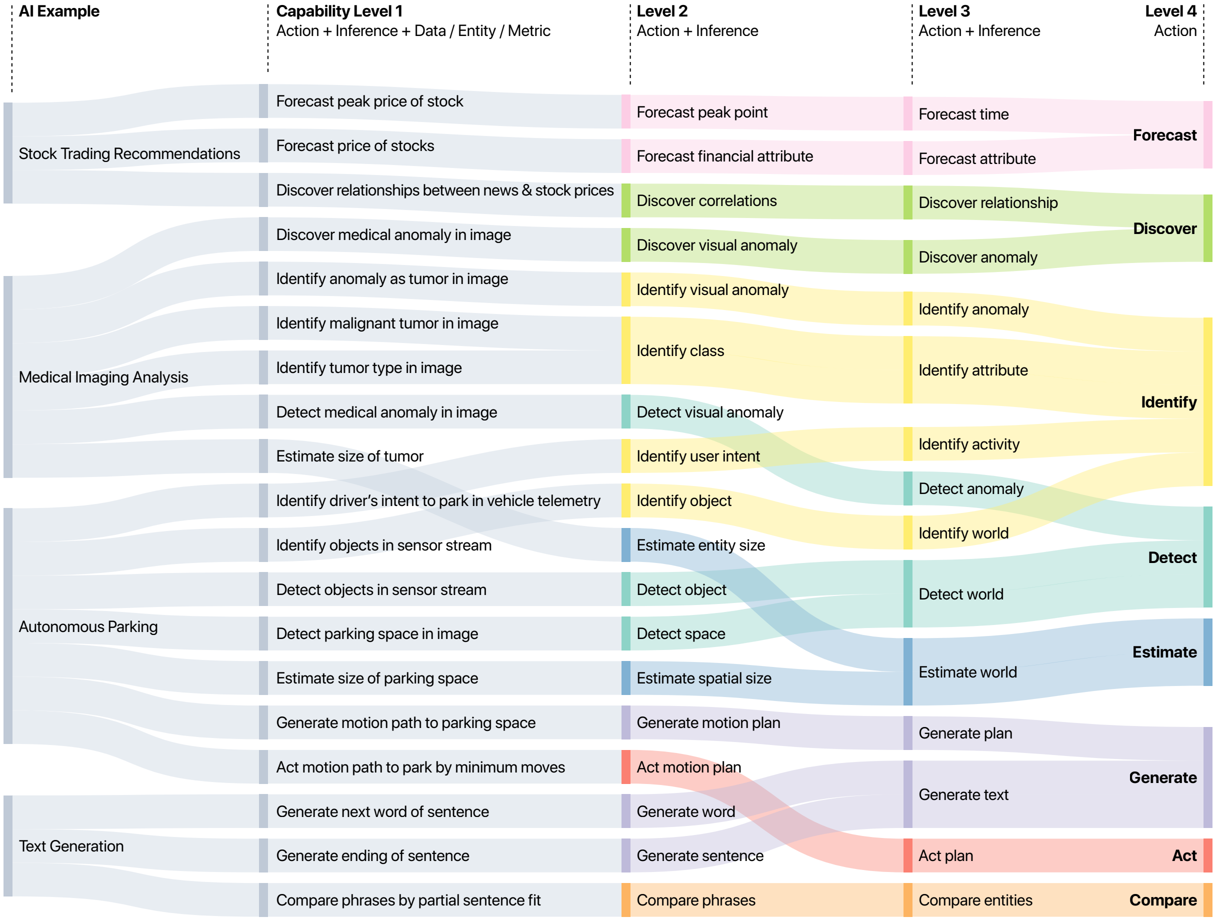


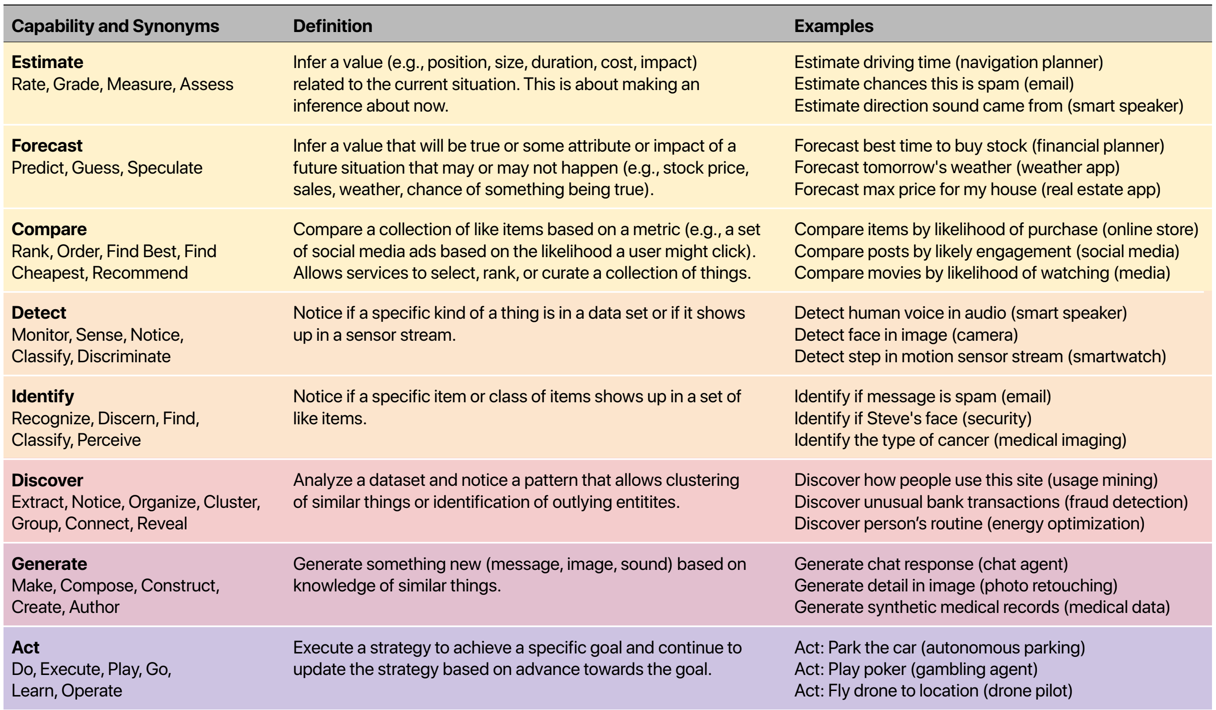
FIGURE 5.2: An excerpt of the examples and four capability levels rendered as a Sankey diagram.

Figure 5.2 provides an excerpt of the AI example-capability relationships rendered as a Sankey diagram. A complete Sankey diagram can be found at aidesignkit.github.io along with definitions for all of the Actions, Inferences, and Data/Entity/Metrics. A description of the eight high-level capabilities can be found in Table 5.2.

5.1.2 Reflection

This design experiment produced several artifacts that collectively provide a resource of AI capabilities and examples. These are captured in the following documents:

1. A detailed list of all AI examples documenting the example description, service type (B2B or B2C), how they co-create value for customers and services,



Capability and Synonyms	Definition	Examples
Estimate Rate, Grade, Measure, Assess	Infer a value (e.g., position, size, duration, cost, impact) related to the current situation. This is about making an inference about now.	Estimate driving time (navigation planner) Estimate chances this is spam (email) Estimate direction sound came from (smart speaker)
Forecast Predict, Guess, Speculate	Infer a value that will be true or some attribute or impact of a future situation that may or may not happen (e.g., stock price, sales, weather, chance of something being true).	Forecast best time to buy stock (financial planner) Forecast tomorrow's weather (weather app) Forecast max price for my house (real estate app)
Compare Rank, Order, Find Best, Find Cheapest, Recommend	Compare a collection of like items based on a metric (e.g., a set of social media ads based on the likelihood a user might click). Allows services to select, rank, or curate a collection of things.	Compare items by likelihood of purchase (online store) Compare posts by likely engagement (social media) Compare movies by likelihood of watching (media)
Detect Monitor, Sense, Notice, Classify, Discriminate	Notice if a specific kind of a thing is in a data set or if it shows up in a sensor stream.	Detect human voice in audio (smart speaker) Detect face in image (camera) Detect step in motion sensor stream (smartwatch)
Identify Recognize, Discern, Find, Classify, Perceive	Notice if a specific item or class of items shows up in a set of like items.	Identify if message is spam (email) Identify if Steve's face (security) Identify the type of cancer (medical imaging)
Discover Extract, Notice, Organize, Cluster, Group, Connect, Reveal	Analyze a dataset and notice a pattern that allows clustering of similar things or identification of outlying entitites.	Discover how people use this site (usage mining) Discover unusual bank transactions (fraud detection) Discover person's routine (energy optimization)
Generate Make, Compose, Construct, Create, Author	Generate something new (message, image, sound) based on knowledge of similar things.	Generate chat response (chat agent) Generate detail in image (photo retouching) Generate synthetic medical records (medical data)
Act Do, Execute, Play, Go, Learn, Operate	Execute a strategy to achieve a specific goal and continue to update the strategy based on advance towards the goal.	Act: Park the car (autonomous parking) Act: Play poker (gambling agent) Act: Fly drone to location (drone pilot)

TABLE 5.2: Eight high-level AI capabilities with synonyms, definitions, and examples.

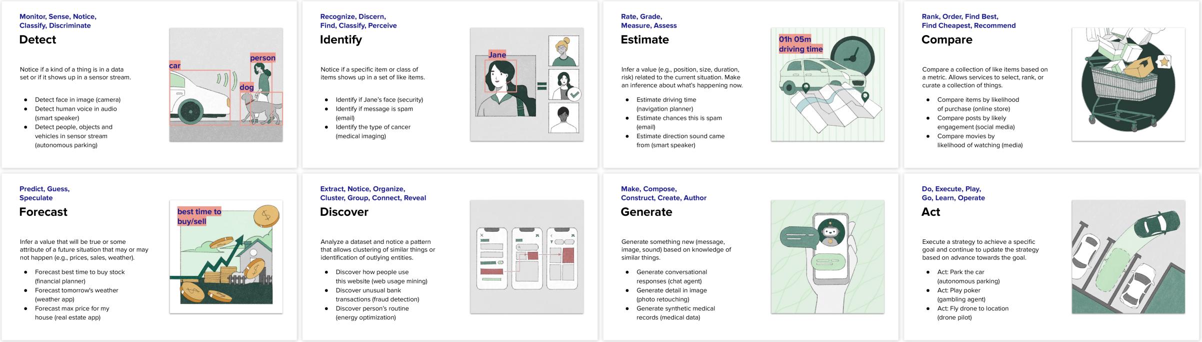
2. Detailed definitions of all the *Actions*, *Inferences*, *Data types*, *Entities*, and *Metrics*. These definitions make it easier to add new AI examples and to recognize when new examples will require the creation of new AI capabilities,

3. A Sankey diagram that visualizes how the examples connect to specific AI capabilities and how the capabilities abstract across four levels.

4. A Github repository hosting the resource files along with a dedicated project website[1].

I viewed this resource of examples and capabilities as a stable initial version that offered *good enough* coverage of what AI can reasonably be expected to do. I felt the collective content could aid non-data scientists in understanding AI capabilities, in ideating new product/service concepts that leverage AI capabilities, and in collaborating with data scientists. Developing this resource surfaced many new research questions. *What communicative forms might make these examples and capabilities useful and actionable in support of ideation? How could we integrate this resource in a design process? How does access to this resource impact ideation? How can we assess the impact on ideation?* These reflections led me to a new design experiment to explore these challenges.

[1] aidesignkit.github.io.



Monitor, Sense, Notice, Classify, Discriminate

Detect

Notice if a kind of a thing is in a data set or if it shows up in a sensor stream.

- Detect face in image (camera)
- Detect human voice in audio (smart speaker)
- Detect people, objects and vehicles in sensor stream (autonomous parking)

Recognize, Discern, Find, Classify, Perceive

Identify

Notice if a specific item or class of items shows up in a set of like items.

- Identify if Jane's face (security)
- Identify if message is spam (email)
- Identify the type of cancer (medical imaging)

Rate, Grade, Measure, Assess

Estimate

Infer a value (e.g., position, size, duration, risk) related to the current situation. Make an inference about what's happening now.

- Estimate driving time (navigation planner)
- Estimate chances this is spam (email)
- Estimate direction sound came from (smart speaker)

Rank, Order, Find Best, Find Cheapest, Recommend

Compare

Compare a collection of like items based on a metric. Allows services to select, rank, or curate a collection of things.

- Compare items by likelihood of purchase (online store)
- Compare posts by likely engagement (social media)
- Compare movies by likelihood of watching (media)

Predict, Guess, Speculate

Forecast

Infer a value that will be true or some attribute of a future situation that may or may not happen (e.g., prices, sales, weather).

- Forecast best time to buy stock (financial planner)
- Forecast tomorrow's weather (weather app)
- Forecast max price for my house (real estate app)

Extract, Notice, Organize, Cluster, Group, Connect, Reveal

Discover

Analyze a dataset and notice a pattern that allows clustering of similar things or identification of outlying entities.

- Discover how people use this website (web usage mining)
- Discover unusual bank transactions (fraud detection)
- Discover person's routine (energy optimization)

Make, Compose, Construct, Create, Author

Generate

Generate something new (message, image, sound) based on knowledge of similar things.

- Generate conversational responses (chat agent)
- Generate detail in image (photo retouching)
- Generate synthetic medical records (medical data)

Do, Execute, Play, Go, Learn, Operate

Act

Execute a strategy to achieve a specific goal and continue to update the strategy based on advance towards the goal.

- Act: Park the car (autonomous parking)
- Act: Play poker (gambling agent)
- Act: Fly drone to location (drone pilot)

FIGURE 5.3: Slides meant to communicate AI capabilities and examples to help designers ideate.

5.2 Design Experiment 2: Making the Taxonomy Useful

I wanted to explore if this collection of AI examples and capabilities might help designers envision concepts that are buildable and valuable. The design experiment included three main activities:

1. **New Forms.** I developed new forms to make the resource useful for ideation.

2. **Assessing Impact.** I conducted an informal assessment to gain insight on how access to the resource transformed ideation. This forced me to consider how to assess the quality of the concepts created during ideation.

3. **Reflection.** I reflected on why ideation with the support of this resource did not change ideation in the way I expected: it did not produce more buildable ideas. This reframed the problem, and it offered insights on what makes some concepts easier or more difficult to build.

5.2.1 Exploring communicative forms

The taxonomy of AI capabilities and examples provided a hierarchical, extensible structure. However, the collection of artifacts making up this resource seemed too abstract and overwhelming to effectively sensitize designers to AI capabilities. To jump start the process of exploring different forms, I first created a one-page table (Table 5.2). This functioned as a sort of cheat sheet for thinking about new forms. The table lists the eight high-level capabilities along with synonyms commonly used. They are organized in subgroups, using color to visually group similar capabilities. For example, *Detect* and *Identify* both address how AI can classify things. Next, the table holds a brief, high-level definition that describes the types of inferences a capability might make. Finally, it holds a small set of examples, illustrating common forms this capability takes in current products and services.

Next, I sketched various communicative forms. I explored making a deck of capability cards, an interactive website where visitors could explore the connections between capabilities and examples, mood boards, high-level capability posters, and slides. Based on recent research that documented practitioner-created AI design resources in the form of playbooks and slide decks [163], I decided to focus on slides. The set of slides included capability definitions (Table 5.2) and each high-level capability as a slide within a 10-page slide deck (see Figure 5.3).

5.2.2 Assessing the Impact on Ideation

With my research team, I discussed what it meant to improve ideation, and different ways of measuring the impact of the AI capability slides. I focused on the general idea of envisioning "better" AI concepts. As the discussions progressed, three specific criteria emerged:

- **Breadth.** Researchers noted that designers learning to work with AI often had a very limited range of ideas. Many seemed to consider only familiar applications, such as chatbots or recommenders [157]. More effective ideation produces a diverse range of alternatives and solutions [21, 44]. I wanted to assess if access to the slides helped designers envision concepts that drew upon more of the capabilities.

- **Effort.** Designers tend to envision AI concepts that cannot be built, and they fail to notice situations where simple, low-risk inferences co-create value for customers and service providers [42, 158]. I wanted to assess how much effort would be needed to create the envisioned AI concepts. While the AI capability taxonomy did not capture any information about development effort, I felt that the choice to limit examples to things that had been commercially viable could guide designers towards more buildable ideas.

- **Impact.** One of the main reasons AI initiatives fail is that they do not generate enough value for the service provider; they do not generate more revenue than it costs to develop and deploy [146, 74, 46]. Similarly, AI initiatives also fail when they do not generate enough value for users, and users do not accept and use the technology as intended [146, 163]. I wanted to assess the impact of an envisioned concept. How much value might it co-create?

I designed a within-subjects study to assess the impact of the slides on ideation. I asked designers to first ideate solutions to a design challenge without the slides. Next, they ideated solutions to a different design challenge with the slides. I chose within-subjects over between-subjects for two reasons. First, prior literature evaluating design resources with between-subjects studies noted that it is challenging to control the variance between experiment and control groups [43]. Second, I wanted designers to compare and reflect on their experiences after brainstorming. One limitation of the within-subjects approach is the session order: I could not switch the order of the conditions. Once designers had seen the slides with the capabilities, they would not be able to forget this when ideating without the slides.

I created two similar design briefs: designing AI-enabled interactions for a ride hailing service and for a vacation rental service. I chose to focus on designing for predefined services – as opposed to imagining new service concepts from scratch – as it more closely resembled the majority of day-to-day design practice. I selected the services based on people's familiarity with them as users. I did not want to select a service that would require additional domain expertise, such as healthcare. I created a single slide for each brief that detailed the available data that could drive the potential AI-enabled features. It also listed a set of pain points, something that typically drives a human-centered design process.

I conducted a literature review to gain insights into the needs and pain points. I looked at the needs of drivers and riders (ride sharing) as well as the needs of hosts and travelers (vacation rental). I prepared personas and user journey maps for each design brief, detailing current experience (e.g., *before, during, and after a ride*). I created a Figma workspace, displaying the design brief, persona, and the user journey as well as sticky notes for ideation.

Before running a full study with professional designers, I first conducted a pilot with 10 HCI students. With my research team, I conducted 2-hour ideation sessions consisting of a brief study introduction, two consecutive ideation sessions, and a post study interview. Participants were asked to generate as many ideas as they could for each phase of the user journey (20 min), and select and refine five concepts (10 min). Next, a member of the research team introduced the slides and the next brief for participants to ideate using slides. After this session, we interviewed participants about their experience, probing on whether they felt the slides impacted their ideation.

I analyzed the interviews, observation notes, and the AI concepts pilot participants generated using affinity diagramming. I specifically looked at breadth and quality (impact and effort). To assess breadth, I compared the capabilities in the concepts during the first and second sessions. To assess quality, I created impact-effort matrices [60], a standard prioritization tool commonly used in innovation [161]. I looked at the five concepts delivered at the end of each session, and rated how difficult they would be to make (effort) and how much value they might co-create (impact). I paid attention to the availability and reliability of data, and how easy it would be to produce good enough inferences. I considered the relevance and usefulness of the AI for the user, then worked to map the concepts on the matrix.

5.2.3 Pilot Findings

Access to the slides seemed to increase the breadth of AI capabilities incorporated into the concepts participants generated. Their interviews echoed this finding. Almost all shared that the slides helped them come up with a larger variety of ideas. Several participants shared that the structure (i.e. *action + inference*) helped them to both generate and communicate concepts. Most found the detailed capabilities (Level 1) the most useful.

Surprisingly, I saw no real difference in the quality of concepts between the two sessions. Almost none of the concepts were easily buildable. The impact-effort matrices showed mostly high effort-low impact ideas: things that are difficult or impossible to build with unclear value co-creation. Interestingly, participants who had the most experience with AI were more able to ideate low effort-medium impact concepts for both sessions.

I noticed that most concepts were created without an awareness of whether the data needed was available. Concepts also generally focused on difficult problems, situations where AI would not likely perform well, and near-perfect performance was needed for an AI system to be valuable. Interestingly, two participants shared that the examples in the slides sensitized them to consider situations where AI would still be useful with moderate model performance. They noted that AI could make things faster with moderate performance and still create value. I found this observation interesting.

I observed that a human-centered design approach — the inclusion of the design brief, persona, and journey map — seemed to conflict with effective AI ideation. Most participants gave their greatest attention to user needs and spent less time considering what AI can do and do well. For example, several participants came up with the idea of predicting rider or traveler reliability (i.e., whether they will cancel) based on historical data. This pain point captured in the user journey would not be easily addressed with an AI prediction. It has too much uncertainty. The human-centered materials seemed to push participants to think of AI as magic and to ignore the value it might generate for users that was not specifically documented in the materials. Similar to recent literature that reported tensions between user-centered design (UCD) and AI development process [163, 167], some participants reflected that the ideation process felt different compared to UCD as they had to consider both AI capabilities and human needs.

5.2.4 Reflection

This design experiment exemplifies Krogh et al.'s claim that RtD is often about *drifting with intention*, the idea that experiments often challenge and change research questions more than they answer them [87]. On the surface, the pilot study failed. I did not get designers to generate more buildable AI concepts. However, it revealed the importance of *AI model performance* and the tensions between UCD and AI ideation as two unarticulated challenges that we need to overcome. Prior work investigating the challenge of engaging with AI as a design material noted that designers struggle to understand AI capabilities [42], and that they seem to focus on situations where there is both great uncertainty around a capability and great complexity in the output of an AI system [158]. This design experiment managed to get a bit more below the surface of this problematic situation.

AI Model Performance

The discussions within my research team about the pilot study led us to an interesting metaphor used by Google for training product teams on how to search for AI use cases [86]. Their internal course asks teams to think of "AI as an island of drunk people". AI can do things quickly and handle an inhuman quantity of information, because there are a lot of people. But drunk people can make mistakes, so teams should not expect a lot of intelligence. This motivated me to go back and re-examine the examples in the taxonomy.

I noticed that many AI examples did not have excellent model performance, but they were still valuable to users and service providers. For instance, *Smart Speaker Question Answering* captured that AI can detect human speech and convert the speech into words. But it did not capture that the generated transcript has errors. Automatic speech recognition has typically about 90-95% accuracy [10], so around one word per sentence will be incorrect. However, this was good enough to find an answer the user wanted from a corpus of pre-written answers [102]. Applications such as voicemail transcripts or

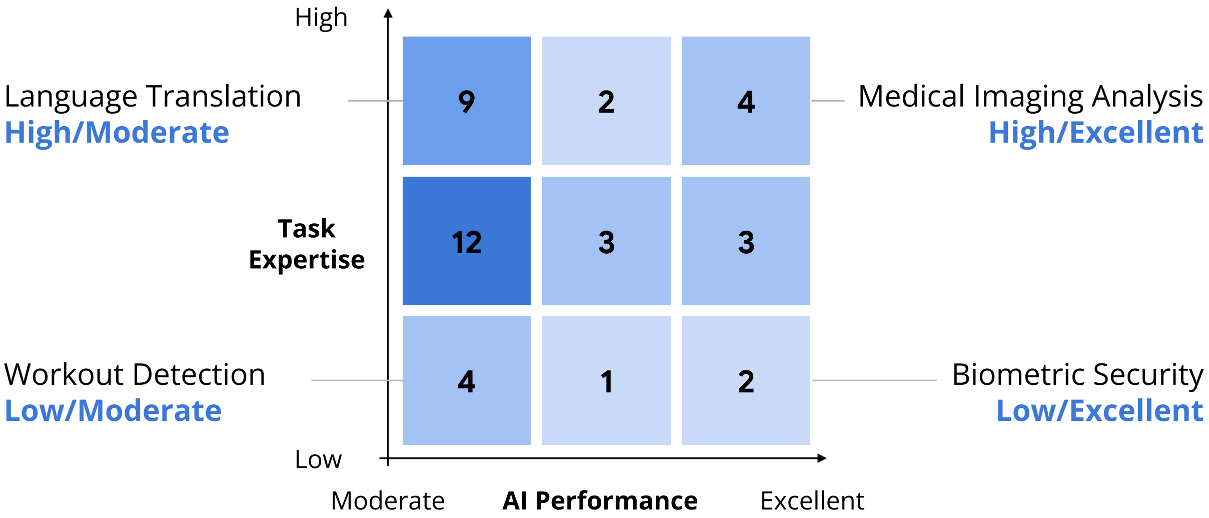


FIGURE 5.4: The Task Expertise-AI Performance matrix analysis of 40 AI examples.

video captions provided other examples where moderate model performance is good enough. These are situations where there is currently no person performing the task, so a moderate quality transcript is better than no transcript. I realized that our resource did not capture how well the AI system needs to perform for co-creation of value.

