Designing for Reliability in Algorithmic Systems

By

Samantha Robertson

A dissertation submitted in partial satisfaction of the

requirements for the degree of

Doctor of Philosophy

in

Engineering — Electrical Engineering and Computer Sciences

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Professor Niloufar Salehi, Co-chair
Professor Aditya Parameswaran, Co-chair
Professor Sarah E. Chasins
Dr. Mark Díaz

Fall 2023

Designing for Reliability in Algorithmic Systems

Abstract

Designing for Reliability in Algorithmic Systems

By

Samantha Robertson

Doctor of Philosophy in Engineering — Electrical Engineering and Computer Sciences

University of California, Berkeley

Professor Niloufar Salehi, Co-chair

Professor Aditya Parameswaran, Co-chair

As we introduce complex algorithmic systems into decision-making in high-stakes domains, system designers need principled approaches to help people set their expectations of these systems, and give them mechanisms for recovery when these systems fail. For example, a doctor using a machine learning-based system in clinical care needs to know when they can expect the model to perform well, and when they should not rely on its output. However, supporting these kinds of judgments is difficult when stakeholders have conflicting needs and goals, or cannot directly assess the quality of a system's output. This dissertation examines two such contexts: matching algorithms for assigning students to public schools; and machine translation systems, which use machine learning to translate between natural languages. I take three approaches to design for reliability in these contexts: first, teaching users what a system can and cannot do; second, aligning system evaluations with people's actual use cases, needs, and goals; and third, helping users recover from failures. By making it easier for users to understand what kinds of inputs are supported by a system, and how they can express their intent within that supported scope, we can increase users' agency to make appropriate decisions about how and when to use these systems in high-stakes settings.

For my parents

Contents

List of Figures

List of Tables

Acknowledgments

This dissertation would not exist without the support of Niloufar Salehi, who agreed to advise me as a lost first year graduate student with no experience in HCI. Thank you, Niloufar, for making me a better writer, researcher, colleague, and helping me learn the value of my own energy.

Many other faculty and staff at UC Berkeley EECS have supported me to this point. Thank you Moritz Hardt, Gireeja Ranade, and Marti Hearst for welcoming me to the department. Shirley Salanio, thank you for answering my endless questions and helping me find my own path through the web of requirements and expectations. I will never forget the relief I felt when I was close to tears talking to you about prelims and you said, "You don't *have* to do this right now." Lastly, thank you Aditya Parameswaran and Sarah Chasins for your incisive feedback and support as committee members.

The first half of this dissertation focuses on student assignment algorithms for matching students to public schools. I am grateful for the contributions and support of Niloufar Salehi, Tonya Nguyen, Afshin Nikzad, Cathy Hu, and Catherine Albiston as collaborators and mentors. Much of this work would not have been possible without the guidance and support of Moonhawk Kim, as well as countless others at San Francisco Unified School District, Oakland Unified School District, and Oakland Enrolls. Lastly, thank you to the many people who contributed to this research as formal and informal interview participants.

The second half of this dissertation is about machine translation. It was a privilege to collaborate with Mark Díaz, Timnit Gebru, Meg Mitchell, Samy Bengio, Michal Lahav, Dan Liebling, Kat Heller, and Wesley Deng in the early days of this project. Thank you, Timnit, for helping me secure my first internship, even as you were dealing with your own mistreatment. Mark, I never tired of hopping on a call with you every morning, and it never ceased to amaze me the time and care you put into our work, even while hosting another intern on top of doing your own research. Thank you for continuing to support me as a committee member after my internship.

Fred Hohman and Mary Beth Kery took me on for my second internship, much to my surprise following a cold Twitter DM to Dr. Hohman. Thank you for taking a chance on me. I learned so much from you both, as well as from Jay Wang and Dominik Moritz. Jay, your humility and generosity are stunning for a person of such incredible talents. I am so glad to have had the chance to work with you. This internship changed my life professionally and personally, and I cannot thank you all enough.

Elaine Khoong, Ana Milisavljevic, and Sachin Shah helped me formulate my thinking about reliable machine translation with a concrete application in healthcare. Thank you for your time, support, and insight. Karan Bains and Katrin Jaradeh, your enthusiasm and willingness to help has been invaluable. Lastly, thank you to the graduate and undergraduate students who have contributed to this work: Nikita Mehandru, Kabir Goel, Sylvie Venuto, and Pauline Padrul.

A third line of work on menstrual cycle tracking didn't make it into this dissertation, but nevertheless shaped my graduate school experience and my next steps. Thank you to

Kim Harley and Amanda Shea for helping me get started working with cycle tracking data, and for your patience and support as I formulated my ideas. Thank you Fred Hohman for championing my ideas and finding a way for me to come back to Apple to build an interactive study update for the Apple Women's Health Study – a project no closer to your research focus than to mine. Chris Curry, thank you for your encouragement and support of that work, and for making it as much a fun educational experience for us as it will hopefully be for our readers. Thank you Halden Lin for sitting on countless hours of video calls teaching me how to build something beautiful. I am so appreciative for your time, patience, and your ability to quietly fix everything I broke with a smile on your face, without ever letting me feel like my contributions were any less. Your vision made our work so much better; I'm sorry for the audible groans every time you said "refactor." Fred and Halden, I will take what you both taught me into my career for many years to come.

Thank you to the teachers and mentors who have helped me reach this point. Paul Nuyujukian introduced me to research and helped me see my own potential. Thank you, Paul, for helping me carve out a place in a field where I have not always felt welcome. Bill Behrman shaped my career with Data Challenge Lab, helping me find a love for data work and visualization, and ensuring I never miss an opportunity to annoy my collaborators by using R. Bill, you were right about grad school, but I guess I am very stubborn. Dr. Jane Sloan will be a lifelong inspiration and role model to me, far more than she may know. Karen Pittard believed in me and pushed me to believe in myself more than I realized for many years.

Claire, thank you for your unwavering support from an ocean away. It has been a privilege to watch each other grow, laugh, cry, and thrive from the cup holders of each other's cars. I have been held up and held together through these years by countless other friends from undergrad and grad school: you know who you are. And to my lab mates – Sijia, Tonya, Liza, Sabriya, Angela – we have seen each other through so much. I am forever grateful for the mid-pandemic Zoom catch ups and the slightly-too-long tea breaks in Soda Hall. Look at us now. I can't wait to see what you all have in store next.

Of course, this dissertation is for my parents, who have supported me in every way through 24 consecutive years of formal schooling. Thank you for taking me seriously, while at the same time not letting me take myself too seriously. Dad, thank you for insisting on reading every paper I've ever written, even when you claim not to understand half of it (which, by the way, just means I need to improve my writing... I guess the learning isn't over yet).

Chapter 1

Introduction

Algorithmic systems guide decisions and shape outcomes in a range of high-stakes domains, like education [295, 366], healthcare [165, 349], and the legal system [22, 3, 23]. These systems produce complex outputs, making it difficult for people to detect, understand, and rectify problems, sometimes with severe consequences [349, 48, 22]. The complexity and uncertainty of these systems opens new questions for designing reliable systems: what can we expect from these unpredictable systems? How can do we recover from failures when neither users nor systems can detect them? Answering these questions is a first step towards principled approaches to designing for reliability in algorithmic systems.

Many threats to the reliability of algorithmic systems arise from the interaction between a system and the social environment in which it is embedded. For example, an AI system for detecting diabetic retinopathy had specialist-level accuracy in offline evaluations, but faced new and unexpected challenges when deployed in low-resource clinical settings, e.g., because of difficulties capturing high resolution images [41]. Gaining a full understanding of these types of challenges and developing solutions requires bringing the theories and methods of human-computer interaction (HCI) to the design and evaluation of algorithmic systems. Human-centered approaches are especially critical when there is no shared, clear definition of how a system should behave. My research has examined systems where stakeholders have conflicting goals, and where it is difficult for end-users to directly verify that outputs are correct. My approaches to designing for reliability in these contexts have covered three areas: teaching users what a system can and cannot do; aligning system evaluations with people's actual use cases needs, and goals; and helping users recover from failures when they occur.

The first part of this dissertation focuses on student assignment algorithms: matching algorithms that are used to match students to schools based on students' preferences, schools' priorities, and capacity constraints. School districts have introduced these algorithms with the goal of providing equitable and flexible access to high quality education, but many have struggled to realize these outcomes in practice. Families have been dissatisfied with time-consuming and confusing application processes, as well as disappointing individual and collective-level outcomes. In Chapter 4, I explore the misalignment between the modeling assumptions that underpin claims about these systems' capabilities, and people's actual needs and constraints in the real world. I find that such misalignment can lead to perceived breakdowns even when systems are functioning correctly, which can undermine users' trust and be difficult for system operators to understand and resolve. In Chapter 5, I demonstrate that the solutions to such challenges in sociotechnical systems may often be more social than technical, especially in low-resource settings. Lastly, in Chapter 6, I extend my analysis to other types of systems that elicit and aggregate individual preferences to make collective decisions, arguing that this approach is not a catch-all solution to reliable and equitable algorithmic systems.

The second part of this dissertation focuses on machine translation (MT) models, which translate input from one natural language to another. People typically use these models to translate to or from a language they don't know, making it very difficult for them to identify translation errors. My work has explored how we can help people avoid miscommunication with MT, from the perspectives of both end-users and developers. In Chapter 7, I show that it is very difficult for users to identify and recover from errors when using machine translation models. Users rely on broad perceptions of the strengths and limitations of these models based on past experience, which may not be accurate across different language pairs, or as models are updated over time. It is easy for MT errors to go undetected in communication, with tangible social consequences. In Chapter 8, I explore how we can build tools for MT model developers that encourage them to prioritize model evaluation resources based on how people use their model. Finally, in Chapter 9, I argue that retrieval-augmented machine translation is one way to help users more easily understand a translation model's strengths and limitations, and use the model within its capabilities. I demonstrate the value of this approach for translating hospital discharge instructions.

In Chapter 10, I bring together insights from these two very different systems to identify common requirements and design patterns that can improve reliability. First, we can only assess a system's reliability with respect to a clearly defined task or goal in a specific context. Defining a system's purpose is critical to designing systems with clear capabilities appropriate for that purpose. With a clearly defined purpose and capabilities, it becomes more possible to guide users towards well-supported use cases and inputs. Another way to improve reliability is to define an intermediate intent specification that acts as a shared language between the user and the system. This specification allows users to refine their goals, and then the system must be able to transform this specification into the output language to meet the specified goals with high certainty. This approach is especially useful when it is difficult for a user to directly verify model outputs or rely on their own judgment when a system fails. By making it easier for users to understand what kinds of inputs are supported by a system, and how they can express their needs and goals within that supported scope, we can increase their agency to appropriately rely on that system. This could mean deciding whether to rely on a specific prediction, or contesting systems at a higher level, for example, deliberating the goals of a system, defining expected behavior, determining how resources should be prioritized to improve or extend a system, or questioning whether that system should exist at all.

Chapter 2

Related Work: Designing for Reliability

This chapter defines my interpretation of reliability in algorithmic systems and briefly reviews related approaches to designing for reliability in the human-computer interaction (HCI) and machine learning (ML) literature. This chapter provides a high-level overview of related work in several areas; each chapter following reviews work specifically related to its content in more depth.

The IEEE Standards define reliability as *the ability of a system or component to function under stated conditions for a specified period of time* [221]. A system does not need to have perfect performance to be reliable; rather, reliability is about maintaining an acceptable level of failures for a given use case, in terms of frequency and severity [27].

Therefore, assessing a system's reliability requires three tasks: first, understanding how the system will be used; second, estimating the system's performance in those use cases; and third, deciding if that performance is acceptable for the use case. Misconceptions in any of these steps can lead to actual or perceived system failures. For instance, reliability failures can arise because system performance was measured on use cases that are not representative of actual real world use, or because the reliability requirements were not reflective of people's actual risk tolerance.

Different stakeholders may assess reliability at different times with access to different information, which can lead to different perceptions of reliability for the same system. For instance, end-users assessing reliability may have a very in-depth understanding of their own use case, but their beliefs about system performance may be based on limited information. On the other hand, developers may have sophisticated methods for evaluating model performance, but a misunderstanding about how people will actually use the model.

Many reliability failures do not (primarily) involve humans, e.g. hardware failures. As a human-computer interaction researcher, my interest is in reliability failures that involve end-users, either where users have an inaccurate perception of system performance, or where developers have an inaccurate perception of users' needs and behaviors.

My research has taken three main approaches to addressing these kinds of reliability challenges. The first approach is to *teach users what a system can and cannot do.* This can help set appropriate expectations, and guide decisions about when to use the system. For example, in Chapter 4, I explore perceived failures of student assignment algorithms that arose from misaligned expectations of what matching algorithms can achieve. In Chapter 7, I argue that interactive guidance is needed to help users understand and work within the strengths and limitations of machine translation models. The second approach is to *align system evaluations with people's actual use cases, needs, and goals.* Chapter 8 proposes developer tooling for machine translation that prioritizes evaluation resources based on how people use the system. The third approach is to *help users recover from failures*, which can reduce the consequences of those failures, and in turn increase people's tolerance for failures. In the student assignment context, this requires non-technical interventions, for instance, offering high quality educational opportunities in every neighborhood so that even worst case outcomes are not catastrophic for students. In Chapter 9, I argue that trading off flexibility in order to more clearly specify a system's capabilities could improve error recoverability and allow people to use MT more safely in higher-stakes settings.

Each of these three high-level approaches has been subject to increasing interest in the AI research community, particularly in the Fairness, Accountability and Transparency (FAccT) community and at the intersection of AI and HCI. The following sections review research from these areas related to each of the three approaches.

2.1 Teaching users what a system can and cannot do

Over time, as they interact with a system, people develop a mental model of how that system works. This mental model helps them build appropriate trust and understand how to interact with the system to achieve their goals [345]. A useful mental model need not be entirely accurate or complete, but should allow users to predict how a system will perform under different conditions, i.e., assess its reliability, and control its behavior. However, the unpredictability and complexity of AI systems makes it difficult for users to build useful mental models [356, 17, 449]. The first approach helps users develop useful mental models of system capabilities. Some work aims to help people understand a system's overall performance, while other approaches aim to help people assess individual system outputs.

One way to improve people's overall awareness of system capabilities is to provide more detailed and consistent documentation for models, datasets, and development processes [163, 321, 407, 128, 380, 219, 372, 346]. This documentation can guide decision-making about when and how to use datasets and models, although it is usually directed towards downstream application developers and people with the power to make system adoption decisions for organizations, rather than end-users.

Another line of work aims to improve people's overall AI literacy. One competency within AI literacy involves identifying AI's strengths and weaknesses, and using this understanding to determine when to rely on AI [297]. One way to help users gain such an understanding is to provide onboarding materials specific to a particular system [88, 89]. Cai et al. [89]

found that the process of developing onboarding materials also helps development teams better understand end-user requirements and evaluate their model with respect to those requirements (Section 2.2).

ML explainability and intelligibility research has also aimed to help users understand how systems work and detect model errors [432, 116]. However, it is unclear what kind of information is useful for end-users, or how that information impacts user behavior [462, 293, 319]. Empirical evidence suggests that explainability interventions are more effective for shaping people's overall perceptions of system performance and strategies for using the system, rather than helping them identify specific errors [382, 289]. In practice, explainability methods have seen most success as debugging mechanisms for ML developers, and it remains an open question whether and how these techniques will be useful for a broader audience [52].

My work explores how we can help users set their expectations of system performance overall, and identify and recover from individual failures, even when they cannot directly evaluate and override system outputs. This is especially challenging when there is disagreement about how to define good performance (student assignment algorithms), or when the user cannot understand the system's output (machine translation).

2.2 Aligning system evaluations with people's actual use cases, needs, and goals

The growing list of examples of algorithmic bias and discrimination in deployed systems (e.g., [76, 349, 342, 263, 125, 58]) has demonstrated that traditional machine learning evaluation practices have struggled to capture system behaviors of real-world importance. Hutchinson et al. [218] point out that machine learning evaluations typically treat all model failures the same, despite the fact that different kinds of failures on different parts of the input distribution may have very different real world implications and harms. The second approach aims to ensure that system evaluations reflect realistic use cases and account for people's risk tolerance in those contexts.

One increasingly common practice that partially addresses this concern is disaggregated analysis, which separately measures model performance on different demographic groups or other covariates of interest [76, 35]. Still, available evaluation datasets may not represent the real world distribution of inputs to a model [386]. Developing datasets that are more reflective of real-world use cases requires understanding a model's usage context including how inputs are generated and how outputs are used [218, 41].

Ultimately, a system's impact is heavily shaped by how people interpret and act on its outputs. Field studies and observational studies are an important but still under-utilized methodology for understanding how models fit into realistic workflows and impact outcomes for those who use or are otherwise impacted by a system [41]. Internal and external auditing is another way to improve the quality and trustworthiness of evaluations of systems in deployment settings [388, 449, 387]. A thread through much of this work is that truly rigorous and comprehensive evaluations must be grounded in a deep understanding of the use case for a system, and the associated risks [218, 388].

2.3 Helping users recover from failures

When systems perform unexpectedly (a reliability failure), principles of good user interface design suggest that the system should help users quickly recognize, diagnose, and recover from the breakdown [340]. Researchers have adapted established interaction design principles for AI systems, and common guidance includes allowing users to refine incorrect outputs, and encouraging feedback from users that can improve the system over time [356, 17, 211, 142]. However, achieving these goals in practice is a major challenge for AI user interface designers because model failures can be nuanced and easily go undetected by both users and systems [356, 211, 17, 142]. In some tasks, users have a back-up option when the model is wrong, e.g., they can rely on their own judgment [378]. However, this is not always the case, for example, when people use machine translation to communicate in a language they don't know. The third approach aims to reduce the consequences of failures by helping users easily recover from errors and continue making progress towards their goals, even when they cannot directly verify system output. When errors are more easily recoverable, users may have a higher tolerance for failures, meaning that a system can still be reliable for a given task even with relatively more frequent failures.

One relevant line of work has developed methods for algorithmic recourse. The goal of this work is to give people tools to contest and change unfavorable decisions made by an algorithm [239]. However, for complex (or "truly inscrutable") models Selbst and Barocas [432] warn that such methods could be dangerous because they encourage local tinkering with parameters that will never reflect the system's true underlying complexity, and could instead foster the development of misleading mental models. In Chapter 9, I explore how different model architectures, e.g., retrieval-augmented machine translation, make designing interaction patterns for recoverability easier by making system failures more predictable. In some cases, developers may face a trade-off between system flexibility on the one hand and error predictability and recoverability on the other, which must be balanced on a case-by-case basis given users' risk tolerance for a given task.

Part I

Matching Algorithms for Public School Student Assignment

Chapter 3

Introduction and background

In most public school districts in the U.S., students are assigned to schools based on where they live. Therefore, racial segregation and economic inequalities result in segregated and unequal schools. Increasingly, school districts have been introducing school choice systems that allow students to apply to schools across the district. Students submit a ranked list of schools they would like to attend and the district uses an algorithm to match students to schools based on those preferences. Many school districts implemented these systems for their potential to advance equitable access to high quality education, create more diverse classrooms, and provide flexibility to families [240]. However, school districts have encountered practical challenges in their deployment. In fact, San Francisco Unified School District voted to stop using and completely redesign their student assignment algorithm because it was frustrating for families and it was not promoting educational equity. In my work, I take a human-centered approach and study parents and policy-makers to gain a deeper understanding of their values, attitudes, understandings, and uses of these student assignment systems in practice.

In this chapter, I summarize my contributions, then review existing literature on matching algorithms for student assignment, largely from economics. I conclude by briefly introducing the two contexts in which I have conducted research: San Francisco Unified School District (SFUSD) and Oakland Unified School District (OUSD).

In Chapter 4, I analyze the student assignment system in San Francisco Unified School District using a Value Sensitive Design approach, drawing on 13 interviews with parents and public policy documents, and find that one reason values are not met in practice is that the system relies on modeling assumptions about families' priorities, constraints, and goals that clash with the real world. These assumptions overlook the complex barriers to ideal participation that many families face, particularly because of socioeconomic inequalities.

One of these barriers involves unequal access to information about the schools and enrollment process. Governments and non-profits have invested in providing more information about schools to parents, but we know little about what information is actually useful for historically marginalized and underserved families. In Chapter 5, I present findings from interviews with 14 families and advocates in Oakland focused on the challenges they faced navigating an online school choice and enrollment system. My findings highlight the value of personalized support and trusting relationships to delivering relevant and helpful information. I contrast this against online information resources and dashboards, which tend to be impersonal, target a broad audience, and make strong assumptions about what parents should look for in a school without sensitivity to families' varying circumstances.

Finally, in Chapter 6, I combine the data across San Francisco and Oakland to study how families and school districts use students' preferences for schools to meet their goals. I find that the design of the preference language, i.e. the structure in which participants must express their needs and goals to the decision-maker, shapes the opportunities for meaningful participation. I define three properties of preference languages – expressiveness, cost, and collectivism – and discuss how these factors shape who is able to participate, and the extent to which they are able to effectively communicate their needs to the decision-maker. Reflecting on these findings, I offer implications and paths forward for researchers and practitioners who are considering applying a preference-based model for participation in algorithmic decision making.

3.1 Matching Algorithms for Student Assignment

Economists in the field of market design have developed matching algorithms to find optimal assignments between two sides of a market based on each side's preferences [155, 441]. These algorithms have since been applied to numerous real world markets, such as university admissions, organ donation, and public school student assignment [409]. Abdulkadiroğlu and Sönmez proposed two variants of matching algorithms[1] for assigning students to public schools [2]. The incoming students and available school seats are the two sides of the market. Students report their preferences by ranking the schools, and schools can define priority categories for students, such as priority for younger siblings of continuing students or priority for students living in the school's surrounding neighborhood. These algorithms have promising theoretical properties that should ensure a fair and efficient allocation of seats. For example, they are strategy-proof, meaning students cannot misrepresent their preferences to guarantee an improved outcome. They are also *student-optimal* in the sense that they are optimized to satisfy student preferences as efficiently as possible,[2] subject to each school's capacity constraints [2].

Student assignment systems based on matching algorithms have been championed for their potential to advance equitable access to high quality education, create more diverse classrooms, and provide more flexibility to families compared to a traditional neighborhood

[1] Deferred Acceptance (DA) [155] and Top-Trading Cycles (TTC) [441] are both used for student assignment. Student-optimal DA finds the stable matching that most efficiently satisfies student preferences, while TTC finds a matching that is Pareto-efficient in the satisfaction of student preferences but is not guaranteed to be stable.

[2] For more details about properties of matching mechanisms and matchings, such as strategy-proofness, and trade-offs between stability and efficiency, see Abdulkadiroğlu and Sönmez [2]. For the purposes of this research, it is most important to keep in mind that the primary goal of these algorithms is to satisfy student preferences.

system [240]. However, as these systems have been implemented in the real world they have faced new types of challenges, such as confusion for families and decreasing classroom diversity. Pathak [366] noticed that early theoretical literature overlooked or oversimplifed challenges of practical importance like strategic incentives, transparency, and coordinating offers to improve the efficiency of waitlists. Economists have since applied empirical and experimental methods to understand strategic behavior [194, 238, 395, 135, 134, 179, 357, 177], information needs [105, 201, 178, 105, 117, 78, 21, 300, 195], and diversity constraints [273, 12, 182, 339, 196, 167, 353].

However, researchers have raised concerns that tinkering with the technical implementation of these algorithms is insufficient to improve the enrollment systems overall. Kasman and Valant [240] discussed the strong political forces shaping how these algorithms are used, understood, and accepted in school districts. They argued that these algorithms are easily misunderstood by stakeholders, and that adoption will depend more on how people interact with these systems than their underlying theoretical properties. Hitzig [204] pointed out that matching algorithms' emphasis on efficiency makes strong implicit assumptions about the optimal distribution of assignments, namely that the ideal outcome is the one where every student is assigned to their first choice school. This is often framed in economics as objectively optimal rather than only one of many ways to distribute resources, and one which may not align with school districts' collective goals.

Prior work in HCI has studied human values with respect to matching algorithms in experimental settings [277, 281]. Central concerns for participants included the algorithms' inability to account for social context, the difficulty of quantifying their preferences, and the lack of opportunities for compromise [277]. My research has built on this work, studying stakeholders' values and real-world experiences with respect to high-stakes matching algorithms that has been in use for almost a decade to assign students to public schools in San Francisco and Oakland, California.

San Francisco Unified School District San Francisco has a long history of heavily segregated neighborhoods which has resulted in segregated schools when students attend their neighborhood school [186]. In 2011, San Francisco Unified School District (SFUSD) introduced a student assignment system based on a matching algorithm[3] in the hopes of promoting equitable access to educational opportunity and diverse classrooms [440]. However, by 2018 the Board had voted to redesign the system in response to widespread dissatisfaction among families and clear evidence that the system was not serving the district's goals [186]. Under this system, families could apply to any school in the district, and could list as many schools as they wanted to in their application. The algorithm gave priority to students living near schools that performed poorly on statewide standardized tests [440]. Although this system, in theory, offers equitable access to all of the district's schools, the algorithm had been unable to promote diverse classrooms and equitable access to education in practice, largely due to racial and socioeconomic disparities in participation rates and segregation in

[3] SFUSD uses a variant of Top-Trading Cycles.

the preferences of those families who do participate [439]. By 2018, however, diversity in schools had instead decreased and parents were frustrated by an opaque and unpredictable process [186]. In fact, many schools were now more segregated than the neighborhoods they were in [438]. The algorithm had failed to support the values its designers had intended and the San Francisco Board of Education voted for a complete overhaul and redesign of the system [186]. My work in San Francisco focused on understanding why this system did not meet expectations (Chapter 4).

Oakland Unified School District Oakland Unified School District (OUSD) also has an open enrollment system, where families can apply to any public school in the district and the district uses a matching algorithm to assign students to schools. Unlike San Francisco, families can only rank up to 6 schools on their application, and the algorithm gives top priority to students living in each school's surrounding attendance zone. Almost 90% of OUSD students are students of color, and over 70% are eligible for free and reduced price lunch,[4] but schools have remained segregated and unequal, with several of the highest resource schools serving almost 50% white students. In response to this problem, OUSD is piloting a new priority category to make more space for students who live in poorer parts of the city to attend these high-resource schools. The district is also concerned about families' access to information about schools and the application process, particularly for parents who are new to the district and/or have limited English proficiency. My work in Oakland focused on learning about these participation challenges (Chapter 5).

[4] Source: OUSD Fast Facts 2021-2022

Chapter 4

Modeling Assumptions Clash with the Real World: Transparency, Equity, and Community Challenges for Student Assignment Algorithms

Across the United States, a growing number of school districts are turning to matching algorithms to assign students to public schools.[1] The designers of these algorithms aimed to promote values such as transparency, equity, and community in the process. However, school districts have encountered practical challenges in their deployment. In fact, San Francisco Unified School District voted to stop using and completely redesign their student assignment algorithm because it was frustrating for families and it was not promoting educational equity in practice. In this chapter, I analyze this system using a Value Sensitive Design approach and find that one reason values are not met in practice is that the system relies on modeling assumptions about families' priorities, constraints, and goals that clash with the real world. These assumptions overlook the complex barriers to ideal participation that many families face, particularly because of socioeconomic inequalities. I argue that direct, ongoing engagement with stakeholders is central to aligning algorithmic values with real world conditions. In doing so we must broaden how we evaluate algorithms while recognizing the limitations of purely algorithmic solutions in addressing complex socio-political problems.

4.1 Introduction

Algorithmic systems are increasingly involved in high-stakes decision-making such as child welfare [421, 71], credit scoring [265], medicine [165, 349], and law enforcement [23]. Documented instances of discriminatory algorithmic decision-making [23, 109, 349, 14] and biased

[1] Tonya Nguyen and Niloufar Salehi contributed to the research presented in this chapter, which was published at ACM CHI 2021: `https://doi.org/10.1145/3411764.3445748`.

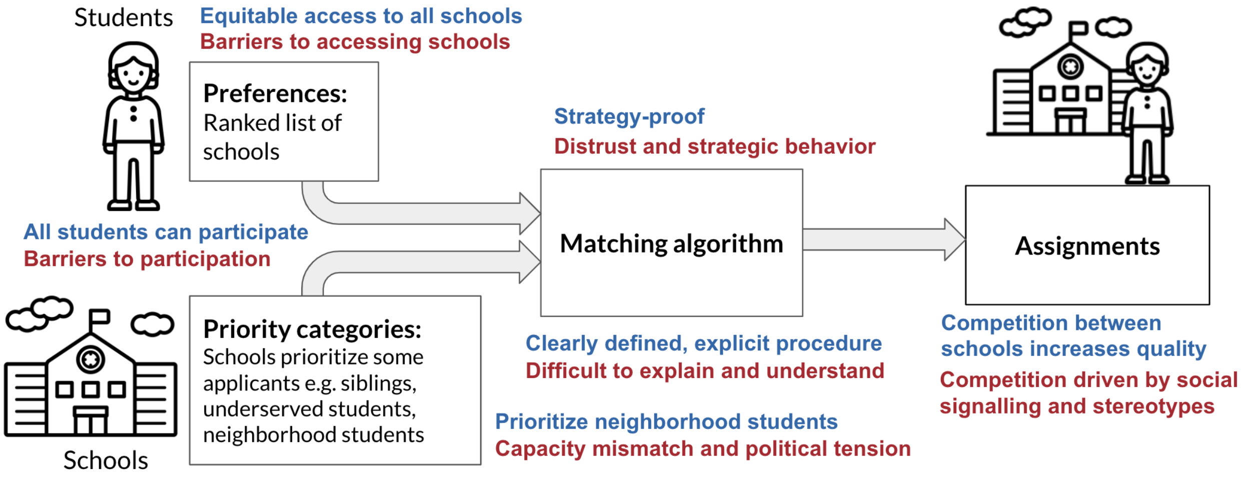

Students
Equitable access to all schools
Barriers to accessing schools
Preferences: Ranked list of schools
Strategy-proof
Distrust and strategic behavior
All students can participate
Barriers to participation
Matching algorithm
Assignments
Priority categories: Schools prioritize some applicants e.g. siblings, underserved students, neighborhood students
Schools
Clearly defined, explicit procedure
Difficult to explain and understand
Prioritize neighborhood students
Capacity mismatch and political tension
Competition between schools increases quality
Competition driven by social signalling and stereotypes

Figure 4.1: Student assignment algorithms were designed meet school district values based on modeling assumptions (blue/top) that clash with the constraints of the real world (red/bottom). Students are expected to have predefined preferences over all schools, which they report truthfully. The procedure is intended to be easy to explain and optimally satisfies student preferences. In practice however, these assumptions clash with the real world characterized by unequal access to information, resource constraints (e.g. commuting), and distrust.

system performance [76, 342, 474, 59] have prompted a growing interest in designing systems that reflect the values and needs of the communities in which they are embedded [154, 11]. However, even when systems are designed to support shared values, they do not always promote those values in practice [506]. One reason why an algorithmic system may not support values as expected is that *these expectations rely on modeling assumptions about the world that clash with how the world actually works.* In this chapter, I examine such a breakdown with the San Francisco Unified School District's student assignment algorithm, to study where and how those clashes occur and to offer paths forward.

In this chapter, I follow a Value Sensitive Design approach to answer two central questions: 1) What values were designers and policy-makers hoping this algorithm would support? 2) Why were those values not met in practice? To answer these questions I first analyzed the school district's publicly available policy documents on student assignment and conducted a review of the relevant economics literature where matching algorithms for student assignment have been developed. To answer the second question, I conducted an empirical investigation into how the algorithm is used in practice. I conducted 13 semi-structured interviews with parents in San Francisco who have used the assignment system and performed content analysis of 12 Reddit threads where parents discussed the algorithm. I complement my qualitative findings with quantitative analysis of application and enrollment data from 4,594 incoming kindergartners in 2017. This triangulation of methods enables us to paint a richer picture of the whole ecosystem in which the algorithm is embedded.

I found that the algorithm failed to support its intended values in practice because it's theoretical promise depended on modeling assumptions that oversimplify and idealize how families will behave and what they seek to achieve. These assumptions overlook the complex barriers to ideal participation that many families face, particularly because of socioeconomic inequalities. Additionally, the system designers vastly underestimated the cost of information acquisition and overestimated the explainability and predictability of the algorithm. In contrast to expectations that the algorithm would ensure an transparent, equitable student assignment process, I find widespread strategic behavior, a lack of trust, and high levels of stress and frustration among families.

Student assignment algorithms promise a clear, mathematically elegant solution to what is in reality a messy, socio-political problem. My findings show that this clash can not only prevent the algorithm from supporting stakeholders' values, but can even cause it to work against them. Human-centered approaches may help algorithm designers build systems that are better aligned with stakeholders' values in practice. However, algorithmic systems will never be perfect nor sufficient to address complex social and political challenges. For this reason, I must also design systems that are adaptable to complex, evolving community needs and seek alternatives where appropriate.

4.2 Related work

In this work I build on two major areas of related work: work in economics on designing and evaluating matching algorithms for student assignment (Chapter 3, Section 3.1); and literature in HCI on Value Sensitive Design (VSD). In this section, I review the VSD literature, then end with a review of literature that examines the role of modeling assumptions in algorithmic systems. I use the term "algorithmic system" or "student assignment system" to broadly refer to the matching algorithm as well as the district's processes and families' practices that make up a part of the application and enrollment process.

Value Sensitive Design

Value Sensitive Design (VSD) is a theoretically grounded methodology to identify and account for stakeholders' values in the design of new technologies [152]. In Value Sensitive Design, "values" are broadly defined as "what a person or group of people consider important in life," although values with ethical import are considered especially important [154]. VSD is a tripartite methodology, involving conceptual, empirical and technical investigations in an iterative and integrative procedure [152]. In the conceptual stage, designers identify stakeholders' relevant values. Empirical investigations examine stakeholders' interactions with the technology and how they apprehend values in practice [124, 153]. Technical investigations explore how the properties and mechanisms of a particular technology support or hinder values. VSD takes a proactive stance: values should ideally be considered early on and throughout the design process [124]. However, VSD can also be applied retrospectively to evaluate deployed systems with respect to human values [152]. I apply VSD methodology to understand what values San Francisco Unified School District's assignment algorithm was designed to support, and why it has not supported those values in practice, leading to its redesign.

Zhu et al. adapt the VSD framework to the design and analysis of algorithmic systems through "Value-Sensitive Algorithm Design" (VSAD) [11]. VSAD emphasizes the need to evaluate algorithms based on whether they are acceptable to stakeholders' values, whether they effectively address the problem they were designed for, and whether they have had positive broader impacts [11]. This is in contrast to traditional evaluation procedures for algorithmic systems, which depend heavily on narrow, quantitative success metrics [11]. Subsequent work has applied the VSAD framework to reveal stakeholder values in the context of a machine learning algorithm used to predict the quality of editor contributions on Wikipedia [453]. The authors emphasize the need to integrate values not only into the design of the algorithm itself, but also into the user interface and work practices that form a part of the algorithmic ecosystem [453]. This is consistent with the interactional principle in VSD, which dictates that "values are not embedded within a technology; rather, they are implicated through engagement" [124].

As VSD has been developed and more widely adopted, researchers have encountered some challenges [124, 60, 275]. One challenge is resolving value conflicts, both between stakeholders with different beliefs [149] and between competing values [447]. However, even when stakeholders agree on important values, it can be difficult to predict whether a technology that supports a value in theory will actually uphold that value when the system is deployed in the real world. Zhu et al. apply VSAD to design and evaluate an algorithm to recruit new editors to Wikipedia communities [11]. They found that their algorithm was acceptable and helpful to the community, but also discovered unanticipated shortcomings. For instance, only more experienced newcomers increased their contributions in response to the recruitment outreach [11]. Ames offers another example of values breakdown, contrasting the intended values of the One Laptop Per Child project, such as productivity, with the consumptive values that were enacted in practice [20].

Researchers have identified various causes of breakdowns between intended values and values in practice. Ames's work highlights the importance of understanding local needs in the context where a technology is to be deployed. Manders-Huits argues that problems can arise when designers misinterpret stakeholders' values, or because stakeholders' values changed over time [311]. Similarly to this work, Voida et al. find that tension arises from a misalignment between how a computational system operationalizes a value and how the people who use the system understand that value [506]. I build on these findings by examining a clash between algorithmic logics and real-world goals and practices. I connect these challenges to emerging work studying the role of modeling assumptions and abstraction in algorithmic breakdown.

Modeling Assumptions in Algorithmic Systems

All algorithmic systems rely on an implicit model of the world in order to compute on it. Any model is a simplified abstraction of reality but the simplifying assumptions often go unstated [64]. For example, Selbst et al. describe the *algorithmic frame* in supervised machine learning, in which each observation in labelled training data represents an abstraction of some real-world entity, often a human being [433]. The authors warn that algorithmic systems can break down if they rely on abstractions that do not capture important aspects of the interactions between technical and social systems. Researchers have documented challenges both when assumptions are too broad, and when they are overly narrow. For instance, Chancellor et al. identified significant inconsistency in how researchers conceptualize and model humans when using machine learning to predict mental health [101]. In contrast, Saxena et al. found an overly narrow focus on risk prediction in the U.S. child welfare system that oversimplifies the complexity of the domain's needs [421].

In the student assignment context, Hitzig identified how matching algorithms rely on an abstraction of the world that makes strong, unstated normative assumptions regarding distributive justice [204], or the appropriate distribution of benefits and burdens in a group. The matching paradigm assumes that the ideal outcome is the one where every student is assigned to their first choice school. Hitzig points out that this emphasis on efficiency may not align with school districts' goals, but is often framed in economics as objectively optimal rather than only one of many ways to distribute resources.

This work demonstrates how unstated, erroneous modeling assumptions about the world can break an algorithmic system. Baumer argues that this breakdown can occur when an algorithm's designers and stakeholders do not share a common understanding of the system's goals and limitations [38]. I expand on this work by exploring how the designers of matching algorithms for student assignment relied on certain modeling assumptions about the world in order to justify their designs with respect to values like equity and transparency. I analyze the breakdown of the student assignment algorithm in San Francisco as a case study of what happens when these assumptions clash with stakeholders' real world goals and constraints.

4.3 Methods

My goal in this research is to understand the values that San Francisco Unified School District's (SFUSD) student assignment system was designed to support and compare and contrast these to parents' experiences in practice. Following Value Sensitive Design methodology [154], I begin with a conceptual investigation drawing on prior literature in economics and SFUSD policy documents to identify the values the system was intended to promote. Then, I conduct a mixed-method empirical investigation to understand why the system ultimately did not support those values and needed to be redesigned.

Data Collection

I collected data from three sources to understand the district's policy goals (how the system was *intended* to work) and parent experiences (how the system has *actually* worked).

District Policies

I collected two official documents from SFUSD to understand the district's policy goals, their justification for their original design in 2011, and the reasons they voted for a redesign in 2018. I accessed the official policy describing the existing assignment system [440] and the resolution that approved the ongoing redesign [186] from the enrollment section of SFUSD's website.[2]

Parent Perspectives

I collected parent experiences in two primary formats: through interviews with parents, and from public online discussions on social media. The interviews allowed us to ask questions and prompt parents to reflect on and dig deeper into their experiences with the assignment system. The online discussions provide potentially less filtered reflections shared without the presence of researchers and reveal how parents seek and share information online. I supplement this data with a presentation titled "Reflections on Student Assignment" by the African American Parents Advisory Council (AAPAC) [6], which was also downloaded from the enrollment section of SFUSD's website.

I conducted semi-structured interviews with 13 parents who have used the student assignment system to apply for elementary schools in SFUSD. I recruited parents through four parenting email and Facebook groups by contacting group administrators who shared a brief recruitment survey on my behalf. During the interview, I asked participants to describe their application and enrollment experiences, and to reflect on their understanding of the assignment algorithm. Interviews were 45 minutes and participants received a $30 gift card. All interviews were conducted over the phone in English between February and August 2020.

12 parents completed a demographic survey. Parents reported their income as low income (1), middle income (5), and upper-middle to high income (4) and identified their race or ethnicity as white (4), Asian (3), Chinese (2), white and Hispanic (1), white and Middle Eastern (1), and Vietnamese (1). The 12 respondents reside in six different zip codes in the city. In all 12 households one or more parents had a Bachelor's degree and in nine households the highest level of education was a graduate degree. To preserve participant privacy, I identify participants in this chapter by unique identifiers P1 through P13.

I supplement the interview data with twelve Reddit threads posted on the r/sanfrancisco subreddit[3] between 2016 and 2020. These threads were selected by conducting a comprehensive search of r/sanfrancisco using the search term "school lottery," as it is commonly known

[2] `https://www.sfusd.edu/schools/enroll/ad-hoc-committee`. Accessed April, 2020.

[3] `https://reddit.com/r/sanfrancisco`

to parents.[4] Each post was reviewed to ensure that it was a discussion of the current SFUSD assignment algorithm. From the twelve threads made up of 678 posts and comments, I manually coded content where the author demonstrated first-hand experience with the assignment algorithm, resulting in a final dataset of 128 posts from 83 contributors. Excluded posts were those that were off topic or presented the author's political view rather than their personal experiences with the system. I paraphrase this content to protect the users' privacy.

Application and Enrollment Data

I complement my qualitative data about parent experiences with publicly available, de-identified kindergarten application data from 2017 to understand higher-level trends in how parents use the system.[5] For each of the 4,594 applicants, the data includes their ranked list of schools, the school they were assigned to, and the school they enrolled in. It also includes the student's zipcode, race, and whether the student resides in a census tract with the lowest performing schools (CTIP1 area), which makes them eligible for priority at their preferred schools. Applicants are 28% Asian or Pacific Islander, 24% white, 23% Hispanic and 3.2% Black. 21% declined to state their race. Approximately 15% of applicants were eligible for CTIP1 priority, 45% of whom are Hispanic. 11% of CTIP1-eligible students are Black, which is 53% of all Black applicants.

Limitations

I recruited interview participants through convenience sampling online and complemented the interviews with existing online data, which biases my data towards those who have the time and motivation to participate in research studies, online discussions, and district focus groups. My dataset lacks sufficient representation of low-income families and Black and Hispanic families. It is important that future work addresses this limitation, particularly considering that integration is a key goal for the school district, and that these families are underrepresented in existing discourses. In future work I will focus on understanding the experiences of historically underserved families with student assignment algorithms, specifically families of color, low-income families, and families with low English proficiency.

Data Analysis

In order to understand the district's values for student assignment and the reasons why the assignment algorithm has not supported these values, I conduct inductive, qualitative content analysis [318] and quantitative data analysis.

[4] Search conducted using the PushShift Reddit repository at https://redditsearch.io/

[5] The data was collected as part of a public records request by local journalist Pickoff-White for a story about how parents try to game the system [370]. The data is available at `https://github.com/pickoffwhite/San-Francisco-Kindergarten-Lottery`.

Qualitative Analysis

My qualitative dataset was made up of district policy documents and community input, interview transcripts, and Reddit content. I performed an open-ended inductive analysis, drawing on elements of grounded theory method [102]. I began with two separate analyses: one to understand the district's values and policies; and a second to understand parent experiences and perspectives. The authors met regularly throughout the analysis to discuss codes and emerging themes. In both analyses I began by conducting open coding on a line-by-line basis using separate code books [102]. I then conducted axial coding to identify relationships between codes and higher level themes. In the axial coding stage for the SFUSD policy documents, I identified three high level codes relevant to my research questions: Values: What are the district's values and goals for student assignment?; Mechanism: How was the district's current system expected to support their values?; and Challenges: Why did the district ultimately decide to redesign the system?. Next, I analyzed parent perspectives from the community input documents, interview transcripts, and Reddit content. I conducted two rounds of open coding. First, I focused only on these three data sources. I identified codes that included "priorities," "algorithmic theories," and "challenges." Then, I linked the open codes from the first round to the challenges identified in the policy documents. I found that challenges parents described in my parent perspectives dataset were relatively consistent with those described in the policy documents and I reached theoretical saturation after approximately ten interviews.

Quantitative Analysis

I linked the application dataset to publicly available school-level standardized test results in order to understand how families use the system to access educational opportunities. I accessed third grade results in the California Smarter Balanced Summative Assessments in 2017-2018, provided by the California Department of Education.[6] I conducted exploratory data visualization to investigate trends in preferences. I measure variation in preferences by race and CTIP1 priority status in order to gain insight into if and how participation varies across groups differently impacted by structural oppression and historical exclusion from high quality education. I present quantitative findings using visualizations to include all students. When comparing summary statistics I use the bootstrap[7] method to estimate statistical significance [65]. For this analysis I used third grade standardized test results as a rough estimate of resources and opportunities at each elementary school. I recognize that there are

[6] Data available at urlhttps://caaspp-elpac.cde.ca.gov/caaspp/ResearchFileList. I link the school achievement data to the applications by state-level (CDS) code. The preference data contains only school numbers, a district-level coding scheme. SFUSD has published a document linking these district school numbers to the school name and state-level (CDS) codes `http://web.sfusd.edu/Services/research_public/rpadc_lib/SFUSD\%20CDS\%20Codes\%20SchYr2012-13_(08-20-12).pdf`.

[7] I use percentile intervals to estimate confidence intervals and the bootstrapped t-test to estimate p-values for differences in means using 10,000 re-samples, following [65]. Groups (race and CTIP1) are re-sampled independently.

many ways in which schools provide value to children that are not reflected in standardized test results.

4.4 Student Assignment in San Francisco: Intended Values

In this section, I present my findings on the values that San Francisco Unified School District (SFUSD) intended their student assignment system to support. In the next section I analyze why this system did not realize those values in practice.

SFUSD has been utilizing different choice-based systems to address educational inequality in the district for almost forty years [438]. Although the mechanism for assigning students to schools has changed significantly over time, SFUSD has been consistent in their values and goals for student assignment. Their current policy designates three primary goals:

1. "Reverse the trend of racial isolation and the concentration of underserved students in the same school;

2. Provide equitable access to the range of opportunities offered to students; and

3. Provide transparency at every stage of the process." [440]

In addition, they emphasize the importance of efficiently utilizing limited district resources, ensuring predictability and ease of use for families, and creating robust enrollments at all schools.

In SFUSD's current assignment system [438], students or their parents apply for schools by submitting their *preferences*: a ranked list of schools they would like to attend (Figure 4.2). To increase flexibility and access to opportunities, students can rank any school in the district and there is no limit on the number of schools they can rank. The district also defines priority categories. Elementary schools give top priority to siblings of continuing students and then to underserved students. Underserved students are defined as those living in neighborhoods with the schools that have the lowest performance on standardized tests, known as *CTIP1* areas. The matching algorithm[8] then takes student preferences and school priorities and produces the best possible assignments for the students subject to the schools' priorities and capacity constraints. Importantly, the resulting assignments from this algorithm are guaranteed to efficiently satisfy *student preferences* not school priorities. School priorities are only used to determine which students are assigned to over-demanded seats. The matching algorithm is also strategy-proof, meaning that it can be theoretically proven that families do not benefit from manipulating their preferences to game the system.