I revisited each AI example. I decided that in addition to capturing the model performance, the human expertise required for the task also needs to be captured. Through discussion, my team and I broke each of these dimensions into three bins. For model performance, we chose to categorize examples as excellent (e.g., above 99% accuracy), good (e.g., 90-99% accuracy), or moderate (e.g., below 90% accuracy). In creating these bins, my focus was not on capturing the maximum quality an AI system might produce, nor on the technical assessment of performance using certain metrics (e.g., precision, recall, F1 scores, etc). Instead, I captured performance from a UX perspective to understand "the minimum quality needed for users to experience AI as useful" [102]: *What is the minimum amount of accuracy or performance needed for this to be acceptable?* Similarly, I captured how much expertise each task would require for people to perform. Based on *the drunk island metaphor*, I ignored issues of speed and scale. I assessed whether the task required more expertise than a typical adult (e.g., *diagnosing cancer*); expertise of a typical adult (e.g., *parking a car*); or less expertise than a typical adult, meaning a child could complete the task (e.g., *recognizing the exercise someone was doing*). I added task expertise and model performance to the description of each example.

To gain new insight on the taxonomy, I developed the Task Expertise-AI Performance matrix (Figure 5.4), viewing this as AI's opportunity space. *When ideating, do designers come up with ideas that cover the entire space, or do they largely focus on envisioning things that are difficult tasks and need near-perfect model performance?* The vertical axis represents the level of expertise, not counting issues of speed and scale. The horizontal axis represents how well the AI system must perform in order to co-create value. The upper left region holds AI applications such as *Language Translation*. These are tasks that require people to have high expertise, and moderate quality output has proven useful (while highly context dependent, often better than nothing). The upper right holds examples such as detecting cancer in a medical image. This requires a highly trained professional, and the performance must be excellent for AI systems to be useful. The lower right holds examples such as Biometric Security. This is fairly easy for people (match a person's face to their driver's license photo), and the model performance must be high for things like unlocking someone's phone. The lower left holds examples like smartwatch step counters. A child could count someone's steps (if they could maintain their attention). The quality only needs to be good enough to compare days. It does not need to be accurate to the individual step.

When I mapped all forty AI examples as a heat map, it revealed that only a few examples were in the upper right corner (Expertise-High/Performance-Excellent). Most examples (25 out of 40) appeared on the left side (Performance-Moderate). This suggested we needed a new approach to brainstorming, one that encourages people to envision situations where moderate model performance creates value.

Tension between AI Ideation and User-Centered Design

Reflecting on the struggles participants had in ideating AI concepts, I realized that user-centered design brainstorming does not work well when the solution must utilize AI. Needs uncovered in user research most often point to issues where AI will not help. In addition, this approach does not privilege what AI can do or what it does well. I considered matchmaking [16], a technology-centered innovation approach that starts with a technical capability and systematically searches for the best customer across many domains. However, work on AI innovation almost always focuses on a single domain, as the dataset that is available points to a specific set of users and contextual issues. This pre-selection of users and contexts seemed to conflict with matchmaking.

I realized that a new innovation approach was needed, one that blends user-centered design and matchmaking. My prior research revealed that this hybrid process blending user-centric and tech-centric innovation is already emergent in industry best practices [163, 161]. I began to rethink the role of design in this innovation process. I drew insight from research showing communication breakdowns between data scientists, domain experts and product managers [113, 104]. Instead of asking designers to envision AI concepts in isolation, I considered designers as experts in ideation who could "facilitate ideation between data scientists and domain experts" [161]. I wondered if priming teams with examples of AI capabilities where moderate performance creates value would lead to better concepts, low-risk yet high-value opportunities.

5.3 Discussion

The AI capability taxonomy provided a resource that captured what AI can do in non-technical terms and made it explicit through example use cases. To make this resource readily available, I put together these artifacts (i.e., eight high-level capabilities with definitions and examples, slides detailing each example, and assessment matrices) as **the AI Brainstorming Kit**[2] – a workshop kit that can be integrated into early phase AI development.

[2] aidesignkit.github.io.

Following the development, I wanted to put this resource into practice to explore its impact on the process of ideation, problem formulation, and assessment of AI concepts. I had three research questions:

1. **Innovation process:** How can ideation blend user-centered design and matchmaking?
2. **Role of HCI:** Can HCI experts effectively scaffold data scientists and domain experts in brainstorming and identifying use cases that broadly cover the problem-opportunity space?
3. **Finding AI use cases:** Does priming ideation with examples of moderate model performance help to generate concepts that are lower-risk in terms of technical feasibility yet still high-value?

Next three chapters present case studies where I joined innovation teams across industry and academia to operationalize this innovation approach and investigate the above questions.

Chapter 6

Case Study 1: Exploring AI Use Cases in Intensive Care

This chapter presents a case study of early phase AI innovation in intensive care. My team and I collaborated with a research team composed of critical care clinicians and data scientists. The team had access to a rich dataset collected across 39 intensive care units (ICUs) from 18 hospitals. We wanted to broadly explore AI's problem-opportunity space in intensive care, especially to identify low-risk, high-value use cases where *moderate AI performance* can create value for critical care practice. We engaged in an iterative design process, moving from ideation to problem formulation, concept assessment and prioritization.

Below, I first describe the design process and methods. I then detail each phase, capturing the challenges and lessons learned on facilitating cross-disciplinary AI concept ideation. Finally, I share my reflections on new practices and processes for effective AI innovation.

6.1 Overview of Design Process

The project team (n=22) included 6 HCI, 6 data science, and 10 healthcare experts. The HCI researchers had backgrounds in interaction design, service design, and data visualization; they brought expertise in human-AI interaction and ideation. The data science members had backgrounds in data analytics, healthcare analytics, and AI research; they brought expertise in AI capabilities and what could be built with the dataset. The healthcare members all had experience in critical care medicine and included 4 attending physicians, 2 fellows, 2 nurses, and 2 non-clinical healthcare experts. They brought expertise in clinician needs.

We engaged in an iterative, reflective design process [21, 169, 116] to explore AI opportunities for the ICU, particularly to search for use cases that leveraged our ICU dataset. We conducted a two-phase study. The first phase focused on brainstorming; we conducted two ideation workshops within our team to generate AI concepts. The second phase focused on problem formulation; we conducted a design workshop to detail a subset of 12 concepts. Each workshop had 15-17 participants involving at least one participant from each role. Below, I provide a brief overview of the ICU dataset our team had access to. I then present each phase in subsequent sections, unpacking the research goals, design activities, and insights gained.

6.1.1 The ICU Dataset

The dataset consisted of two parts: electronic healthcare records (EHR) and staffing metadata. Similar to the publicly available MIMIC dataset [73], the EHR data included patient level variables, such as hospitalization (e.g., age, gender, race, discharge disposition, admission and discharge dates, etc.); diagnosis and procedure codes, comorbidities; medications; clinical events, mechanical ventilation; and others with a total of 15 variables. The staffing metadata included the transformation of patient level variables to anonymously identify the unique care providers across different roles (i.e., physicians, nurses, respiratory therapists) who provided primary care for each patient at a shift-level. The creation of this additional dataset was motivated by prior literature that indicated whether and how long individual care providers had worked together in the same team impacts the quality of care in the ICU [39]. The dataset was collected across 39 ICUs from 18 hospitals on the East Coast of the United States between 2018 and 2020.

6.2 Phase 1: Brainstorming AI Concepts

In the first phase, our goal was to rapidly and broadly explore the problem-solution space to identify clinically relevant and buildable AI concepts that can improve intensive care medicine.

6.2.1 Method

Workshop 1: User-centered approach

The first workshop followed a traditional user-centered approach. We used “how might we” prompts to drive ideation (e.g., *How might we help clinicians in orchestrating a sequence of tasks? How might we support the adoption of evidence-based practice? How might we reduce clinicians’ burden with documentation tasks?*). Inspired by design thinking methods [71], we set our objectives as ‘thinking outside the box’ and ‘deferring judgment’ to let go of thinking about the limits of technology.

We conducted a 2-hour in-person workshop. The workshop agenda included the introduction of goals (10 min), two consecutive ideation sessions with a short break in between (30 min), impact-effort assessment of concepts (30 min), and a short debriefing and reflection (10 min). During the ideation sessions, each team member reviewed the how-might-we prompts to first ideate individually. They next shared concepts within the group to brainstorm collectively. We used large papers, sticky notes, and markers to note down concepts. At the end of the session, we selected a subset of concepts based on the team’s interest, and placed these on a large impact-effort matrix [60] by getting group consensus on whether the concept was relevant and useful to critical care (impact) and if it would be easy or difficult to implement (effort). After the workshop, I mapped the concepts to the task expertise-AI performance matrix to assess the coverage of design space.

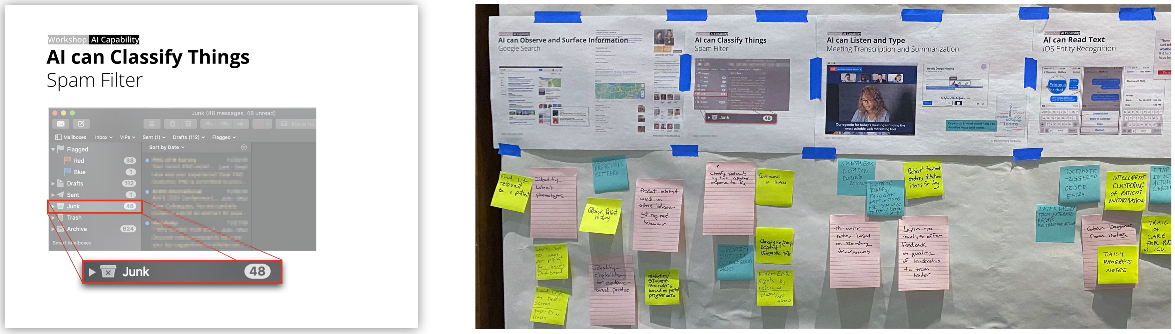


FIGURE 6.1: An AI capability abstraction and example (left), poster printouts to prompt ideation across each capability (right).

Workshop 2: User-centered and tech-centered approach

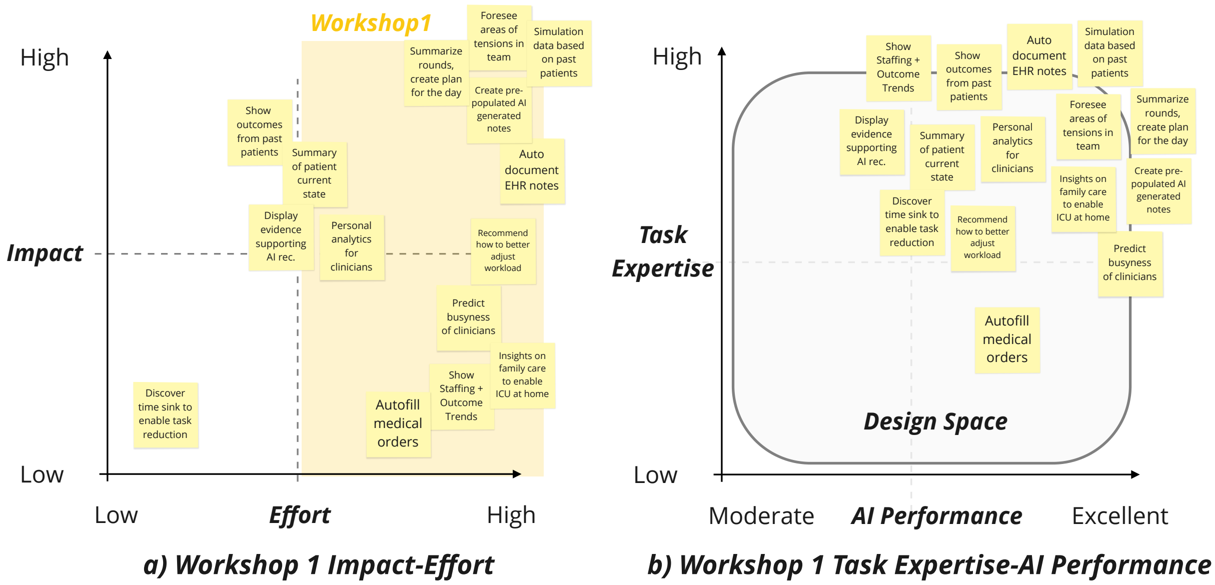
Following the first workshop, I had concerns that our concepts mostly focused on places where near-perfect AI performance was needed for the use cases to be valuable – a pitfall we experienced in the AI Capability taxonomy pilot study. In the second workshop, I decided to bring elements from the matchmaking method [16] to blend user-centered and tech centered innovation approaches. Prior to the workshop, we selected a subset of AI capabilities and examples from the AI Brainstorming Kit [162]. Hoping to move away from envisioning use cases that required high AI accuracy, we mostly selected examples where moderate performance and imperfect AI capabilities produced value.

The capabilities and examples included *observe and surface information* (contextual web search); *classify things* (spam filter); *listen and type* (real-time meeting transcription); *read text* (text message entity recognition); *predict text* (email sentence completion); *cluster similarities* (online shopping recommender system); *discover patterns* (smartwatch activity trends) [see Figure 6.1]. Selection and curation of capabilities were not meant to be exhaustive; similar to industry best practice, our goal was to have a *good enough* subset to inspire ideation.

We conducted a 2-hour in-person workshop following the same structure as in the first workshop. This time, we started by reviewing the AI capabilities and examples we had prepared in the form of slides during the introduction session (10 min). I used the slides as poster printouts to prompt ideation across each specific capability. For instance, talking about “email spam filter” as an example of binary classification (spam or not spam), I probed if we could envision use cases where classifying things as important or not important, or as urgent or not urgent could be useful. Ideation sessions were followed by impact-effort assessment and debriefing, as in the initial workshop.

Data Collection and Analysis

Workshops were audio and video recorded, and transcribed. The analysis included reviewing (1) the transcripts using interpretation sessions, and (2) workshop outcomes using affinity diagramming [78,


Workshop1
High
Impact
Low
Low
Effort
High
Foresee areas of tensions in team
Simulation data based on past patients
Summarize rounds, create plan for the day
Create pre-populated AI generated notes
Show outcomes from past patients
Summary of patient current state
Auto document EHR notes
Display evidence supporting AI rec.
Personal analytics for clinicians
Recommend how to better adjust workload
Predict busyness of clinicians
Insights on family care to enable ICU at home
Show Staffing + Outcome Trends
Autofill medical orders
Discover time sink to enable task reduction
High
Task Expertise
Low
Moderate
AI Performance
Excellent
Show Staffing + Outcome Trends
Show outcomes from past patients
Auto document EHR notes
Simulation data based on past patients
Display evidence supporting AI rec.
Summary of patient current state
Personal analytics for clinicians
Foresee areas of tensions in team
Summarize rounds, create plan for the day
Insights on family care to enable ICU at home
Create pre-populated AI generated notes
Discover time sink to enable task reduction
Recommend how to better adjust workload
Predict busyness of clinicians
Autofill medical orders
Design Space

a) Workshop 1 Impact-Effort

b) Workshop 1 Task Expertise-AI Performance

FIGURE 6.2: Our first workshop resulted in concepts that were technically difficult, some of which were clinically relevant.

96], and the task expertise-model performance matrix. The analysis focused on identifying key themes for the concepts, challenges in collaboration, and the impact of design activities on workshop outputs.

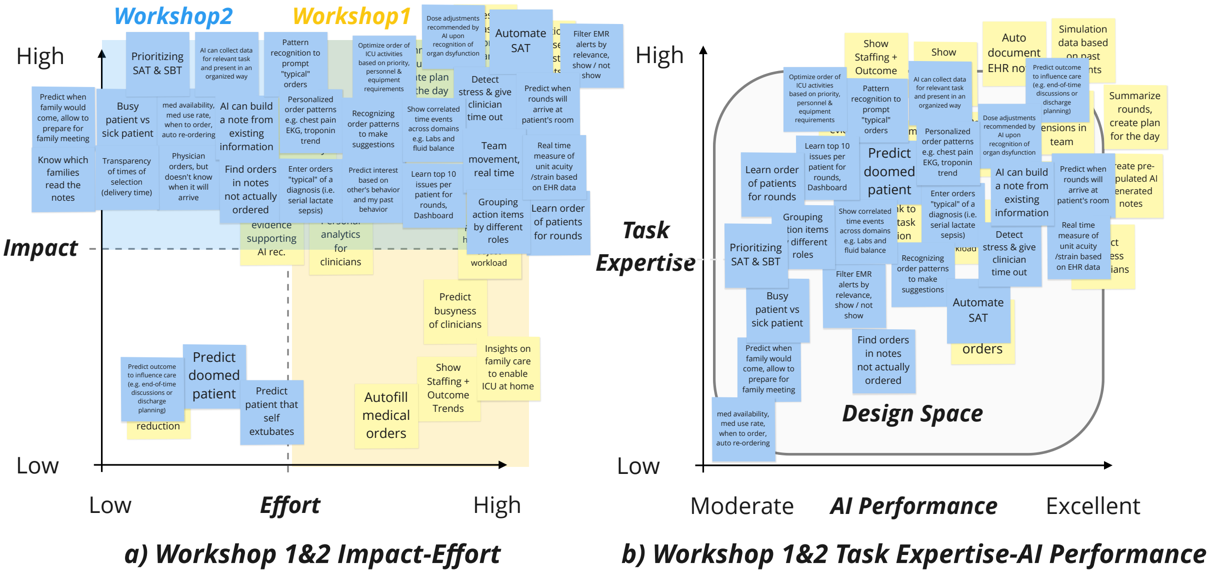
6.2.2 Findings

In this section, I present workshop results by describing (1) outcomes detailing the quality of concepts, and (2) our reflections on what worked, what did not work, and what was unexpected.

Workshop 1 Outcome

The first workshop was effective at getting all members of the team to ideate. However, the outcomes seemed to cover a narrow space. The impact-effort assessment showed that the majority of our concepts were difficult to build, while only about half seemed relevant and useful for critical care medicine (Figure 6.2a). The analysis of high-level brainstorming themes also indicated a lack of breadth: more than a third of concepts focused on clinical decision making, and another third described systems that automated documentation. A few of the concepts described new AI-enabled interactions. One concept described a system that forecasts expert disagreement. For example, it might predict that a nurse would not perform a specific assessment because they viewed the patient as not qualifying while the physician would want the assessment to have been performed. Another described an AI assistant that listens and transcribes conversations between clinicians.

Overall, our team collectively felt that the concepts were not very novel. Most of the concepts addressed existing interactions instead of proposing new ways of working. Concepts often described latent desires around trust, feedback, and explainability (e.g. *AI can take feedback on why it is wrong*); human-AI interaction forms (e.g. checklist, chatbot, recommendation system, conversational assistant);


Workshop2
Workshop1
High
Impact
Low
Low
Effort
High
Prioritizing SAT & SBT
AI can collect data for relevant task and present in an organized way
Pattern recognition to prompt "typical" orders
Optimize order of ICU activities based on priority, personnel & equipment requirements
Dose adjustments recommended by AI upon recognition of organ dsyfunction
Automate SAT
Filter EMR alerts by relevance, show / not show
Predict when family would come, allow to prepare for family meeting
Busy patient vs sick patient
med availability, med use rate, when to order, auto re-ordering
AI can build a note from existing information
Personalized order patterns e.g. chest pain EKG, troponin trend
Recognizing order patterns to make suggestions
Show correlated time events across domains e.g. Labs and fluid balance
Detect stress & give clinician time out
Predict when rounds will arrive at patient's room
Know which families read the notes
Transparency of times of selection (delivery time)
Physician orders, but doesn't know when it will arrive
Find orders in notes not actually ordered
Enter orders "typical" of a diagnosis (i.e. serial lactate sepsis)
Predict interest based on other's behavior and my past behavior
Learn top 10 issues per patient for rounds, Dashboard
Team movement, real time
Real time measure of unit acuity /strain based on EHR data
Grouping action items by different roles
Learn order of patients for rounds
evidence supporting AI rec.
analytics for clinicians
workload
Predict busyness of clinicians
Insights on family care to enable ICU at home
Show Staffing + Outcome Trends
Autofill medical orders
Predict outcome to influence care (e.g. end-of-time discussions or discharge planning)
Predict doomed patient
reduction
Predict patient that self extubates
High
Task Expertise
Low
Moderate
AI Performance
Excellent
Design Space
Show Staffing + Outcome
Show
Auto document EHR no
Simulation data based on past
Summarize rounds, create plan for the day
ensions in team
Predict doomed patient
orders

a) Workshop 1&2 Impact-Effort

b) Workshop 1&2 Task Expertise-AI Performance

FIGURE 6.3: In the second workshop, concepts moved towards (a) low-effort and high-impact; (b) from high expertise-excellent performance to medium expertise-moderate performance.

desired system behaviors (e.g. *recommendation is not intrusive, recommendation comes when ICU team is together*); and pain points (e.g. *placing orders is a burden; I want to eliminate and delegate tasks*).

Similar to the impact-effort assessment results, the task expertise-AI performance analysis showed that most of the concepts mapped to the upper right region (high expertise-excellent performance), missing the larger design space (Figure 6.2b). Concepts often required near-perfect AI performance to be useful. For instance, anticipating clinician disagreement or predicting if a nurse will not perform an assessment can be useful *only* if the AI system can correctly capture 9 cases out of 10. The system would not be useful if it incorrectly flags situations or can only catch cases correctly once in a while. Concepts also seemed too focused on situations with high uncertainty where the task is difficult even for highly trained experts (e.g., clinical decision making, anticipating potential disagreements).

Post-workshop reflection. The first brainstorming workshop was successful in that our team generated many concepts for AI use cases. Data science and healthcare team members found the brainstorming exercise novel, as they had not previously engaged in formal, structured brainstorming or human-centered design perspectives. However, the concepts were not of the quality we wanted. Our process was not generating any concepts that were easy to develop; *low hanging fruit* where moderate AI performance could generate value in the ICU. Some concepts did not require AI, and several called for data that does not exist. Reflecting on the outcomes, we set a new goal to move ideation towards *situations where moderate AI performance could still generate value.*

Workshop 2 Outcome

The second workshop led to concepts that mapped to a broader set of themes. Examples included AI systems that would improve coordination between clinicians (e.g. *generate a schedule for nurses and respiratory therapists for extubation*); systems that improved logistics and resource allocation (e.g. *predict which medications would be needed based on current patients and pre-order from pharmacy*); systems that inferred workload and effort, possibly in support of dynamic staffing (e.g. *classify patients as sick or busy*); systems that better support attention management (e.g. *classify alerts as urgent or not urgent*); systems that improve efficiency, particularly around data entry and documentation (e.g. *predict and recommend orders typical for diagnosis*); systems that anticipate needed information (e.g. *learn relevant information based on patient conditions*).