[8] SFUSD uses a variant of the Top Trading Cycles algorithm [441]. See [2] for a technical analysis of Top Trading Cycles in the student assignment context or [409] for a more broadly accessible introduction to market design.

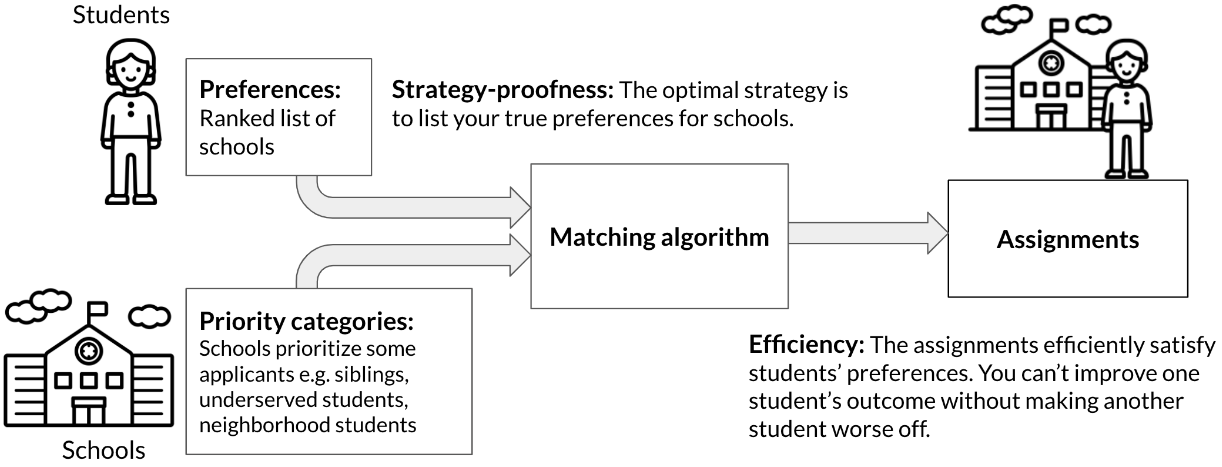

Students
Preferences: Ranked list of schools
Strategy-proofness: The optimal strategy is to list your true preferences for schools.
Matching algorithm
Assignments
Priority categories: Schools prioritize some applicants e.g. siblings, underserved students, neighborhood students
Schools
Efficiency: The assignments efficiently satisfy students' preferences. You can't improve one student's outcome without making another student worse off.

Figure 4.2: The matching algorithm takes students' preferences over schools and schools' pre-defined priority categories as inputs and outputs the most efficient assignment of students to schools.

I consolidated the school district's stated goals for student assignment into four high-level values: (1) transparency, predictability and simplicity; (2) equity and diversity; (3) quality schools; and (4) community and continuity (Table 4.1). In this section, I described the system that was expected to support these values. In the next section, I explore why these expectations were not met in practice.

4.5 Algorithmic Breakdown: Values in Practice

In December 2018, San Francisco Board of Education determined that the the algorithm was not working as intended [186]. While the number one stated goal of the algorithm was to "reverse the trend of racial isolation and the concentration of underserved students in the same school," the Board found that segregation had *increased* since the algorithm was introduced and there was widespread dissatisfaction amongst parents [439, 440]. The assignment algorithm had failed to respect the values that it was designed to support and the Board voted to stop using it and to design a new system. In this section I present my findings that help explain why.

For each of the district's four high-level values for student assignment (Table 4.1), I first review the theoretical properties and promises of the algorithm related to that value: why would economists and district policy-makers expect that the system would respect that value? Next, I analyze what implicit modeling assumptions those expectations depend on. Finally, I explain how families' needs, constraints, and values in the real world clashed with system designers' assumptions about them, which prevented the algorithm from meeting its theoretical promises and enacting the district's values in practice.[9]

Transparency, Predictability, and Simplicity

Theoretical promises

Matching algorithms are clearly and explicitly defined procedures. This differentiates them from assignment systems based on imprecise admissions criteria, which have historically been more difficult to justify and have led to legal disputes [2]. If a student wants to understand why they did not receive an assignment they were hoping for, the algorithm's decision can be explained. Matching algorithms are also provably strategy-proof. That is, students cannot guarantee a more preferable assignment by strategically misrepresenting their preferences. Strategic behavior requires time and effort, so preventing strategic advantages is critical not only for simplicity and efficiency, but also for ensuring that all families can participate equally.

Modeling assumptions: families will accept their assignment as fair and legitimate as long as the algorithm's logic is explained to them.

This assumes that the school district provides an accessible, comprehensible explanation and that families would seek out, understand, and trust this explanation. Families have known preferences for schools and recognize that they should report those preferences truthfully.

Real world challenges

In practice, families find the assignment system difficult to navigate and struggle to find relevant, clear, and consistent information. Some parents engage in strategic behavior to try to improve their child's assignment, contrary to theoretical incentives. Rather than seeking and accepting an explanation, families who are dissatisfied with their assignment seek to change it. Families' trust in the system is eroded by the lack of clear information and the belief that some parents are able to game the system.

Parents face a significant information cost to understand the various opportunities available across the city. There are 72 elementary school programs in SFUSD [186]. Parents indicated that researching schools is a burdensome time-commitment. In-person school visits are a popular source of information when forming preferences, but these visits are time-intensive and logistically difficult.

> *[. . . I]t's like a full time job doing all the school tours. (P7)*

[9] In this work I identify the school district's values and draw on families' experiences to explain why they haven't been supported. The district's values may not completely align with families' values. I assume that satisfying families is one of the district's priorities, and I find substantial overlap between the four district values and what parents in my sample find important. I leave a detailed analysis of families' values to future work.

Table 4.1: consolidated the San Francisco Unified School District's goals for student assignment into four overarching values. Assignment algorithms have theoretical properties aligned with these values. However, the San Francisco assignment algorithm's theoretical promises have not been realized because they rely on modeling assumptions that clash with real world challenges.

Value	*Promises and Properties*	*Modeling Assumptions*	*Real World Challenges*
Transparency, predictability, and simplicity	Algorithm has a clearly defined procedure. Assignments are explainable.	The district provides accessible, clear information. Families want and understand explanations. Families do not try to game the system.	Finding and understanding information is difficult. Some parents try to game the system. There is a lack of trust: assignments are perceived as unpredictable and unfair.
Equity and diversity	Any student can apply to any school. Underserved students are given priority access.	All families participate equally and the all-choice system offers identical opportunities to all families.	Time, language and economic constraints create participation barriers for lower resourced families.
Quality schools	Competition for applicants will drive up the overall quality of schools in the district.	Families base their preferences on accurate estimates of school quality. Schools can respond to competitive pressures.	Competition is driven by social signalling and negative stereotypes. Underserved schools lack resources to attract applicants.
Community and continuity	Priority for siblings and students in the school's attendance area.	Schools have sufficient capacity to handle demand from siblings and neighborhood children.	A lack of guaranteed access to local schools frustrates families living in neighborhoods with very popular schools.

Online information is another widely used source, but school information is not centralized, nor is it consistent across schools. A number of parents mentioned the difficulty of navigating online district resources:

> *[. . . F]inding and gathering the information about the schools from the district is a mess. (P11)*

None of the parents interviewed felt that they had a clear understanding of how the algorithm works. The algorithm is colloquially known to parents as "the lottery." Although the algorithm has only a small lottery aspect to break ties between students with the same priority, many believe it is mostly or entirely random.

> *I'm not really that confident in their actual lottery system. It could be bingo in the background for all I know. (P4)*

This leaves families feeling a lack of agency and control over their child's education.

> *I mean, the word itself, lottery, most of it is random. I don't feel like we can do anything at all. (P5)*

Confused and frustrated by district resources, parents frequently seek advice from other parents online and in-person. Reddit users sought and shared complex strategies, sometimes relying on substantial independent research. This is consistent with prior work showing that advice sharing in social networks can encourage strategic behavior [135, 134]. Advice from other families is often conflicting and unclear, further exacerbating confusion about the system.

> *[W]e also got different advice from different parents. They're very, very different from each other. Some people say, "Put in as many schools as possible," and some people say, "No, just put two schools that you really wanted, and then you have a higher chance of getting those." (P5)*

The 2017 application data indicates that strategic behavior may be more widespread amongst more privileged families. On average, families who were eligible for the CTIP1 priority for underserved students ranked 5.5 schools in their application (95% confidence interval (CI): 5.0–6.2 schools), while families in other areas of the city ranked an average of 11.6 (95% CI: 11.2–12.1 schools; difference in means: $p = 0.00$) (Figure 4.3). 96% of families eligible for CTIP1 priority were assigned their first choice, so this difference may reflect these families' confidence that they will get one of their top choices. On the other hand, it may reflect disparities in access to the time and resources needed to research schools and strategies. White students submitted especially long preference lists (mean = 16.5; 95%

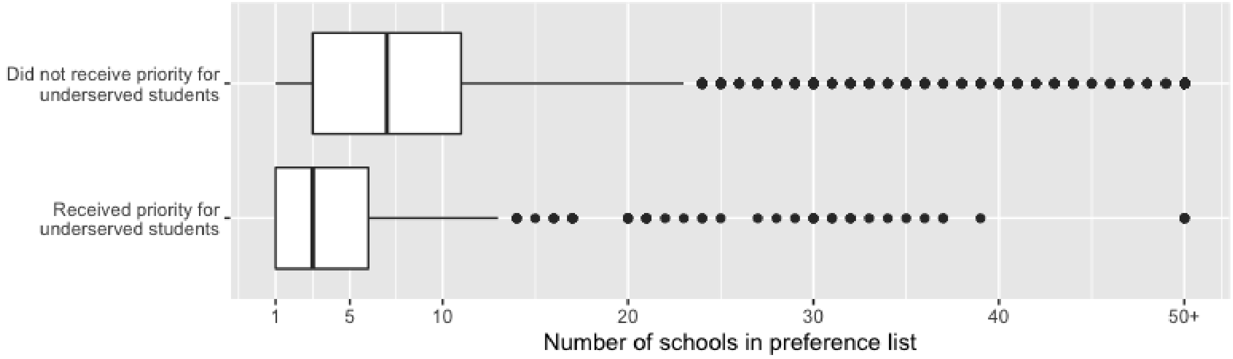


Figure 4.3: Families who were eligible for priority for underserved students ranked fewer schools on average (mean: 5.5; 95% CI: 5.0-6.2) than other families in the city (mean: 11.6; 95% CI: 11.2-12.1; difference in means: p=0.00). This may suggest that stategic behavior is more widespread amongst higher resource families.

CI: 15.6–17.6),[10] a further indication that strategic behavior is more popular with families with more structural advantages.

Receiving an unfavorable assignment was a major concern for families in my sample. The district offers multiple rounds of the assignment algorithm, which many parents participate in if they are dissatisfied with their child's assignment. However, this process can be long, uncertain, and frustrating. Some parents received the same assignment every round with no further explanation or assistance.

> *[. . . T]he first announcement that we got [. . .], I actually wasn't that upset. I said, "You know what, there's more rounds. [. . .] We could stick it out." But I was really upset at the second one because there was literally no change. And that really had me questioning, "I'm just trying to play by the rules. Should I not trust this any more than it's going to work out?" (P9)*

Parents on Reddit recommended unofficial avenues for recourse, many of which require substantial time and resources. These include going in person to the enrollment office repeatedly to request reassignment, remaining on waiting lists up to ten days into the school year, and opting out of the public school system altogether.

Overall, a complicated algorithm together with a shortage of transparent and accessible information has fostered distrust and frustration amongst parents in the district. Distrust is fuelled by perceptions that the system is random and unscientific, and that it allows parents with more time and resources to gain an unfair advantage.

[10] Differences in means between white students and Black, Asian or Pacific Islander, and Hispanic students is highly statistically significant, even with conservative adjustments for multiple hypothesis testing.

> *It's definitely convoluted. It's definitely multilayered, it's complex. And that favors people who have the time and the wherewithal to figure it out. [. . . T]he complexity invites accusations of [corruption] and does not inspire trust (P9)*

Diversity and Equity

Theoretical promises

The assignment system is an all-choice system with unrestricted preference lists, so any student can apply to any school in the district. Compared to a neighborhood system, or even more restricted choice systems, this design has the potential to enable more equitable access to educational opportunity. In an effort to promote equitable access to education and diverse schools, SFUSD has added the CTIP1 priority category, which gives priority admission at over-demanded schools to students from neighborhoods with under-performing schools.

Modeling assumptions: all families participate equally in the system and the all-choice system offers identical opportunities to all families.

CTIP1 students prefer to attend over-demanded schools if they can access them. Applicant pools reflect the racial and socioeconomic diversity of the city.

Real world challenges

Although an all-choice system offers greater flexibility than a neighborhood system, my results show that families with fewer resources face significant barriers to ideal participation in SFUSD's choice system. Although families can rank any school on their application, some families are not able to choose the schools that offer the greatest opportunities. Preferences are segregated by race and income, preventing the algorithm from creating diverse assignments.

My results indicate that the all-choice system does not offer identical opportunities to all families. Every family can apply to any school, but that does not mean that every family can actually access every school. For example, transportation logistics can be a significant challenge. When choosing a kindergarten for their child, P1 met with an education placement counselor at SFUSD to understand the special education services offered across the district. P1 recalled their response to one of the counselor's suggestions:

> *So, you are telling me this school is [. . .] three blocks uphill and we're supposed to do that with a kindergartner and no car? [. . .] There's no way that on my worst day that I would be able to drag my kindergartner with special needs uphill in the rain. (P1)*

The CTIP1 priority is potentially a useful advantage for underserved students. In 2017, 96% of students who were eligible for this priority were assigned their first choice school, compared to 58% of students without this priority. However, CTIP1 priority is only useful for

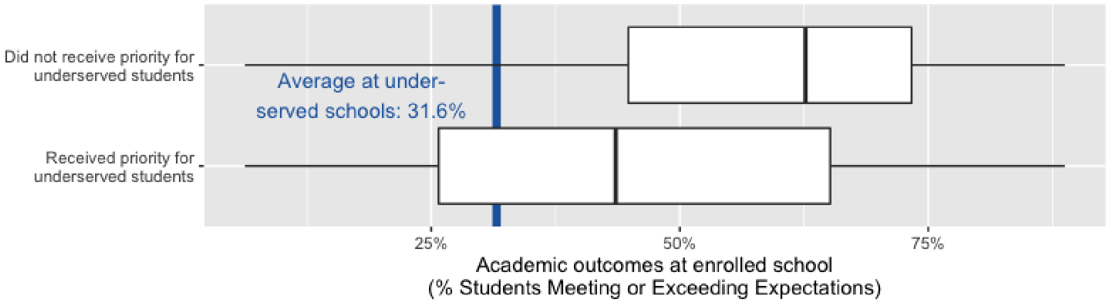


Figure 4.4: The priority for underserved students helps those students access educational opportunity, but there remain inequities that priority enrollment cannot address. Students with priority enrolled in higher performing schools (mean: 45.0% of students met or exceeded expectations on standardized tests; 95% CI: 43.2% – 46.7%), than their average neighborhood school (mean: 31.6%). However, they still enrolled in lower performing schools on average than students who were not eligible for priority (mean: 57.2%; 95% CI: 56.5%–57.9%) (difference in means: p = 0.00). Academic outcomes are measured as the percentage of third grade students at the enrolled school who met or exceeded expectations in the 2017-18 statewide assessments.

advancing educational equity if these students can actually use it to enroll in well-resourced schools. In 2017, students with CTIP1 priority enrolled in schools with lower academic outcomes than other students (Figure 4.4). On average, underserved students enrolled in a school where 45.0% of third graders met or exceeded expectations in the English Language Arts/Literacy exams[11] (95% CI: 43.2% – 46.7%), compared to 57.2% (95% CI: 56.5% – 57.9%) of students at the average school that other students enrolled in (difference in means: p = 0.00). This difference points to persisting inequities in access to higher resource schools that priority assignment is insufficient to address. CTIP1 priority cannot, for example, help students access schools that are physically inaccessible for them. Social factors may also influence choice patterns. For instance, the African American Parent Advisory Council (AAPAC) has raised concerns that Black students in San Francisco continue to face racism and biases in racially diverse classrooms [6].

These findings are consistent with prior work showing that while proximity and academics are important to most families, more privileged parents tend to put more emphasis on a school's academic performance [196, 1, 79], while parents from low-income or racialized backgrounds may be more likely to prioritize proximity [273] or representation of students from a similar background [196]. As a result of differences in students' preferences, applicant pools at schools across the city are segregated by race and income. This prevents the algorithm from creating diverse assignments [186, 273].

[11] Qualitatively similar results to those presented in this section hold for Mathematics results.

Quality Schools

Theoretical promises

System designers have suggested that choice systems indirectly improve school quality. For instance, Pathak argues that matching mechanisms create competition between schools, which pushes under-demanded schools to improve in order to attract applicants and sustain their enrollment [366]. In addition, Pathak points out that an algorithmic system based on student preferences creates a useful source of demand data for the district to target interventions or closures at underenrolled schools [366].

Modeling assumptions: a competitive market will drive up the overall quality of offerings.

This assumes that demand is driven by accurate estimates of school quality.

Real world challenges

Unfortunately, competition in SFUSD has not resulted in an improvement in educational opportunities and outcomes across the district [186]. My findings reveal that parents base their preferences on noisy signals of school quality. Still, some students depend on under-demanded schools and are harmed by under-enrollment and school closures.

My results suggest that parents' preferences are strongly shaped by social learning and stereotypes. Many parents reported using other parents' opinions and experiences of schools to inform their preferences. Some feel that a few schools are disproportionately regarded as the "best" schools in the city. Parents on Reddit attested that many good schools are unfairly dismissed by more advantaged parents, sometimes on the basis of thinly veiled racist and classist stereotypes. Standardized test scores or aggregate scores like those reported by greatschools.org are another popular source of information. Though seemingly more objective, these measure are heavily correlated with resources and demographics at schools [34], further exacerbating preference segregation. In the presence of these types of competitive pressures, well-resourced schools are heavily over-demanded while under-resourced schools struggle to maintain robust enrollments [186]. SFUSD believes the algorithm has created "unhealthy competition" between schools, resulting in schools ranging in size from 100 to nearly 700 students [439].

While Pathak argues that choice patterns are useful in determining which schools to close and which to support and expand [366], this overlooks the correlation between demand patterns and existing patterns of inequality. Under-enrollment and school closures can seriously harm the communities at those schools, which often serve predominantly poor students of color [173, 145]. SFUSD has acknowledged the need to more equitably distribute resources, but it can be politically difficult to direct resources to schools with low demand and enrollment [440].

Community and Continuity

Theoretical promises

SFUSD's sibling and attendance area priority categories are designed to encourage a sense of community and cohesion for families. In addition, students attending PreK or Transitional Kindergarten in the attendance area are given priority to ensure continuity for students.

Modeling assumptions: schools have sufficient capacity to handle demand from siblings and neighborhood children.

Real world challenges

Many families are dissatisfied by a lack of access to their local schools. In many neighborhoods there is a mismatch between demand for the attendance area school and its capacity. In fact, current attendance area boundaries are drawn such that some schools do not have the capacity to serve every student in the attendance area [438]. As a result, the attendance area priority does not provide an acceptable level of predictability for those who want to enroll in their local school.

For parents living in neighborhoods with popular schools, access to their attendance area school is far from guaranteed. One Reddit user expressed frustration after they found out that they may not be able to enroll their child in their local school. Due to their family's circumstances, they feared it would be impossible to get their child to a school further from home.

Parents in my sample value access to local schools for convenience and a sense of community. Under the existing system, two children who live close to each other may attend schools on opposite sides of the city. There are even neighborhoods in San Francisco where students are enrolled across all 72 elementary school programs [186]. Some parents felt that this dispersion undermines the educational experience for children:

> *[I]t is really important for our children to bond and build relationships in their community. And they really connect to their education and their educational environment very differently [when they do]. (P1)*

By underestimating the mismatch between demand for neighborhood schools and capacity at those schools, the assignment system has generated significant dissatisfaction among parents who live near popular schools. These parents are increasingly pushing for a return to a traditional neighborhood system. However, this would restrict flexibility and access to educational opportunities for many families across the city who use the system to enroll their children in schools other than their neighborhood school.[12]

[12] A district analysis showed that 54% of kindergarten applicants did not list their attendance area school anywhere in their preference list for the 2013-14 school year [438]. This is especially true of underserved students: according to the 2017 application data, around 75% of students who received CTIP1 priority

4.6 Design Implications for Student Assignment

In the previous section I showed how incorrect or oversimplified modeling assumptions have played a role in the breakdown of the student assignment algorithm in San Francisco. In this section I draw on these findings to present four design implications for student assignment systems: (1) provide relevant and accessible information; (2) (re)align algorithmic objectives with community goals in mind; (3) reconsider how stakeholders express their needs and constraints; and (4) make appropriate, reliable avenues for recourse available. I emphasize that student assignment is a complex, socio-political problem and my results and recommendations are my first step to better understanding it. In the future, I will continue this work focusing explicitly on the needs of underserved students. In the next section I discuss broader implications of this work for the design of algorithmic systems.

Provide relevant and accessible information

When looking for a school for their child, parents need to find schools that meet their needs, and then understand how to apply. My research shows that information acquisition is very difficult, which leaves families with a sense of distrust and perceptions of randomness, unpredictability, and unfairness. However, more information is not always better. Information about algorithmic systems should be congruent with stakeholder needs and interests and should be limited to the most relevant information in order to minimize cognitive load [130]. In the student assignment setting, I found the most salient information for families is information about the schools available to them that best meet their needs. Relevant, accurate information should be easy to find and navigate. San Francisco Unified School District has recognized this need and has committed to making this information available in a variety of languages [186]. Further work is needed to understand what kind of information about schools will be relevant and helpful without exacerbating negative stereotyping and preference segregation.

Transparency information about the algorithm itself may also reduce stress and increase trust in the system, but only if this information is clear and useful [341, 269, 106]. The algorithmic information most relevant to parents in my sample is their chances of receiving a particular assignment. This information is currently difficult to find, in part because these probabilities depend on others' preferences. However, this information may reduce stress and increase predictability. One concrete goal moving forward could be to ensure that information about schools and admission probabilities are easily available.

enrolled in an elementary school outside of the CTIP1 census tracts. Schools in CTIP1 census tracts were determined according to the definition updated for the 2014-15 school year `https://archive.sfusd.edu/en/assets/sfusd-staff/enroll/files/Revising_CTIP1_for_2014_15_SY.pdf`.

(Re)Align algorithmic objectives with community goals in mind

SFUSD expected their assignment system to satisfy individual preferences *and* promote community-level goals like equitable access to education and diverse classrooms. However, the system has had limited success in promoting educational equity, and racial and economic segregation has worsened since it was introduced [186]. One reason for this breakdown is that the primary objective of matching algorithms is to efficiently satisfy students' preferences, and in San Francisco students' preferences are already heavily segregated by race and income [186]. This indicates a breakdown between community goals and what the algorithm is optimizing for.