In addition to these new themes, we generated concepts that built on the themes from the previous workshop, including decision support (e.g. *predict if the patient is eligible for extubation*); documentation (e.g. *generate a draft patient note based on available information*), and automation of menial tasks (e.g. *recommend best billing code based on the patient note*). Table 6.1 lists the high level themes and example concepts from each round of workshops.

Using AI capabilities and examples served as a springboard for our team to recognize situations where a capability could be useful to then effectively transfer that capability to a use case. For example, a nurse practitioner envisioned classifying patients into two groups, sick patients and busy patients. This mirrored the *classify things* capability. Sick patients typically require more attention. Busy patients included patients who needed many time-consuming procedures: *"Is this a busy patient? Or is this a sick patient? It would be useful for managing nursing tasks to tell the difference between a patient who's incredibly sick, but doesn't have a lot of tasks. … [versus] if they have a lot of weeping wounds or something like that, that can make for a very busy patient." (Nurse 2, H8)* This concept hinted at the potential for more dynamic staffing or could be used to balance work difficulty and staff expertise across an ICU. Another capability, *observing and surfacing information*, spurred the concept of learning what EHR screens and information clinicians looked at based on patient condition in order to prefetch or highlight relevant patient history information on a dashboard.

In impact-effort assessment, our concepts moved towards the upper left quadrant: we were able to identify concepts that required low implementation effort with potentially high-impact (Figure 6.3a). The task expertise-model performance assessment also revealed that the concepts moved from high expertise-excellent performance to medium expertise-moderate performance (Figure 6.3b). For example, generating an ordered list of patients for rounds based on the uncertainty of what to do seemed relatively low-risk. A moderate quality, draft triage list is still better than no prioritization; the clinical team will still attend to all the patients in the ICU. Interestingly, in expanding the solution space towards situations where moderate AI performance could be useful, we moved beyond high-stakes situations with great uncertainty (e.g., clinical decision making) and produced concepts for relatively underexplored places (e.g., coordination, managing workload, anticipatory information retrieval).

Phase	Theme	Idea
W1	Decision support	Show outcomes from recent past patients
	Documentation	AI assistant that listens to clinician conversations
	Information retrieval	Summary of patient current state
	Patient-centric care	Insights on family care to enable ICU at home
	Personal informatics	Fitbit for clinicians: how am I doing?
	Team dynamics	AI recommendation system foresees areas of tension
	Workload management	Recommend how to better adjust workload
W2	Automation	AI suggests best billing code based on the patient note
	Coordination	Generate a schedule for nurses and respiratory therapists for extubation
	Decision support	Classify potential discharges based on vitals and most recent progress note
	Documentation	Recognize discrepancy in notes, i.e. doc A says X, doc B says not X
	Eligibility for EBP	Predict if patient is eligible for extubation
	Information retrieval	Learn what clinicians look at based on condition, prefetch to dashboard
	Patient-centric care	Predict when family would come, allow to prepare for family meeting
	Personal informatics	Listen to rounds, offer feedback on quality of leadership to team leader
	Reducing errors	Find orders in notes that are actually not ordered
	Resource planning	Predict what meds would be needed, pre-order from pharmacy
	Task acceleration	Predict and recommend orders typical for diagnosis
	Workload management	Classify patient as a busy patient or a sick patient

TABLE 6.1: High level themes and example concepts from first and second ideation workshops.

Post-study Reflection

Discussing specific AI capabilities and examples prior to the workshop seemed to have a significant impact on the outcomes of ideation, yielding a broader design space where a mediocre, imperfect AI model would still provide enough value for clinicians. Explicitly talking about AI capabilities also provided our team with a shared language. Unlike the first round, most sticky notes described interaction concepts starting with capability verbs (e.g. *detect, recognize, classify, notice, predict, generate...*). Using this language, clinicians probed data scientists about technical possibilities. *"Can AI notice the sequence of orders? ... Can AI cluster tasks?"* Ideation became a collective conversation to discuss what would be doable, how that would produce value for users, and whether any relevant data was captured.

Although the quality of the concepts improved, we still encountered challenges. First, while the concepts were grounded in what's technically possible, only a few of them were implementable using our specific ICU dataset. Most concepts required additional data collection or instrumentation (e.g. tracking clinician clicks in UI to learn and pre-fetch information to dashboards). In some cases, the data existed but it was not in our dataset (e.g. unstructured text from clinical notes), rendering our concepts infeasible unless we sought out more and different data. Overall, the ideation exercise was valuable for informing data collection for future implementations, but we were ignoring the value of our ICU dataset in our ideation.

Our team had a tendency to attribute familiar interaction forms, such as alerts, to specific capabilities and concepts based on past experiences. For instance, while we liked the concept of classifying patients, we always seemed to imagine this as an alert or a reminder. Given the well-known research on alert fatigue and clinician burnout [28], this seemed problematic. Our fixation on existing forms bound to a capability posed a threat to ideation, as the team would dismiss concepts based on known problems with the familiar forms. As prior research reported [155], we found ourselves trying to separate the inference (e.g., predicting that a patient would need a scan) from the interaction (e.g., recommending the action to a clinician or proactively ordering a scan).

Relatedly, rapid ideation resulted in surface level concepts that require further exploration. For instance, clinicians liked the concept of having a ranked list of patients to visit during rounding. However, the criteria needed to prioritize patients was not clearly defined: should it be based on sickness level (see sickest patients first) or patient uncertainty (patients where it was least obvious what to do)? In order to more effectively assess the concepts and select candidates for development, we needed more detail on what the concept was and how it might work in terms of data requirements and the form of the AI output clinicians would encounter.

6.3 Phase 2: Problem Formulation

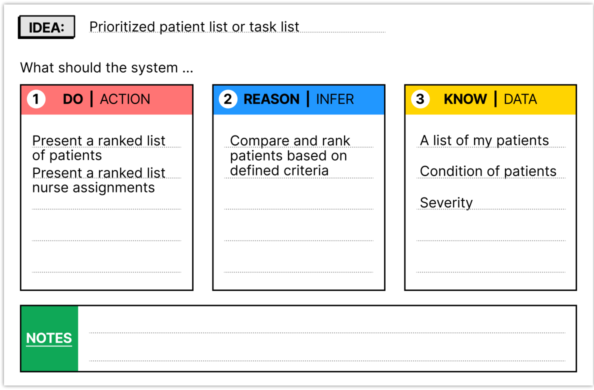
As we moved from ideation to problem formulation, we set three goals. First, we wanted to leverage the unique properties of our dataset, and ground our concepts in what we could realistically build. Second, we wanted to separate interaction form and AI inference when discussing concepts. Third, we wanted to deeply explore some of the concepts to have more mature conversations on their feasibility, desirability, and potential implications.

6.3.1 Method

We conducted a 2-hour in-person workshop for detailing a subset of 12 concepts. Concepts were selected based on data availability, breadth, and match with our team's research interests and expertise. These included: anticipatory pre-ordering of medications; predicting medication time-to-delivery; generating a prioritized list of nurse assignments; identifying sick or busy patients; forecasting unit acuity; generating an ordered list of patients to see for rounds; predicting the eligibility of patients for extubation from mechanical ventilators; generating a coordinated schedule for extubation; identifying clinician workload patterns; identifying bias in clinical orders; predicting typical orders for diagnoses; and discovering the sequence of tasks.

In preparation for the workshop, I created a new abstract representation: *the Do-Reason-Know worksheet* (Figure 6.4). Each section respectively captures the interaction (do), model reasoning and inference (reason), and data requirements (know). I pre-populated the worksheets for each concept based on what was discussed in prior workshops.

The workshop kicked off with a short review of the worksheet and the 12 concepts we pre-selected (15 min). Then, we divided into two groups, where each group collectively discussed and detailed 6 concepts (90 min). We used worksheet printouts as a starting point and detailed each section by adding sticky notes. For instance, when deliberating on *predicting whether a patient might need a certain procedure (e.g. surgery, intubation)*, we discussed if the time of a procedure is documented and whether there were relevant actions or events we could use as proxies (e.g. bleeding prior to surgery). We concluded with a brief reflection and discussion on the next steps (10 min).



IDEA: Prioritized patient list or task list

What should the system ...

1 DO | ACTION	2 REASON | INFER	3 KNOW | DATA
Present a ranked list of patients Present a ranked list nurse assignments	Compare and rank patients based on defined criteria	A list of my patients Condition of patients Severity

NOTES

FIGURE 6.4: The Do-Reason-Know worksheet enabled us to detail each concept in terms of model reasoning, data, and interaction form.

Data Collection and Analysis

I audio and video recorded and transcribed the workshop. I documented the worksheet printouts, and analyzed the transcripts and artifacts using the same methods as in Phase 1 (see section 6.2.1).

6.3.2 Findings

Workshop Outcome

One of our goals was to focus on low-risk, medium-value concepts. Throughout the workshop, we reworked our concepts in a way that reduced the required model performance to help us identify relatively simple, low-risk AI concepts. We repeatedly asked *"Is there a simpler, dumber version of this concept that is still 'good enough' to produce value?"* Below, I share two concepts to illustrate how this approach helped us effectively formulate concepts.

Predicting if a mechanically ventilated patient is eligible to receive a breathing trial, instead of predicting if the patient should be extubated. Liberation from mechanical ventilation is a complex process that requires coordinated actions of nurses, respiratory therapists, and physicians. It involves two integrated actions. Typically, the nurse assigned to a specific patient will perform a *Spontaneous Awakening Trial (SAT);* they will cut off a patient's sedation and observe if they can tolerate being awake. Next, the respiratory therapist, who is typically in charge of making changes to the ventilator settings, will perform a *Spontaneous Breathing Trial (SBT).* They will suspend breathing support and observe how well patients take over their own breathing. These assessments allow the team to decide if a patient can be extubated (liberated from a ventilator).

Remaining on a ventilator is associated with several adverse outcomes including delirium, pneumonia, and heart damage; however, extubating the patient and taking them off the ventilator too soon leads to another host of problems [75, 100, 63]. When one of the steps gets missed (SAT and SBT), then the clinical team lacks the information to make a decision about extubation, meaning the patient remains on the ventilator for another day.

Our initial concept around patient extubation focused on *predicting if a patient will successfully extubate* and *discovering the right amount of sedation for a patient on a ventilator*. These are hard problems that need excellent model performance and very high quality healthcare data, which may not exist. During our discussions, clinician team members shared that physicians can become risk averse when extubations fail. They speculated that this might result in patients remaining on a ventilator longer than needed.

With this in mind, we turned our attention to the execution of SAT/SBT procedures instead of the clinical decision making for patient extubation. This led the concept towards *predicting a patient's eligibility to receive SAT/SBT*. This is a comparatively low-risk, moderate-performance, and medium-value concept, as it focuses on an intermediate, safe-to-perform action rather than a critical decision.

Predicting medication availability and anticipatory ordering. One of the promising concepts that emerged from our ideation was predicting what medications would be needed based on the patient conditions in the unit. The concept was inspired by Amazon's anticipatory shipping [132] –an AI capability and example that came up during capability-based ideation workshop– where the AI system would keep track of the inventory and pre-order medications to reduce time and uncertainty.

During problem formulation, clinician team members shared that this would be really useful for custom mixed antibiotics: *"Sometimes you say 'Antibiotics. Now!' and two hours later it still hasn't arrived." (Physician 1, H1)* They noted that delays are more likely to happen in busier wards, which can be deadly [54]. However, clinicians were also cautious as the incorrect predictions might lead to unused medications, and therefore waste.

We broke down this concept into several lower-risk concepts. First, instead of preordering, the predictions could be used only to inform the pharmacists so that they have a sense of what to expect. Second, we could instead predict time-to-medication to provide feedforward to the clinical team when placing orders. Third, a simpler approach could check for antibiotic dosing errors to prevent delays:

> ***Physician 2:*** *"I want this antibiotic for my patient. When the pharmacist finally gets to it, they say, you ordered the wrong dose. Because this patient is this size, this weight and has this renal function. Something smart would be able to figure that out, like smart dosing." (H2)*
>
> ***Data Scientist 1:*** *"That's a lot easier to do. We have that history of conditions, and what was given to [patients], so maybe these kinds of predictions." (DS1)*

Use of the Do-Reason-Know Worksheet

The worksheet helped to scaffold conversations around data dependency, model behavior, and interaction behavior. It allowed us to express concepts in a more refined way as we moved from sticky note concepts to more fleshed out problem formulations. It prompted us to further probe each concept in terms of how it would generate value for clinicians, and which features in our dataset could drive it, if at all possible. For instance, when discussing what *patient priority* means:

> ***Physician 4:*** *It's a two by two table. There are sick people that if you do the things you need to do, they're going to be just fine. And then there's the sick people who are uncertain. I need*

to pay attention to this patient in the next four hours because if I don't, six hours from now, they might be dead. ... [It would be great if] it was clear who those patients were, and you didn't have to take 15 minutes to figure that out. (H4)

HCI Researcher 1: *What information helps you determine which category that patient falls into? (HCD5)*

Physician 4: *I look at what drips they're on, what's their vent settings. You'd be looking at the amount of drip titration, certain kinds of orders, certain kinds of labs, maybe some radiology findings. I think you can observe some of that in the data. (H4)*

HCI Researcher 1: *How accurate do you feel like your rankings are after you spend fifteen minutes? (HCD1)*

Physician 2: *There can be surprises, but I'm relying on my team to give me a better idea. (H4)*

HCI Researcher 4: *Do you think it would be useful? At which point this would be most useful? (HCD4)*

Physician 2: *The idea is to reduce the cognitive load on the physician. That's probably most useful at the beginning of the day, maybe at the end of the day when we switch shifts, handing off to the other person. If there was a tool there, I might check it once or twice throughout the day like, has anything changed? (H2)*

Data Scientist 1: *Presumably in the algorithm, we could do it every four hours. (DS1)*

Describing the concept with this level of detail made it clear this would function as two separate two-class classifiers. Each patient would be classified as *not-sick* or *sick*, and they would be classified as *certain of what to do* or *uncertain of what to do*. Interestingly, as the model capability and reasoning became clear, our discussions moved towards:

1. **Model performance:** How accurate or robust do the predictions need to be?
2. **Point of interaction:** When, where, and how the inference should be delivered to produce value? (e.g. *are predictions available 15 minutes before or the night before?*)
3. **Risk:** What are the consequences of errors? (i.e. *false positives and false negatives*)
4. **Data quality:** Is the training data trustworthy? Is it likely to introduce bias?

Specifically, the worksheet helped with the three challenges we previously encountered. First, it allowed us to collectively define and formulate AI experiences in a way that is grounded in our dataset. Second, it allowed us to free up our concepts from existing forms by separating the interaction, AI capability, and data. Third, it informed our design deliberation and supported a deeper discussion of the concepts before starting model building and prototyping. For example, when discussing the concept *predicting typical orders for diagnoses*, one physician likened this to a personalized contacts list in email clients, where typing upon a contact name would present the most frequent contact at the top. The personalization aspect raised some concerns: would the medication orders be based on an individual clinician's previous orders or based on a group of clinicians' orders? Physicians seemed to prefer a personalized system, which seemed more complex and costly (both in terms of model building and continuous learning). These deliberations helped us weigh cost-value tradeoffs throughout problem formulation.

The problem framing workshop had an additional, unexpected benefit: our discussions helped us reveal existing or potential problems in our dataset. For instance, one of the concepts was around predicting patient eligibility for extubation from a mechanical ventilator to help clinicians plan for extubation. While exploring potential features in our data, we discussed whether we could use Riker scores, a numeric score for documenting the level of a patient's sedation level and consciousness. When discussing this concept, healthcare members shared that Riker score data were not trustworthy. The scores nurses entered into the EHR did not always reflect the actual level of sedation. This problematic data did not impact the quality of care as clinicians looked at the patient before making a decision. They did not make sedation decisions based on what was captured in the EHR. Thus, they never fixed this data entry problem. Interestingly, this issue is neither reported nor speculated in medical literature. Uncovering this insight early on in the process helped our team avoid using data features that clinicians did not trust.

Post-study Reflection

The problem formulation workshop with the focused worksheet activity helped us detail our concepts for further development. The clinical team lead found the workshop series valuable from a portfolio building and de-risking point of view: *'In [healthcare ML research] there is a lot of inertia towards low-risk, low-reward areas that doesn't move the needle in a meaningful way. This exercise is really valuable because people can replicate these methods to identify lower-risk yet high-reward ideas that are worth doing. Every research portfolio should have a mix of those.' (H1)* Reflecting on how the exercise can be improved, some clinicians shared that involving a broader set of stakeholders would be more helpful: *'It might be useful to have in the room like somebody from hospital management, somebody from pharmacy… to help fill in some of the gaps, [as we have] been making some assumptions.' (H2)* Finally, all data science team members expressed that they found the problem formulation workshop the most beneficial. It seemed to help them to gain a deeper understanding of clinical domain knowledge in relation with the data: *"It's great to hear how and where the data is coming from." (Data Scientist, DS2*).

Following this workshop, we further worked on the concept around predicting the eligibility of mechanically ventilated patients to help nurses and respiratory therapists coordinate SAT and SBT. Our process involved concurrent model development and interaction design, starting with a field study at three hospitals to understand the clinical context [165] and designing a new SAT/SBT dashboard to elicit initial feedback from clinicians [164].

6.4 Discussion

This case study provided preliminary evidence that sensitizing data scientists and domain experts to AI capabilities and example applications can scaffold effective early phase ideation. It also surfaced tensions between user-centeric and AI-centric innovation processes. Below, I discuss the implications of this study on developing new, hybrid innovation processes, and resources that support these processes.

6.4.1 Towards a New Innovation Process for Early Phase Ideation

We started our ideation process following a traditional user-centered design approach. However, the team felt tensions between user-first and technology-first thinking. Our team was not able to produce high quality concepts. I suspect that my initial approach – asking clinicians what would be most valuable – has unintentionally led the team to produce technically challenging, high-risk ideas. Additionally, traditional rules of brainstorming, such as letting go of technical limitations, seemed to result in a breakdown in AI innovation, leading us to high uncertainty situations where AI technologies might not be best suited to the problem at hand.

A complementary approach, sensitizing clinicians to AI capabilities, led to a more concrete perspective and enhanced ideation. It enabled clinicians to understand what AI can do, and recognize situations where AI can improve their workflows. Consequently, ideation became a conversation between domain experts, data scientists, and HCI/design experts, probing what is useful and what is doable. Moreover, having clinicians share many situations and possible points of interaction sensitized data scientists and HCI/designers to the lived experience of clinicians and the data. In terms of ideation outputs, this approach moved the team towards more technically achievable, low-risk concepts and towards a broader exploration of the solution space.

I suspect that this unexpected shift towards lower-risk concepts stemmed from the selection of the capabilities and examples: commercially successful human-AI interactions often bring users closer to their goals by accelerating their tasks or enhancing their performance of work, instead accomplishing their goals for them [161]. More research is needed to understand the impact of the selection of examples and capabilities on ideation.

In addition to ideation, assessing our concepts as a group was beneficial in establishing a shared set of values, which are otherwise assumed implicitly. The assessment discussions worked like a *design crit* [21]. It enabled our team to reflect on action [123], to tease out what was and was not working with an idea. Throughout these discussions, we gained new information that informed our next steps for each workshop.

Upon reflection, this case study revealed an emergent design process that blends user-centered and technology-centered innovation processes. I suspect that explicitly talking about AI capabilities, examples, and data source will scaffold ideation between data science and domain experts. I anticipate that generating multiple concepts and assessing them collectively could help AI innovation teams in identifying the right human-AI problem to solve. I suggest that HCI and design innovators on interdisciplinary teams take this hybrid approach and start with a review of AI capabilities in order to surface user needs that best fit what AI can actually and effectively do. Do the challenges we encountered generalize to different contexts within healthcare or other domains? Would these emergent processes and approaches apply to other contexts? Our community would benefit from well-documented case studies and design experiments.

6.4.2 Resources for Scaffolding Ideation

Throughout the ideation process, I used several resources and artifacts: a set of AI capability abstractions and examples; a worksheet capturing the interaction, model reasoning, and data; and the task expertise-AI performance matrix. These resources and artifacts ended up playing a twofold role: they helped our team members who did not have a background in AI (i.e. designers, domain experts) in understanding and engaging with AI; and they exposed our healthcare and data science team to the human-centered thinking and to the challenges faced in critical care medicine.

Moving forward, I see a great opportunity for researchers to develop resources, artifacts, and boundary objects that can scaffold design, HCI, data science, and domain expertise collaboration. Which AI capabilities and examples might effectively scaffold ideation? How should these be curated and presented? What details and dimensions should be captured in artifacts for ideation and assessment? I suspect that AI innovation teams can benefit from a modified prioritization matrix that captures the AI task difficulty, acceptable AI performance, and risk of errors and bias, in addition to typical assessment metrics, such as feasibility, viability, and desirability. Future research should investigate the specific needs of key stakeholders from such artifacts, especially in the early stages of AI development.

Chapter 7

Case Study 2: Exploring Vision-Language Model Use Cases in Radiology

This chapter presents a case study of AI innovation in the context of radiology. I collaborated with a group of AI researchers, radiologists, and clinicians to better understand the design space of Vision-Language Models (VLMs) – multimodal foundation models that combine large language models (LLMs) with vision capabilities. While recent advances showcase impressive new capabilities, such as generating a radiology report from a medical image (e.g., [11, 18, 153, 68, 141]), the clinical utility of these advances remain unclear. *What might be some clinically relevant use cases for VLMs to enhance radiology workflows for radiologists and clinicians?*

With my research team, I set out to explore these questions through an iterative design process. The first phase involved in-depth discussions and brainstorming sessions within our team. We discussed how radiologists interpret images and write reports, and how clinicians review these to make patient care decisions. Using VLM capabilities and examples, we brainstormed and sketched scenarios and wireflows to identify use cases that would be useful and acceptable. In the second phase, we selected four use cases to further design: *Draft Report Generation, Augmented Report Review, Visual Search and Querying*, and *Patient Imaging History Highlights*. In the third phase, we sought feedback from 13 radiologists and clinicians outside of our team to investigate if and how these concepts might be useful in clinical practice.

In the following sections, I present a reflective account of our design process. I describe how we identified high-value use cases that required high-expertise, near-perfect performance (e.g., report generation) as well as medium-expertise, moderate-performance (e.g., patient history highlights). I reflect on how the search for moderate AI performance expanded our search space. Finally, I discuss the implications of our modified innovation approach blending user-centered and technology-centered innovation.

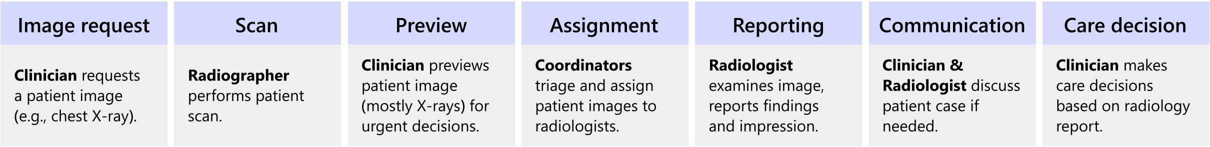


Image request	Scan	Preview	Assignment	Reporting	Communication	Care decision
Clinician requests a patient image (e.g., chest X-ray).	**Radiographer** performs patient scan.	**Clinician** previews patient image (mostly X-rays) for urgent decisions.	**Coordinators** triage and assign patient images to radiologists.	**Radiologist** examines image, reports findings and impression.	**Clinician & Radiologist** discuss patient case if needed.	**Clinician** makes care decisions based on radiology report.

FIGURE 7.1: Overview of the radiology workflow.