The focus on satisfying students' preferences can also obscure other problems. For example, if I look only at preference satisfaction, then underserved students appear to have a strong advantage in the current system. 96% of incoming kindergartners who were eligible for priority for underserved students received their first choice school in 2017, compared to only 58% of other students. However, underserved students continue to enroll in lower resourced schools and an opportunity gap persists between underserved students and others in the district. Due to the limitations of my sample, I cannot conclusively explain the reasons for segregated and unequal preferences. Nevertheless, these two challenges suggest that technical system designers need to work closely with policy-makers and community members to ensure that their algorithm's objectives and evaluation metrics are aligned with higher-level goals and values.

Reconsider how stakeholders express their needs and constraints

Another way to make progress towards community goals is to reconsider how families express their values, needs, and constraints. Matching algorithms model families as independent, self-interested agents with some inherent preferences over schools. Schools are assumed to be merely "objects to be 'consumed' by the students" [2]. However, my findings highlight that preferences are based on limited information and are strongly shaped by social context. Schools are also important communities for children and their families. Researchers have found that matching algorithms for group decision-making do not give participants the space to understand each others' concerns and arrive at compromises that might be natural in a negotiation amongst humans [277, 278]. One avenue for future work is to develop alternative methods for eliciting students' preferences that better reflect their needs and allow for compromise and community building. For example, families could submit their weighted priorities over factors like language programs or proximity to their home. In my interviews I found that parents already make these types of comparisons frequently when researching schools. Such an approach might help shift families' focus from how high their assigned school was in their personal ranked list to how their assigned school meets their needs and constraints and contributes to progress towards community-level goals.

Make appropriate, reliable avenues for recourse available

Because there is limited space at popular schools, some students will receive a disappointing assignment. There are multiple rounds of the algorithm for students who wish to appeal their assignment. However, my results suggest that this process can be frustrating and unpredictable. One concrete recommendation is to improve communication with parents throughout the process about their application status and their available next steps. My findings also suggest that privileged stakeholders will continue to seek unofficial channels to achieve their goals. Therefore, future work developing fair processes for recourse should prioritize the needs of lower resourced stakeholders and design low cost appeals processes.

4.7 Discussion and Future Work

In the previous section, I suggested ways to improve student assignment algorithms to better support stakeholders' values. In this section, I discuss the implications of my work for algorithm design more broadly and identify opportunities for future work.

This work presents an example of how incorrect assumptions can prevent a system from supporting intended values in practice. Direct engagement with stakeholders early on in the design process may help system designers identify incorrect or oversimplified modeling assumptions. For example, economists initially assumed that matching algorithms would be easy to explain to families and that the procedure would be perceived as fair. A value sensitive approach would have encouraged designers to engage with stakeholders early in the development process to gauge their perceptions and acceptance of the technology [11]. Economists may have discovered that stakeholders' acceptance of matching algorithms for student assignment would depend heavily on social and political factors, such as pre-existing institutional trust in the school district.

Even with improved methods to align algorithm design with stakeholders' values, unanticipated challenges will arise because algorithmic systems must rely on *some* abstractions and assumptions that will always be an imperfect approximation of the real world [64, 433]. Crawford analyzed sites of conflict between algorithms and humans, and has warned of the danger of understanding algorithmic logics as autocratic [119]. Instead, algorithmic systems should be accountable to community values beyond the formal design process and stakeholders should have ongoing opportunities to voice concerns, even after the system has been deployed [11]. Future work is needed to design algorithmic systems that are adaptable and flexible in response to this feedback.

In advocating for ongoing engagement with stakeholders, it is important to grapple with differences in power and participation among them [11, 130]. We need to design mechanisms for participation that are equitable and low-cost for lower resource families to voice their concerns [190]. In the student assignment setting, found that convenience sampling strongly skewed my sample of parents towards higher resource parents with the time and motivation to voice their concerns. While building a system that serves all stakeholders is ideal, trade-offs are inevitable when systems impact a large number of stakeholders with diverse perspectives and needs [11, 130]. Avenues for participation should encourage deliberation of trade-offs and include safeguards to prevent powerful stakeholders from compromising important community values in order to design a system that better serves their own interests.

Designing systems while taking into account stakeholders with conflicting values and priorities will require a broader view of algorithmic performance. The research literature on matching algorithms has typically emphasized theoretical guarantees, such as whether assignments are efficient or stable. A human-centered analysis of algorithmic performance would involve evaluating the system in its real world context, along dimensions such as acceptance from stakeholders and broader impacts [11]. This is in contrast to typical practices in algorithmic fields such as machine learning, where algorithms are developed and evaluated with respect to narrow, quantitative metrics such as efficiency. A broader view of algorithmic performance may identify challenges that are central to stakeholders' experiences with the system if not directly related to the algorithm's design, such as the difficulty of forming a preference list.

Finally, we cannot expect that every algorithmic system can support community values if only the right design choices are made. Demand for a technology in the first place is often closely tied to particular politics, which may necessitate certain values and preclude others. For example, education researcher, Scott argues that modern school choice programs reflect a neoliberal ideology focused on empowering parents as consumers of educational opportunities for their child [427]. Advocates claim that school choice promotes educational equity by enabling underserved students to attend a school other than their neighborhood school. Assignment algorithms can support this approach to equity with technical features like priority categories or quota systems. However, this is not the only approach to educational equity. In fact, it offers limited benefits to those who do not have the time or resources to exercise informed choice [429]. A redistributive principle, on the other hand, would prioritize providing underserved students with educational opportunities in their own communities and protecting local students' access to those resources [6]. Assignment algorithms cannot effectively support such an approach: increasing enrollment at under-demanded schools using an algorithm would require violating some students' preferences and may be disruptive and harmful to the existing communities at those schools [6, 174]. Therefore, student assignment algorithms exist within and to uphold a political ideology that privileges individual choice sometimes at the cost of other values, such as democracy, resource equality, and desegregation [429]. This example shows why it is important not only to consider how certain design choices *might* support the values that stakeholders find salient, but also what values a technology *necessitates or precludes* based on the implicit politics of its existence. Value Sensitive Design does not provide an explicit ethical theory to designate what kinds of values *should* be supported [311, 60]. Therefore, in addition to an understanding of implicit values and politics, our analysis must include a commitment to justice [118] and accept refusal as a legitimate way of engaging with technology [111].

4.8 Conclusion

In this chapter I conducted qualitative content analysis of parent experiences and district policies, and quantitative analysis of elementary school applications to understand why the student assignment system in place in San Francisco Unified School District has not supported the district's goals and values. I identify four values that the system was intended to support: (1) transparency, predictability and simplicity; (2) equity and diversity; (3) quality schools; and (4) community and continuity. I identify how the algorithm's theoretical promises to uphold these values depend on assumptions about how stakeholders behave and interact with the system, and explore the ways in which these assumptions clash with the properties and constraints of the real world. I discuss the implications of this work for algorithm design that accounts for complex and possibly conflicting values and needs.

Chapter 5

Not Another School Resource Map: Meeting Underserved Families' Information Needs Requires Trusting Relationships and Personalized Care

Public school districts across the United States have implemented school choice systems that have the potential to improve underserved students' access to educational opportunities.[1] However, research has shown that learning about and applying for schools can be extremely time-consuming and expensive, making it difficult for these systems to create more equitable access to resources in practice. A common factor surfaced in prior work is unequal access to information about the schools and enrollment process. In response, governments and non-profits have invested in providing more information about schools to parents, for instance, through detailed online dashboards. However, we know little about what information is actually useful for historically marginalized and underserved families. We conducted interviews with 10 low-income families and families of color to learn about the challenges they faced navigating an online school choice and enrollment system. We complement this data with four interviews with people who have supported families through the enrollment process in a wide range of roles, from school principal to non-profit staff ("parent advocates"). Our findings highlight the value of personalized support and trusting relationships to delivering relevant and helpful information. We contrast this against online information resources and dashboards, which tend to be impersonal, target a broad audience, and make strong assumptions about what parents should look for in a school without sensitivity to families' varying circumstances. We advocate for an assets-based design approach to information support in public school enrollment, which would ask how we can support the local, one-on-one support that community members already provide.

[1] Tonya Nguyen and Niloufar Salehi contributed to the research presented in this chapter, which was published at ACM CSCW 2022: `https://doi.org/10.1145/3555207`.

5.1 Introduction

Technology increasingly mediates low-resourced and marginalized people's access to social and economic resources, like employment [131, 350, 226, 418], transportation [133], healthcare [191, 416, 192, 175], political power [143, 188, 129], and education [470, 371, 478]. Research has found that factors like financial cost, digital literacy, and trust in online platforms make it difficult to build technology that effectively promotes more equitable access to resources [132]. Another such factor is access to information: people need to be aware of the resources and services that are available to them, and understand how to access them. In this research, I study how low-resourced families access information about public schools and enrollment in a U.S. public school district.

In the United States, low-income students and students of color face systemic barriers to educational resources. In many cities across the U.S., neighborhoods are heavily segregated based on race and income, which leads to educational segregation when students go to their neighborhood school. In response to this problem, a growing number of public school districts have opened up public schools to the entire district and have implemented policies that allow students to apply to whichever public schools they want to attend. Although these systems have the potential to provide lower-resourced students with greater access to high quality educational opportunities, school districts have run into challenges realizing this in practice. In a number of districts including New York City, San Francisco, and Oakland, schools remain segregated and unequal, even with concerted effort and resources committed to maintaining and improving enrollment systems [99, 454, 170]. My work in Chapter 4 and prior research has shown that one source of these challenges is unequal access to information about the available schools due to time and resource constraints [195, 21, 300]. In response, governments and non-profit organizations have developed new sociotechnical infrastructure to provide information to families, particularly through online information and data dashboards.

In this research, I seek to understand whether and how these kinds of interventions benefit low-resourced families. I studied how families navigate the public school enrollment system in Oakland, California. Oakland Unified School District district serves over 50,000 students, of whom 90% are students of color, and 72% are from low-income families [348]. I conducted 14 semi-structured interviews, 10 with parents of color and low-income parents in the district, and 4 with people who work at the school district, individual schools, or local non-profits, and who have supported parents during the enrollment process ("parent advocates"). By bringing together parents' and advocates' perspectives, I gained insight into families' needs and goals, as well as how community members currently support them towards those goals.

I found that the parents in my sample were seeking a high quality education for their child, but faced challenges in reaching this goal, including finding useful and relevant information about the schools available, and making choices between schools that balanced access to well-resourced schools against other considerations like safety, inclusion, and convenience. Consistent with prior work on information support in education [514, 515], I found that parent advocates built trusting relationships with parents to provide personalized support. To provide relevant information, advocates leveraged their deep knowledge of the school system, as well as their familiarity with the challenging circumstances families were facing. In some cases, this included connecting enrollment support to other resources not typically considered in conversations about enrollment, such as food, housing, and healthcare. To support families choosing between schools, advocates avoided overly narrow assumptions about what makes a school a good fit for a family, and pushed for longer-term social change outside of school choice so that every school offers the education and resources its students need.

My findings underscore the importance of recognizing the work that already happens in communities to support people's access to information and resources in order to understand where technology could play a useful role. I advocate for an assets-based design approach to information support, prioritizing interventions that amplify and utilize assets already present in the community [515, 369]. Viewing parent advocates' practices and relationships as assets, an assets-based design approach would ask not how we can design interventions that reduce or replace the work of advocates, but how we can *support and amplify* it [369, 90, 224].

5.2 Background: School Choice Policies and Technology

School choice policies in the United States have been promoted on the premise that families should be able to choose a school that they believe best meets their child's needs, rather than being assigned to a school based on proximity or desegregation plans [427]. In this section, I provide a brief background on this approach, then introduce the public school choice system in Oakland.

The public school system in the United States is shaped by the history of legal and de facto racial segregation in schools and housing [410]. School choice has been closely related to desegregation efforts, first as a way for white families to avoid integration [202], and later as an alternative to centralized redistricting or busing plans [254]. These policies can take many different forms, from subsidized private school vouchers, to charter schools[2], to transfers between public schools within a district or across district boundaries. The movement towards prioritizing parents' individual choice of schools over centralized decision-making is reflective of a broader shift away from the redistributive principles of the civil-rights era towards marketization and individualism, values characteristic of neoliberalism [428, 62].

Given this history, the potential for school choice to advance social and racial justice in education has always been extremely controversial. Today, schools across the country remain heavily segregated by race and class, with large disparities in educational opportunities and resources available [354, 442]. Proponents of school choice argue that it has the potential to improve outcomes for underserved students by allowing them to choose higher-resourced schools outside of their neighborhood, and relying on market pressures to push lower performing schools to improve [195]. Others disagree, arguing that these policies can exacerbate

[2] Charter schools are privately run schools that receive public funding and are free to attend.

segregation [512], drain resources at already struggling schools [145],[3] and place too much responsibility on individual parents for the quality of their child's education [429]. In this work, I focus on an existing choice plan, and study how the technology designed to support this plan enables low-resourced and marginalized families' participation.

Case Study

In this chapter I study the public school enrollment system in Oakland, California. Oakland Unified School District has an intra-district "open enrollment" system, which means that families can apply to any public school in the district. Oakland also has more than 30 charter schools, which have a similar enrollment process to district schools. This type of school choice policy is very common in cities across the U.S., for instance, similar processes exist in New York City, Chicago, New Orleans, and San Francisco.[4] The school district serves a student population that is 90% students of color (48% Latinx and 22% African American) and 72% low-income[5] [348]. Around 40% of low-income students are learning English [170]. As is the case in many other major U.S. cities, schools remain segregated and unequal across racial and socioeconomic lines. Low-income students are overrepresented at over half of the schools in the district, and the majority of Black and Latinx students attend one of these high-poverty schools. Meanwhile, several schools in wealthier neighborhoods have close to 50% white students.

To apply for schools,[6] families submit a ranked list of the schools they are interested in attending. The school district assigns seats using an algorithm that is designed to maximally satisfy students' preferences, determining how to assign overdemanded seats using the school district's priorities, which include priority for siblings of current students and students who live near the school [1]. Students are not guaranteed admission to their preferred school, and schools give priority to students living in a surrounding zone, meaning that the most popular schools admit very few students from other neighborhoods. Recently, the school district piloted a new equity-oriented priority category at three heavily overdemanded schools, which prioritized applicants living in areas with majority low-income African American or Latinx families.

For this system to work towards equitable outcomes, historically underserved families need to be fully informed about their options, choose to apply to higher-resourced schools than their neighborhood school, and be admitted to one of those schools [195, 203]. The school district has invested in infrastructure to support families participating in this process.

[3] In the U.S., public schools receive funding per student.

[4] Allowing choice among district schools has in part been forced on school districts by a trend in federal and state policies towards increased standardized testing and reliance on mandated school transfers as an accountability mechanism [203]. For example, under the No Child Left Behind Act of 2001, students attending schools that underperformed on standardized tests three years in a row were eligible to transfer to a higher performing school in their district [203].

[5] Percentage of low-income students based on enrollment in the Free and Reduced Lunch program.

[6] I describe the application process for district-run public schools. The application for charter schools is separate but very similar to the process for district-run schools.

For example, the frequently asked questions page on the district's website encourages families to research the available schools using a dashboard-style school finder website, similar to that provided by GreatSchools.[7] The dashboard shows the available schools on a map, and then for each school includes a blurb about the school, statistics on the school's size, demographics, and academic performance, as well as a list of special programs offered, start and end times, and information about how to apply (Figure 5.1). Families can also get personalized recommendations by answering a brief survey about their child's incoming grade, home address, priorities for offerings like before or after school care, language immersion, special interest programs (e.g. STEM, arts, sports), and learning styles, and whether they care more about proximity or program offerings. Once families have narrowed down the schools that they are interested in, they are encouraged to look for more information on the school's website, or attend a school tour. Finally, families can submit an application online by submitting a ranked list of up to 6 schools. The school district also has a dedicated office that provides enrollment support in-person or over the phone. In this work, I seek to explore how this socio-technical system facilitates or inhibits low-resource families' access to educational resources.

5.3 Related Work

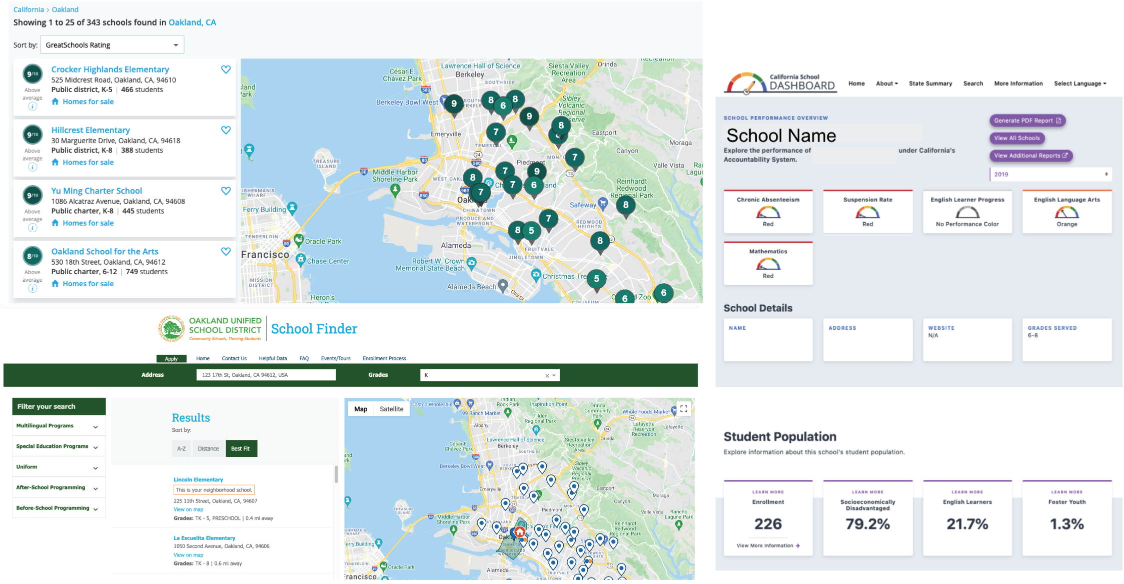
By studying the challenges that low-resourced and marginalized families face in navigating public school choice systems, I build on a body of work in HCI and CSCW that aims to advance social justice by broadening people's access to resources. After an overview of the broader context of HCI for social justice, I focus on challenges and opportunities related to information access in lower-resourced communities, and how this manifests in the context of school choice.

HCI for Social Justice

Increasingly, HCI researchers have been interested in applying their design and research methods to address large scale social problems [137, 37]. This work is especially challenging because its central questions are complex, political, and often have no correct or definitive answers. Research has shown the importance of designing with the specific needs and constraints of low-resourced people in mind. For example, ridesharing technologies have great potential to actually improve people's mobility, and by extension their access to employment and other resources. However, Dillahunt et al. found that lower-resource users face significant barriers to fully realizing these benefits, such as cost, low digital literacy, and a lack of trust in the platform [133]. Similarly, prior work has shown that technologies to support job seekers often fail to serve people with limited resources or education [132, 350].

Even technologies that were intentionally designed to improve people's social, economic, or political position can fall short of this goal. Prior work has explored how technology can

[7] https://www.greatschools.org/



Showing 1 to 25 of 343 schools found in Oakland, CA

9/10 Above average — Crocker Highlands Elementary
525 Midcrest Road, Oakland, CA, 94610
Public district, K-5 | 466 students

9/10 Above average — Hillcrest Elementary
30 Marguerite Drive, Oakland, CA, 94618
Public district, K-8 | 388 students

9/10 Above average — Yu Ming Charter School
1086 Alcatraz Avenue, Oakland, CA, 94608
Public charter, K-8 | 445 students

8/10 Above average — Oakland School for the Arts
530 18th Street, Oakland, CA, 94612
Public charter, 6-12 | 749 students

Address: 123 17th St, Oakland, CA 94612, USA
Grades: K

Lincoln Elementary
This is your neighborhood school.
225 11th Street, Oakland, CA, 94607
Grades: TK - 5, PRESCHOOL | 0.4 mi away

La Escuelita Elementary
1050 Second Avenue, Oakland, CA, 94606
Grades: TK - 8 | 0.6 mi away

SCHOOL PERFORMANCE OVERVIEW
School Name
Explore the performance of under California's Accountability System.
2019

Chronic Absenteeism	Suspension Rate	English Learner Progress	English Language Arts
Red	Red	No Performance Color	Orange

Mathematics
Red

School Details

NAME	ADDRESS	WEBSITE	GRADES SERVED
		N/A	6-8

Student Population
Explore information about this school's student population.

Enrollment	Socioeconomically Disadvantaged	English Learners	Foster Youth
226	79.2%	21.7%	1.3%

Figure 5.1: Three online dashboards available to families in Oakland. Top left: GreatSchools displays schools on a map, ranked based on a proprietary score out of 10. Each school has a separate page with more detailed information including the breakdown of the overall score into an academic progress score and an equity score, student demographics, and test score data disaggregated by race, income, and disability. Right: The California Schools Dashboard provides statistics on every public school in California, including standardized test scores, suspension and attendance rates, and student demographics. Bottom left: The Oakland SchoolFinder gives families the option to find their neighborhood school, filter by language programs, special education programs, uniforms, and before and after school care, and view all schools on a map.

support crowd workers to increase their income and improve their working conditions [226, 418], help low-income people and immigrants share information and emotional support [214, 228], and offer avenues to broaden civic engagement and participation in local politics [143, 188]. One common theme in this work is that technological systems are limited in their ability to foster trust, build community, and shift power relations, all of which are critical to lasting social change. For instance, Erete and Burrell found that communities used a variety of technologies to engage in local politics, but existing structural inequalities shaped the impact of their practices [143]. The authors found that wealthier, white communities were more likely to have their voices heard by local politicians than communities of color, proving that technology alone cannot equitably empower citizens. I build on this work by studying how lower-resourced families use a system that was intended to support their access to educational resources.

Unequal Access to Information

Information costs are one central barrier that low-resourced people face when accessing resources and support [228]. A large body of work in HCI has studied how people use technologies to search and navigate information (e.g., see [258] for an overview). However, research with lower-resourced people has shown that making information more easily available is not necessarily sufficient to support their information-seeking needs. For instance, Israni et al. found that social norms and a lack of institutional and interpersonal trust can deter low-income people from participating in an online social network to find and share information, even when that network was a closed community of other low-income people run by a trusted organization [228]. Their findings were in line with prior theories of information sharing in marginalized communities, which suggest that low-resourced people prefer to seek information and resources from trusted, close relationships, such as family or friends, because these people are more likely to understand their context and less likely to judge or patronize them [103, 244].

As digital technologies like parent portals and free online learning resources have become more popular in education, research has studied how these tools impact lower resourced parents' engagement in their children's education [136, 515, 514, 305, 516, 107, 248]. For example, DiSalvo et al. found that the increasing prominence of online educational tools and social networks has increased inequality between children of lower and higher socio-economic status, due to differences in information seeking practices [136]. One potential source of these differences is that lower-resourced and marginalized people employ culturally-specific strategies that may not align with dominant approaches to improving educational outcomes [515]. Indeed, schools often standardize communication and materials to project a sense of equality [514], but the same resources do not benefit families equally based on their social, economic, and cultural position [515, 136]. Wong-Villacres et al. explored how parent liaisons carefully craft patchworks of information to include and engage Latinx immigrant families in U.S. schools, leveraging their bi-cultural knowledge and social networks [514]. I build on this work by exploring how these dynamics play a role in public school enrollment, and how centering the work of parent liaisons (or advocates in this work) can contribute to more effective support for low-resourced and marginalized families.