7.1 Overview of Radiology Workflows

Radiology workflows unfold across many clinician roles (Figure 7.1). First, *referring clinicians* request an imaging study for a patient (e.g., a chest X-ray). Next, *radiographers* perform patient scans, and *radiology coordinators* may prioritize and assign patient images to radiologists. Next, *radiologists* examine patient images, and document their *findings* – descriptions of normal or abnormal observations, such as lesions or nodules – and their *clinical impression* – a summary that synthesizes the findings and suggest possible causes or further tests. Referring clinicians then review the radiology report, and may consult radiologists for further questions or clarifications before making care decisions.

A radiology report typically consists of a *Background* section that describes the patient information and the clinical question that referring clinicians seek to answer, and *Findings* and *Impression* sections that communicates radiologists' interpretation [76]. Different imaging modalities have different workflows. For instance, plain (2D) imaging, such as X-rays, are high volume and fast-paced, taking minutes to review [33]. Complex (3D) imaging on the other hand, such as CTs and MRIs, take more time (10-20 minutes) and cognitive effort [33]. Reports are often in the form of prose (sometimes called *narrative report*), while there is also research that calls for structured reporting approaches (e.g., short, bullet-point style sentences) for improved clarity [52]. Reports are usually written using voice dictation, often utilizing templates or draft reports produced by radiology trainees (interns or residents in the US context) in hospital settings. A major challenge within the radiology workflow is the sheer volume of scans, leading to a backlog of unreported images [119]. Wait times might be a few days to a week for radiology reports [105].

The majority of human-centered AI research on radiology imaging has focused on mechanisms to explain AI outputs to domain experts [8, 25, 24, 109], such as explaining the diagnostic outputs for specific chest X-ray findings (e.g., cardiomegaly) by highlighting what feature changes in the medical image would lead the AI system to give a different diagnosis [8]. Other work explored AI acceptance or the impact of using AI systems on radiologist diagnostic performance [13, 24, 23]. Relatively little work investigated current radiology workflows or asked radiologists where they needed support [145, 152, 109]. Xie et al.'s work presents a rare example of an early phase needfinding and design study, where they conducted a three-phase design process to explore opportunities for AI-assisted radiology in the context of X-rays [152]. This work builds on this existing body of research by investigating radiologists' and clinicians' current needs and desired futures for VLM-assisted radiology workflows.

Table 7.1: Participants in user feedback sessions. ‘Consultant’ denotes a senior doctor with specialist training (the equivalent title in the US is ‘physician’.) (*) denotes clinical trainees (interns or residents in the US context).

ID	Professional Role	Exp.	AI Familiarity
R1	Emergency Care Radiologist	12yr	Very familiar
R2	Pediatric Radiologist	15yr	Very familiar
R3	Uroradiologist	10yr	Somewhat fam.
R4*	Gastrointestinal Radiologist	4yr	Somewhat fam.
R5	Cardiothoracic Radiologist	10yr	Very familiar
C1	Intensive Care Consultant	10yr	Very familiar
C2*	Intensive Care Fellow	1.5yr	Somewhat fam.
C3	Intensive Care Consultant	8yr	Very familiar
C4	Public Health Physician	11yr	Somewhat fam.
C5	Internal Medicine Consultant	7yr	Somewhat fam.
C6	Cardiothoracic Consultant	20+yr	Not familiar
C7	Consultant Oncologist	20+yr	Very familiar
C8	Pediatrician	19yr	Somewhat fam.

7.2 Method

We had two high-level research questions: **(RQ1)** *What might be the clinically relevant use cases for vision-language model capabilities in radiology?* **(RQ2)** *Whether, how, and in what situations these use cases might provide value for radiologists and/or clinicians?* Our study unfolded in three-phases:

1. **Phase 1: Brainstorming VLM Use Cases** We conducted seven 30-minute in-depth discussions with our clinical team members (four sessions with a cardiothoracic radiologist (R1F); three sessions with a general practitioner clinician (C1F)) to probe pain points in current radiology workflows (e.g., How do radiologists read a medical image?) We also conducted four 1-hour brainstorming sessions involving team members with clinical expertise, AI expertise, and HCI expertise. The sessions also involved reviews of VLM capabilities from recent literature (e.g., [130, 18]). We created sketches, scenarios, and wireflows and ranked concepts based on their clinical relevance, feasibility, and data requirements.

2. **Phase 2: Sketching VLM Concepts** We narrowed our focus to four specific use cases: *Draft Report Generation, Augmented Report Review, Visual Search and Querying*, and *Patient Imaging History Highlights* and translated each use case into design concepts by sketching click-through Figma [49] prototypes. We then populated the prototypes with relevant images and reports from the open source MIMIC-CXR X-ray dataset [73], and validated the plausibility of each design concept with a radiologist team member.

3. **Phase 3: User Feedback Sessions** We recruited 13 clinical stakeholders across eight hospitals in the UK and the US (5 radiologists, 8 clinicians, 12 male, 1 female) who had not been involved in our design process. Participants represented a range of clinical specialties including: intensive

care, emergency care, pediatrics, family medicine, and other domains. Table 7.1 provides an overview of our participants' clinical roles and experience. Following the capture of demographic information, we probed the perceived usefulness of either radiologist- or clinician-facing use cases based on the participant's role.

7.2.1 Data Collection and Analysis

We audio and video recorded and transcribed all brainstorming sessions using video conferencing software. We analyzed the data using a combination of affinity diagramming [96], interpretation sessions [66], and service blueprinting [15]. The feedback sessions lasted 1 hour and was conducted remotely via video conferencing software. We audio and video recorded the sessions. The data was analyzed using affinity diagramming [34] to iteratively generate codes for participant utterances, which were then synthesized into high-level themes related to specific use cases; including concerns and desires for additional support.

7.3 Phase 1: Brainstorming VLM Use Cases

Our discussions and brainstorming sessions surfaced many challenges, ranging from requesting a patient scan to prioritization, reporting, and assessment. Our team generated many ideas for improvement (some of which are discussed in prior literature [124]), such as detecting redundant scan orders; detecting poor quality images at the time of scan to reduce re-scans; and optimizing image triage and assignment based on patient urgency and provider subspeciality. I provide a broad overview of these challenges and opportunities using a customer journey map of the radiology workflow.

In this section, I detail our insights into VLM-specific use cases, mainly around radiology reporting and report review due to our team's research focus. Where relevant, I provide direct quotes from our clinical team members that were involved in in-depth discussions (R1F, C1F) and brainstorming sessions (R2F, C2F) – denoted with *F (formative study)* to distinguish clinical team members from the user feedback study participants.

7.3.1 Use Cases for Draft Report Generation

In considering how VLM capabilities can support radiology image review and reporting, we discussed whether an AI-generated draft report might provide any value. Interestingly, our radiology team members likened these to reports they receive from their trainees: *"I would treat it as a draft report coming from my trainee." (R2F)* As to how much effort was involved in reviewing and editing these reports, R2F shared: *"Junior trainees' reports will require more work. Depending on how good it is, I might dictate from scratch … Senior trainees, I usually look at [their reports] and sign. I'll just say 'I agree'. I'm not going to correct a typo. I might do small edits to say 'there is also this' … If I disagree, I will say "My interpretation is this… " I will dictate if it's a few sentences or type a few words here and there."*

Elaborating on what makes a radiology report 'good', we teased out three aspects: the report is (1) accurate (i.e. findings are correct); (2) complete (i.e., there are no missing findings); and (3) error-free (i.e. report does not have typos). This led us to further probe the value proposition AI might bring into radiology in the form of improved report quality and reduced reporting time. Radiology team members pointed out that they often prioritize speed over quality; they had to work really quickly due to the large number of images waiting to be reported. A team member asked whether AI-generated findings in the form of bullet points would provide any value if radiologists still had to dictate the report by themselves (to reduce the risk of errors). Radiology team members pushed back, noting that the system would not save them time in reporting, thus it would provide little value. They recalled instances where the voice recognition system introduced transcription errors, and stressed that they do not want to spend additional time correcting an AI system's errors: *'[recounting an incorrect transcription of 'abdominal viscera' as 'animal viscera'] It was embarrassing. It should be able to correct these, so that I can sign without having to read what I dictated." (R2F)* These discussions hinted at time savings as a key design requirement for clinician acceptance.

Finally, our conversations brought up the question: Should a draft report be shown to clinicians? R2F reflected that this may lead to tensions in terms of responsibility and radiologist acceptance: *"There is an issue of responsibility. Radiologists might think they're out of the loop" (R2F).* Both clinicians and radiologists proposed that AI-generated findings could be used for triage and early flagging of critical findings without presenting too much detail. This became one of the central themes of exploration in our later study.

7.3.2 Use Cases for Visual Search and Querying

When reviewing visual question-answering capabilities, both clinicians and radiologists brought up that they regularly perform web searches to look for similar images or clinical information relevant to the patient case. These included medical databases and clinical guidelines (e.g., nice.org.uk – The National Institute for Health and Care Excellence guidelines), as well as websites that provide peer-reviewed patient cases (e.g., gpnotebook.com, radiopaedia.org, radiologyassistant.nl, uptodate.com). R2F described two scenarios where searching similar images was helpful. The first case included situations where she would suspect that there is a pattern in the patient image, but cannot be sure what anomaly it might be: *"I know there is a pattern but I don't know what it is."* She would use search queries that described the pattern (e.g., glass opacities CT lung) to find similar images to help with diagnostic assessment. The second case was having diagnostic uncertainty about the suspected pattern: *"I think this is crazy paving, but I haven't seen crazy paving in a while."* She would search for a certain pattern in trusted websites (e.g., *"crazy paving chest ct radiopaedia"*) to see examples of that particular pattern to help disambiguate possible interpretations.

Both radiologist and clinician team members indicated forming search queries with the abnormality and imaging modality to find similar cases with an overview of pathologies listing common causes: *"I'll look at the differential diagnoses [listed] … [which makes me think] I haven't considered that, but knowing what I know about the patient, yeah that makes sense." (R2F)* We discussed how radiologists might perform visual searches if they had the ability to query a region in a patient image, for instance, drawing a bounding box and typing ‘is this normal or abnormal’ (image query, text query, or image and text query). R1F shared that text query might be preferable: *“I would prefer text, because if I’m selecting a lump, anything might look like a lump.”* R2F however preferred the following search query type: *“If I could snip a region ... so that I don’t have to translate that to a text query.”*; suggesting variations in search preferences.

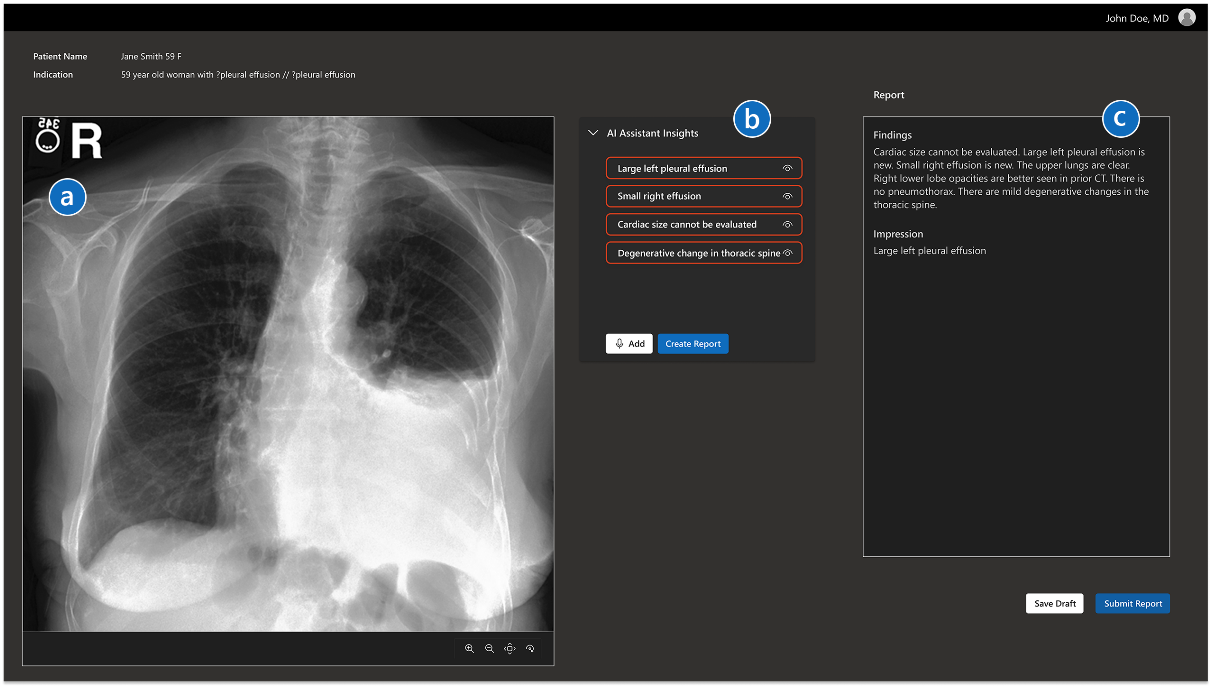
Our discussions also touched on clinician-radiologist interactions, and the types of questions asked. Clinicians shared that they might ask clarifying questions for less visible findings: *“You said in the image [there is this] ... Where is it? Is this normal?” (C2F)* Both radiologists and clinicians noted that image annotation tools were part of the reporting software, yet were rarely used. Clinicians also sought information on next steps: *“Do you think we need to act on this? What [additional] imaging should we order? Who should we call about this?” (C2F)* Radiologist team members shared that such clarification interactions can be overwhelming: *“Sometimes clinicians want to hear from their favorite radiologists that they’ve built a trust relationship over the years, which can be overwhelming for the radiologist.” (R2F)* We discussed that visual annotations and image search capabilities might reduce some of the back and forth.

7.3.3 Use Cases for Longitudinal Imaging

VLM capabilities enable the comparison of a patient’s prior images for longitudinal assessment, a core practice in radiology reporting [3, 127]. Reflecting on situations where this capability could be useful, R2F spoke of the challenge of tracking the size of nodules over time: *“It might look like the size hasn’t changed much [compared to the most recent image], but actually it’s grown 5 millimeters compared to two years ago.”* We envisioned that a system could summarize past images and reports to provide key highlights, such as chronic events, operations, and the trajectory of abnormalities.

7.4 Phase 2: Sketching VLM Concepts

We identified four VLM use cases to further design and investigate: Draft Report Generation *(radiologist only)*, Augmented Report Review *(clinician only)*, Visual Search and Querying *(clinician & radiologist)*, and Patient Imaging History Highlights *(clinician & radiologist)*. In selecting these, we sought to cover use cases where moderate AI performance may still provide utility (e.g., visual search, summarizing prior patient reports) in addition to concepts that are higher-risk and require near-perfect AI performance (e.g., report generation). This section details our design goals and strategies in selecting each VLM use case, and elaborates on their design.



Patient Name Jane Smith 59 F

Indication 59 year old woman with ?pleural effusion // ?pleural effusion

- Large left pleural effusion
- Small right effusion
- Cardiac size cannot be evaluated
- Degenerative change in thoracic spine

Findings

Cardiac size cannot be evaluated. Large left pleural effusion is new. Small right effusion is new. The upper lungs are clear. Right lower lobe opacities are better seen in prior CT. There is no pneumothorax. There are mild degenerative changes in the thoracic spine.

Impression

Large left pleural effusion

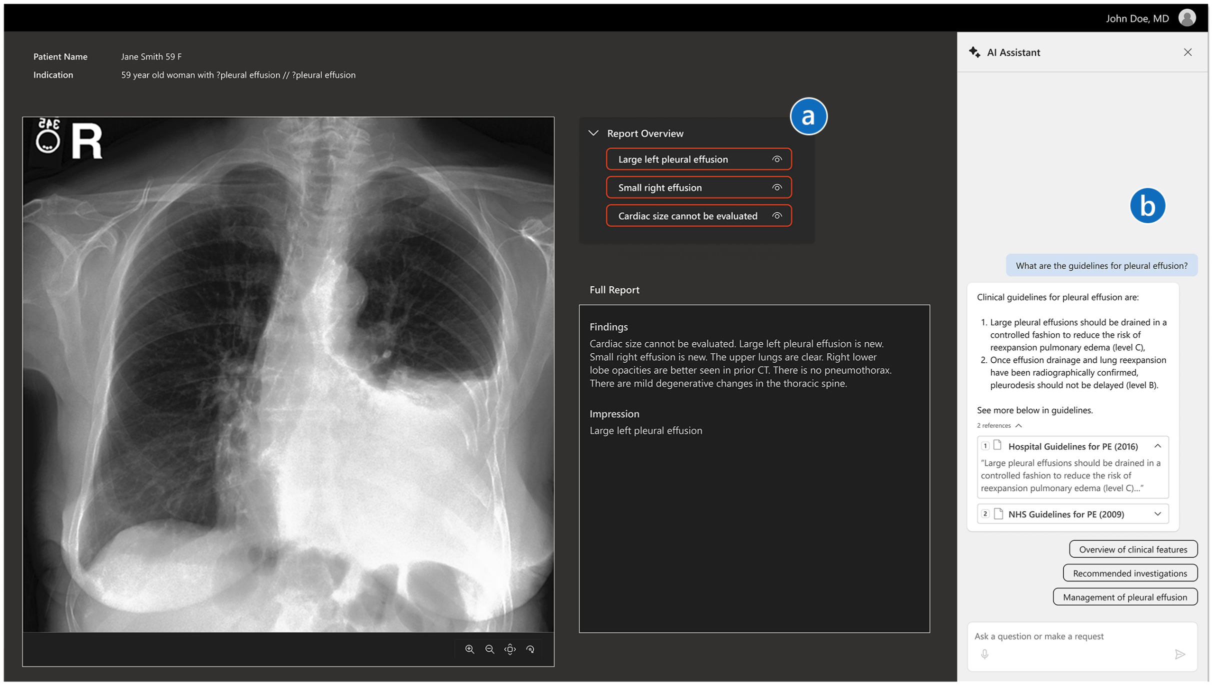
FIGURE 7.2: The Draft Report Generation (radiologist only) concept displayed (a) a chest X-ray image with patient information and clinical indication, (b) an AI-generated report in bullet point form, and (c) a narrative report created using the bullet points.

7.4.1 Design Concepts

Draft Report Generation

Motivated by the insight that radiologists are accustomed to working with draft reports from their trainees, the first concept explored the idea of an AI-generated radiology report as a 'draft'. The Draft Report Generation concept (Figure 7.2) displayed (a) a chest X-ray image with patient information and clinical information, (b) an AI-generated report in short sentence form, and (c) a narrative report created using the short form report. It demonstrated a scenario where the radiologist could review the findings to see annotations in the image, and edit the draft in short form (e.g., crossing out, editing, or adding bullet-point style sentences). The short form text – illustrated as bullet-points – was sought to assist in spotting mistakes and enables linking the outputs to source materials (e.g., referencing to previous scans or reports, localizing text findings in the image). Our goal was to explore the balance between *introducing friction* and slowing down radiologists by having them verify the report, and yet still have them achieve time savings overall.

The concept aimed to explore the following questions: (1) When and how would radiologists want to interact with an AI-generated draft report, if at all? (2) Could there be utility to having a short form report (e.g., bullet points)? (3) Should the draft report be available to clinicians? If so, in what level of detail? (4) What is considered as 'good enough' AI performance for draft reports to be useful?


John Doe, MD
R
a
b
AI Assistant
Report Overview
Large left pleural effusion
Small right effusion
Cardiac size cannot be evaluated
Overview of clinical features
Recommended investigations
Management of pleural effusion
Ask a question or make a request

Patient Name Jane Smith 59 F

Indication 59 year old woman with ?pleural effusion // ?pleural effusion

Full Report

Findings

Cardiac size cannot be evaluated. Large left pleural effusion is new. Small right effusion is new. The upper lungs are clear. Right lower lobe opacities are better seen in prior CT. There is no pneumothorax. There are mild degenerative changes in the thoracic spine.

Impression

Large left pleural effusion

What are the guidelines for pleural effusion?

Clinical guidelines for pleural effusion are:

1. Large pleural effusions should be drained in a controlled fashion to reduce the risk of reexpansion pulmonary edema (level C),
2. Once effusion drainage and lung reexpansion have been radiographically confirmed, pleurodesis should not be delayed (level B).

See more below in guidelines.

2 references

1 Hospital Guidelines for PE (2016)

"Large pleural effusions should be drained in a controlled fashion to reduce the risk of reexpansion pulmonary edema (level C)..."

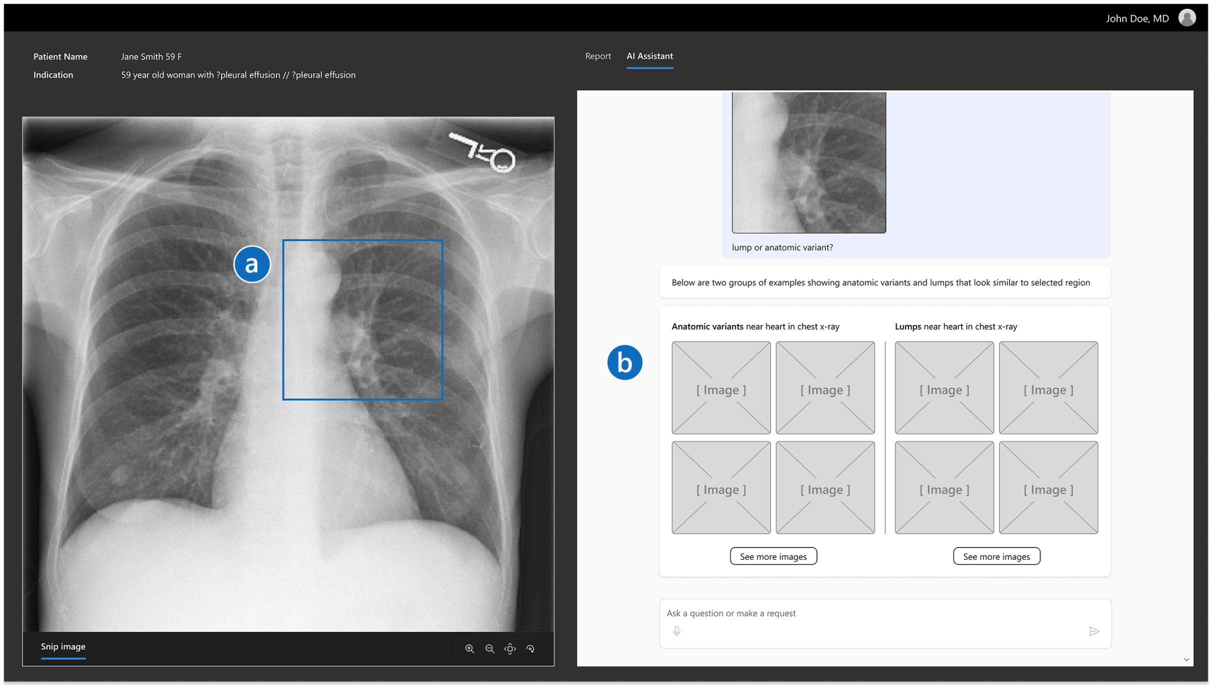
2 NHS Guidelines for PE (2009)

FIGURE 7.3: The Augmented Report Review (clinician only) concept displayed (a) a report overview feature above the full report, and (b) an AI assistant feature.

Augmented Report Review

The Augmented Report Review concept (Figure 7.3) had two main features: (a) a report overview feature shown above the full report, and (b) an AI assistant feature. The report overview displayed a list of abnormal findings extracted from the report that can be visually highlighted in the patient image to facilitate its localization (e.g., *large right pleural effusion*). The AI assistant showcased numerous prompts inspired by clinician questions (e.g., *Given the image-based findings, what are the clinical guidelines for pleural effusion?*). For this concept, a critical design consideration was around latency: vision-language models are currently slow and costly. We speculated whether contextual queries can be pre-run prompts, where answers could be displayed immediately. As an alternative, we also sketched the AI assistant feature as a chatbot with a text input field to provide contrasting options. The prototype displayed example prompts (e.g., guidelines, suggested investigations) as conversation starters to help clinicians envision what might be useful.