Information and School Choice

There is growing recognition that participating in school choice is costly for families, who may face an overwhelming number of choices, often with little prior knowledge of the available schools or what they should be looking for in a school [105, 117]. A major concern with these systems is that information barriers may disproportionately impact lower-resource families, thus limiting the potential for choice policies to improve educational outcomes or promote integrated schools [117, 78].

Researchers, primarily in economics, have thus been interested in how information interventions could influence families' applications to schools. Several experimental studies have found that when low-income and non-English speaking families are provided with more information about schools with high test scores or graduation rates, they are more likely to apply to and enroll in one of those schools [21, 117, 300, 195, 105, 49]. This kind of data about schools is increasingly available to parents, as government and non-profit organizations have produced tools to make school accountability data available on a larger scale, e.g., GreatSchools.org[8] or the California Schools Dashboard[9] (Figure 5.1). However, it is not clear whether these tools can actually broaden participation in school choice. One concern is that more well resourced families use and benefit from any available information more than lower resourced families [117, 425], as has been the case with other open educational resources [136]. Another is that emphasizing quantitative measures, especially test scores, disadvantages schools that serve more low-income students and English learners [34]. This could exacerbate segregation if more privileged families use this information to avoid these schools [55, 49].

While these concerns have been particularly prominent, neither centers the needs and experiences of lower resourced families. Qualitative and quantitative evidence shows that although parents across social groups value academics when choosing schools, on average, low-income families and families of color have fewer high quality options practically available, leading them to choose apparently lower performing schools [79, 1]. Conflicting factors include proximity to home [195, 78], representation of peers from similar ethnic or socioeconomic backgrounds [79, 425], and likelihood of admission at heavily oversubscribed schools [117]. For example, Hastings and Weinstein ran a field experiment where they sent families information about schools with higher test scores than their current school, and found that families were more likely to apply to a school with higher test scores if there were options close to their home [195]. Corcoran et al. gave students at high poverty middle schools a list of nearby high schools with above median graduation rates, and included information about the students' chance of admission. While the students, on average, did not apply to schools with higher graduation rates, they did use the information to choose schools where they had a higher likelihood of admission, reducing their chances of matching to a school with a very low graduation rate [117].

Taken together, this work highlights the gaps remaining in our understanding of how lower-resourced and marginalized families navigate public school choice, and how technologies, especially relating to information about schools, could best support them. In this work I aim to further this goal by bridging economists' analyses of the role of information in school enrollment and research in HCI and CSCW on marginalized parents' information seeking.

5.4 Methods

My goal in this chapter is to understand how low-resourced and marginalized families navigate the public school enrollment process. In particular, I was interested in understanding the

[8] https://www.greatschools.org/

[9] https://www.caschooldashboard.org/

challenges that parents face in finding information about schools, completing the application, and enrolling in a school.

Research Partnership

I began this research in collaboration with staff at the school district and an organization that runs charter school enrollment. I met with these partners weekly to identify research questions and develop a recruitment strategy over the course of several months. My collaborators were especially interested in how their information infrastructure could better support families who are harder to reach through their existing outreach strategies (e.g., online feedback surveys, outreach to parents at school sites, or hosting in-person outreach events), specifically low-income families of color and parents/guardians with limited English proficiency.

Data Collection

I conducted semi-structured interviews with 10 parents and 4 parent advocates. The parent advocates were people who worked to support parents through the enrollment process as staff at a school, the school district, or a non-profit. Three of the four parent advocates had also enrolled their own children in public schools using the same enrollment system, giving them a mix of parent and advocate perspectives. To preserve participant privacy, I identify participants in this paper by unique identifiers: P1 through P10 for parents, and A1 through A4 for parent advocates.

Recruitment

In order to target lower resource families, I distributed a brief screener survey in English and Spanish that asked for the respondent's race, zipcode, the year they most recently enrolled a student in a public school, and how easy or difficult they found the enrollment process (5-point Likert scale).[10] My recruitment materials also included a phone number that prospective participants could text for more information. The charter school enrollment organization and a number of community-based non-profits shared a recruitment flyer on my behalf through their social media accounts.

P1, P2, and A1 were recruited through this recruitment survey, but responses to this survey were very slow, and most parents who filled it out did not respond when I reached out to set up an interview. In light of this, I turned my attention to building stronger connections with people who work with families in my target groups. I met people who worked with families in a wide variety of capacities, from people who specifically provide enrollment support, to school leadership and staff, to education-related non-profits and activist organizations. Through this process I met and interviewed A2-A4, and I had informal conversations with several other people. While my initial target sample was parents or others

[10] Since responses were sparse, my only exclusion criteria was applied to white parents. The exception to this criteria was A1, who I recruited for his unique perspective as a school principal.

who participated in enrollment, I quickly realized that these community members played a key role in informational support. Thus, my analysis of families' experiences navigating this process is enriched by the perspectives of those who support them. Throughout the paper I refer to this group of participants as "parent advocates" for simplicity. However, while there are some commonalities across their roles and perspectives, each brings a unique perspective.

Interview protocol

During the interview, I asked parents to describe their experiences applying for schools and enrolling their children. I was particularly interested in what parents wanted in a school, how they found information about schools, and their experiences submitting the application and receiving results, but I encouraged participants to speak about whatever part of the enrollment process was most salient for them. I asked parent advocates to describe their work, the challenges they face, and how they felt the enrollment system could better serve the needs of the families they work with. Interviews were between 30 and 50 minutes and participants received a $30 gift card. Interviews were conducted over the phone in English and Spanish (with an English-Spanish bilingual interpreter) between September 2020 and March 2021.

At the end of the interview, I asked parents a set of open-ended demographic questions. Parent demographics are shown in Table 5.1. As the research proceeded I adjusted the demographic survey in an effort to respect participants' comfort level while also ensuring I reached a diverse set of lower resource participants. For this reason, the information in Table 5.1 is not consistent for every participant. For instance, in earlier interviews I asked for participants' home zipcode as a proxy for income, while in later interviews I asked participants for their highest level of education, whether they were employed and, if they were employed, whether their income was above or below the city's median. Table 5.2 shows parent advocate demographics and a description of their role working with parents.

Limitations and Opportunities

My sample in this paper is small because I spent significant time and effort recruiting participants and building relationships with community groups. Certainly, my data does not represent all of the different experiences of families of color, low-income families, or immigrant families in the school district. By interviewing parent advocates and some of the parents they have worked with, my sample represents the perspectives of families who needed more support and were able to seek that out. My findings should not be taken to indicate that all families rely heavily on parent advocates. Despite these limitations, I believe the findings provide detailed insight into a subset of families' experiences, and are indicative of some challenges that lower-resourced and marginalized families can face. Further, my findings are consistent with prior research into marginalized families' experiences and information practices in other areas of public education [514].

Table 5.1: Parent demographics. Interviews with participants marked with † were conducted in Spanish and English with an interpreter. Incomes marked with * are estimates based on median income in home zip code, all other fields are self-described.

	Race/Ethnicity	**Income**	**Education**
P1	African American	Low*	-
P2	Filipino	Low*	-
P3	African American/Black	Low*	-
P4	-	-	-
P5†	Honduran	Low*	-
P6†	Latino	Below median	Middle school
P7†	-	Unemployed	None
P8†	Guatemala	Unemployed	2nd grade
P9	-	-	-
P10	African American	Below median	Some college

Table 5.2: Parent advocate roles and the target groups they serve. Race or ethnicity was self-described, the authors inferred role descriptions and target groups from the interview transcripts. Three of the four advocates (marked with ‡) have also enrolled their own children in public schools, and shared perspectives in their interview both as a parent and a parent advocate.

	Race/Ethnicity	**Role**	**Target groups**
A1‡	White	School principal	School is 50% low-income students
A2‡	-	Non-profit staff	Primarily Black and Latinx families
A3‡	Latino	School-Parent liaison	Primarily Spanish-speaking families
A4	White	School district staff	Recently arrived refugee families

Future work should engage with families from a wider range of backgrounds, including people who speak other languages, and people who did not receive enrollment support. I use the term "parent" throughout this paper because it accurately describes my participants, but there are other people who participate in the enrollment process, like foster guardians and older students. School districts wishing to replicate this work should define target groups of families and engage with families in each one in order to understand the range of local needs.

Data Analysis

I analyzed the interview transcripts using inductive, qualitative analysis [318]. Working from prior literature, the first author developed a code book containing 32 codes identifying sources of information that parents use and challenges they face when enrolling their students. These codes mostly fell into three higher-level categories: information (e.g. "online," "school visits"), factors of consideration for schools (e.g. "test scores," "convenience," "safety"), and broader concerns about the process (e.g. "finding a good school," "systemic racism"). Next, I conducted open coding on a line-by-line basis [102]. The first two authors worked together to code two transcripts this way, resolving disagreements through discussion. Each author then analyzed half of the remaining transcripts. The initial codebook was used as a reference, but codes were adjusted, added, and removed as necessary to best fit the data. At the end of this process, the two authors discussed their findings and again resolved disagreements. The final codebook contained 39 codes. There were two natural groupings of codes ("finding information" and "considering priorities"), which contained 25 of the 39 codes. Other codes included "building relationships," "worrying about availability/scarcity," and "voicing concern." Next, I conducted axial coding to identify relationships between codes and extract higher level themes.

Parents' experiences varied widely based on their prior experiences, access to social and economic resources, and the level of support they received from other people, and were further shaped by intersecting identities of class, race, language, and education. I do not mean to imply that "low-resourced" or "marginalized" are fixed or stable categories. To the contrary, many parents I interviewed faced a mix of experiences of privilege and marginalization at different points when participating in school choice [120]. I strive to avoid artificial distinctions when presenting my results, and instead aim to present an authentic account of participants' experiences, using their own words as much as possible.

Research Approach and Researcher Positionality

In this work I sought to engage low-resourced and marginalized parents to understand how they navigate the public school enrollment process. Prior work has discussed the challenges for researchers engaging with low-resourced populations in equitable and just partnerships. For example, Harrington et al. discussed how collaborative design workshops with underserved communities can cause harm when researchers are not sensitive to the communities' historical context and experiences of oppression [190]. In order to minimize these kinds of harms, researchers have called for new approaches and methodologies that account for historical contexts and systems of oppression (e.g. [137, 118, 190, 225, 351, 33, 288]). In this section, following feminist methodology, I reflect on my approach to this research and the ways in which my position in the world, my goal, and my belief shape my interpretation of the findings [33].

Prior work has emphasized the importance of working with trusted community members to ensure that research with marginalized communities is conducted in culturally appropriate ways, and to gain legitimacy and trust with participants [190, 118]. The online screener survey approach was not effective, possibly because parents did not trust an impersonal online flyer, did not have the time to fill out a screener survey, or were hesitant to participate in research they felt would not serve them. I was also sensitive to the additional burden of the ongoing COVID-19 pandemic, during which many parents had increased childcare responsibilities, and which had a disproportionate and devastating impact on Black and Latinx communities in the United States [271, 517]. Reflecting on this process, it would have been better to begin by engaging not only with the school district, who hold significant institutional power and whose interests may not represent those of marginalized community members, but with various community leaders to develop research questions and methods. I acknowledge that had I taken this approach, this research may have looked very different in terms of central questions and methods.

Neither myself nor my closest collaborators in this work come to this work with first-hand experience of a public school choice system, neither as parents nor students. I am a white woman from a class privileged background and does not experience the forms of marginalization that I discuss in this chapter. One collaborator is a first-generation college student and comes from a family of Vietnamese refugees. The other collaborator is an immigrant and woman of color who attended public school at her home country. My lack of shared context with participants likely shaped my approach to this work, the kinds of experiences and insights that participants shared, as well my interpretation of those that they did share. Our team has a range of disciplinary backgrounds, but primarily adopt an HCI lens to study a specific sociotechnical system situated within much broader conversations about education and social justice.

Certainly, my perspective has been influenced by working in partnership with staff members at the school district and charter school enrollment system, although I have not consciously shaped the findings in any way to meet their expectations or censor families' negative experiences. I have shared the findings back to the district partners and parent advocates to verify my interpretations and contribute to ongoing discussions regarding improvements to their enrollment process that could promote more equitable participation and access to schools.

5.5 Results

Next, I present my findings about low-resource families' information needs in the school enrollment process. I organize my findings based on the two major stages of the application process: finding information about the available schools, and forming a ranked list of schools. First, I summarize the challenges that parents faced, then I highlight parent advocates' strategies to provide support.

Available information lacks relevance and personalization

When they are first participating in school choice, parents need to find out how the system works and learn about their options. My findings indicate a mismatch between the information families need and the information that is easily available. Further, families are faced with a potentially overwhelming number of options, which exacerbates the challenge of finding relevant and helpful information. To address this challenge, parent advocates build trusting relationships with families and learn about their specific needs and circumstances. This enables them to make personalized recommendations and provide more holistic support for families.

Challenges

Families face challenges learning about the choice process and why they may want to participate, and finding useful information relevant to their priorities. On top of this, they are faced with a potentially overwhelming number of options to learn about.

Before they even begin to evaluate different schools, families need to be aware that they have a choice to apply to different schools, and then find information about their options. If parents are not aware of the resource disparities between schools in their district, they may not be motivated to participate in school choice. As A2 told us, *"you assume that people who have a degree or people who are educated are going to educate your babies."* In fact, several parents I interviewed only learned of the choice system after they had enrolled their children in their neighborhood school. For example, P8 heard from other parents at her child's school that *"there are way better schools.... They say far away schools ... are better, but I don't know, I'm not sure about that."*

Next, once a family has decided to participate, they need to find information about the available options. However, I found that available information often lacked relevance and utility to families. For example, school websites were generally poor quality except at high resource schools, and some parents found the information available to be tailored to the interests of higher resource parents. For those schools,

> *I mean, you know everything, you got teacher bios, walkathon, the hundreds of thousands of dollars you can raise, how you can pay it, sign up for afterschool. ... it's geared towards parents who are looking for certain things. It's geared toward a middle-class, upper middle-class aesthetic, because the other schools, that's not where they're putting their resources, understandably. (P1)*

As discussed in section 5.3, concerns about information access have driven the growth of dashboard-style websites like GreatSchools[11] or the California School Dashboard[12], which provide school statistics, like test scores and demographics, in a digestible format. While some parents in my sample found this kind of data useful, others pointed out that it is still limited, especially if you are concerned that your child might face more adversity than the average student. P3 tried to use test scores disaggregated by race to find a school that would be a good fit for her daughter. However, she pointed out that average scores are only rough indicators of how a school will support your child:

[11] https://www.greatschools.org/

[12] https://www.caschooldashboard.org/

> *[After looking at test scores] you are thinking, "Oh, my child's here," and then you take the test and they're like, "Well, actually, they're behind." Now, are the school going to help me get her up to what she needs to be? Do they offer or have those resources available? (P3)*

The district encourages families to use online information to narrow down the schools they are interested in, and then attend in-person tours to learn more. Many participants agreed that tours provide a valuable opportunity for parents to assess whether their family will be safe and included at a school. For example, A2 tells parents to visit the schools that they are interested in to make sure that they *"feel welcome."* However, prior work has found that school tours are time-consuming and often inconvenient for parents with inflexible work or childcare schedules (Chapter 4). Participants in this study discussed additional challenges with school tours that were specific to families of color. For P1, *"it felt very isolating to be looking for schools as a middle-class Black parent in [city name]. . . . We went to the [school name] tour for my five year old, lines out the door, full cafeteria, people asking questions like, "What's your track record on where you get kids into college?""* A1 recalled, *"my wife was the only non-white person on the tour. You know, like they had all this like Spanish translation available [but] nobody needed it."* Despite efforts to include families of color, for example, by providing interpreters, A1 found that parents still struggled to get useful information specific to the experiences of children of color during school tours:

> *People have real questions, . . . especially around identity safety, whether or not there are other kids like theirs at the school, things like that. . . . I went on some tours and I saw people asking questions towards that question and the tour guide either kinda didn't pick up the gist of the question, the nuance of it, or they did and they answered it in a like not real way. (A1)*

Overall, participants identified a mismatch between the information that is easily available to families, which assumes a level of understanding of how the system works and a certain set of interests, and the information that parents need, particularly regarding how lower-resourced students are included and academically supported at each school.

Strategies for support

Parent advocates provided more relevant and concise information to families by first building a trusting relationship with them, learning about their personal circumstances and priorities, then connecting them with the right information or resources. There are over 80 district schools and more than 30 charter schools in Oakland, so much of this work involved reducing information overload for families. In addition, this support often extended beyond enrollment to address other survival needs like housing, food, and healthcare.

When A2 starts working with a parent, the first step is *"a conversation of getting to know the parent and what it is that they're ultimately looking for."* This conversation is not simply eliciting a set of criteria that the parent already has in mind, but is a more personal process of *"learning about the family and what it's going to take for that child to do well and be successful."* A2 brings together parents' expertise on their family's circumstances, with her own expertise on schools and education, to recommend schools that will meet the family's needs within their constraints.

A4 works with refugee families who have recently arrived in the United States. Rather than providing detailed information about all of their school options, he accounts for these families' difficult circumstances when providing support.

> *Usually when families are coming to enroll with me . . . they're totally overwhelmed with the whole process and the information overload so the idea of like sitting there and talking through like ups and downs, all the ins and outs of each school it's like, no, they don't want that, they want a couple of options, and to feel like I am a trustworthy person . . . that cares about their students and going to give them the right information. (A4)*

He emphasized that this kind of support relies on trust and care:

> *My absolute nightmare is somebody coming to me and being like, "You know what we need is like a resource map or like a resource guide so we can start giving more information out to families." I'm always like that is the last thing we need, we need more people to explain things to people in person or one-on-one, and answer their questions. Because, especially if you're new here, or don't have a lot of education yourself, the idea of navigating a website and filling out forms. It's just doesn't. . . it's. . . it's ridiculous. So, I think that the best way is just more personalized. (A4)*

In order to build trust with these families and provide relevant and helpful support, advocates considered their circumstances holistically, and where necessary, connected families to other resources. One reason that families might be looking for a new school is that they are experiencing some form of instability, for instance, they have recently immigrated and/or they have unstable housing. In these circumstances, enrolling their child in school may be only one of many challenges they are facing. For instance, when I spoke to A4 he was working with a mom who had very recently arrived in the United States. He not only helped her enroll her son in a school, but also organized for him to receive the required immunizations and signed him up for free meals through the school district. He also provided her with financial support and was helping her find free legal representation. P6 also worked with A4 to enroll her son in school. She still frequently reaches out to him for support, and he has helped her, *"know about the food they give out in school, . . . connect with the internet service and . . . finding the router, and then he has helped me with my son's computer."* A3, who works at a school, also connected families with different sources of support, such as financial aid, mental health care, housing, and legal aid.

I found that this type of support could also be helpful for other low-resource families, but that it was difficult to find. For example, P4 was frustrated by a lack of support from her son's school to find permanent housing.

> *I was searching for honestly like resources for housing because like when I enrolled him they were saying how if you need any assistance in anything [they could help.] . . . They did sign him up for free meals to be delivered, so that was one thing they helped with. . . . But as far as anything else it's not really a great experience honestly. (P4)*

Typically, discussions of information support for school choice focus on information about schools. This assumes that families can afford to spend significant time and effort researching different school options, and overlooks other elements of the process, like ensuring students have the immunizations and technology they need to participate in school activities, as well as the food and housing security they need to focus on learning. Parent advocates' strategies focused on personal relationship building and trust and highlighted the importance of attending to individual families' circumstances and priorities.

Ranking schools involves difficult trade-offs

Families apply for schools by submitting a list of schools ranked in the order of their preference. I found this process can be challenging for families, even aside from the challenges of finding relevant information outlined above, because they need to know what to look for in a school, and often face difficult trade-offs in deciding which schools would be the best fit for their family. To support families making these decisions, advocates avoid overly narrow assumptions about what makes a school a good fit for a family, and push for longer-term change so that every school offers the education and resources its students need.

Challenges

Parents may not know what to look for in a school, and even those who both know what they want and have relevant information available may face complex trade-offs between those priorities.

School choice has created an unusual situation in which parents are expected to have substantial knowledge about what makes for a quality education. Some participants, particularly those who had recently migrated to the U.S. and/or who had fewer years of formal education themselves, felt unsure about what to look for in a school. For example, when I asked P6 to describe the ideal school for her son she told us, *"I wouldn't know because I don't really know a lot of schools."* Some parents found it easy to pick their top choice school, as they had specific criteria or personal connections, for instance, P6-P8 prioritized proximity to home, and P5 wanted to enroll her children in the school where their former teacher had recently transferred. However, many of them did not have back-up options in case they were not assigned to their top choice.

By claiming that offering students the choice to leave their neighborhood school increases educational equity, open enrollment policies can give the impression that some schools are desirable, while others are not. As A1 put it, *"Why would we have school choice, unless we are saying that there are some schools that are not good enough for your kids?"* However, even with choice, higher resource schools are not practically available to every family. Due to a long history of systemic racism in education and housing, higher resourced public schools remain concentrated in affluent neighborhoods and serve more white students. As a result, low-income families and families of color who invest in school choice face complicated trade-offs to find a school that provides resources (e.g. experienced teachers, high quality equipment and facilities, and smaller class sizes) but also offers safety, inclusion, and convenience.

For example, P1 repeatedly described her experience of finding a school for her children as a *"balancing act."* For P1, the most important factor about a school was *"achievement, but,*

> *I wouldn't just say test scores, . . . I also was looking at resources, . . . I wanted a school where I felt like, as an African American family, or as a Black kid, my child was not going to be the only [one]. And I also wanted a school where there was going to be some socioeconomic diversity, and where I was going to fit in with the parents. And so, it was a balancing act of wanting to be at a place where I felt like... because for Black kids, there's so much literature, there may be a school that's a good school, but that doesn't mean it's the best school for my kid. (P1)*

Economists have sought to reduce information overload by prioritizing information about schools that perform well on one or more quantitative measures of school quality, such as test scores, chronic absenteeism, or graduation rates [21, 117, 195]. However, this approach makes strong assumptions about which schools families *should* prefer, overlooking these kinds of trade-offs as well as other factors that contribute to a student's experience at a school.

A1 is a principal at an elementary school where around half the students are from low-income families. He didn't believe that high resource schools are unequivocally better for students from low-resource backgrounds:

> *I don't think they really know what it means to support a child that's, you know, catching the bus from [low-income neighborhoods] and has experienced severe trauma, you know, maybe has an unstable living situation, stuff like that. . . . There's so much growth that has to happen in that community before that kid even really has a chance at a school like that. (A1)*

A4 encourages newcomer refugee families to attend schools where there is an existing community from the same country, so that students have *"their community members there already and . . . the teachers already have done professional development to learn the background of students from these places."*

P1 pointed out that political power dynamics at the school district level are another factor to consider. After her son had *"a really rough experience ... with a teacher who just* *really was very anti-Black and racist,"* she was forced to move him to another school. This time she chose a school in a wealthier neighborhood, despite it having fewer Black families, partly because she felt that *"the resources and the privileges there are going to mean that when there's a problem that [the school district] is going to listen."* Unfortunately, three other parents (P3, P10, A3) in my sample reported similar stories where their child was in an unsafe classroom environment due to treatment by an elementary school teacher.

Families can rank up to six schools on their application, and are strongly encouraged to submit complete applications to maximize their chances of being assigned to at least one of their preferred schools. Some parents have a specific school in mind, for instance, the one closest to their home, in which case ranking schools is easier, as long as that school is not oversubscribed. However, many families face difficult trade-offs between priorities like academics, convenience, and inclusion, induced by the inequitable underlying distribution of resources across the city.

Strategies for support

Advocates considered school quality broadly and accounted for families' circumstances when making recommendations or helping them navigate trade-offs. Most importantly, several of the people I interviewed viewed their work as a fight for long-term social change rather than exclusively enrollment support, and sought to ensure students had the resources they needed no matter which school they ended up at.

The level of information advocates provided to families depended on their circumstances and priorities. For example, A2 narrows the set of choices and recommends schools for families based on how they've described their priorities, but always leaves the final decision to the family.

> *We just kind of let them know, "Hey, out of the schools, based on the information that you gave me, here are some of the schools that when we look at data and information, these are like your top eight schools," and we always give them at least five to be able to choose from. [...] We never tell the person, "this, this, this, this, this." You rank it how, you know, works for your family.* (A2)

A4 made stronger recommendations to families, but was sensitive to their constraints and helped them work within them. For example, sometimes the school with the strongest existing community of immigrants from a newcomer's country of origin is further from home. Proximity is often the top priority for these families (P5, P6, P7, P8), so he works outside of the formal enrollment system to arrange free bus tickets for the students to ease this trade-off.

Ultimately, school choice policies can only address inequalities in resources across schools by providing students from underserved communities the opportunity to access higher resourced schools. There are a limited number of seats available at higher resourced schools, so there will always be some students who cannot access those schools, even if they apply.

> *"There's not a whole lot of [...] quality schools that are in the city, [...] and the ones that are high demand, quality schools, everybody else know about them, too. So what are my chances of being able to actually get into those schools?" (A2)*

This means that supporting families to participate in the choice system is a limited and insufficient avenue to promote longer-term justice in education. Unlike resources focused solely on choice, parent advocates were able to balance between helping families access resources within the existing system while also pushing for a future in which the choice system isn't necessary. For example, part of A2's work involves advocacy at the district-level, as well as grassroots organizing with parents. Another strategy is targeting resources and support to students who enroll in lower resourced schools. For instance, at a school where several families that A4 worked with attend, they were able to,

> *"set up English classes for parents at that site and then summer school classes for students and concurrently ESL [English as a second language] classes for the parents. And [...] when community-based organizations were looking for office space at that time we managed to get them space in the school and in offices close to that school." (A4)*

In contrast, broader discussions of school choice and research into low-resource families' preferences for schools too often assume that the best way to move towards a more equitable education system is to nudge these families towards historically higher resource schools, denying the complexity of the landscape that they are forced to navigate and the limits of individual choice for promoting long-term justice.

5.6 Discussion

As school choice has expanded as a way to increase access to high quality public education, so has the concern that these systems require too much time and effort from families. Time and resource constraints disproportionately exclude underserved families, who were positioned as the main beneficiaries of choice to begin with. In an effort to reduce these barriers to participation, researchers, governments and non-profits have developed online tools and informational resources that rank and sort schools. The broad success of sites like GreatSchools indicates that these kinds of technologies are helpful for many families. However, an abundance of examples within and beyond HCI research have shown that even technology that is intended to improve underserved people's access to information and resources can struggle to do so in practice [418, 214, 228, 143]. My goal in this work was to understand the extent to which a primarily online enrollment system supported lower-resourced families participating in open enrollment, with a focus on technologies that provide information about schools and how to apply. My findings provide insight into the information mismatches and difficult trade-offs that families face in this process. Consistent with prior work [514], my findings highlight how parent advocates provide personalized support to fill the gaps in the information infrastructure. In this section, I discuss the implications of my findings for online informational resources and their potential to address participation barriers in public school choice. Then, I discuss paths forwards, advocating for an assets-based approach that recognizes and resources the work of community members that already goes on to support families and promote longer-term social justice.

The limits of informational resources for promoting educational equity

I began this research hoping to identify ways in which the online enrollment system could be improved, for instance, what kind of information dashboards should contain in order to be relevant to low-resourced families. However, my findings highlight three specific limitations of improving informational technologies for promoting more equitable participation in enrollment:

- First, I found that *web-based informational resources lack personalization and nuance*, which are invaluable for lower resourced and marginalized families. How could an online dashboard tell a family whether they will feel safe, welcomed, and heard in a school community? Providing relevant, helpful information requires understanding a family's circumstances and goals, as well as nuanced and detailed information about the community and resources at each school.

- Second, *prioritizing information about schools assumes that families* should *gather as much information about schools as they can.* However, for some families it is entirely reasonable not to commit time and resources to searching for a school [429]. Some families are not aware that they need to apply for schools, and thus do not even seek out information to begin with. Even those who are aware of their options may not know what they should look for in a school. My interviews highlighted that parents are navigating an inequitable system, in which they have a low probability of admission at higher resourced schools, and where those schools, if they are admitted, may expose their children to isolation and racism. Meanwhile, they may be facing much more urgent survival needs. Providing relevant and useful support in some cases may involve deprioritizing information about schools and connecting families to other critical resources like food, housing, or healthcare.

- Third, *online informational resources are generally targeted to individual families, which does not build community or trusting relationships, or promote longer-term social change.* Being in community with other people experiencing similar challenges creates opportunities for families to form a shared understanding of those challenges and find ways to exert collective political power to make change [456]. Further, technologies designed for individuals can overlook the social, economic, and political context that produces an inequitable distribution of needs and burdens in the first place [137].

My findings in this research remind us to remain wary of purported tech “fixes” to long-standing social problems [44, 486, 37], and to think critically and reflexively about how the methods and approaches of HCI and computing more broadly can contribute to social change [4, 500, 137]. Ultimately, new or improved tools and online resources focused on enrollment would benefit some families, but it would also create more work for parent advocates to support the lowest-resource families in accessing and utilizing these tools. New technology rarely *reduces* the overall amount of work that needs to happen to make systems run smoothly, rather it displaces that work, often onto the most marginalized and in ways that are invisibilized and under-resourced [471, 63].

Certainly, some technical improvements to the system may still prove worthwhile. In fact, many parents I spoke to strongly appreciated the district’s introduction of an online enrollment option, including two parents who did not have stable housing and were able to complete the entire process on their mobile phone. However, we should not expect new technology to easily solve participation barriers in enrollment, like a lack of time to invest in the process, or conflicting factors like proximity or inclusion barring access to higher resourced schools. The cost of the substantial labor and resources needed in order for new technology to be inclusive should be factored into decisions about what new technologies to pursue and where to invest resources. To conclude, I advocate for an assets-based design approach to guide the development of interventions that support low-resourced families.

An assets-based approach to enrollment support

Assets-based design is an approach to designing interventions that are sustainable and useful to communities in the long-term, especially in low-resourced settings [369, 515, 107, 313, 266, 70]. This approach is grounded in a deep understanding of people’s capacities and assets, rather than looking to solve their needs or deficits [266], and has an established history in education research [328, 515, 530, 189]. One way of doing this in HCI is to design interventions that leverage and amplify existing resources and practices in a community and minimize technical novelty [369]. Allowing parents to enroll over the phone or a mobile-friendly website, for instance, did not require sophisticated technical innovation, but allowed many of the parents in my sample to participate using a device they had easily available and practices they were already familiar with. In this work, parent advocates offered rich insight into their successful practices for supporting families through enrollment both within and outside of the constraints of the existing system. Advocates did not simply nudge parents to access existing information and resources, but augmented those resources with personalized informational and emotional support, and worked in partnership with families to meet their specific needs. Viewing these relationships and practices as an asset in the community, an assets-based design approach would ask not how we can design interventions that reduce or replace the work of advocates, but how we can *support and amplify* it [369, 90, 224].

A direct way to support the work of parent advocates is to provide resources for this type of work, e.g. by hiring more people who are trained to provide personalized, culturally relevant, long-term support to more families. Many parents in my sample worked with a parent advocate because of my snowball sampling approach, but in general most people do not have access to someone with both deep knowledge of the school system and the time to provide personalized support. Although the district's enrollment center provides support in person and over the phone, staff at this center serve the entire district, and thus may lack the time and specialized knowledge (e.g. about resources specific to newcomer refugees) to provide in-depth and personalized support. In addition to district staff, it may also be helpful to train trusted community members, e.g. at community centers or places of worship, to either help families directly or get them in contact with a more specialized parent advocate.

A complementary approach could be to build systems that support parent advocates' work. Identifying such opportunities was not my focus in this work, but my data points to some promising directions for further investigation. For example, community advocates in the district recently showed that the enrollment system admitted very few non-neighborhood students to popular, high-resource schools, and contributed to a new priority system that will make more space for low-income students of color at those schools. One way to support this work could be to design more transparent systems that make it easier for stakeholders to identify harmful aspects of the system's design and suggest improvements [500]. Another challenge that advocates in my sample faced was connecting families to resources beyond enrollment, like mental health care and housing. This suggests the potential for tools that help advocates build, maintain, and share networks of support and resources in their area. This echoes similar findings in other domains, for instance, environmental justice [24].

Finally, while support during enrollment can be useful, it is important to be wary of interventions that are too narrowly centered on school choice and enrollment. Advocates and many parents agreed that choice is a limited avenue to justice, and the top priority should be ensuring that every school is high quality and well resourced. For instance, while admitting more low-income students to historically overserved schools is an improvement, there are still a limited number of seats at those schools and many more students who will not be admitted. While advocates for school choice position it as the best way to "empower" parents and promote equity in education, it is far from the only option. In fact, Scott points out that the individualistic, market-based principles underlying school choice are not only in tension, but in direct conflict, with the redistributive principles that defined 20th century civil rights movements and continue to mobilize grassroots movements for educational equity [428]. Across the country, students, parents, and teachers have organized for smaller class sizes, experienced teachers, community engagement in school governance, and more equitable funding structures, in an effort to guarantee a high quality education for historically underserved students without forcing them to move schools [392, 230, 173, 428]. My participants surfaced tension between using the existing enrollment process to access resources *today*, and working towards a more just arrangement in the future. In considering how technologies could support families and advocates, we should design to surface, rather than obscure, these tensions and avoid entrenching the dominance of individual choice over grassroots strategies to promote educational justice.

5.7 Conclusion

In this work, I found that low-resourced families and families of color faced challenges navigating a public school enrollment system that was intended to improve their access to educational resources. In particular, the information provided about schools often lacked relevance and did not account for the difficult trade-offs that families must navigate when choosing between schools. I found that parent advocates provide personalized, community-based support that cannot be emulated by impersonal and generic online resources. I reflect on this research as an example of the importance of an assets-based design approach to supporting communities' ongoing work to advance social justice.

Chapter 6

Expressiveness, Cost, and Collectivism: How the Design of Preference Languages Shapes Participation in Algorithmic Decision-Making

Emerging methods for participatory algorithm design have proposed collecting and aggregating individual stakeholders' preferences to create algorithmic systems that account for those

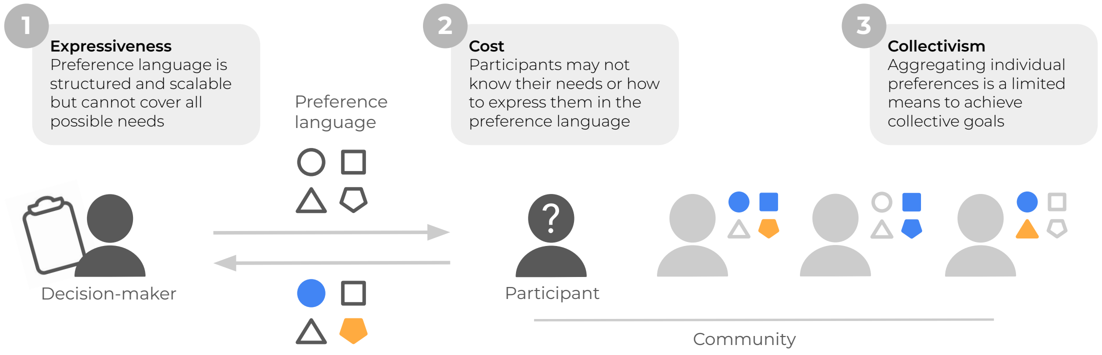

1
Expressiveness
Preference language is structured and scalable but cannot cover all possible needs
2
Cost
Participants may not know their needs or how to express them in the preference language
3
Collectivism
Aggregating individual preferences is a limited means to achieve collective goals
Preference language
Decision-maker
?
Participant
Community

Figure 6.1: Preference-based systems consist of a preference language, in which participants express their needs and goals to a decision-maker, and an aggregation algorithm, which aggregates individuals' preferences into a collective decision. I identify three ways that preference languages shape opportunities for meaningful participation in algorithmic decision-making: **1) expressiveness**, the range of needs that participants can communicate; **2) cost**, the effort it takes for participants to express their needs and goals in the preference language; and **3) collectivism**, the extent to which aggregating individuals' preferences can achieve collective goals.

stakeholders' values.[1] Drawing on two years of research across two public school districts in the United States, I study how families and school districts use students' preferences for schools to meet their goals in the context of algorithmic student assignment systems. I find that the design of the preference language, i.e. the structure in which participants must express their needs and goals to the decision-maker, shapes the opportunities for meaningful participation. I define three properties of preference languages – expressiveness, cost, and collectivism – and discuss how these factors shape who is able to participate, and the extent to which they are able to effectively communicate their needs to the decision-maker. Reflecting on these findings, I offer implications and paths forward for researchers and practitioners who are considering applying a preference-based model for participation in algorithmic decision making.

6.1 Introduction

Algorithmic systems increasingly impact peoples' lives by mediating their access to resources and by making high-stakes decisions in domains like education, employment, healthcare, and child welfare [165, 385, 347, 243, 71]. Documented issues of discrimination [23, 109, 349, 14], biased system performance [76, 342, 474, 59], and dissatisfaction among key stakeholders [464, 366] have increased pressure to improve these systems by accounting for the values and needs of those who use or are affected by them. Building on social choice theory [245], emerging methods for participatory algorithm design have proposed collecting and aggregating individual stakeholders' preferences to create algorithmic systems that represent the values and goals of those stakeholders [343, 237, 282, 280, 150, 146, 233, 529, 523, 80, 268, 144]. In this chapter, I study how the *design of the preference language*, i.e. the language in which participants are asked to express their preferences, shapes the opportunities for meaningful participation.

Matching algorithms for student assignment are one example of an area in which participant preferences are incorporated into algorithmic decision-making [409]. School districts reason that compared to neighborhood-based assignments, these preference-based assignment systems provide more flexibility to families, create more diverse classrooms, and promote educational equity [440]. However, many school districts have found that the algorithms do not meet these expectations in practice (See previous chapters in this part).

In this chapter, I bring together my analysis of student assignment in San Francisco (Chapter 4) and Oakland (Chapter 5) to understand how families and school districts use preferences to meet their goals. This case study offers insight into the challenges and limitations of the preference-based approach to incorporating stakeholder participation in algorithmic decision-making. My data includes 27 semi-structured interviews with parents and with community members who helped parents through the enrollment process (e.g.

[1] This chapter was written in collaboration with Tonya Nguyen, Cathy Hu, Catherine Albiston, Afshin Nikzad, and Niloufar Salehi, and published at ACM CHI 2023: `https://doi.org/10.1145/3544548.3580996`.

district, school, and non-profit staff), in addition to several informal conversations with relevant stakeholders over the course of two years.

I find that the design of the preference language defines the opportunities for participation (Fig. 6.1). In student assignment systems, participants submit a ranked list over schools, possibly of limited length. Other common preference languages can include pairwise comparisons between real or hypothetical options, and providing weights over features of the decision outcome [280]. I define three properties of preference languages – **expressiveness**, **cost**, and **collectivism** – and discuss how the design of the preference language with respect to each of these factors can shape and limit meaningful participation. Preference languages are often designed to be structured and scalable so that large numbers of participants can be involved at low cost. This means that a preference language cannot cover all possible needs, and some participants may be able to express their needs more than others (**expressiveness**). Second, it takes time and effort for participants to translate their complex and often vague needs into the structured preference language (**cost**). This can be especially costly for participants who do not already know what they need, or how to express their needs in the preference language. These costs create disparities between people based on their access to time and resources. In the school assignment case, parents with fewer resources face greater barriers to convey all of their needs and goals through their ranked list of schools. This is partly due to the cost of gathering information about the schools (See Chapter 5). Finally, I find that aggregating individual preferences is often a limited means to achieve complex collective goals (**collectivism**). In the case of student assignment, school districts' goals such as integration and educational equity have been very difficult to achieve through the aggregation of individual preferences, which can only express families' self-interested priorities.

Reflecting on these findings, I offer implications and paths forward for researchers and practitioners who are considering applying the preference-based model for participation in algorithmic decision making. First, I discuss opportunities to improve expressiveness and reduce costs associated with a preference language. These include improving the resources available, providing support (e.g. information support) for stakeholders to use the preference language, and simplifying or re-designing the preference language to make it a more natural representation of how people already think about their needs. Second, I discuss paths forward for engaging community members in deliberation to co-define collective goals, drawing insight from procedural justice theory. Procedural justice theory considers how the process through which a decision is made impacts satisfaction, perceptions of fairness and legitimacy, and compliance with that decision [494, 495]. I also propose technical mechanisms for better aligning preference aggregation mechanisms with community-defined collective goals, even while only eliciting self-interested preferences. Finally, I discuss the limitations of preference-based systems, and the potential for their emphasis on individualism and free market values to cause harm. Ultimately, I argue that preference elicitation and aggregation mechanisms will never reach the ideals of participatory methodology [330] without careful attention to how the preference language, and sociotechnical infrastructure supporting it, enable equitable and meaningful participation.

6.2 Related Work

In this chapter, I use student assignment algorithms as a case study to examine the challenges for using individual preference aggregation as a form of participation in algorithmic decision-making. In this section I provide an overview of participatory algorithm design and the preference-based approach to algorithm design.

Stakeholder Participation in Algorithmic Decision-Making

Growing awareness that algorithmic decision-making may harm marginalized people has driven researchers to seek methods for directly involving stakeholders in the design of algorithmic systems. Incorporating direct stakeholder participation is intended to bring diverse knowledge and perspectives into design, and to build technologies that have a more positive influence on people's lives [126, 98, 54]. Recent work has drawn on methods and practices from frameworks like participatory design [331], human-centered design [345], and value-sensitive design [11]. Common methods include design workshops [71, 15, 361] and interviews [279, 453, 10, 531, 420]. Delgado et al. [126] catalogued 9 different approaches to increasing stakeholder participation in algorithmic systems, noting that they vary substantially in terms of the degree of power different stakeholders are afforded, and when in the design and deployment of a system they are afforded that power. Birhane et al. [54] developed standards for evaluating whether particular approaches are aligned with the goals and values of participatory AI, such as the degree of reciprocity and participant empowerment.

Prior work in participatory and human-centered design methods foreshadows some of the challenges that arise when attempting to engage stakeholders to build more beneficial and just algorithmic technologies. For instance, it can be very difficult to subvert power dynamics, both between designers and participants and among participants, to promote genuine and equal participation [126, 192, 205]. The format and outcomes of participation are also often constrained in a way that presumes that there must be a technical solution and only allows participants to tinker around the edges of it, rather than offering meaningful decision-making power about what the system should do and whether it should exist at all [126, 205]. In the worst case, seemingly participatory practices can offer a guise of legitimacy to harmful technologies, making it more difficult to challenge those systems in the long run [452, 44, 138, 54].

Even if a participatory process effectively and meaningfully engages marginalized stakeholders, challenges remain. For example, researchers or designers often have to aggregate input from many participants and translate this input into a technical design specification, leaving room for unequal representation, misinterpretation, or even disregard of stakeholders' views. Some researchers have tried to bridge this gap by allowing participants to play a more direct role in building algorithmic systems. For example, the ORES system allows Wikipedia editors to directly specify and request machine learning models that meet their editing needs [183]. In this chapter I focus on another line of work, which draws on social choice theory [245] to develop what I refer to as "preference-based systems." This approach incorporates direct input from stakeholders by eliciting individuals' preferences over some available alternatives, and then aggregating those preferences using an algorithm to make a decision. Preference-based systems are appealing because they easily scale to allow a large number of people to directly contribute their views, and there exist a range of well studied aggregation procedures to translate this input into a decision [66]. I next discuss how this class of algorithms has been used to increase participation in algorithmic decision-making, and what challenges remain.

Preference-based Algorithmic Systems

There are two main stages of any preference-based system, *preference elicitation* and *preference aggregation*. In the preference elicitation stage, a decision-maker asks participants to quantify their relative value for some available options in a structured *preference language*. A preference language consists of a set of features and a way for participants to express their priorities across those features, typically either by ranking them, making pairwise comparisons, or providing weights over features [245]. Features may be a set of real alternatives (e.g., a list of schools in a school district), or attributes of those alternatives (e.g., a school's location, start time, or language programs).

Once the decision-maker has collected preferences from stakeholders, they then need to aggregate those preferences to make a decision. Researchers in social choice theory[2] have developed various aggregation procedures that are designed to make decisions that satisfy some normatively justified axioms or formal definitions of optimal assignment [245, 66]. For example, matching algorithms are guaranteed to produce assignments that are stable — no two students can swap assignments in a way such that both the students and the schools are better off — or Pareto efficient — there is no way to improve one student's assignment without making another worse off [2]. In other contexts, researchers have advocated for voting procedures on the basis of properties like robustness to noisy preference information [237, 161, 42].

Systems based on preference elicitation have long been used in market and mechanism design [409] and participatory democracy [84]. More recently, researchers have explored how these methods could incorporate stakeholders' perspectives and values more directly into the design and function of algorithmic systems. Some of this work falls under the umbrella of computational social choice [66]. For example, Noothigattu et al. [343] proposed collecting people's preferences in hypothetical scenarios, using that data to build personalized preference models, and then aggregating predicted preferences to make decisions in new situations. The authors suggested that this approach could be used to build autonomous vehicles that align

[2] A substantial body of work in social choice theory across economics and philosophy seeks to understand how individual preferences should be aggregated to inform decisions on behalf of a group. Kenneth J. Arrow [245] provides an overview. In this work I apply human-centered methods to understand experiences with a matching algorithm in practice and compare these experiences to the ideals of participatory design. I leave a detailed comparison of these findings to theoretical results and philosophical discussions of preferences to future work.

emergency decision-making with people's ethical preferences. Researchers have subsequently explored how this approach can improve efficiency and fairness in distributing donations for a non-profit organization [282], align shift scheduling with workers' and managers' preferences [280], and account for citizens' ethical preferences in automated flood management decisions [146]. Freedman et al. [150] suggested a related approach, modifying a matching algorithm for organ transplants to weight matches according to peoples' preferences about which patients should be prioritized. In experiments, their algorithm improved outcomes for typically under-demanded patients who are currently less likely to receive an organ match. Other areas where researchers have studied how individuals' preferences can be elicited and aggregated to guide algorithmic decision-making include patient triage in hospitals [233], selecting appropriate performance trade-offs for ML models [529, 523, 80], allocating public budgets and resources [84, 268], defining diversity quotas for elections [144], selecting student volunteers for conferences [362], dividing goods and labor among groups of people [281, 277], aligning recommender systems with users' values [468], and balancing conflicting perspectives in content moderation [169].