The concept aimed to explore: (1) Would clinicians want to review AI-generated annotations? If so, which findings are helpful to highlight for different image modalities (e.g., CT)? (2) Could there be any utility to having contextual information when reviewing images? (3) What would clinicians query? What would they never query? (4) Would there be a need for follow up queries (e.g., a chatbot style interaction that can maintain context)?



Patient Name: Jane Smith 59 F

Indication: 59 year old woman with ?pleural effusion // ?pleural effusion

lump or anatomic variant?

Below are two groups of examples showing anatomic variants and lumps that look similar to selected region

Anatomic variants near heart in chest x-ray

Lumps near heart in chest x-ray

FIGURE 7.4: The Visual Search and Querying concept displayed (a) a visual selection tool that enabled image search or image and text queries, (b) an AI assistant that returned query results without providing an interpretative answer.

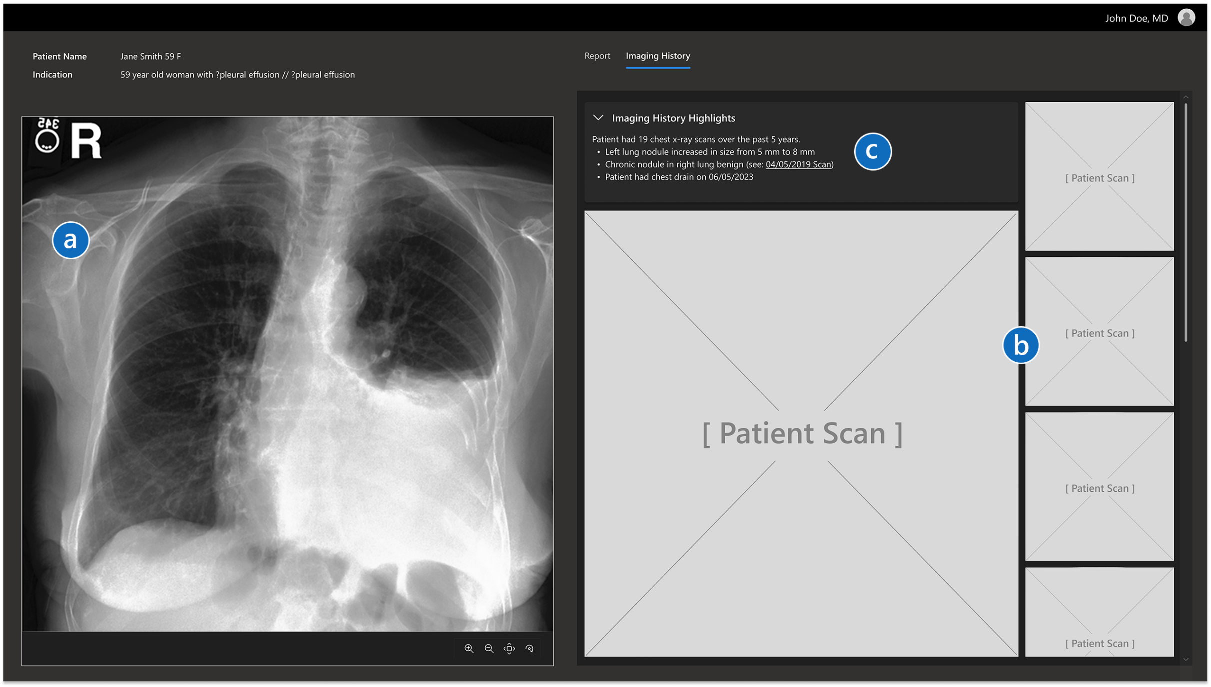
Visual Search and Querying

Building on the insight that radiologists and clinicians perform image searches online, the Visual Search and Querying concept (Figure 7.4) explored potential utility by displaying: (a) a visual selection tool that enabled image search (e.g., *Find similar images that look like this region*) or image and text queries (e.g., *"Is this lump or anatomical variant?"*). In line with recent literature showing clinicians look for evidence rather than explanations [156], we envisioned this concept to return groups of similar images instead of providing an interpretative answer (e.g., *"Below are two groups of examples showing anatomic variants and lumps that look similar to the selected region."*) (Figure 7.4b).

The concept aimed to explore: (1) What would clinicians and radiologists visually query? (2) Could there be utility in performing image *and* text queries? (3) Would clinicians prefer to have an answer along with image examples (e.g., *"Region likely normal"*)? (4) What might be the data requirements for finding similar images (e.g., past images and reports from a hospital database)?

Patient Imaging History Highlights

Given that clinicians and radiologists commonly review patients' prior images, the Patient Imaging History Highlights concept explored extracting and highlighting key insights across a patient's image history. The prototype (Figure 7.5) displayed: (a) a new X-ray scan, (b) prior images, and (c) an AI-generated summary of prior images and/or reports. Example highlights included changes in abnormalities (e.g., *Left lung nodule increased in size from 5 to 8 millimeters*); chronic conditions (e.g., *Chronic*



Patient Name Jane Smith 59 F

Indication 59 year old woman with ?pleural effusion // ?pleural effusion

Imaging History Highlights

Patient had 19 chest x-ray scans over the past 5 years.
- Left lung nodule increased in size from 5 mm to 8 mm
- Chronic nodule in right lung benign (see: 04/05/2019 Scan)
- Patient had chest drain on 06/05/2023

Figure 7.5: The Patient Imaging History Highlights concept displayed (a) a new X-ray scan, (b) prior patient images, and (c) an AI-generated summary of prior images and/or reports.

nodule in right lung benign, see image reference); and operations (e.g., *Patient had chest drain on this date*).

The concept aimed to explore: (1) What is relevant to highlight in a prior imaging summary? (2) Would a summary based only on reports be still useful; what is the least AI can do? (3) Would clinicians query prior images? If so, how (e.g., *"Show me only abdomen CTs"*)? (4) How would clinicians envision prior imaging summary to best be presented?

7.5 Phase 3: Eliciting User Feedback

This section reports participants' feedback on each design concept, capturing perceived benefits and suggestions for improvement.

7.5.1 Draft Report Generation

Expectation of near-perfect AI performance: All radiologists expressed that having an AI-generated draft report would be valuable as long as the model performed really well; with high sensitivity and specificity. Describing how AI reporting errors could add burden, one radiologist explained: *"If it misses something, I've got to say that. If it's false positive, I either have to click to remove it from the report entirely, or I have to change something."* (R2) To better understand what would be considered as good enough AI performance for this use case, we asked *"Out of 10 reports, how many are you willing to correct?"*. Almost all replied *"1 out of 10"* (R1, R2, R3) or *"5 to 10 out of 100"* (R5); suggesting the need for near-perfect performance for AI-generated draft reports to provide real utility. Only one radiologist, a trainee, responded *"3 out of 10"*, noting that the system could make them more confident even if it did not reduce their workload: *"It [would be] getting stuff right enough for me to feel comfortable just to edit the 30% of cases where it's going to be wrong." (R4)*; suggesting potentially added benefits for trainee learning.

Accounting for fast-paced practice & high workload: Echoing our initial findings, radiologists noted that their practice is fast-paced and high volume: *"It is literally going as fast as humanly possible. Scrolling through things, looking at image, saying whatever I can, go over the spellchecks. Make sure I didn't say anything really wrong and then sign and get on the next one. ... I just need to get my job done fast. I don't get paid more for quality."* (R2). Consequently, participants mainly spoke of value as time savings, especially when reading multi-slice images such as those captured by CT that take significantly longer to review and report than i.e. X-rays, and images that are outside of their subspecialty (R1, R2, R3, R5): *"I might be a seasoned reporter for lung or cardiac, but as every week it happens, we'll get a neck CT ... when you're not doing it day in day out, it's extremely difficult. You would love an AI which is at least giving you the salient findings." (R5)* This suggests a draft report may reduce risks of key clinical observations being missed and could assist with image interpretation confidence. Apart from time savings, participants also mentioned potential benefits in reduced cognitive burden. For simpler X-ray images, R2 for example mentioned: *"I can do [X-rays] in 10 seconds... [but] there's the cognitive burden. Having to say the words and go through it all is painful."* R4, who was a trainee, reflected that the main benefit of the system would be reducing reporting time rather than the time spent for image interpretation: *"Regardless of what the system says, I'm still going to go through my same search patterns for the findings and interpreting those ... the only area where it's going to be saving time is in creating that draft [prose] report because then I don't have to worry about the wording and if I've missed something"*.

Preference for short, standardized reporting: Interestingly, when probed whether short form sentences could be useful, all radiologists shared that they prefer to work with bullet point style findings instead of prose text. Several participants highlighted the literature on *structured reporting*, which is proposed as a solution for improving report quality and consistency [52]:

> *"The idea of a narrative report happened in 1898 and we've not moved on from it. It's full of hedging, it's full of weird language that only radiologists use: 'likely to be', 'cannot exclude'. [This is] what we should be moving away from rather than using the technology to reverse engineer the future into what we got." (R3)*

Commenting on how the bullet list findings in the prototype were presented, R1 reflected *"My reporting style is much more telegraphic. So I'll say 'large right pleural effusion', that's exactly how I'd phrase. I wouldn't say 'there is' or 'is seen' or all those kinds of phrases. I don't think [they] are helpful, especially for findings."* Similarly, R3 advocated for structured findings for consistency and objectivity: *"Rather than saying 'suspected mild cardiomegaly', you say 'heart is enlarged' or 'heart enlarged', which is a statement. It may be right or wrong, but it's objective."* All these suggest a preference for concise, accurate and consistent reporting over the historic use of more ambiguous prose text, something that AI reporting could assist in standardizing.

Favoring prioritized findings & confidence indications to assist image interpretation: Additionally, radiologists described the benefits of having findings structured by their clinical relevance and the systems' confidence in the generated outputs. For example, a systems capability to compare a current study to a patient's prior image enables ordering report findings by: what is new, what has changed or is unchanged, which gives important context to aid image interpretation and subsequent clinical action. For example, the sudden 'new' appearance of a pneumothorax would require urgent clinical attention whilst a reduction in consolidation in the patients chest upon pneumonia diagnosis may suggest that antibiotic treatment is working. Furthermore, all participants (R1, R2, R3, R5) suggested having confidence intervals to communicate AI uncertainty: *"Rather than using 'likely to be', 'unlikely to be', 'possibly' ... 'Likely prostate cancer 4 out of 5', [which is] more robust and easier to interpret." (R3)* One radiologist suggested displaying the model confidence and ranking findings on this basis: *"[Say for a finding] I don't totally agree, I don't disagree. But if it's confidence is only like 56%, I'm just going to knock that out." (R2)*

Impressions present key interpretative work: While short form, structured reporting was preferred for findings, some radiologists (R1, R3) shared that having unstructured, prose text is more appropriate for the impression section which is the *"non-objective, doctor bit" (R3)*: *"The main focus of communication between us and the team taking care of the patient is that impression part of the report. So it's really important to me to have that correctly crafted." (R1)* R5 reflected that findings could be useful, yet the impression will be more difficult to get right: *"We get a lot of [outsourced] reports from teleradiology, which just tell you what the findings are. A clinician will want to know the clinical impression. ... Is a report better than no report? I think it is fine if it gets the findings right, even if it doesn't do all the synthesis clinically."* Given the importance of the impression section and its broader interpretative work that may include additional contextual information, the feedback from our participants suggests that clinicians may want to remain in charge of this task; positioning AI's role closer to the extraction of relevant findings from an image rather than its overall clinical interpretation.

Broading uses of (prose) draft reports: When asked how an AI-generated draft report should be presented, all radiologists suggested having both bullet points and prose report presented together whereby bullet points serve to assist the review, and prose for clinical communication: *"I could just get rid of [a bullet point] and it takes it out of the report, that's great. Because editing at that level is so much easier than editing on the report." (R2)* A few radiologists noted that a patient-facing report could also be generated based on the list of findings (R1, R3); suggesting additional use cases and user groups.

In response to making an AI-generated draft report available to clinicians, all radiologists thought the AI-generated report could be useful for triage purposes, especially in situations where clinicians could escalate cases – as long as it did not look *too final*: *"The subtlety there is that a draft report sounds too final in the health culture. But a 'prelim' or a 'wet read', that's a very rough, not final thing. The clinicians would take that information and use their judgement to call the radiologist or wait for the report." (R2)* Alongside legal, regulatory and other organizational requirements to approve any such AI use, this requires a system design that appropriately communicates and clearly discloses the nature of preliminary AI-generated contents.

7.5.2 Augmented Report Review

Locating image findings & their prioritization by clinical relevance: Exploring how VLM capabilities could be utilized to augment the experiences of clinicians when reviewing the radiology report, all described finding image annotations helpful, especially for complex images like CTs. Most clinicians shared that they do not receive training to read CTs: *"I look at CT scans, but I'm not trained to look at CT scans. I'm trained to look at X-rays." (C5)* Some (C3, C6, C7) noted that they are comfortable reading CTs mainly within their subspeciality: *"[In a brain scan] I would 100% be able to localize where things are. But if it was a report of a liver I would struggle." (C7)* They pointed out that for such multi-slice images, current systems require them to manually navigate to the image slice indicated in the report to view abnormalities. Having "clickable" findings, either on the report itself or in an overview section, that would direct them to the image location of relevance, was perceived valuable to save time and make it easier to differentiate what is in the image: *"[Looking at a CT scan that had multiple areas of edema infarction] As a clinician, you're like, well, this must be the bit that's bleeding, but this must be the inflamed bit. But they look similar to me." (C1)* Clinicians additionally described several abnormalities that can be difficult to interpret: *"Lymph nodes are the thing that people often miss on chest X-rays. Small pneumothoraces are difficult to see. The difference between a pneumothorax and a bullae [is] a common problem with the misreading of chest X-rays." (C6)* As such, they ascribed value to AI image annotations in aiding their understanding of the reported findings. Furthermore, similar to radiologists' feedback, clinicians reflected that an overview section could highlight the most important and actionable findings: *"Report overview would work best if you constrain it to show the top 6 salient features. We can get a lot of information overload if there are 25 of them." (C7)*

Building an appropriate mental model of the AI: When discussing more broadly how AI assistance could feature within workflows, one clinician differentiated for example a radiology assistant from a clinical assistant, whereby the former is embedded within the image viewer for radiology-specific tasks, whereas the latter –which is conceived as answering broader clinical questions– would be expected to sit within the EHR system: *"If I've got a radiologist at my fingertips, I'd restrict to asking it the kind of questions I might be asking the radiologist. Therefore it belongs in [the radiology] screen, whereas some of the other things like, how should I treat this patient? I think that belongs in the main body of EHR rather than in this radiology reporting system." (C4)* This commentary highlights the importance of workflow integration for building an appropriate mental model of the AI's likely purpose and capabilities.

Cautioning about chat format & too complex queries: In response to the AI assistant embodied as chatbot, several clinicians (C1, C3, C5, C7) commented that they were unlikely to use an assistant in chat form due to time-demands and lack of trust in generated, potentially high-risk responses: *"I don't need a chatbot function where I'm talking and stuff. I haven't got the time for it." (C5)* Some clinicians raised concerns about responsibility in clinical decision making: *"I'm not all of a sudden going to ask*

ChatGPT 'What am I going to do with the brain tumor?' I'm going to ask my friend who's a specialist of this. There's a question of responsibility. " (C1) Similarly, in answers to questions what clinicians would not want to use an AI assistant for (whether in chat or any other form), C7 – an oncologist – emphasized that he would not use it as a prognostic tool: *"The radiology assistant shouldn't be used to make predictions. It's not a radiomic analysis in that sense."* Similarly, a cardiothoracic physician indicated that she would not ask what's unknowable: *"You wouldn't ask things that are impossible to know. Things that are too complicated, like [the patient is] on six other drugs, how are they going to interact in combination? I wouldn't bother asking, I wouldn't trust the answer cause it's too individualized." (C6)* Another concern was around the reinforcement of radiology observations that present negative findings. Here, clinicians stressed that they weigh positive findings more than negatives: *"[If someone asks] 'Can you confirm there really isn't a small pneumothorax on this?' Then the answer from the assistant should be 'No, you can't'." (C7)* In other words, clinicians cautioned the uses of AI for more ambitious, high-risk VLM use cases involving prognosis, more complex patient cases, or a definite negation of abnormalities – given more likely chances of errors and their negative implications on patient care.

Focusing on task- and patient-specific, functional queries: However, clinicians described an array of rather functional, task-specific queries where they could imagine AI to assist by either connecting them to, or extracting information on their behalf. For example, clinicians envisioned the AI assistant to perform image-based quantifications such as calculations of the cardiothoracic ratio (calculated by measuring the maximum diameter of the heart and thoracic cavity); Mirels' score (indicating the risk of bone fracture); sarcopenia index (muscle-fat ratio to track weight loss in cancer patients); and waist-to-hip ratio in CT scans. All of these are currently calculated manually, often using phone apps: *"It would be perceived added value if it could be quickly extracted from [an image] read, as you wouldn't calculate it unless you needed." (C7)* In keeping with these more functional tasks, participants often envisioned AI interactions in familiar forms, such as tool buttons, alerts or reminders for specific conditions and workflows; thereby describing expectations of the AI being designed as a workflow tool. One clinician expressed: *"I almost would want the prompt 'Have you thought about this?'" (C5)* whilst simultaneously cautioning that such prompts could easily become annoying: *"[For guidelines] I want to be able to click [on a finding], guidance, then it searches and brings it up for me. I don't want pop-up fatigue." (C5)*

Furthermore, clinicians described how such practical, patient-specific AI functionality could be achieved even more effectively if VLM capabilities were combined with patient EHR data: *"You want it to give you, here's their allergies, here's their weight, here's their renal function, here's their swallow plan. Do they have a cannula in place? And here's their other medications that could interact with that medication. If it can pull from the system that type of information, excellent, you're saving me a huge amount of time." (C5)*

Criticizing many of the more generic information that were probed in our concept sketch (e.g., clinical features, differential diagnoses), clinicians emphasized the benefits of including additional EHR data to provide patient-context relevant information: *"I don't need [it to remind me] the 10 common causes* *of pleural effusion. What will be really helpful is for it to know that actually in this context, hypothyroidism becomes not the 29th thing, but actually upping [that to] your top five you should be considering ... because this patient's got some other clues or signs." (C3)* Similarly, surfacing a patient's eligibility for clinical trials or surfacing specific hospital or NHS level guidelines were described useful (C1, C2, C5, C6, C7); re-emphasizing the need for information retrieval specific to each patient's context.

7.5.3 Visual Search and Querying

Aiding interpretation via comparison with relevant patient cases: All clinicians and radiologists shared that they perform web searches to find similar images, though not too frequently (e.g., 1/week). For this concept, being able to visually search radiology images and reports within the context of their hospital and patient population was valued the most: *"Often you look at a CT scan on [internet] and you go 'my CT scans don't look anything like that' [because it was a different generation CT scanner]. So it's very important to visualize the abnormality in the context of the type of imaging you would see in your center." (C7)* Most clinicians and radiologists wanted to query what is normal, or queries with age and sex: *"Recently we had a big debate: What does a 16 year old thymus look like normally?" (C6)* An intensive care unit (ICU) clinician also described the difficulty of assessing rare conditions where they overlap with other abnormalities, because such cases are too infrequent and unfamiliar: *"Nasogastric (NG) tubes in the wrong place on a chest X-ray on someone in ICU with pneumonia is even less common [than misplaced NG tubes alone]. So people have to simulate abnormalities in their head and compare the X-ray with their simulation. Showing [cases] similar to your patient would be useful." (C1)*

All this suggests potential benefits of VLM use in retrieving or simulating other patient cases that enable comparative image assessments for either rare and complex (e.g., querying 'NG tube' + 'pneumonia'), or normal cases to assist interpretation. For such uses, participants positioned the AI system as a tool for extracting, searching or filtering information rather than as a conversational interface: *"I'd have it as a tool that I can work with, and not conversation." (R1)* Describing how they would use queries to refine image search, one clinician added: *"To then be able to type in pneumonia for example, and then the other [search results] go away. 'Just female patients' or 'I'm only interested in people over 75'." (C7)*

AI insights to provide reassurance to 'human' interpretation: Reflecting on *when* in their workflow visual search and query capabilities could be useful, some clinicians suggested their use for follow-up questions about the radiology report: *"Radiologist might have looked at it, but just not commented on it. I just want the reassurance, is that normal or not? Is it a nodule? Is it a mass? Is it a piece of consolidation? Same goes with head scans. Does this look like quite a full brain? Does the patient have hydrocephalus or not?" (C5).* Yet, other clinicians reflected that even with AI functionality to retrieve i.e., similar images, they might still want to ask a radiologist to be assured: *"Would I be reassured if it flashed up a whole load of other people's chest X-rays and said, this was reported as normal and this was reported as normal, for yours is probably normal. I'm not sure that I would, but maybe." (C6)* Interestingly, none of the participants expected the system to provide an answer, and preferred example patient cases to inform their decisions: *"Here's a bunch of pictures, you decide. And that's reasonable, right? I'm not asking some kind of segmentation to then take responsibility for the decisions." (C1).*

7.5.4 Patient Imaging History Highlights

Reducing laborious information gathering: All radiologists and clinicians highly valued having a summary of a patient's prior images highlighting key events and chronic conditions. Recognizing the potential for time savings: *"Half of my life is kind of spent chasing notes and pre-existing conditions. A sentence or two, just about the radiology, would save me a lot of time." (C1)* Some clinicians (C3, C7) spoke of a time-reward trade-off: *"The problem with image interpretation is, how far back do you look when interpreting for change?" (C7)* They expressed feelings of guilt as they mostly look through recent reports, but not images, due to lack of time. Radiologists, on the other hand, shared they take a thorough look at past images, yet expressed desires for an automated summary: *"That is a pretty standard practice already for radiologists, but certainly being able to more easily get at that imaging history is going to be a help." (R1)*

Facilitating relevant patient information access: Probing what would be useful to highlight, participants mainly described the historical status of the patient, such as the baseline lung architecture before a patient had pneumonia. Examples included past operations (e.g., *Do they have a collapsed lung?*), key events (e.g., *When their pacemaker first appeared or their sternotomy wires first went in?*) and changes in abnormalities (e.g., *New masses, fluid consolidation, rib fractures, are they old or new?*). When asked whether a text summary would be still useful in comparison to more multimodal, VML capabilities (e.g., text summary of key events along with image annotations), most participants commented that linked reports and visual highlights could aid verification: *"If you clicked on it [for it to show you annotated images], then you can corroborate." (C6)* Finally, a few clinicians (C3, C6) pointed out that unlike radiologists, the interface they use to review prior reports only presents a list view without images. They thought AI would still be useful if it could point them, at least, to important reports to guide their navigation to the relevant image: *"I have to click on each one individually, wait for it to load. ... Even if I had a little red flag next to it saying 'open this one, this has got money in it'." (C3).* This suggests that the utility may be achieved with simpler AI capabilities.

7.6 Discussion

Similar to the previous chapter, this case study explored the use of AI capabilities and examples for ideating, prioritizing, and sketching VLM concepts with clinician and AI collaborators. Due to the project's focus on interaction level research questions, my team and I were able to take some of the concepts forward to gain insights into what predictions clinicians would find valuable. In this sense, this case study demonstrates how HCI research can inform data collection and model building efforts in the early stages of AI development.

Going back to the high-level research questions I seek to answer in my dissertation, below I reflect on how considering moderate AI performance impacted our solution space. I also expand on key learnings on our modified innovation approach blending user-centered design and matchmaking.