This research highlights the potential for these kinds of systems to align algorithmic decision-making with people's needs and values at scale. The process of quantifying one's preferences can help people better understand their own needs [282, 280], and this approach makes explicit to decision-makers the pluralism of priorities and values in a group of stakeholders [169, 268]. However, it is difficult to design effective and easy-to-use preference languages and aggregation procedures. For example, Lee and Baykal [277] conducted an experiment where a group of people used a matching algorithm to divide up tasks according to their preferences. Participants found it difficult to quantify their preferences using the given preference language. As a result, they relied on error-prone cognitive heuristics to simplify the task. Individual preferences also cannot account for cooperative and altruistic behavior. In the same study and in a related experiment involving goods division [281], participants wanted to deliberate and cooperate to find more acceptable compromises than the algorithm's allocation.

Matching algorithms are a kind of preference-based system that have been used in real world markets, such as organ transplant matching, assigning medical students to residency programs, and assigning students to public schools, for decades [408]. In this time, researchers have been able to observe how these systems function in the real world. This has provided insight into how to build effective systems, as well as what issues remain unsolved. My goal in this work is to draw lessons from the student assignment context that illuminate key considerations for building effective, equitable preference-based systems. I also draw on this newer body of work developing other kinds of preference-based systems to inform the design of student assignment algorithms. In particular, I study how the *preference languages* in these different systems shape who can participate, what they can communicate about their needs, values, and priorities, and how collective goals can be achieved using the submitted preferences. I conclude this section with an introduction to student assignment algorithms.

Student assignment systems are a useful case study of preference-based algorithmic systems because they have been used in school districts across the country for several decades, giving researchers the opportunity to observe challenges in practice, such as confusion for families and decreasing classroom diversity [409]. In this chapter, I build on this prior literature by bringing together the interview data I collected in San Francisco and Oakland, California to discuss the role of the preference language in shaping opportunities for participation. I compare and contrast this case study to other kinds of preference-based systems to identify broader implications for the design of preference languages and propose key paths forward for the emerging field of participatory algorithm design.

6.3 Methods

This work draws on the qualitative data collected as part of separate, but closely related, research projects that I conducted in collaboration with two neighboring school districts, San Francisco Unified School District (SFUSD) and Oakland Unified School District (OUSD) (See Chapters 4 and 5)

Data Collection

In this chapter, my analysis draws on all 27 semi-structured interviews with parents (13 SFUSD; 10 OUSD) and staff in schools (2 OUSD), district offices (1 OUSD), and community-based organizations (1 OUSD) to understand how families use their preferences over schools to communicate their needs and constraints to school districts, and how the districts aggregate those preferences to satisfy parents' needs and meet district goals such as school diversity. For complete data collection methodology, refer to Chapter 4, Section 4.3 for the SFUSD interviews, and Chapter 5, Section 5.4 for the OUSD interviews.

Data Analysis

I conducted inductive, qualitative analysis on the interview transcripts [318]. First, I conducted open coding on a line-by-line basis [102] to understand how parents use the system and what challenges they face in meeting their goals. I then conducted axial coding to identify relationships between codes and higher level themes. I first analyzed the data from San Francisco, before conducting interviews in Oakland. At this stage, the first author conducted open coding and identified a common set of parents' priorities (e.g., "test scores," "resources," and "travel logistics") and challenges (e.g., "stress," "information needs") when participating in school choice. Next, I analyzed the data from Oakland, building on the codes I used when analyzing the San Francisco data. The first two authors worked together to code two transcripts, then discussed findings and resolved misaligned interpretations through discussion. Each author then analyzed half of the remaining transcripts. The initial codebook from San Francisco was used as a reference, but codes were adjusted, added, and removed as necessary to best fit the data. I then conducted axial coding, and grouped 25 of my 39 codes into two higher level groups, "finding information" and "considering priorities." Other codes included “building relationships,” “worrying about availability/scarcity,” and “voicing concern.” Finally, I conducted a final round of coding on the full dataset to compare and synthesize findings across the two districts.

Researcher Positionality

My collaborators and I recognize that our personal and professional backgrounds and contexts shape our approach to this research, our interactions with participants, and our interpretations of the findings [33]. The first and second authors conducted recruitment and interviews. Neither author has first-hand experience of a public school choice system, nor is either author a member of the communities in San Francisco and Oakland with whom we conducted this research. Throughout the course of this work, we shifted from direct engagement with parents to working with trusted community leaders, finding that this was a better way to show respect towards the relationships and work already happening in those communities (Chapter 5). My interpretation of the findings is certainly shaped by both my position as an outsider, as well as our team’s disciplinary backgrounds, which between the authors include human-computer interaction, sociology, law, and economics.

6.4 Designing Preference Languages

In this section, I discuss how the design of a preference-based system shapes (and can limit) participation in algorithmic decision-making. I argue that a core consideration should be the design of the *preference language*, which consists of a set of features and a way for participants to express their priorities across those features. First, I discuss the implications of the preference language for individual participants: **expressiveness** and **cost**. Then I explore the implications of the preference language for achieving collective goals (**collectivism**). For each of these three factors, I illustrate relevant challenges using examples from the student assignment context, then discuss how those challenges appear and are addressed in other kinds of preference-based systems.

Expressiveness: No preference language can cover all possible needs

Preference-based systems offer participants a fixed set of alternatives over which they can express their preferences. A core part of designing a preference language is selecting what alternatives or factors are available to participants. Because preference languages must be structured and scalable to large groups of people, no preference language will perfectly capture every possible dimension of everyone’s needs and values. Certain preference languages will be more expressive for some participants than others. It is therefore important to attend to not only how expressive a given preference language is, but *for whom* it is more or less expressive.

Student Assignment

School choice systems increase access to educational opportunities by offering each student a wider range of schools to choose from, rather than being directly assigned to their neighborhood school. However, even if students can apply to more schools this does not mean they have access to those opportunities in practice. Most families want a school that is close to home and that will provide their child with a high-quality education [79, 272]. However, in the United States, a long history of racist housing and education policies have concentrated schools with more resources and higher academic performance in high-income, predominantly white neighborhoods [411, 272]. As a result, families with more resources are able to more heavily prioritize academic factors, whereas other families face difficult trade-offs between economic and social factors, such as transportation logistics or their child's safety and sense of belonging [273, 196].

Low-income families are more likely to face this trade-off between school resources and proximity to home, since higher resource schools are mostly located in wealthier areas of the city [195, 78, 1, 272]. Proximity was especially important to parents of younger children, parents who don't have access to a car, and parents who were concerned about the safety of their neighborhood. O12 *"used to walk with him* [to school] *and then I started to learn to drive. Because when I walk or when I wait for the bus, the bus would take so long."* Discrimination and segregation in classrooms also mean that a school that offers high quality opportunities to white students may not provide a safe and supportive environment for students of color [79, 425, 6]. For instance, O14 saw that her son was *"becoming a different person, I couldn't even recognize him,"* due to his treatment by his first grade teacher. Resolving this situation was a slow and stressful process with serious, lasting impacts on her child, *"[the teacher] was fired but I need to build my son's self-esteem back."* When O7's daughter started crying on the way to school every day, he faced a language barrier when raising concerns with the school principal. This experience in part motivates his current work as a bilingual parent liaison: *"I don't want any other parents going through the nightmare that I went through."*

Empowering parents to decide what is best for their child is one of the central arguments in support of school choice, but that empowerment is often elusive in reality [428]. A system that offers parents the ability to rank any of the schools in their district may appear to maximize expressiveness and empowerment.[3] However, this approach assumes that all families receive the same utility from being assigned their first choice school, second choice school, and so on. This makes invisible the trade-offs that families face, and the fact that for some families there may be no school that currently meets their needs. This becomes particularly problematic for evaluating and improving these systems. For example, statistics like the

[3] Note that even a complete ranked list of every school in a school district does not fully specify an individual's preferred decision outcome. The full space of decision outcomes is intractable, roughly n^s, where n is the number of students needing an assignment, and s is the number of schools available. Even this huge space of outcomes does not express any needs that are not met by the existing schools. As I emphasize throughout this work, even preference languages that appear "natural" have been designed and impose a particular set of constraints and costs.

percentage of students assigned to their first choice school can overlook persisting segregation and inequalities in access to resources. Since ranked lists do not explicitly convey what factors families value in schools, it is also difficult to reduce these inequalities. For instance, districts trying to increase the enrollment of underserved students at higher resourced schools may assume that those families simply didn't know about those schools and jump to informational interventions, when really families need better transportation options to make those schools viable.

Other preference-based systems

The expressiveness of a preference language is constrained by its features, and to a lesser extent the format in which participants express their priorities over those features. Designers of preference-based systems have carefully considered how the selection of features over which participants can express their preferences shapes the information that those preferences convey. For instance, researchers have been exploring how to ask users explicitly what kind of online content brings them value to guide recommendation systems, rather than relying on implicit signals like likes and engagement time, which may not be correlated with value and well-being [468]. Park et al. [362] highlight the importance of considering the ways in which relations of power and positionality influence disparities in expressivity between stakeholders. For example, a participant in their study suggested that students who need to secure visas for international conference travel could be prioritized for student volunteer positions, but pointed out that if the organizers had not dealt with this challenge personally it may not occur to them to include this factor in the selection process. Participants in Lee and Baykal [277] observed that the preference language assumed equal total utility for different participants, which did not account for the fact that some participants were indifferent about which task they would have to complete, and were happy to cede their preferred task to someone who had very strong preferences.

In response to these challenges, researchers have acknowledged that participation must begin at the stage of designing the preference language in order to ensure that it is expressive and morally acceptable [146, 144, 362]. For example, Lee et al. [282] and Lee et al. [280] began by conducting interviews with stakeholders to collaboratively define key factors over which they would later be asked their preferences. Freedman et al. [150] used an open-ended survey to elicit factors that people felt were and were not morally acceptable for prioritizing organ transplant recipients. However, developing participatory processes to define the preference language brings with it many of the challenges that preference-based systems aimed to address in the first place. For example, Park et al. [362] and Freedman et al. [150] found that for almost every factor they considered, some participants thought it was very important to consider in decision-making, while others felt strongly that it should not be considered. It is not clear how to resolve these disagreements in designing a preference language. As demonstrated by the student assignment case study, choices at this stage have significant consequences for participants' ultimate ability to express their needs to the decision-maker.

Cost: Participants have to translate their values and needs into the preference language

The assumption underpinning preference-based systems is that each individual has a latent utility function over some space of alternative decision outcomes, and the preference elicitation mechanism is a way of extracting partial information to estimate that function. In reality, people's preferences are constructed in and shaped by their particular social context, and shift over time. The costs of forming preferences depend on each individual's relevant knowledge, the time they have spent reflecting on their preferences, and how directly they can map their conception of their preferences to the constraints of the given preference language. These costs can exclude participants with less time and fewer resources to dedicate to the process, and can lead participants to rely heavily on social learning and heuristics that exacerbate bias and stereotypes in choice patterns.

Student Assignment

In the student assignment setting, researching the available schools, forming a ranked preference list, and submitting it to the school district can be difficult and time-consuming for families. First, families need to be aware that they have a choice to apply to different schools, and then understand what they are looking for in a school. Parents with fewer years of formal education and/or who had more recently arrived in the district often did not know that they could apply to several schools or what they should look for (Chapter 5). Even for parents who know what they are looking for, researching schools to transform those known preferences for *types* of schools into preferences for *actual* schools can be very time-consuming and frustrating (Chapters 4 and 5).

For example, when O5 first enrolled her child in kindergarten, *"it was very hard. A single parent, never done this before, never been in a public school, so I don't know how everything is run. People just assume that you know what you're doing and I had not a clue."* To learn more about the process, she enrolled in *"a week-long class of learning how to do the application process and how to pick schools, and what makes you want to pick a school."* While this helped her figure out what factors to consider (e.g., academics, sports, other extra-curricular activities), and find schools that best met her criteria, this process required a significant time investment.

This challenge is exacerbated when information is not equally available and useful to all families. For example, online information was disorganized and inconsistent across schools (Chapters 4 and 5). Schools with more resources tended to have more detailed and up-to-date websites, and the information available appeared tailored to the interests of higher-resourced families (Chapters4). School tours were often cited as the most useful way to learn about a school. However, these tours can be extremely time consuming and logistically challenging, making them inaccessible to many parents, for instance, those who cannot take time off of work during school hours, or those who cannot secure childcare during the tour (Chapters 4). Beyond logistical challenges, tours may also offer less helpful information to families from marginalized backgrounds if tour guides cannot speak to the experiences of children from those backgrounds (Chapter 5).

As a result of this lack of information, parents rely heavily on others in their network, including family and friends, as well as in online social networks, to inform their preferences. However, in relying on social learning to cope with a lack of authoritative and relevant information, people may fall back on stereotypes and generalizations to make judgments about schools (Chapter 5). For instance, one parent in San Francisco said that,

> *Facebook has been the most helpful because I feel like it's an insider's look into the actual school quality and parent experiences there. Parents there are very biased, but it's better to have information than no information.* (S3)

Until recently, parents in San Francisco needed to submit their preference list in person at the district's enrollment office by a set deadline. On-time participation significantly improves the chances of being assigned a preferable option, as the number of open seats dwindles in subsequent rounds. However, this requires keeping track of deadlines and finding the time to visit the district office, which creates additional information, time, and language barriers [454, 439].

Other preference-based systems

A key challenge in designing preference elicitation mechanisms is balancing the trade-off between how much information is extracted, and the costs imposed on participants (e.g., time and cognitive load). For example, Johnston, Blessenohl, and Vayanos [233] propose a dynamic preference elicitation mechanism with the goal of minimizing the number of pairwise comparisons that each participant is asked to make. Usability issues not only make the task more difficult for participants, but also undermine the informativeness of the preferences they provide. For instance, Lee and Baykal [277] describe how participants turned to error-prone heuristics when they found it difficult to express their preferences using the provided interface. Studies that have crowd-sourced preference information also raise concerns about the quality of those judgments when participants lack information and a sense of personal investment [29].

These concerns may be mitigated when the participants involved already have substantial expertise and experience making similar kinds of judgments. For example, Lee et al. [282] elicited preferences regarding food donation matching from stakeholders who had personal experience with this process: staff at the non-profit that matches donors and recipients, volunteers who transport donations, and staff at donor and recipient organizations. Another approach is to provide participants with additional information to help them form their judgments. Researchers have shown experimentally that information provision can change how people report their preferences in the school choice context [21, 105, 117, 195, 336]. However, this information also must contend with limits on participants' time and cognitive resources, and there is likely always going to be additional context that could influence people's judgments that is not available. For example, Freedman et al. [150] asked people to make pairwise comparisons between potential organ transplant recipients on the basis of factors like whether they had skin cancer, but the researchers admit that there is far more information that someone may need to judge this situation, such as the prognosis for the disease.

The structure of the preference language can impose higher or lower informational and cognitive load on participants. For instance, Lee et al. [280] argue that pairwise comparisons are easier than ranking tasks for people who have not already formalized their preferences. In fact, the process of making pairwise comparisons actually helped participants understand their own perspectives better by forcing them to make concrete judgments and trade-offs. Ultimately, designing low cost preference languages requires considering participants' access to time and resources, their knowledge about the decision domain, and how much they have reflected on their needs and priorities.

Collectivism: Aggregating individual preferences is a limited means to achieve collective goals

The challenges with preference elicitation are further compounded when preferences are aggregated. In addition to satisfying individuals' preferences, decision-makers often have collective-level goals for how they distribute resources. For example, many school districts want to promote diverse classrooms and increase equitable access to educational opportunities [440]. However, optimizing for satisfying participants' *individual* preferences does not necessarily align with collective goals.

Student Assignment

In the school assignment context, preference-based systems circumvent the issue of defining and working towards collective goals by implicitly defining the socially optimal outcome as the one where each individual receives their most preferred option. In the case of San Francisco, this means that regardless of how well the system works, segregated choice patterns can lead to segregated schools. Even if SFUSD were able to assign every student to their first choice school, schools would remain heavily racially and economically segregated, with students from low-income and historically marginalized backgrounds concentrated in under-served schools [186]. In other words, the ultimate decisions remain heavily shaped by individuals' submitted preferences.

Although individuals' preferences shape both individual and collective outcomes, families have no formal avenues to participate in student assignment beyond independently submitting their individual, self-interested preferences. The introduction of school choice policies however, have been closely tied to desegregation efforts, an explicitly collective goal. Most parents in my sample were aware that the system was designed to promote integration and educational equity, and many supported those goals in the abstract. Several parents said that student diversity was an important factor that they looked for in a school (S5, S6, O1, O3). However, student assignment mechanisms model parents as consumers in a market for schools, rather than citizens participating in and negotiating a political process in which their own choices impact other members of their community. Parents are expected to secure a high quality education for their own child, while at the same time accepting that there are not sufficient seats at high-resourced schools for every child to have such an opportunity. As O4 put it, referring to under-resourced schools in Oakland, *"If you don't want to put your baby there, why would it be okay for me to have to put my baby there?"* Over time, these systems entrench the idea that individual choice is the only or best way to provide families access to schools, leaving school districts with fewer politically viable opportunities to promote community-level goals.

Other preference-based systems

In some systems, like assigning students to schools, dividing goods or tasks in a group of people [281, 277], or scheduling shifts [280], elicit preferences from people who will be directly, immediately impacted by the decision. In this case, it is possible to ask each person which outcome they would prefer for themselves. In other cases, preference elicitation is used to gather a wide variety of perspectives about hypothetical decision scenarios. This data is then used to train algorithms to make similar decisions that represent those perspectives in real scenarios, for example, who should receive a transplant [150] or public housing [268]. In the latter examples, people still contribute individual preferences, but those preferences are in regards to how *others* should be treated, rather than what they want for *themselves*.

However, asking people directly about their preferences for collective outcomes is still insufficient for ensuring that decisions meet collective goals or principles. For instance, people hold conflicting ethical views [29, 268, 150], in which case the choice of aggregation procedures can have critical influence over whose views are represented in the final decision [169]. Gordon et al. [169] point out that majoritarianism has silently ruled in machine learning data collection, and argue that making the pluralism of annotators' opinions more explicit can better enable decision-makers to prioritize particularly relevant or informed perspectives in a given decision.

In contexts where a decision-maker elicits people's preferences for their own outcome, we may not expect participants to be willing to trade-off any personal gain to advance collective goals. However, empirical evidence refutes this assumption. For example, in experimental settings, participants wanted to discuss and negotiate the final division of labor and goods after using a splitting algorithm, making room for cooperation and altruism [277, 281]. Giving participants insight into the complexity and constraints of the decision problem can also help them consider what trade-offs they are willing to accept, and why those trade-offs are needed [280, 281]. Further research is needed to understand how negotiation could scale beyond small groups, and how systems could engage participants in considering broader collective outcomes and trade-offs without significantly increasing the time and cognitive costs of engagement.

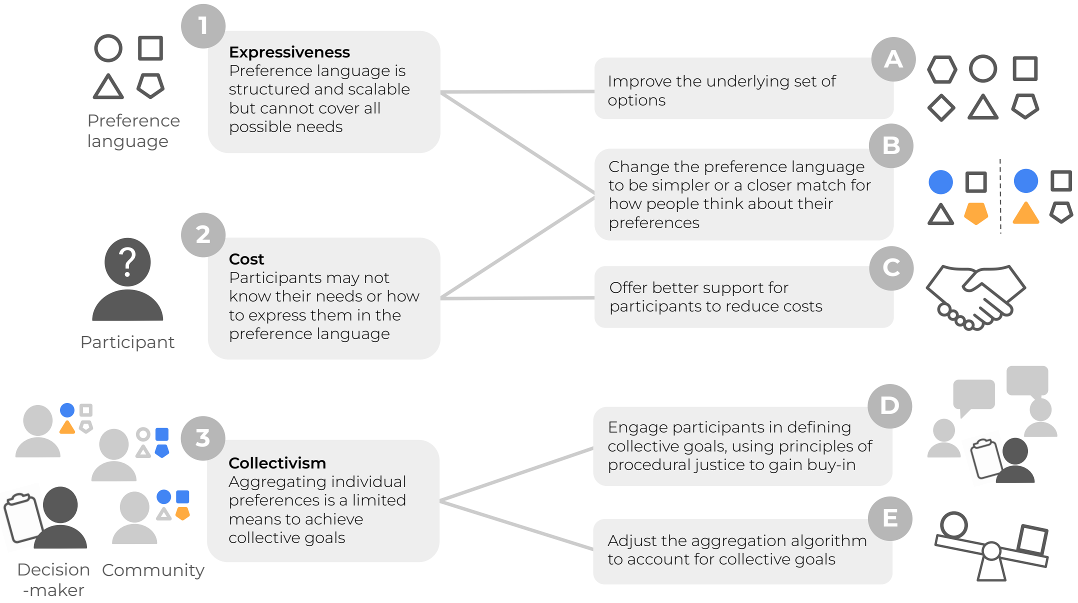

Preference language
1
Expressiveness
Preference language is structured and scalable but cannot cover all possible needs
Participant
2
Cost
Participants may not know their needs or how to express them in the preference language
Decision -maker
Community
3
Collectivism
Aggregating individual preferences is a limited means to achieve collective goals
A
Improve the underlying set of options
B
Change the preference language to be simpler or a closer match for how people think about their preferences
C
Offer better support for participants to reduce costs
D
Engage participants in defining collective goals, using principles of procedural justice to gain buy-in
E
Adjust the aggregation algorithm to account for collective goals

Figure 6.2: In Section 6.4 I identified three considerations for the design of preference languages that shape participation: 1) Expressiveness (6.4); 2) Cost (6.4); and 3) Collectivism (6.4). If the preference language makes it difficult and costly for marginalized stakeholders to communicate their needs, then the system offers unequal opportunities to participate. Systems that only account for individual preferences have limited means to reach collective goals. In Section 6.5 I discuss paths forward for designing preference eliciation mechanisms that are more expressive and less costly (A-C; 6.5) and aligning preference aggregation algorithms with co-defined collective goals (D-E; 6.5).

6.5 Addressing the Challenges with Preferences

In the previous section I described how the design of a preference language shapes what participants can convey through their preferences, how costly it is for them to do so, and the extent to which decision-makers can use those preferences to reach collective goals. Preference-based systems are gaining popularity because preferences can be a cheap and scalable way to quantify the relative value that each person has for some available alternatives, and there are a wide range of well-studied aggregation procedures that can translate individual preferences into a collective decision [66]. When the stakeholders involved are well-informed about the decision context and have relatively equal power and access to resources, this process has been shown to improve both quantitative measures of efficiency and fairness, and stakeholders' perceptions of the decisions [282]. However, problems arise in settings characterized by political tensions, power imbalances among participants, and high-stakes decision outcomes. The student assignment case study illustrates the challenges for designing preference-based systems that support equitable participation and work towards collective goals in addition to meeting individuals' needs. This case study also highlights the ways in which privileging individual choice can cause harm, particularly to those who are marginalized in the process and face greater barriers to participating.

The challenges in the student assignment domain are long-standing and complex, highlighting the need for new approaches to designing such systems. To this end, I discuss the implications of my findings for researchers and practitioners who are considering or implementing preference-based algorithmic systems. Figure 6.2 summarizes my recommendations for addressing each of the three considerations I discussed in the previous section. First, I discuss paths forward in the design of preference elicitation mechanisms that improve expressiveness and reduce costs (Section 6.5). Second, I draw on theories of procedural justice to discuss how to engage participants in co-defining collective goals, and propose technical mechanisms to account for those collective goals in aggregation procedures (Section 6.5). Finally, even with interventions to improve expressiveness, reduce costs, and account for collective goals, preference-based systems can cause harm, and are not appropriate in every setting. I conclude by discussing these risks (Section 6.5).