7.6.1 Finding Moderate Performance AI Use Cases

In this project, our focus was largely on multimodal VLM capabilities to gain insights into how these research advances might translate into clinical practice. Our case study showed that while applications of these capabilities in high-risk scenarios can provide value, clinicians invariably expect a near-perfect performance in high-expertise tasks (e.g., draft report generation). Taking a complementary approach, we asked 'Can there be simpler, dumber versions of these concepts?' This helped us expand our search space to identify medium-expertise, moderate-performance use cases that were perceived as high value. For example, summarizing prior patient reports requires relatively lower expertise –a medical student level task– where having a 'good enough' summary could be still more useful than no summary.

This finding suggests that AI designers and developers should evaluate whether VLM (or other multimodal AI) capabilities are truly needed and appropriate for a task and explore alternatives. The Patient Imaging History Highlights concept demonstrates this approach well: while a multimodal model can summarize rich patient image-report data; text-only models may already create value by summarizing previous report texts or pointing to important reports without text extraction and summarization.

7.6.2 AI Innovation Driven by Technology and Users

Similar to my prior explorations, this case study provided evidence that using AI capabilities and sensitizing domain stakeholders to moderate AI performance yields a broad set of problem-solution matches. Using multiple sketches as *instantiations of capabilities* rather than concrete design proposals, and probing clinicians and radiologists (e.g., *Knowing that AI can do this, can you think of situations where this capability would be useful?*) seemed to work well. Participants envisioned many use cases that can provide clinical utility, such as calculating medical ratios and scores; assessing organ sizes, volumes or density; and retrieving EHR patient data. Future research should expand this emergent approach that blend human-centered and tech-centered innovation processes.

While *sketching* [21] with VLM capabilities scaffolded ideation, separating the underlying capability from the form was a challenge. For example, the AI literature often uses the term 'Visual Question-Answering' to refer to AI tasks around image-to-text or text-to-image capabilities, yet these capabilities do not necessarily require a conversational form. Similarly, we struggled to envision novel VLM interactions that go beyond chatbots, alerts, and recommenders, a well-known challenge in AI design literature [158]. We approached this challenge by framing VLM capabilities as *queries* that can be formed in different ways (e.g., conversational questions, pre-run prompts, alerts, visual annotations, etc). Interestingly, the way participants described VLM interactions resembled *robotic process automation*: AI that fetches data in the background and presents it in an unremarkable manner [159] that can either be included or easily ignored. These findings point to a need for new design patterns beyond current paradigms of LLM or VLM uses as chat or conversational queries – especially in workflow-oriented contexts.

Finally, I see opportunities for design research to investigate how to effectively sketch and prototype VLM interactions. In this project, we utilized click-through sketches to scaffold clinicians' thinking around what AI can do and how the system might behave in specific use cases. One limitation of this approach is evaluation; without functional prototypes, the feedback given on the concepts remain speculative as initial indications into clinician expectations and acceptance. Further research is needed to substantiate, test, and challenge the insights and assumptions that are presented in this work.

Chapter 8

Case Studies with Industry Partners

In the previous two chapters, I explored the use of AI capabilities and examples to scaffold the ideation of AI products and services within two different interdisciplinary teams in healthcare. This work showed considerable promise; it shifted the teams' focus toward high-value, low-risk AI concepts. It also indicated a breakdown in current innovation approaches: neither user-centered nor technology-centered approaches seemed to work well for AI. Instead, integrating aspects of both –starting with capabilities and asking domain experts to think of problems where moderate performance creates value– seemed to be more effective.

This chapter further explores this hybrid innovation approach by branching into other domains. I present two additional case studies with industry partners. *Does brainstorming with AI capabilities and examples result in more effective ideation for identifying AI use cases? How does this innovation approach impact the way teams brainstorm and develop AI products?* To investigate these questions, I collaborated with teams working in (1) insurance and (2) accounting to conduct a series of ideation workshops using the AI Brainstorming Kit.

Below, I first describe the process of establishing these collaborations and provide an overview of the study procedure. I then detail the iterative design process that involved a series of ideation workshops with each industry partner. This process provided valuable insights into how teams currently brainstorm and identify AI use cases and the impact of this approach on their practices. It also revealed several challenges that still need to be addressed.

8.1 Design Process

8.1.1 Establishing Collaborations

I had two criteria for selecting industry teams to partner with. First, I looked for cross-functional teams (e.g., product managers, data scientists, AI engineers, UX designers, etc) that can carry out the AI development work. Second, I considered whether the teams had access to data that could be leveraged for AI projects. I did not filter teams based on their prior experience with AI.

The process for finding partnering teams involved two activities. First, my team and I reached out to our network, emailing direct contacts across over 25 industry and public organizations. Second, we gave research talks at various organizations and practitioner-facing conferences, where we presented our innovation approach and the AI Brainstorming Kit. Through these connections, we initiated relationships with 12 potential partners. We shared a one-page project description detailing the proposed AI ideation workshop activities. Our point of contact circulated this within their organization to identify product teams who might be interested in collaborating. In parallel, we developed non-disclosure agreements (NDAs) with each partner and obtained IRB approval within our institution. This process lasted over a year and led to partnerships with an insurance company and an accounting company. Overall, the selection of companies was based on their availability.

8.1.2 Procedure

The study procedure had three main parts: pre-workshop meeting, brainstorming workshops, and post-workshop interview.

Pre-workshop Meeting

Once we identified a product team, we set up an introductory meeting (30-45 minutes) with team leads to discuss workshop details and logistics. Each team lead described their team's product, the customers of these products, how their team is structured, and current AI-based product features (if applicable). The academic team shared the research goals, proposed activities, and expected outputs. Collectively, we discussed potential starting places to explore AI opportunities during the workshops. Based on the team's focus and availability, we defined the workshop scope, number of sessions, duration, and team composition.

Brainstorming Workshops

We conducted two consecutive workshops with each industry partner. Each workshop included participants from a product team (e.g., data scientist, product manager, machine learning engineer, etc) and research team (e.g., HCI researcher, data scientist) with a total number of participants between 6-14 people. Workshop sessions lasted between 1-4 hours. Sessions were conducted remotely and were facilitated by the academic team. We used Miro [98] and Mural [103] to create a digital whiteboard for collaborative brainstorming and mapping of concepts.

Post-workshop Meeting

Following the workshops, we set up a post-workshop meeting (45-60 minutes) with team leads for debriefing. We probed what worked well, what could have been better, and whether the teams might use this approach in the future.

8.1.3 Data Collection and Analysis

Data collected included (1) workshop and meeting transcripts and (2) workshop outcomes. We analyzed the collected data using affinity diagramming [61], the impact-effort matrix [60], and the task expertise-AI performance matrix [162].

8.2 Study 1: Insurance

8.2.1 Background

The insurance provider we worked with develops both business-to-business (B2B) and business-to-consumer (B2C) services. On a B2C basis, customers can get quotes and purchase insurance policies (e.g., auto or home insurance) directly without any intermediaries, such as agents. On a B2B basis, insurance companies partner with independent insurance agents to serve customers who may have complex insurance needs (e.g., small businesses looking for specialty insurance). Independent agents represent multiple insurance carriers. They create value for customers by providing personalized guidance when choosing insurance policies as well as ongoing support (e.g., assisting customers with claims).

The product team we partnered with was responsible for insurance agent-facing systems (e.g., system for agents to get quotes) and agent support systems (e.g., system for sales teams to manage relationships with agents). In the pre-workshop meeting, the team lead shared an overview of their domain, pain points, and potential areas where AI might create value. These included things like better personalization towards customers and end users, gaining better insights into agent behavior, and helping agents with insurance recommendations. On a higher level, the team lead noted that their innovation approach must balance business objectives (e.g., customer satisfaction, growth, retention) with risks and regulations. This initial conversation provided us with a shared understanding of the domain and set the stage for the workshops.

8.2.2 Method

Participants

The workshop sessions included 12 participants from the product team (i.e., 6 product managers, 3 data scientists, 3 machine learning engineers) and 2 participants from the research team (i.e., 2 HCI researchers). The second workshop included two additional participants from the product team (i.e., a data scientist and a software engineer from the IT department). All participants had more than 10 years of professional experience.

Workshop 1: User-centered Innovation

The first workshop followed a user-centered approach. Based on our discussions in the pre-workshop meeting, we focused on the insurance agent as the target persona, and defined the workshop goal as "Exploring AI opportunities to improve agent quoting". In preparation for the workshop, we created "how might we" prompts as starting places for ideation *(e.g., How might we better personalize our systems towards agents? How might we help agents better understand their customers? How might we help novice agents in recommending insurance options?)*

The workshop session lasted 4 hours (two 2-hour sessions with one hour break in between). The agenda included introductions and workshop overview (15 min), a lightning talk session where the product team provided an overview of the agent workflow, jobs to be done, and pain points (20 min), two consecutive ideation sessions (30 min each), discussion (20 min), sorting and voting (30 min), impact-effort assessment (60 min), and debrief (20 min). In ideation sessions, participants first generated concepts individually using sticky notes in a Mural board, then shared out with the group for collective brainstorming. Following ideation, we sorted ideas using affinity diagramming [61]. Next, we voted concepts using dot stickers and mapped this subset onto an impact-effort matrix [60]. After the workshop, the research team further analyzed the concept using task expertise-AI performance matrix [162].

Workshop 2: User-centered and Tech-centered Innovation

The second workshop followed a process blending user-centered and tech-centered innovation. In the first workshop, we encountered challenges assessing the value of the generated concepts with product managers as proxies for end users (i.e., insurance agents). In the second workshop, we decided to shift the focus towards internal use cases. We defined the workshop goal as "Exploring AI opportunities to support work and productivity", and asked participants to reflect on their work practices and envision ways AI can support their work.

The capabilities and examples included *summarize text* (review summaries); *format data* (converting text to table); *classify things* (spam filter); *robotic process automation* (querying data to perform actions); *discover anomalies* (banking fraud detection); *cluster similarities* (online shopping recommender system); *forecast demand* (anticipatory shipping); *discover individual or group patterns* (smart home thermostat); *estimate breakdowns* (predictive maintenance); and *estimate task duration or difficulty* (driver-rider matching). We conducted a 4-hour workshop following the same structure as in the first workshop. This time, instead of starting with a lightning talk on pain points, we reviewed the AI capabilities and examples we had prepared as slides (15 min) and placed these in the Mural board to prompt ideation.

8.3 Findings

8.3.1 Workshop 1 Outcome

The first workshop produced outcomes that cover a broad set of themes. Examples included systems for gaining insights into agents (e.g., predicting agent behavior); gaining insights into customers (e.g., customer segmentation); gaining insights into competition, fraud detection and verification; improving customer support (e.g., personalizing content for agents); improving agent decision making; streamlining agent workflow (e.g., automated data entry); upskilling agents; and optimizing operational costs.

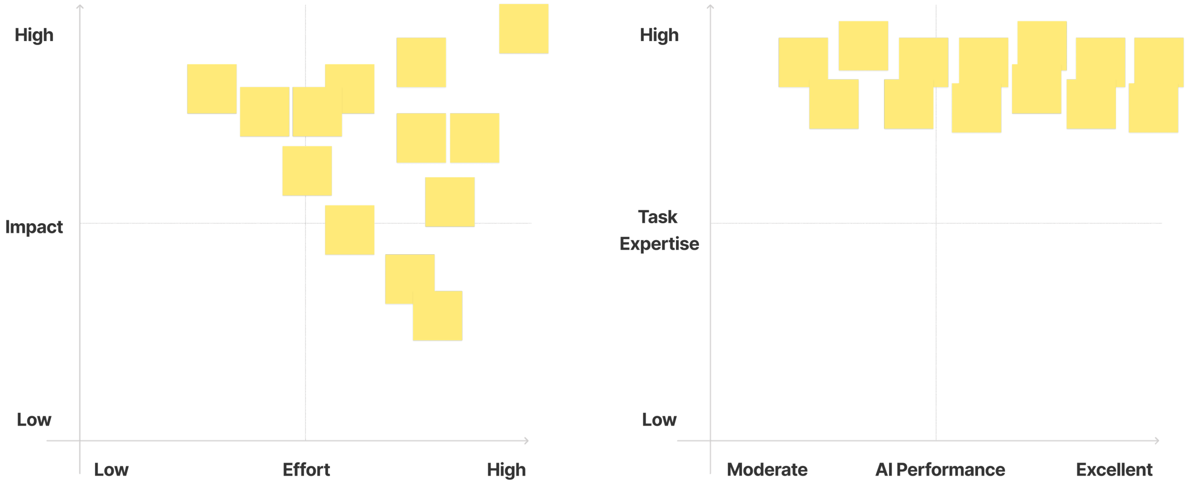


FIGURE 8.1: The first workshop resulted in concepts that were mainly *high-impact*, *high-effort*. The level of AI performance needed ranged from *moderate* to *excellent*.

The impact-effort assessment showed that the concepts were mainly *high-impact, high-effort* along with some *high-impact, low-effort* and *low-impact, high-effort* concepts (Figure 8.1). The task expertise-AI performance matrix revealed that all concepts were focused on tasks that required *high-expertise* (e.g., customer segmentation). The level of AI performance needed for concepts ranged from *moderate* (e.g., data mining to gain customer insights) to *excellent* (e.g., recommending custom insurance plans).

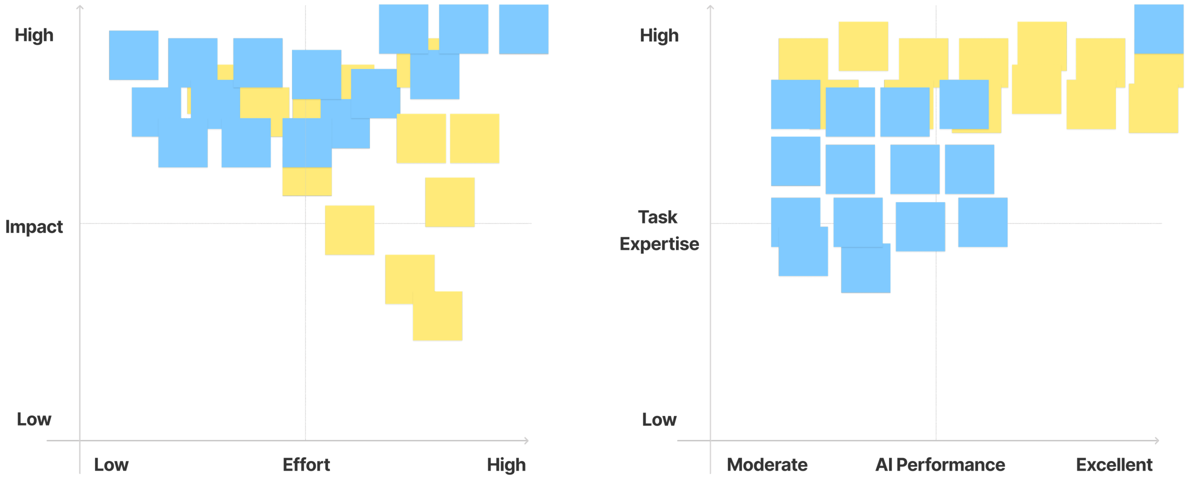


FIGURE 8.2: The second workshop resulted in concepts that were *high-impact* and *low* to *high-effort*. The concepts were largely *moderate-performance*, with task-expertise ranging from *medium* to *high*.

8.3.2 Workshop 2 Outcome

The second workshop produced a similar set of themes, along with new themes that focused on internal use cases rather than agents. Examples included systems for streamlining employee workflows (e.g., generating text for internal forms); supporting information seeking and sensemaking (e.g., summarizing new regulations); increasing developer productivity (e.g., converting code between languages), and optimizing staffing (e.g., forecasting staffing needs).

The impact-effort assessment showed that all concepts were *high-impact* and implementation effort ranged from *low* to *high* (Figure 8.2). The task expertise-AI performance matrix mapping revealed that concepts were largely *moderate-performance*, with task-expertise ranging from *medium* (e.g., summarizing customer support feedback) to *high* (e.g., predicting customer lifetime value).

8.3.3 Post-workshop Reflections

Both workshop sessions were successful in producing many AI use cases that are feasible and valuable. Workshop outputs were similar in terms of quality and breadth, however, the concepts shifted towards moderate performance in the second workshop. When we asked the team members to reflect on each process, all participants expressed a preference for the approach that used AI capabilities and examples as a starting place: *"Having technology helps better scope and narrow down the problem space. [Talking about a capability] like text extraction, then saying there is an unmet need where we can use that ... Moderate level performance is really valuable, that's good intuition." (MLE)* Additionally, participants found the exercise valuable for sensitizing the team members to AI capabilities: *"It's not just the output of the brainstorming sessions. There is the education angle, people getting more exposure to what AI can help with." (PM)*

Reflecting on what worked well, several things stood out that made the sessions with insurance innovators different than the healthcare work. First, the team members were familiar with the data pipeline. They also sporadically conducted brainstorming sessions as a regular part of their innovation work. A PM shared a framework that they use to explore potential AI use cases that captured the data *(What data is available to drive predictive insights from?)*, prediction *(What can we predict using this data?)*, action *(What action(s) can we take based on this prediction?)*, and value *(What is the value generated by this?)*. Second, several team members reflected that the culture really matters; they felt that they had an environment where team members felt safe bringing up ideas. They implied this had not been true in other places they had previously worked. Finally, the team members had a good grasp of regulations and constraints. They did not envision any "moonshot" use cases. Instead, they thought of incremental improvements that would mainly improve the data quality and the accuracy of predictions, which resulted in largely feasible concepts. In speculating on their lack of risk-taking during ideation, we wondered if this was the result of many years of effort trying to innovate in a heavily regulated industry.

Our post-workshop discussions also revealed some remaining challenges. Following ideation, we encountered challenges in assessing and prioritizing concepts. The team reflected that a follow-up session could be helpful to detail some of the concepts as sketches or wireframes. Currently, it remains difficult to capture concepts in a way that is comparable for rapid assessment and prioritization. Another challenge was moving concepts to production. The team shared that they typically create a proof of concept to pull resources for further development and deployment. However, it was unclear how projects get prioritized. Several DS and MLE team members spoke of challenges in integration (e.g., cost, dependencies), noting that modeling was often the easier part. Finally, the team reflected that it is challenging to get a sense of whether a concept will generate more revenue than its development and integration costs, a challenge previously reported in the literature [161].

8.4 Study 2: Accounting

8.4.1 Background

Accounting companies often operate within the B2B space by providing services to other companies to help businesses manage their finances and comply with regulatory requirements. Accounting services typically include auditing, tax advisory, consulting, and risk management. Our collaborators in the accounting company included two teams: an AI innovation team and a business team. The AI innovation team consisted of members with AI expertise who had experience working within specific application domains (e.g., tax services). These domain specific subgroups served as a central resource to provide innovation support to product teams, who submitted the AI use cases they identified to the innovation team. These included a brief description of the use case, expected value (e.g., time and cost savings), and details about the data requirements (e.g., data availability, storage, dependencies, etc). The innovation team worked with individual teams to scope, assess, and prioritize use cases. Based on this assessment, the selected use cases were assigned resources (i.e., data science and development team) for further design, development, and deployment of a minimum viable product (MVP).

The business team we partnered with provided accounting and tax services to B2B clients (e.g., preparing and filing corporate tax returns, ensuring compliance with tax regulations). The team included accounting professionals who were interested in developing AI solutions within their domain. In the pre-workshop meeting, the product team lead shared an overview of their process and pain points, starting from the initial project proposal phase to staffing, technology setup, client data ingestion, reviewing accounts, preparing tax forms, quality assurance, and filing tax returns. We discussed areas of interest for potential AI use cases. These included things like supporting tax experts in their daily tasks, improving the business process workflow, and personalizing services for clients. Similar to the first case study, team leads shared that they work in a highly regulated industry with clearly defined rules for data use. These initial conversations helped with level-setting as we prepared for the workshops.

8.4.2 Method

Participants

The workshop sessions included four participants from the AI innovation team (i.e., 1 product manager, 2 AI engineers, 1 business analyst), two participants (i.e., accounting professionals) from the business team, and three participants from the research team (i.e., 3 HCI researchers). All participants had more than 5 years of professional experience.

Workshop 1 & 2: User-centered and Tech-centered Innovation

With this team, we conducted workshops that followed our proposed approach, blending user-centered and tech-centered innovation. We conducted two 1-hour workshops, focusing on brainstorming and concept assessment respectively. Based on initial discussions, we defined our goal as "Exploring AI use cases for tax workflows".

The first workshop agenda included introductions and workshop overview (10 min), review of AI capabilities and examples (15 min), ideation (30 min) and debrief (5 min). We used the same selection of capabilities and examples as outlined in the previous section and created a Miro board for brainstorming and sorting concepts. The second workshop built on the concepts generated in the first workshop. We selected a subset of 8 concepts (one per each high-level theme) to further detail and assess. We then rated each concept in terms of its feasibility, desirability, and viability. The workshop agenda included workshop overview (5 min), concept assessment (45 min, about 5-6 min per concept), discussion and debrief (10 min). After the workshop, the research team mapped the concepts on impact-effort and task expertise-AI performance matrices based on the team discussion.

8.4.3 Findings

Workshop 1: Brainstorming Outcome

The brainstorming session produced concepts covering a large set of themes. Examples included systems for assessing client data quality (e.g., detecting inconsistencies or abnormalities); improving data processing (e.g., smart data formatting); evaluating rules and requirements (e.g., recommending relevant regulations); accelerating information retrieval (e.g., document summarization and triangulation); automating form processing (classifying forms); gaining insights into clients (e.g., discovering client characteristics); and supporting dataset creation (e.g., synthetic data generation).

Workshop 2: Concept Assessment Outcome

The concept assessment session mainly focused on feasibility (e.g., data availability, level of AI performance needed), desirability (e.g., value proposition), and viability (e.g., time savings or other benefits). The question of "how well does the AI need to perform" was at the crux of our discussion as it defined both the user experience and technical requirements. To search for simpler applications, we probed *"Is*

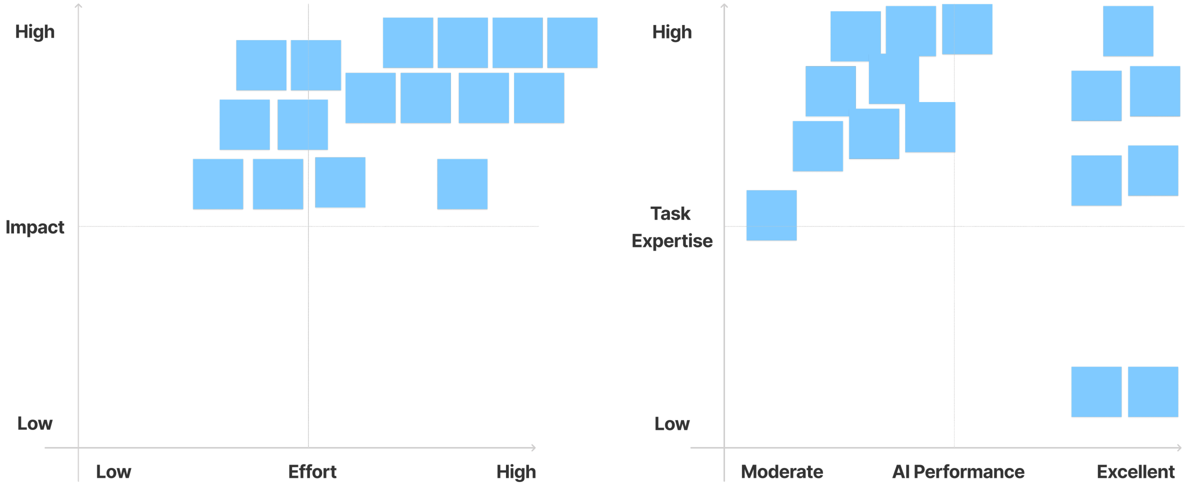


FIGURE 8.3: The workshop resulted in concepts that were largely *high-impact* and *medium-effort* to *high-effort*. The task expertise-AI performance matrix mapping revealed clusters around *high-expertise/moderate-performance* tasks along with tasks that required *excellent-performance*.

there a simpler, moderate AI version of this concept that is still valuable?" for each use case. To assess desirability and viability, we probed domain experts for back-of-the-envelope calculations. For example, for a use case around automated form classification, domain experts deliberated on the task workflow and estimated the time savings as total employee hours per year.

The impact-effort assessment showed that the concepts were largely *high-impact* and *medium-effort* to *high-effort* (Figure 8.3). The task expertise-AI performance matrix mapping revealed clusters around high-expertise/moderate-performance tasks (e.g., detecting data inconsistencies or abnormalities) along with tasks that required excellent-performance. The level of expertise needed for these ranged from *low* (e.g., classifying forms) to *medium* and *high* (e.g., smart data formatting). Overall, the concepts were largely skewed towards tasks that required high domain expertise.

Post-workshop Reflections

The workshop sessions produced several high-value, medium-to-high effort AI use cases. Reflecting on each session, the team members shared that they did not struggle with brainstorming. Domain experts (accounting professionals) expressed that they were self-taught in AI and automated-decision making. They repeatedly recognized AI opportunities and initiated the development of several internal AI applications. The innovation team, on the other hand, had a backlog with over a hundred AI use cases submitted by teams across the company. The innovation team typically sets up a scoping call with each team to understand the use case and score it based on the value proposition, cost, effort for development and testing, and scalability within the company. Interestingly, the use cases were considered case by case as they were submitted; there was no holistic comparison of use cases. Both the business and innovation teams found the concept assessment workshop more useful as it allowed them to deliberate on, assess, and prioritize a subset of use cases.

We noted a few strategies that worked well. First, the focus on moderate performance helped us reduce the complexity of use cases. Several team members reflected that there are limitations in terms of data availability due to operating in a sensitive domain, therefore they prioritized use cases where performance requirements could be lower: *"We work in a highly regulated industry where data is sensitive. We have to be able to accept a lower quality of accuracy." (PM)* Second, doing back-of-the-envelope calculations with domain experts helped the team gain a deeper understanding of user value and business value. Finally, focusing on internal use cases with domain experts helped us come up with largely high-impact concepts.

In terms of challenges, we mainly struggled with problem formulation and concept assessment. There seemed to be a lot of nuance in our discussions detailing what a concept is, which turned out to be difficult to capture. This made it difficult to compare and rank concepts against each other. Additionally, similar to the previous case study, the team found it challenging to assess the feasibility of concepts, as the development effort was often tied to the complexity of problem formulation. Finally, the workshop outcomes largely covered high-expertise tasks, only a few concepts were around low-expertise tasks. The team reflected that this might be a useful lens to introduce during brainstorming to broaden the consideration of tasks: *"I liked in the framework that you're thinking explicitly about Task Expertise — that's not an attribute we call out here very often." (AI Engineer)*

8.4.4 Discussion

I started this project with a focus on improving the ideation process: *How can we help teams identify high-value, low-risk AI use cases, places where moderate AI performance can be useful?* Results showed that both case studies produced a broad set of high-value AI use cases where required AI performance ranged between moderate to excellent. Many team members found the exercise useful and educational; they expressed interest in adopting this innovation approach and resources for future use. Notably, the idea of looking for situations where moderate model performance is 'good enough' was new and perceived as valuable by both teams. On these levels, the workshops were effective in sensitizing teams to broadly and rapidly exploring AI's problem-solution space.

On the other hand, it remains difficult to attribute this success solely to the workshop sessions. Both teams had expertise in AI innovation; they seemed comfortable identifying AI use cases and collaborating with team members from other disciplines. Both organizations seemed to encourage brainstorming to foster innovation. While there was no systematic approach to ideation, teams brainstormed sporadically to identify use cases within their product area towards building a proof-of-concept and requesting resources for development. They did not seem to struggle with brainstorming AI concepts. Instead, echoing the findings in previous chapters, the challenge was in problem formulation, concept assessment, and prioritization. In particular, we struggled to capture concepts in a way that made them less ambiguous to be able to discuss associated responsible AI risks.

Reflecting on these insights, I see new research opportunities in this underexplored research space:

1. **Problem formulation:** How to support teams in formulating AI use cases? Can pushing concepts towards a "moderate performance version" yield better (low-risk, high-value) outcomes?

2. **Concept representation:** How to capture AI use cases in detail in a way that makes concepts easier to understand and compare? What level of granularity is needed to be able to assess risks, especially around responsible AI?

3. **Concept assessment:** How to rapidly assess and prioritize a set of AI use cases? How to scaffold the discussion around the assessment criteria, including but not limited to feasibility, desirability, viability, and responsible AI considerations?

Future research should explore these questions to better understand the current state of art and improve the early phase AI design and development.

Chapter 9

Conclusion and Future Directions

Design is about making an advance towards a preferred future. Design inquiry generates knowledge as a proposal to reframe and re-understand a problematic situation through making and substantial reflection [168, 123, 53]. The research goal and questions change over the course of design experiments, leading research programs to *drift with intention* [171, 87]. This dissertation is an example of such drift.

I started my research journey by expanding on the work that framed *AI as a design material* [158], asking *"Can we make designers better at envisioning novel AI products and services? Would that lead to better, more human-centered AI innovations?"* With each step I took towards this goal, my understanding of the problem space has changed, leading me to reframe my goal and ask new questions. Below, I outline these major reframings to unpack the advance I made towards the preferred future of AI innovation:

- **Making AI innovation better is about making teams better at working together. Human-centered approaches have a tremendous potential to reduce AI failures by facilitating early phase ideation and problem formulation between interdisciplinary team members.** My initial focus on improving the ability of designers to innovate with AI involved a few assumptions. First, it assumed design practitioners ideating on their own to come up with AI concepts. Second, it implied that design and HCI experts should drive AI innovation. My formative exploration of how experienced designers work with AI revealed these assumptions to be flawed. This design-centric view was not reflective of the collaborative and interdisciplinary nature of AI innovation. This shifted my focus from making designers better to making innovation teams better at envisioning AI use cases. It surfaced that AI innovation should not be design-led. Instead, designers and HCI experts should focus on facilitating ideation and problem formulation between domain experts and data scientists. This is a place where human-centered approaches can significantly influence what gets built with AI.

- **Understanding what AI can do is necessary, but not sufficient. AI innovators should search for "low hanging fruit" – low-risk, high-value use cases where moderate performance creates value.** I started developing the AI Capability taxonomy with the goal of helping designers understand what AI can do, yet soon I realized that this resource could be useful to not only designers but to non-data scientists in general. Putting the taxonomy in practice was a significant learning moment in my research journey; I discovered that AI capabilities are not enough on their own. Capabilities helped designers and domain experts gain a sense of what AI can do, but the concepts they produced were largely infeasible. This design experiment revealed *model performance* and *task expertise* as two critical yet unarticulated dimensions. It led to the creation of the Model Performance-Task Expertise matrix as a representation of AI's problem-solution space. Reflecting on this, I reframed my research goal to explore how to help innovation teams notice situations where moderate performance AI can be useful.

- **User-centered design does not work for AI innovation. We need a new innovation process that blends user-centered and technology-centered approaches.** A key reframing coming from my research is the realization that neither user-centered nor technology-centered approaches work for envisioning successful AI products and services. By focusing too much on technology, we often overlook whether the problem being addressed is worth solving. By focusing solely on users, we disregard what the technology can do well and whether it is suited to the problem at hand. Reflecting on my observations of industry best practices and my own experience designing AI products and services, I noticed a need for an emergent design process that blends user and technology centered innovation. The case studies provided in this dissertation provide a glimpse into this modified innovation process.

- **Thinking about what level we are innovating at helps with goal setting and choosing appropriate design methods.** Prior literature sought to bring UX design expertise into AI innovation. Consequently, my initial research had a narrow focus on interaction and interface level AI innovations. Studying innovation teams across a wide range of settings helped me uncover the larger landscape of AI innovation across project phases. User-centered design methods are most effective in *Versioning*, when adding new AI-based features to existing products. On the other hand, service design and systems thinking methods are more suited for *Visioning* and *Venturing*, when thinking more strategically about new AI products and services for longer horizons. This insight marks an opportune area for future research to develop new methods for ideation and prototyping that can support all levels of innovation.

- **After ideation, problem formulation is the most critical step for designing successful AI products and services.** My dissertation work largely focused on ideation as a first step towards improved AI innovation. Broadly envisioning many AI use cases before selecting what to implement is key to reducing the risk of developing unwanted technologies. Yet, the lack of ideation is not the only challenge. The case studies I undertook revealed that after effective ideation, many challenges remain in formulating, assessing, and prioritizing use cases. This realization again shifted my understanding of the problem space, leading me to ask *How can we support teams detailing and assessing many AI concepts, before choosing what to build?* – a question that marks an open space for future research.

9.1 Future Directions

I want to take a step back and highlight a few promising research directions for AI innovation with the hope of informing future research in this area.

How to innovate with technology as a design material?

My dissertation research underscored the need for a new innovation process for designing AI products and services – one that takes both technology and people as a starting place. A crucial aspect of this process is the comprehensive assessment of various risks, including feasibility *(Can this be built?)*, value co-creation *(Will this generate value for people and service providers)*, and harm *(Will this lead to issues around privacy, bias, ethics, and fairness?)*. This shift is already underway in the industry and public services, as AI practitioners are increasingly aware of these risks and their role in AI failures. Future research should build upon this emerging innovation process. Furthermore, stepping back from artificial intelligence, the question of *How do we effectively innovate with any technological material?* remains open. I anticipate that the innovation process outlined in this dissertation will generalize to a variety of projects exploring the use of emerging technologies. Investigations in this area will ultimately pave the way for more human-centered technology innovation.

How to effectively anticipate responsible AI considerations during the concept ideation and problem formulation stages?

Borrowing from Buxton on why ideation is a cornerstone of any innovation process [21], *"The question is not "Do I want this?", but rather, "Do I want this rather than that, and why?"* My PhD research largely focused on improving the ideation of AI concepts, a piece missing in this work is the selection process. Currently, teams set out to build the first thing they can think of. They discover and address responsible AI concerns only after a system is built, deployed, and has negatively impacted people. Many open questions remain regarding how effective ideation and problem formulation might eliminate AI solutions that are likely to cause unintended consequences and harm. How can we systematically compare AI concepts? What criteria should be used to surface and assess the risks around fairness and harm? My studies with clinicians hinted that medium-task expertise, moderate-model performance use cases may be perceived as lower-risk. This raises the question, could pushing back on the level of model performance needed for a concept reduce the associated responsible AI risks? Future research should develop new methods and practitioner-facing resources to systematically assess risk during ideation and problem formulation.

How to foster data and AI literacy to broaden the participation in AI development?

One of the central themes of my dissertation was the integration of participatory approaches in the early phases of AI design and development. This ongoing challenge is gaining increasing attention within the HCI and AI research communities. My research took an initial step in this direction by using AI capabilities and examples to scaffold domain experts' understanding of what AI can do well. A promising and important research direction is exploring how to foster data and AI literacy among stakeholders: How can technologists and domain experts design datasets with downstream applications in mind? Creating intentional datasets for specific tasks and workflows will be especially important for applications of foundational models. This line of work has the potential to fundamentally shape how data *can* and *should* be defined.

9.2 Conclusion

Artificial intelligence poses unique challenges for innovating new products and services. Our current knowledge and methods barely scratch the surface of what is possible and what might be problematic with AI products. This knowledge gap manifests as a breakdown in innovation: AI projects fail at a higher rate than any other technology. In my dissertation, I explore this problematic situation through design research and take steps toward reframing and re-understanding it to describe a preferred future.

I began my research by investigating the current practices of industry teams in the early product development phase with the suspicion that a lack of effective ideation might lead teams to select suboptimal innovations to pursue. My formative studies with practitioners confirmed this hunch and uncovered some emergent best practices for *discovering the right things to design with AI*. Acting on these insights, I set out to explore whether changes to ideation might improve the AI innovation process. I developed new resources and methods to help non-data scientists better understand what AI can effectively do. I demonstrated the potential impact of these resources and methods with cross-functional innovation teams working on early phase AI applications. This exploration enabled me to better outline AI's problem-solution space by explicitly discussing a concept's model performance and task difficulty. My work with academic and industry innovation teams contributed rich case studies to this under explored area.

On a higher level, my research lays the groundwork for understanding and improving the AI innovation process through the lens of human-centered design. I see many parallels between today's AI development process and the early days of software development, where many products failed to reach product-market fit due to a lack of human concerns. Moving forward, our community can greatly benefit from human-centered methods and processes for harnessing the potential of AI while simultaneously mitigating its risks and implications. This dissertation is a step toward this future.

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*, pages 1–18, 2018.

[2] Accenture. Accenture logistics platform: win the last mile, 2020. URL: https://www.accenture.com/gb-en/services/public-service/accenture-logistics-platform.

[3] Uwa O Aideyan, Kevin Berbaum, and Wilbur L Smith. Influence of prior radiologic information on the interpretation of radiographic examinations. *Academic Radiology*, 2(3):205–208, 1995.

[4] Christopher Alexander. *A pattern language: towns, buildings, construction*. Oxford university press, 1977.

[5] Amazon. Alexa voice design guide, 2018. URL: https://developer.amazon.com/fr/designing-for-voice/.

[6] Saleema Amershi, Dan Weld, Mihaela Vorvoreanu, Adam Fourney, Besmira Nushi, Penny Collisson, Jina Suh, Shamsi Iqbal, Paul N Bennett, Kori Inkpen, et al. Guidelines for human-ai interaction. In *Proceedings of the 2019 chi conference on human factors in computing systems*, pages 1–13, 2019.

[7] Apple. Human interface guidelines: machine learning, 2019. URL: https://developer.apple.com/design/human-interface-guidelines/technologies/machine-learning/introduction/.

[8] Matan Atad, Vitalii Dmytrenko, Yitong Li, Xinyue Zhang, Matthias Keicher, Jan Kirschke, Bene Wiestler, Ashkan Khakzar, and Nassir Navab. Chexplaining in style: counterfactual explanations for chest x-rays using stylegan. *arXiv preprint arXiv:2207.07553*, 2022.

[9] Jennifer Aue. Ai design & practices guidelines, 2018. URL: https://medium.com/design-ibm/ai-design-guidelines-e06f7e92d864.

[10] Microsoft Azure. Test accuracy of a custom speech model, 2022. URL: https://learn.microsoft.com/en-us/azure/cognitive-services/speech-service/how-to-custom-speech-evaluate-data?pivots=speech-studio.

[11] Shruthi Bannur, Stephanie Hyland, Qianchu Liu, Fernando Perez-Garcia, Maximilian Ilse, Daniel C Castro, Benedikt Boecking, Harshita Sharma, Kenza Bouzid, Anja Thieme, et al. Learning to exploit temporal structure for biomedical vision-language processing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15016–15027, 2023.

[12] Rachel KE Bellamy, Kuntal Dey, Michael Hind, Samuel C Hoffman, Stephanie Houde, Kalapriya Kannan, Pranay Lohia, Jacquelyn Martino, Sameep Mehta, Aleksandra Mojsilović, et al. Ai fairness 360: an extensible toolkit for detecting and mitigating algorithmic bias. *IBM Journal of Research and Development*, 63(4/5):4–1, 2019.

[13] Michael H Bernstein, Michael K Atalay, Elizabeth H Dibble, Aaron WP Maxwell, Adib R Karam, Saurabh Agarwal, Robert C Ward, Terrance T Healey, and Grayson L Baird. Can incorrect artificial intelligence (ai) results impact radiologists, and if so, what can we do about it? a multi-reader pilot study of lung cancer detection with chest radiography. *European Radiology*:1–7, 2023.

[14] Sarah Bird, Miro Dudík, Richard Edgar, Brandon Horn, Roman Lutz, Vanessa Milan, Mehrnoosh Sameki, Hanna Wallach, and Kathleen Walker. Fairlearn: a toolkit for assessing and improving fairness in ai. *Microsoft, Tech. Rep. MSR-TR-2020-32*, 2020.

[15] 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.

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

[17] Sussane Bodker. Creating conditions for participation: conflicts and resources in systems development. *Human–computer interaction*, 11(3):215–236, 1996.

[18] Benedikt Boecking, Naoto Usuyama, Shruthi Bannur, Daniel C Castro, Anton Schwaighofer, Stephanie Hyland, Maria Wetscherek, Tristan Naumann, Aditya Nori, Javier Alvarez-Valle, et al. Making the most of text semantics to improve biomedical vision–language processing. In *European conference on computer vision*, pages 1–21. Springer, 2022.

[19] Virginia Braun and Victoria Clarke. Thematic analysis. 2012.

[20] Tim Brown et al. Design thinking. *Harvard business review*, 86(6):84, 2008.

[21] Bill Buxton. *Sketching user experiences: getting the design right and the right design*. Morgan kaufmann, 2010.

[22] Carrie J Cai, Samantha Winter, David Steiner, Lauren Wilcox, and Michael Terry. Onboarding materials as cross-functional boundary objects for developing ai assistants. In *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems*, pages 1–7, 2021.

[23] Francisco Maria Calisto, João Fernandes, Margarida Morais, Carlos Santiago, João Maria Abrantes, Nuno Nunes, and Jacinto C Nascimento. Assertiveness-based agent communication for a personalized medicine on medical imaging diagnosis. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, pages 1–20, 2023.

[24] Francisco Maria Calisto, Carlos Santiago, Nuno Nunes, and Jacinto C Nascimento. Breastscreening-ai: evaluating medical intelligent agents for human-ai interactions. *Artificial Intelligence in Medicine*, 127:102285, 2022.

[25] Francisco Maria Calisto, Carlos Santiago, Nuno Nunes, and Jacinto C Nascimento. Introduction of human-centric ai assistant to aid radiologists for multimodal breast image classification. *International Journal of Human-Computer Studies*, 150:102607, 2021.

[26] Pascale Carayon, Peter Hoonakker, Ann Schoofs Hundt, Megan Salwei, Douglas Wiegmann, Roger L Brown, Peter Kleinschmidt, Clair Novak, Michael Pulia, Yudi Wang, et al. Application of human factors to improve usability of clinical decision support for diagnostic decision-making: a scenario-based simulation study. *BMJ quality & safety*, 29(4):329–340, 2020.

[27] Michelle Carney, Barron Webster, Irene Alvarado, Kyle Phillips, Noura Howell, Jordan Griffith, Jonas Jongejan, Amit Pitaru, and Alexander Chen. Teachable machine: approachable web-based tool for exploring machine learning classification. In *Extended abstracts of the 2020 CHI conference on human factors in computing systems*, pages 1–8, 2020.

[28] Jared J Cash. Alert fatigue. *American Journal of Health-System Pharmacy*, 66(23):2098–2101, 2009.

[29] Hao-Fei Cheng, Logan Stapleton, Ruiqi Wang, Paige Bullock, Alexandra Chouldechova, Zhiwei Steven Steven Wu, and Haiyi Zhu. Soliciting stakeholders' fairness notions in child maltreatment predictive systems. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pages 1–17, 2021.

[30] Ed Chi. A twist on loss functions as boundary objects, 2020. URL: https://www.youtube.com/watch?v=jHTL_ysgetE&t=758s/.

[31] Eric S Chung, Jason I Hong, James Lin, Madhu K Prabaker, James A Landay, and Alan L Liu. Development and evaluation of emerging design patterns for ubiquitous computing. In *Proceedings of the 5th conference on Designing interactive systems: processes, practices, methods, and techniques*, pages 233–242, 2004.

[32] Design Council. The 'double diamond'design process model. *Design Council*, 2005.

[33] Ian A Cowan, Sharyn LS MacDonald, and Richard A Floyd. Measuring and managing radiologist workload: measuring radiologist reporting times using data from a r adiology i nformation s ystem. *Journal of medical imaging and radiation oncology*, 57(5):558–566, 2013.

[34] Rikke Friis Dam and Teo Yu Siang. Affinity diagrams: how to cluster your ideas and reveal insights, 2022. URL: https://www.interaction-design.org/literature/article/affinity-diagrams-learn-how-to-cluster-and-bundle-ideas-and-facts.

[35] Fernando Delgado, Solon Barocas, and Karen Levy. An uncommon task: participatory design in legal ai. *Proceedings of the ACM on Human-Computer Interaction*, 6(CSCW1):1–23, 2022.

[36] Fernando Delgado, Stephen Yang, Michael Madaio, and Qian Yang. Stakeholder participation in ai: beyond" add diverse stakeholders and stir". *arXiv preprint arXiv:2111.01122*, 2021.

[37] Wesley Hanwen Deng, Manish Nagireddy, Michelle Seng Ah Lee, Jatinder Singh, Zhiwei Steven Wu, Kenneth Holstein, and Haiyi Zhu. Exploring how machine learning practitioners (try to) use fairness toolkits. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, pages 473–484, 2022.

[38] Alicia DeVos, Aditi Dhabalia, Hong Shen, Kenneth Holstein, and Motahhare Eslami. Toward user-driven algorithm auditing: investigating users' strategies for uncovering harmful algorithmic behavior. In *CHI Conference on Human Factors in Computing Systems*, pages 1–19, 2022.

[39] Aaron S Dietz, Peter J Pronovost, Pedro Alejandro Mendez-Tellez, Rhonda Wyskiel, Jill A Marsteller, David A Thompson, and Michael A Rosen. A systematic review of teamwork in the intensive care unit: what do we know about teamwork, team tasks, and improvement strategies? *Journal of critical care*, 29(6):908–914, 2014.

[40] Accenture The Dock. Designed intelligence, 2020. URL: https://youtu.be/pkFOcgD8AnA?t=551.

[41] Accenture The Dock. Man with machine: designing a new future for humans and ai, 2019. URL: https://medium.com/accenture-the-dock/man-with-machine-designing-a-new-future-for-humans-and-ai-4f43e9a94517.

[42] 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*, pages 278–288, 2017.

[43] Steven Dow, T Scott Saponas, Yang Li, and James A Landay. External representations in ubiquitous computing design and the implications for design tools. In *Proceedings of the 6th conference on Designing Interactive systems*, pages 241–250, 2006.

[44] Steven P Dow, Alana Glassco, Jonathan Kass, Melissa Schwarz, Daniel L Schwartz, and Scott R Klemmer. Parallel prototyping leads to better design results, more divergence, and increased self-efficacy. *ACM Transactions on Computer-Human Interaction (TOCHI)*, 17(4):1–24, 2010.

[45] Reaktor Education. Elements of ai, 2018. URL: https://www.elementsofai.com/.

[46] Tatiana Ermakova, Julia Blume, Benjamin Fabian, Elena Fomenko, Marcus Berlin, and Manfred Hauswirth. Beyond the hype: why do data-driven projects fail? In *Proceedings of the 54th Hawaii International Conference on System Sciences*, page 5081, 2021.

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

[48] 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)*, volume 3, 2010.

[49] Figma. Figma: the collaborative interface design tool. 2016. URL: https://www.figma.com/.

[50] Jodi Forlizzi. Moving beyond user-centered design. *Interactions*, 25(5):22–23, 2018.

[51] Jodi Forlizzi and John Zimmerman. Promoting service design as a core practice in interaction design. In *Proceedings of the 5th International Congress of International Association of Societies of Design Research-IASDR*, volume 13, pages 1–12, 2013.

[52] Dhakshinamoorthy Ganeshan, Phuong-Anh Thi Duong, Linda Probyn, Leon Lenchik, Tatum A McArthur, Michele Retrouvey, Emily H Ghobadi, Stephane L Desouches, David Pastel, and Isaac R Francis. Structured reporting in radiology. *Academic radiology*, 25(1):66–73, 2018.

[53] 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, 2012.

[54] Jennifer C Ginestra, Rachel Kohn, Rebecca A Hubbard, Andrew Crane-Droesch, Scott D Halpern, Meeta Prasad Kerlin, and Gary E Weissman. Association of unit census with delays in antimicrobial initiation among ward patients with hospital-acquired sepsis. *Annals of the American Thoracic Society*, (ja), 2022.

[55] Fabien Girardin and Neal Lathia. When user experience designers partner with data scientists. In *2017 AAAI Spring Symposium Series*, 2017.

[56] Google. Machine learning crash course, 2018. URL: https://developers.google.com/machine-learning/crash-course.

[57] Google. Material design guidelines, 2015. URL: https://material.io/design/guidelines-overview.

[58] Saul Greenberg and Bill Buxton. Usability evaluation considered harmful (some of the time). In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 111–120, 2008.

[59] Nielsen Norman Group. Toggle-switch guidelines, 2018. URL: https://www.nngroup.com/articles/toggle-switch-guidelines/.

[60] Nielsen Norman Group. Using prioritization matrices to inform ux decisions, 2018. URL: https://www.nngroup.com/articles/prioritization-matrices/.

[61] Bruce Hanington and Bella Martin. *Universal methods of design expanded and revised: 125 Ways to research complex problems, develop innovative ideas, and design effective solutions*. Rockport publishers, 2019.

[62] Patrick Hebron. *Machine learning for designers*. Infinite Skills, 2017.

[63] Sean M Hickey and Al O Giwa. Mechanical ventilation. In *StatPearls*. StatPearls Publishing, Treasure Island (FL), 2020.

[64] Jim Highsmith and Alistair Cockburn. Agile software development: the business of innovation. *Computer*, 34(9):120–127, 2001.

[65] Kenneth Holstein, Jennifer Wortman Vaughan, Hal Daumé III, Miro Dudik, and Hanna Wallach. Improving fairness in machine learning systems: what do industry practitioners need? In *Proceedings of the 2019 CHI conference on human factors in computing systems*, pages 1–16, 2019.

[66] Karen Holtzblatt and Hugh Beyer. Field research: data collection and interpretation. In *Contextual Design: Evolved*, pages 11–20. Springer, 2014.

[67] Eric Horvitz. Principles of mixed-initiative user interfaces. In *Proceedings of the SIGCHI conference on Human Factors in Computing Systems*, pages 159–166, 1999.

[68] Stephanie Hyland, Shruthi Bannur, Kenza Bouzid, Daniel C Castro, Mercy Ranjit, Anton Schwaighofer, Fernando Pérez-García, et al. Maira-1: a specialised large multimodal model for radiology report generation. *arXiv preprint arXiv: 2311.13668*, 2023.

[69] IBM. Design for ai, 2019. URL: https://www.ibm.com/design/ai/.

[70] IBM. Everyday ethics for artificial intelligence, 2019. URL: https://www.ibm.com/watson/assets/duo/pdf/everydayethics.pdf.

[71] IDEO. The human-centered design toolkit, 2009. URL: https://www.designkit.org/.

[72] Anna Jobin, Marcello Ienca, and Effy Vayena. The global landscape of ai ethics guidelines. *Nature Machine Intelligence*, 1(9):389–399, 2019.

[73] Alistair EW Johnson, Tom J Pollard, Seth J Berkowitz, Nathaniel R Greenbaum, Matthew P Lungren, Chih-ying Deng, Roger G Mark, and Steven Horng. Mimic-cxr, a de-identified publicly available database of chest radiographs with free-text reports. *Scientific data*, 6(1):317, 2019.

[74] Mayur P Joshi, Ning Su, Robert D Austin, and Anand K Sundaram. Why so many data science projects fail to deliver. *MIT Sloan Management Review*, 2021.

[75] Jeremy M Kahn, Christopher H Goss, Patrick J Heagerty, Andrew A Kramer, Chelsea R O'Brien, and Gordon D Rubenfeld. Hospital volume and the outcomes of mechanical ventilation. *New England Journal of Medicine*, 355(1):41–50, 2006.

[76] Charles E Kahn Jr, Curtis P Langlotz, Elizabeth S Burnside, John A Carrino, David S Channin, David M Hovsepian, and Daniel L Rubin. Toward best practices in radiology reporting. *Radiology*, 252(3):852–856, 2009.

[77] Shivani Kapania, Oliver Siy, Gabe Clapper, Azhagu Meena SP, and Nithya Sambasivan. " because ai is 100% right and safe": user attitudes and sources of ai authority in india. In *CHI Conference on Human Factors in Computing Systems*, pages 1–18, 2022.

[78] Holtzblatt Karen and Jones Sandra. Contextual inquiry: a participatory technique for system design. In *Participatory design*, pages 177–210. CRC Press, 2017.

[79] Anna Kawakami, Venkatesh Sivaraman, Logan Stapleton, Hao-Fei Cheng, Adam Perer, Zhiwei Steven Wu, Haiyi Zhu, and Kenneth Holstein. "why do i care what's similar?" probing challenges in ai-assisted child welfare decision-making through worker-ai interface design concepts. In *Designing Interactive Systems Conference*, pages 454–470, 2022.

[80] Matthew Kay, Cynthia Matuszek, and Sean A Munson. Unequal representation and gender stereotypes in image search results for occupations. In *Proceedings of the 33rd annual acm conference on human factors in computing systems*, pages 3819–3828, 2015.

[81] 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*, pages 1–8, 2019.

[82] Krishika Haresh Khemani and Stuart Reeves. Unpacking practitioners' attitudes towards codifications of design knowledge for voice user interfaces. In *CHI Conference on Human Factors in Computing Systems*, pages 1–10, 2022.

[83] Rochelle King, Elizabeth F Churchill, and Caitlin Tan. *Designing with data: Improving the user experience with A/B testing*. " O'Reilly Media, Inc.", 2017.

[84] Jake Knapp, John Zeratsky, and Braden Kowitz. *Sprint: How to solve big problems and test new ideas in just five days*. Simon and Schuster, 2016.

[85] Rafal Kocielnik, Saleema Amershi, and Paul N Bennett. Will you accept an imperfect ai? exploring designs for adjusting end-user expectations of ai systems. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, pages 1–14, 2019.

[86] Cassie Kozyrkov. Making friends with machine learning: advice for finding ai use cases, 2022. URL: https://kozyrkov.medium.com/imagine-a-drunk-island-advice-for-finding-ai-use-cases-8d47495d4c3f.

[87] P Krogh and Ilpo Koskinen. *Drifting by intention*. Springer, 2020.

[88] Sean Kross and Philip Guo. Orienting, framing, bridging, magic, and counseling: how data scientists navigate the outer loop of client collaborations in industry and academia. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW2):1–28, 2021.

[89] Tianyi Li, Mihaela Vorvoreanu, Derek DeBellis, and Saleema Amershi. Assessing human-ai interaction early through factorial surveys: a study on the guidelines for human-ai interaction. *ACM Transactions on Computer-Human Interaction*, 2022.

[90] Q Vera Liao, Daniel Gruen, and Sarah Miller. Questioning the ai: informing design practices for explainable ai user experiences. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pages 1–15, 2020.

[91] James Lin and James A Landay. Employing patterns and layers for early-stage design and prototyping of cross-device user interfaces. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 1313–1322, 2008.

[92] Eoin Ó Loideáin. Towards algorithmic equality, 2019. URL: https://medium.com/design-voices/towards-algorithmic-equality-118e6ba10b74.

[93] Ryan Louie, Andy Coenen, Cheng Zhi Huang, Michael Terry, and Carrie J Cai. Novice-ai music co-creation via ai-steering tools for deep generative models. In *Proceedings of the 2020 CHI conference on human factors in computing systems*, pages 1–13, 2020.

[94] Michael A Madaio, Luke Stark, Jennifer Wortman Vaughan, and Hanna Wallach. Co-designing checklists to understand organizational challenges and opportunities around fairness in ai. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pages 1–14, 2020.

[95] Ezio Manzini. New design knowledge. *Design studies*, 30(1):4–12, 2009.

[96] Bella Martin, Bruce Hanington, and Bruce M Hanington. Universal methods of design: 100 ways to research complex problems. *Develop Innovative Ideas, and Design Effective Solutions*:12–13, 2012.

[97] Microsoft. Principles of bot design, 2017. URL: https://docs.microsoft.com/en-us/previous-versions/azure/bot-service/bot-service-design-principles?view=azure-bot-service-3.0.

[98] Miro. Miro: online whiteboard for visual collaboration, 2011. URL: https://miro.com/.

[99] ml5js. Ml5js friendly machine learning for the web, 2018. URL: https://ml5js.org/.

[100] Alessandro Morandi, Nathan E Brummel, and E Wesley Ely. Sedation, delirium and mechanical ventilation: the 'abcde'approach. *Current opinion in critical care*, 17(1):43–49, 2011.

[101] Camille Moussette and Richard Banks. Designing through making: exploring the simple haptic design space. In *Proceedings of the fifth international conference on Tangible, embedded, and embodied interaction*, pages 279–282, 2010.

[102] Cosmin Munteanu, Ronald Baecker, Gerald Penn, Elaine Toms, and David James. The effect of speech recognition accuracy rates on the usefulness and usability of webcast archives. In *Proceedings of the SIGCHI conference on Human Factors in computing systems*, pages 493–502, 2006.

[103] Mural. Mural online collaboration, 2011. URL: https://www.mural.co/.

[104] Nadia Nahar, Shurui Zhou, Grace Lewis, and Christian Kästner. Collaboration challenges in building ml-enabled systems: communication, documentation, engineering, and process. *Organization*, 1(2):3, 2022.

[105] Sandeep S Naik, Anthony Hanbidge, and Stephanie R Wilson. Radiology reports: examining radiologist and clinician preferences regarding style and content. *American Journal of Roentgenology*, 176(3):591–598, 2001.

[106] Andrew Ng. Machine learning specialization, 2012. URL: https://www.deeplearning.ai/courses/machine-learning-specialization/.

[107] Jakob Nielsen and Rolf Molich. Heuristic evaluation of user interfaces. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 249–256, 1990.

[108] Donald A Norman and Roberto Verganti. Incremental and radical innovation: design research vs. technology and meaning change. *Design issues*, 30(1):78–96, 2014.

[109] Nazmun Nisat Ontika, Sheree May Sassmannshausen, Aparecido Fabiano Pinatti De Carvalho, and Volkmar Pipek. Pairads: hybrid interaction between humans and ai in radiology. In *HHAI 2023: Augmenting Human Intellect*, pages 395–397. IOS Press, 2023.

[110] 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, 2010.

[111] Google PAIR. People + ai guidebook, 2019. URL: pair.withgoogle.com/guidebook.

[112] Samir Passi and Solon Barocas. Problem formulation and fairness. In *Proceedings of the conference on fairness, accountability, and transparency*, pages 39–48, 2019.

[113] David Piorkowski, Soya Park, April Yi Wang, Dakuo Wang, Michael Muller, and Felix Portnoy. How ai developers overcome communication challenges in a multidisciplinary team: a case study. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW1):1–25, 2021.

[114] Polytopal. Lingua franca: a design language for human-centered ai, 2020. URL: https://linguafranca.polytopal.ai/.

[115] Jon Reifschneider. Human factors in ai, 2021. URL: https://www.coursera.org/learn/human-factors-in-artificial-intelligence.

[116] Eric Reis. The lean startup. *New York: Crown Business*, 27:2016–2020, 2011.

[117] Barry Richmond. System dynamics/systems thinking: let's just get on with it. *System Dynamics Review*, 10(2-3):135–157, 1994.

[118] Eric Ries. *The lean startup: How today's entrepreneurs use continuous innovation to create radically successful businesses*. Crown Currency, 2011.

[119] Abi Rimmer. Radiologist shortage leaves patient care at risk, warns royal college. *BMJ: British Medical Journal (Online)*, 359, 2017.

[120] Samantha Robertson and Niloufar Salehi. What if i don't like any of the choices? the limits of preference elicitation for participatory algorithm design. *arXiv preprint arXiv:2007.06718*, 2020.

[121] Virpi Roto, Jung-Joo Lee, Effie Lai-Chong Law, and John Zimmerman. The overlaps and boundaries between service design and user experience design. In *Proceedings of the 2021 ACM Designing Interactive Systems Conference*, pages 1915–1926, 2021.

[122] Elizabeth B-N Sanders. From user-centered to participatory design approaches. In *Design and the social sciences*, pages 18–25. CRC Press, 2002.

[123] Donald Schön and John Bennett. Reflective conversation with materials. In *Bringing design to software*, pages 171–189. 1996.

[124] Sectra. How radiology can improve communication with referring physicians, 2013. URL: https://sectraprodstorage01.blob.core.windows.net/medical-uploads/2017/09/report-how-radiology-can-improve-communication-with-referring-physicians.pdf. [Accessed 11-22-2023].

[125] Birger Sevaldson. Systems oriented design: the emergence and development of a designerly approach to address complexity, 2013.

[126] Hong Shen, Alicia DeVos, Motahhare Eslami, and Kenneth Holstein. Everyday algorithm auditing: understanding the power of everyday users in surfacing harmful algorithmic behaviors. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW2):1–29, 2021.

[127] C Sherry, M Adams, L Berlin, L Fajardo, G Gazelle, DB Haseman, et al. Acr practice guideline for communication of diagnostic imaging findings. *American College of Radiology*, 2022.

[128] Ben Shneiderman. Human-centered artificial intelligence: reliable, safe & trustworthy. *International Journal of Human–Computer Interaction*, 36(6):495–504, 2020.

[129] Jesper Simonsen and Toni Robertson. *Routledge international handbook of participatory design*, volume 711. Routledge New York, 2013.

[130] Karan Singhal, Shekoofeh Azizi, Tao Tu, S Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, et al. Large language models encode clinical knowledge. *Nature*:1–9, 2023.

[131] Mona Sloane, Emanuel Moss, Olaitan Awomolo, and Laura Forlano. Participation is not a design fix for machine learning. *arXiv preprint arXiv:2007.02423*, 2020.

[132] J Spiegel, M McKenna, G Lakshman, and P Nordstrom. Amazon us patent anticipatory shipping. *Amazon Technologies Inc*, 12, 2014.

[133] Susan Leigh Star and James R Griesemer. Institutional ecology,translations' and boundary objects: amateurs and professionals in berkeley's museum of vertebrate zoology, 1907-39. *Social studies of science*, 19(3):387–420, 1989.

[134] Marc Stickdorn and Jakob Schneider. *This is service design thinking: Basics, tools, cases.* John Wiley & Sons, 2012.

[135] Peter Stone, Rodney Brooks, Erik Brynjolfsson, Ryan Calo, Oren Etzioni, Greg Hager, Julia Hirschberg, Shivaram Kalyanakrishnan, Ece Kamar, Sarit Kraus, et al. Artificial intelligence and life in 2030: the one hundred year study on artificial intelligence. *arXiv preprint arXiv:2211.06318*, 2022.

[136] Simone Stumpf, Vidya Rajaram, Lida Li, Margaret Burnett, Thomas Dietterich, Erin Sullivan, Russell Drummond, and Jonathan Herlocker. Toward harnessing user feedback for machine learning. In *Proceedings of the 12th international conference on Intelligent user interfaces*, pages 82–91, 2007.

[137] Hariharan Subramonyam, Jane Im, Colleen Seifert, and Eytan Adar. Solving separation-of-concerns problems in collaborative design of human-ai systems through leaky abstractions. In *CHI Conference on Human Factors in Computing Systems*, pages 1–21, 2022.

[138] Petra Sundström, Alex Taylor, Katja Grufberg, Niklas Wirström, Jordi Solsona Belenguer, and Marcus Lundén. Inspirational bits: towards a shared understanding of the digital material. In *Proceedings of the SIGCHI conference on human factors in computing systems*, pages 1561–1570, 2011.

[139] Paul Tennent, Joe Marshall, Vasiliki Tsaknaki, Charles Windlin, Kristina Höök, and Miquel Alfaras. Soma design and sensory misalignment. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pages 1–12, 2020.

[140] Jenifer Tidwell. *Designing interfaces: Patterns for effective interaction design.* " O'Reilly Media, Inc.", 2010.

[141] Tao Tu, Shekoofeh Azizi, Danny Driess, Mike Schaekermann, Mohamed Amin, Pi-Chuan Chang, Andrew Carroll, Chuck Lau, Ryutaro Tanno, Ira Ktena, et al. Towards generalist biomedical ai. *arXiv preprint arXiv:2307.14334*, 2023.

[142] Connor W Upton and Fergus R Quilligan. Greybox scheduling: designing a joint cognitive system for sustainable manufacturing. In *CHI'14 Extended Abstracts on Human Factors in Computing Systems*, pages 897–900. 2014.

[143] Philip van Allen. Prototyping ways of prototyping ai. *Interactions*, 25(6):46–51, 2018.

[144] Bart van Dijk and John Zimmerman. Discovering users for technical innovations through systematic matchmaking. In *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems*, pages 1–6, 2019.

[145] Himanshu Verma, Roger Schaer, Julien Reichenbach, Mario Jreige, John O Prior, Florian Evéquoz, and Adrien Depeursinge. On improving physicians' trust in ai: qualitative inquiry with imaging experts in the oncological domain. *BMC Medical Imaging, in review*, 2021.

[146] Joyce Weiner. Why ai/data science projects fail: how to avoid project pitfalls. *Synthesis Lectures on Computation and Analytics*, 1(1):i–77, 2020.

[147] H. James Wilson and Paul R. Daugherty. Creating the symbiotic ai workforce of the future, 2019. URL: https://sloanreview.mit.edu/article/creating-the-symbiotic-ai-workforce-of-the-future/.

[148] Maximiliane Windl, Sebastian S Feger, Lara Zijlstra, Albrecht Schmidt, and Pawel W Wozniak. 'it is not always discovery time': four pragmatic approaches in designing ai systems. In *CHI Conference on Human Factors in Computing Systems*, pages 1–12, 2022.

[149] Richmond Y Wong, Michael A Madaio, and Nick Merrill. Seeing like a toolkit: how toolkits envision the work of ai ethics. *arXiv preprint arXiv:2202.08792*, 2022.

[150] Allison Woodruff, Yasmin Asare Anderson, Katherine Jameson Armstrong, Marina Gkiza, Jay Jennings, Christopher Moessner, Fernanda Viegas, Martin Wattenberg, Fabian Wrede, Patrick Gage Kelley, et al. " a cold, technical decision-maker": can ai provide explainability, negotiability, and humanity? *arXiv preprint arXiv:2012.00874*, 2020.

[151] Austin P Wright, Zijie J Wang, Haekyu Park, Grace Guo, Fabian Sperrle, Mennatallah El-Assady, Alex Endert, Daniel Keim, and Duen Horng Chau. A comparative analysis of industry human-ai interaction guidelines. *arXiv preprint arXiv:2010.11761*, 2020.

[152] Yao Xie, Melody Chen, David Kao, Ge Gao, and Xiang'Anthony' Chen. Chexplain: enabling physicians to explore and understand data-driven, ai-enabled medical imaging analysis. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pages 1–13, 2020.

[153] Shawn Xu, Lin Yang, Christopher Kelly, Marcin Sieniek, Timo Kohlberger, Martin Ma, Wei-Hung Weng, Attila Kiraly, Sahar Kazemzadeh, Zakkai Melamed, et al. Elixr: towards a general purpose x-ray artificial intelligence system through alignment of large language models and radiology vision encoders. *arXiv preprint arXiv:2308.01317*, 2023.

[154] Qian Yang, Nikola Banovic, and John Zimmerman. Mapping machine learning advances from hci research to reveal starting places for design innovation. In *Proceedings of the 2018 CHI conference on human factors in computing systems*, pages 1–11, 2018.

[155] 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*, pages 1–12, 2019.

[156] Qian Yang, Yuexing Hao, Kexin Quan, Stephen Yang, Yiran Zhao, Volodymyr Kuleshov, and Fei Wang. Harnessing biomedical literature to calibrate clinicians' trust in ai decision support systems. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, pages 1–14, 2023.

[157] 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*, pages 585–596, 2018.

[158] 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*, pages 1–13, 2020.

[159] Qian Yang, Aaron Steinfeld, and John Zimmerman. Unremarkable ai: fitting intelligent decision support into critical, clinical decision-making processes. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, pages 1–11, 2019.

[160] 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*, pages 565–576, 2016.

[161] Nur Yildirim, Alex Kass, Teresa Tung, Connor Upton, Donnacha Costello, Robert Giusti, Sinem Lacin, Sara Lovic, James M O'Neill, Rudi O'Reilly Meehan, et al. How experienced designers of enterprise applications engage ai as a design material. In *CHI Conference on Human Factors in Computing Systems*, pages 1–13, 2022.

[162] Nur Yildirim, Changhoon Oh, Deniz Sayar, Kayla Brand, Supritha Challa, Violet Turri, Nina Crosby Walton, Anna Elise Wong, Jodi Forlizzi, James McCann, et al. Creating design resources to scaffold the ideation of ai concepts. In *Proceedings of the 2023 ACM Designing Interactive Systems Conference*, pages 2326–2346, 2023.

[163] Nur Yildirim, Mahima Pushkarna, Nitesh Goyal, Martin Wattenberg, and Fernanda Viégas. Investigating how practitioners use human-ai guidelines: a case study on the people+ ai guidebook. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, pages 1–13, 2023.

[164] Nur Yildirim, Susanna Zlotnikov, Deniz Sayar, Jeremy M Kahn, Leigh A Bukowski, Sher Shah Amin, Kathryn A Riman, Billie S Davis, John S Minturn, Andrew J King, et al. Sketching ai concepts with capabilities and examples: ai innovation in the intensive care unit. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, pages 1–18, 2024.

[165] Nur Yildirim, Susanna Zlotnikov, Aradhana Venkat, Gursimran Chawla, Jennifer Kim, Leigh A Bukowski, Jeremy M Kahn, James McCann, and John Zimmerman. Investigating why clinicians deviate from standards of care: liberating patients from mechanical ventilation in the icu. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, pages 1–15, 2024.

[166] Ming Yin, Jennifer Wortman Vaughan, and Hanna Wallach. Understanding the effect of accuracy on trust in machine learning models. In *Proceedings of the 2019 chi conference on human factors in computing systems*, pages 1–12, 2019.

[167] Sabah Zdanowska and Alex S Taylor. A study of ux practitioners roles in designing real-world, enterprise ml systems. In *CHI Conference on Human Factors in Computing Systems*, pages 1–15, 2022.

[168] John Zimmerman and Jodi Forlizzi. Research through design in hci. In *Ways of Knowing in HCI*, pages 167–189. Springer, 2014.

[169] 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, 2007.

[170] John Zimmerman, Changhoon Oh, Nur Yildirim, Alex Kass, Teresa Tung, and Jodi Forlizzi. Ux designers pushing ai in the enterprise: a case for adaptive uis. *Interactions*, 28(1):72–77, 2020.

[171] John Zimmerman, Aaron Steinfeld, Anthony Tomasic, and Oscar J. Romero. Recentering reframing as an rtd contribution: the case of pivoting from accessible web tables to a conversational internet. In *CHI Conference on Human Factors in Computing Systems*, pages 1–14, 2022.

[172] John Zimmerman, Anthony Tomasic, Charles Garrod, Daisy Yoo, Chaya Hiruncharoenvate, Rafae Aziz, Nikhil Ravi Thiruvengadam, Yun Huang, and Aaron Steinfeld. Field trial of tiramisu: crowd-sourcing bus arrival times to spur co-design. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 1677–1686, 2011.