Preference elicitation: improve the options, simplify the preference language, and provide support

In Section 6.4, I discussed two ways that the design of the preference language shapes participation: it determines which of their needs participants can communicate, and how costly it is for them to do so. As a result, student assignment systems offer some families access to schools they prefer over their neighborhood school, but for others this choice is more elusive than meaningful. In this section I discuss three alternatives to mitigate these issues: improving the underlying set of options, simplifying the preference language, and providing support to reduce costs for participants.

Improving the options

In some cases, the preference language can only express people's preferences for a set of *existing* options. For example, the set of schools in a school district. This makes it less expressive for people who benefit less from that existing set. One path forward is to spend resources improving the underlying options to ensure every participant has a high quality option realistically available to them. A similar argument could be applied in settings like medical triage, where funds could be spent on obtaining more equipment rather than developing more complex methods for rationing the limited existing stock. Of course, this option is usually appealing to all stakeholders, but is often infeasible in the short-term. Regardless, it is worth reiterating that there can be a trade-off between spending time and money on new technology for distributing resources and improving that pool of resources.

Simplifying the preference language

A second alternative is to change the preference language that the system uses to collect participants' preferences to make it simpler or a closer match for how people already conceptualize their needs and values. For example, in the student assignment domain, one option is to reduce the choices available in meaningful ways. Currently, systems rely on each individual participant solving a complex task to reflect their preferences: they need to search among a very large set of potential schools and form a rank-ordered list of a handful of schools as their preference list. This contributes to inequality as the costs are higher for participants with less background knowledge about the school system, and more burdensome for those with less time and resources to spend on the process (as discussed in Section 6.4).

One approach to alleviate the adverse effect of informational asymmetry is offering *choice menus*, i.e., a limited subset of schools offered to each participant from which they can select their preferred schools. The basic idea is that searching in a larger pool puts those with higher search costs at a higher disadvantage compared to those with lower search costs.[4] Such choice menus could be constructed considering an applicant's characteristics (such as background and priorities). They can also act as a lever to balance the population characteristics in a school, by offering that school to a more balanced population of students.

This is in line with the approach that SFUSD has taken in their ongoing redesign of their student assignment system [437]. It is important that decision-makers consider how changes to a preference language reduce or appear to reduce people's agency and choice, and the political ramifications of that decision. For example, SFUSD has engaged families in defining geographical zone and worked to communicate with them about the reasons for the changes.

Providing support

Even with a simple and carefully designed preference language, there will be participants who need support to fully participate. For example, most people are not experts in education, and many first-time parents do not know what to look for in a school. These participants will need informational support, even with fewer options to choose from. Support interventions must account for potential participants' knowledge of the context, technology access, language proficiency, and time constraints. In Chapter 5, I advocated for an assets-based design approach, which asks how we can amplify existing resources and strategies in a community [266]. I emphasized the important of personalized, one-on-one support through trusting relationships. Other options for reducing the costs of using the preference language could include making it easier for people to apply online using devices they already own, e.g., through a mobile-friendly website.

In the next section I expand on the need to engage participants more deeply than through preference elicitation alone in order to define collective goals, and discuss technical mechanisms for working towards those collective goals in practice.

[4] As an extreme case, consider menus of length one, where the asymmetry across search costs has no effect on those with higher costs.

Preference aggregation: co-defining and accounting for collective goals

Preference-based systems have mostly assumed an individual model of preferences, where each person has some self-interested preferences for the available options, which they report to the decision-maker. However, in many distributive decisions, there are collective-level outcomes that are important but not captured in people's individual preferences. For example, many school districts have prioritized racial and socioeconomic diversity in schools, but assignment algorithms offer limited opportunities to trade-off between individual preferences for schools and collective goals.

As discussed in Section 6.4, adapting the preference language to explicitly ask people for their preferences over collective outcomes is likely not sufficient to address this problem. For one, such an approach is likely to make the preference language more complex and increase the costs of participation, in opposition to my recommendations in Section 6.5. Further, people's preferences are likely to be conflicting, leaving the decision-maker to choose explicitly or implicitly (through the choice of an aggregation procedure) which perspectives to prioritize [169, 29]. Instead, decision-makers may wish to enforce certain ethical principles or collective values, even if those do not align with the views of some stakeholders [343]. Therefore, there are two challenges to improving the alignment of preference-based systems with collective goals: defining which collective goals are important, and then ensuring that the aggregation procedure respects those goals.

Co-defining collective goals

Achieving collective goals with preference-based systems is only a problem when those goals do not align exactly with individuals' preferences. Therefore, working towards those goals will require that some stakeholders receive less personally preferable outcomes. In this way, advancing an individual preference-based system to pursue collective goals creates a deep tension between individual outcomes and collective values. It is thus especially important to consider who has a voice in shaping the collective goals, and how the decision-maker can build buy-in to those goals to establish legitimacy and trust. The principles of procedural justice offer paths forward.

Procedural justice theory teaches that the process through which a decision is made is as important as the outcome to satisfaction, perceptions of legitimacy, and compliance with that decision [494, 495]. The term "procedural justice" refers to individuals' perceptions that the procedures used to make decisions are fair. Importantly, procedural justice is a measure of subjective perceptions of fairness, not a measure of the objective fairness or equity of a decision [16]. Empirical determinants of procedural justice include the degree of voice and control individuals have in the process, whether the decision maker was respectful and unbiased, whether individuals identify with decision-makers and their values, and whether they understand the decision-making process [494, 493, 216]. Procedures that incorporate these process elements result in more satisfaction with the decision, greater perceptions of legitimacy of the decision maker, and more compliance with the decision, even when the outcome is unfavorable [491, 492]. At the same time, because procedural justice represents only the perception of fairness, yet mobilizes satisfaction and compliance with negative outcomes, it risks legitimating unfair decisions [166, 236, 473].

Rather than take individual preference satisfaction as the primary goal by hard wiring it into the assignment system, an alternative approach could encourage community deliberation and discussion to define collective goals. This process could acknowledge explicitly that in some instances, individual preferences must give way to collective values, and enable participants to determine appropriate trade-offs. With appropriate voice, participation, inclusion, and information, procedural justice principles suggest that policy makers would have a cushion of support for creating a process in which not everyone gets their first-choice school, but important collective values receive appropriate weight [496]. Procedural justice research suggests that satisfaction with outcomes, perceptions of the legitimacy of the decision makers, and acceptance of the decisions will follow.

Adjusting algorithms to account for collective goals

Once decision-makers have determined their collective goals, ideally in collaboration with the community, they need to ensure that the mechanism will account for those goals in determining outcomes. Most existing preference languages do not account for collective goals. Systems thus can feature undesirable equilibrium outcomes. For example, segregated schools can remain segregated, in part due to their reputation and the absence of means through which the participants can coordinate their decisions. In the context of assigning candidates to pre-military academies, "preference-specification" languages have been proposed that allow academies to express their preferences over the diversity of candidates they admit by specifying lower or upper quota constraints on subpopulations of candidates [12]. Building on this work, I next discuss two approaches to address this issue, using student assignment algorithms as an example: one based on making minimal changes to Deferred Acceptance, and the other based on Mathematical Programming approaches that remove focus from Deferred Acceptance and stability as solution concepts.

School-specific priority scores are often used in the course of Deferred Acceptance to ration seats among students when there is excess demand at a school. One way to shift away from segregated equilibrium outcomes is the dynamic adaptation of these priority scores. When the policy maker has a different target population distribution at a school than at status quo, then higher priority scores can be given to the most absent subpopulations. These scores should be adjusted dynamically as the population distribution gets closer to the target over time. This approach can balance the population distribution induced by the algorithm at a school while preserving stability overall.

There are other technical approaches that could more directly implement collective goals at the expense of dismissing stability. One example is methods based on Mathematical Programming which, e.g., could associate a binary variable to each student-school pair and find an assignment that respects feasibility constraints (such as schools' capacities) while taking other collective goals into account, such as a statistical distance of the population distribution at a school from a target distribution.

While I have offered paths forward for designing preference-based systems that address the challenges discussed in Section 6.4, I do not believe that preference-based systems are always appropriate. I conclude this section by discussing the risks and limitations of preference-based systems as a whole.

Limitations and risks of preference-based systems

In deciding whether a preference-based system is appropriate, one should determine the extent to which the system offers meaningful choice (and to whom), and weigh the costs involved in providing adequate support to participants. It is also important to consider how introducing choice can make alternative or complementary avenues for promoting positive change more difficult.

Preference-based systems can only improve people's access to resources if there are high-quality options *realistically* available to every participant. As discussed in Section 6.4, offering participants a choice among the same set of alternatives does not necessarily offer every participant the same value. Some would argue that offering some degree of choice is better than offering none, even if that choice can only partially reduce disparities in access to resources. For instance, many parents in my sample were able to use the choice system to access a school they preferred over their neighborhood school, while admitting that others in their community were still not afforded that opportunity. While this limitation may not be a reason to forgo a preference-based system altogether, awareness of the realistic capacity for such a system to promote social change and equalize access to resources is important so that decision-makers and advocates can ensure complementary strategies are in place.

In any participatory process there will be unequal costs associated with participating, which if not properly addressed will exclude those already at the margins. As discussed above and Chapter 5, personalized, one-on-one support is crucial for ensuring the full and equal participation of marginalized stakeholders in preference-based systems (6.5). This call to prioritize personalized support over one-size-fits-all solutions is at odds with some of the motivations for preference-based systems, e.g., that they easily scale to large groups of people. However, in order to reduce costs for participants, the decision-maker must be prepared to take on some of those costs on their behalf. To genuinely promote longer-term social change and progress towards distributive justice, we must recognize the necessary frictions and ongoing maintenance required to create inclusive, democratic, participatory processes [452]. Preference-based systems will not function equitably at scale if resources are not provided for this maintenance and support work.

A higher level political risk is that individual preference-based systems, e.g. school choice systems, can entrench the belief that individuals have a *right* to choose, and that the only optimal allocation of resources is that which efficiently maximizes people's individual preferences [204]. These systems resonate with individualist values, belief in markets, and popular neoliberal policy solutions. Collective goals, which may have been the stated motivation for the system in the first place, become tangential to the process. Further, when used to distribute public resources, this ideology shifts responsibility onto individual stakeholders to secure their access to resources to which they would otherwise be entitled, and normalizes and even exacerbates inequality. Once such a system is in place, it becomes increasingly difficult to implement policies that reduce, or seem to reduce, the degree to which the decision-maker respects individuals' preferences.

Alternatives to choice exist. For example, one alternative to ensure equitable opportunities in public education is a redistributive approach, which would create high quality programs in low-resource neighborhoods and protect local students' access to these programs [6]. Parent participation in school governance and strong teacher's unions also have a history of promoting change, for instance, fighting school closures in Black communities [145, 173], and pushing for improvements at schools like smaller class sizes, experienced teachers, and more equitable funding structures [392, 230, 428]. However, offering choice risks reducing the effectiveness or political viability of these alternatives by promoting individualism and free market values, and shifting the responsibility for ensuring the quality of children's education onto their guardians.

Finally, a question remains about whether preference-based systems should ever be considered truly "participatory" systems [330]. My findings highlight that preference elicitation *alone* does not align with the values of participatory design. For instance, in the student assignment setting, families have little to no say over the conditions or structure of their participation, or how the system should work overall. Recent work somewhat addresses this problem by integrating other forms of engagement with participants, such as conducting interviews or surveys to define what the preference language should look like [279, 150]. As discussed above, another opportunity for integrating deeper participation is in defining collective goals that the system should support (Section 6.5). In summary, the specific configuration and conditions of participation are critical factors in determining whether a particular preference-based system is truly increasing people's voice and power in algorithmic decision-making [505].

6.6 Conclusion

A convenient way to incorporate stakeholders' input into algorithmic decision-making is by collecting and aggregating individual preferences. Implementing a preference-based system requires designing a preference language, in which participants will convey their needs and goals to the decision-maker. I used student assignment algorithms as a case study to illuminate three properties of preferences languages that shape opportunities for meaningful participation: expressiveness, cost, and collectivism. When these factors are not appropriately accounted for, preference-based systems can exacerbate inequality and fail to promote collective goals. Based on my findings, I offered implications and paths forward to increase the expressiveness of preference languages, reduce costs for participants, and work towards co-defined collective goals. With these paths forward comes the warning that preference-based systems are not appropriate in every setting, and that preference elicitation alone will never be sufficient to engage stakeholders in meaningful sharing of power and agency in algorithmic decision-making.

Part II

Machine Translation

Chapter 7

Understanding and Being Understood: User Strategies for Identifying and Recovering From Mistranslations in Machine Translation-Mediated Chat

Machine translation (MT) is now widely and freely available, and has the potential to greatly improve cross-lingual communication.[1] In order to use MT reliably and safely, end users must be able to assess the quality of system outputs and determine how much they can rely on them to guide their decisions and actions. However, it can be difficult for users to detect and recover from mistranslations due to limited language skills. In this chapter we collected 19 MT-mediated role-play conversations in housing and employment scenarios, and conducted in-depth interviews to understand how users identify and recover from translation errors. Participants communicated using four language pairs: English, and one of Spanish, Farsi, Igbo, or Tagalog. We conducted qualitative analysis to understand user challenges in light of limited system transparency, strategies for recovery, and the kinds of translation errors that proved more or less difficult for users to overcome. We found that users broadly lacked relevant and helpful information to guide their assessments of translation quality. Instances where a user erroneously thought they had understood a translation correctly were rare but held the potential for serious consequences in the real world. Finally, inaccurate and disfluent translations had social consequences for participants, because it was difficult to discern when a disfluent message was reflective of the other person's intentions, or an artifact of imperfect MT. We draw on theories of grounding and repair in communication to contextualize these findings, and propose design implications for explainable AI (XAI) researchers, MT researchers, as well as collaboration among them to support transparency and explainability in MT. These directions include handling typos and non-standard grammar common in

[1] This chapter was written in collaboration with Mark Díaz and published at ACM FAccT 2022: `https://doi.org/10.1145/3531146.3534638`.

interpersonal communication, making MT in interfaces more visible to help users evaluate errors, supporting collaborative repair of conversation breakdowns, and communicating model strengths and weaknesses to users.

7.1 Introduction

To use machine learning (ML) systems reliably and safely, end users must be able to assess the quality of system outputs and determine their reliability to guide decisions and actions. This is challenging when users lack information about how the system works, or its strengths and limitations. Machine translation (MT) is one ML system that has the potential to improve cross-lingual communication. However, limited language skills in either the source or the target language makes it difficult for users to determine when the model is wrong and recover from the error.

MT can help speakers of minority languages within a given society communicate with others and access resources. For example, 9% of people living in the United States have limited English proficiency [36], which can make it more difficult for them to access critical resources including housing [168], employment [36], and healthcare [513]. While MT has the potential to help, unexpected and undetected errors can cause confusion, frustration, and embarrassment [290]. When the stakes of an interaction are higher, the consequences can be far worse; instances have been recorded when MT systems produced harmful or threatening language from benign source inputs and vice versa, with grave consequences when used by police or content moderators [484, 48, 465]. For speakers of low-resource languages, these problems stand to be more frequent due to weaker machine translation support [235].

Conversational communication is an important use case for MT [290], and casual text-based communication mediated by MT is only likely to become more widespread as MT features are embedded into messaging apps[2] and social media sites.[3] Prior research has shown that people can have successful conversations across languages using imperfect MT [187, 524], but it is unclear why users are able to identify and recover from some errors, while they are misled by others. As MT systems improve and produce increasingly fluent translations, it is especially important to understand when users are likely to be misled and how systems might intervene to promote reliable use of MT.

To this end, we collected 19 MT-mediated dyadic text conversations and in-depth debrief interviews. During the conversations, participants role played high-stakes employment and housing scenarios. In each conversation pair, one participant wrote in English, while the other participant, who was bilingual with English, wrote in one of Spanish, Persian/Farsi, Igbo, or

[2] Examples include Microsoft Translator's integration into SwiftKeys (https://web.archive.org/web/20210329164205/https://support.swiftkey.com/hc/en-us/articles/360001314546-How-to-use-Microsoft-Translator-with-your-Microsoft-SwiftKey-Keyboard) and the Google Pixel 6 Live Translate feature (https://web.archive.org/web/20211021223350/https://www.xda-developers.com/pixel-6-live-translate-messages-captions/)

[3] Facebook (https://www.facebook.com/help/509936952489634) and Twitter (https://help.twitter.com/en/using-twitter/translate-tweets) offer users the option to translate content using MT.

Filipino/Tagalog. During the conversation, participants annotated confusing translations, and in the debrief interview we showed participants their conversation transcript along with source messages and their machine translations. This allowed us to document not only users' perception of conversation quality as it unfolded, but also to identify instances of misunderstanding and unnoticed miscommunication. We conducted qualitative analysis of the conversation and interview transcripts to understand user challenges, strategies for recovery, and the kinds of translation errors that proved more or less difficult for users to overcome.

Our findings show that users have difficulty identifying translation errors, particularly when translations are fluent and might reasonably make sense in context. As a result, several participants were unaware they had misunderstood parts of their conversation until the debrief interview. Uncaught mistranslations have the potential for serious harm in real world contexts; for example, if critical information is unknowingly misunderstood, or if erroneously rude translations are attributed to a person. In addition, participants' strategies for repair often hindered achieving common understanding. Finally, some users tried to avoid translation errors by adjusting their writing style and choices, but this proved difficult to achieve in practice and risked negatively impacting the social dynamics of the conversation.

Drawing from theories of grounding and repair in communication as well as prior work in explainable AI (XAI) and FAccT, we identify promising paths forward to support users in identifying, recovering from, and avoiding translation errors. We highlight opportunities for interface design, model development, and interdisciplinary collaborations that bridge natural language processing (NLP), explainable AI (XAI), and human-computer interaction (HCI).

7.2 Related Work

Prior work has shown that people face challenges using MT in conversational settings because it is difficult to assess the quality of individual translations and, when users can identify a low quality translation, it is difficult to efficiently repair communication. To provide relevant, helpful, and actionable user support, we need to understand how users assess translation quality, and when it is particularly difficult to identify and recover from translation errors. In this section we first connect our work to the field of explainable AI (XAI) and the broader FAccT community. Then, we review related literature studying MT in conversational settings.

Transparency, Explainability, and Trust in Machine Learning

In many ML systems it can be difficult for an end-user to discern whether an output is correct or reliable. In the HCI and FAccT communities, scholars have explored trustworthy design for ML models and AI as a whole [485, 479] as well how to calibrate user trust in individual system outputs [88, 256, 534]. Research on explainable AI cites supporting appropriate user trust as a core goal [141, 293, 399]. To use an AI system reliably, a user must consider both their trust in the general design and technical underpinnings of a technology to function as expected, while also considering individual scenarios or outputs that may be more or less reliable for their needs.

In this chapter, we engage primarily with user evaluations of MT reliability in conversational settings. Investigations of trust calibration in ML systems often focus on contexts in which the user, such as a domain expert, can rely on alternative assessments if trust is questioned, such as their own judgment [355]. Often people use MT because they need to understand a language they do not know and for which no one is available to interpret or translate [290]. Jacovi et al. [229] describe distrust in AI as a mode of mitigating risk, which must be present for trust or distrust to manifest [301], but it is unclear how MT users with limited language abilities assess and mitigate the risk of inadequate translations, or calibrate their trust in MT. For example, a pilot study by Martindale and Carpuat [312] suggests that users' trust in MT was more impacted by encounters with disfluent translations than with inadequate ones. We build on this work by understanding how users assess translation quality, what information they seek when they believe a translation is poor quality, and in what circumstances they are unable to identify poor quality translations. This understanding is key to identifying future directions for human-centered explainable AI [140] for MT.

Machine Translation-Mediated Communication

As their availability has increased, MT systems have become a convenient option for people who need to communicate across language barriers [290]. Researchers have conducted user studies to understand how MT impacts communication [187, 448, 525, 526, 92, 156], and to evaluate new interface designs for conversational MT systems [522, 157, 323, 445]. In these studies, participants typically engage in text-based or spoken conversations mediated by MT to complete a given task, for example, collaborative storytelling [187], idea brainstorming [157], and simple games designed to force participants to develop shared referring expressions [526, 525, 156, 445]. Some researchers have also studied MT in more realistic settings. For example, Shin et al. observed participants using MT over the course of a four week clinical role-play [448], and Calefato et al. conducted a controlled experiment to study the impacts of MT in software requirements meetings [92].

Researchers have analyzed MT-mediated communication through the lens of grounding theory, which frames communication as a collaborative process of establishing shared understanding [112]. In this model, contributing to a conversation involves both producing an utterance, and verifying that it has been sufficiently understood by the addressee(s) [68]. Yamashita and colleagues showed that it is challenging for people to maintain grounding in MT-mediated conversation because it is difficult to know what parts of your utterance have been understood by the other person [525, 526]. These challenges are exacerbated by inconsistent and asymmetric translations, which make it difficult to maintain consistent referring expressions or make reference to an earlier part of a conversation [525].

Prior work has also found that as users interact with an MT system, they develop adaptive strategies like simplifying their language, repeating and rephrasing, and guessing the meaning of confusing translations [187, 448]. Even with these adaptations, errors can make communication frustrating, cognitively burdensome, and imprecise [187, 290, 92, 448]. Users may make incorrect guesses, may not know how to rephrase in a way that improves the translation, or they may not even realize a translation error has occurred [524]. These challenges are particularly concerning in higher stakes settings and settings with a power difference between communicators. For instance, Liebling et al. [290] describe the story of a woman who lost a job after migrating to the United States because she did not speak English and her employer found it too difficult to communicate with her via a mobile MT app. Our work builds on prior user studies with a novel focus on understanding *when and why users have difficulty identifying translation errors, and how these difficulties impact communication.* A second goal of this work is to investigate these challenges in high-stakes conversations that reflect real-world power differentials.

7.3 Method

To understand how users identify and recover from errors in MT, we collected 19 MT-mediated text-based conversations in three realistic role play scenarios across four language pairs: English-Spanish (5), English-Farsi (5), English-Tagalog (5), and English-Igbo (4). In this section we describe our user study and our approach to data analysis.

Participant recruitment

We recruited English-speaking participants and bilingual participants who knew both English and one of Farsi, Tagalog, Igbo, or Spanish through dscout, an online user research platform. Due to anticipated challenges recruiting participants fluent in low-resource languages, we also recruited from an active employee resource group of Farsi speakers at a large technology company. Prospective participants filled out a screener survey that asked for their reading and writing proficiency in English and their other language on a scale from 1 (not well at all) to 5 (very well).[4] We only recruited respondents who rated their reading and writing proficiency in both English and (for bilingual participants) their other language at least 4 out of 5.[5] For bilingual respondents, we asked which dialect of the language they knew, how they learned it, and how they use it in their life (see supplementary material). We used the answers to these questions to verify participants' experience with the language.

All but three bilingual participants rated their reading and writing proficiency in their non-English language a 5 out of 5; the other three rated their reading a 5 and their writing a 4. Five of the bilingual participants rated their English reading a 5 and writing a 4; one rated

[4] A similar scale is used in the American Community Survey English-Ability question `https://www.census.gov/content/dam/Census/library/working-papers/2015/demo/SEHSD-WP2015-18.pdf`

[5] Note: there was one exception who we recruited before adding an English proficiency question to our recruitment survey. After participating in the study, this person reported their English writing proficiency a 3 in a post-survey, and thus does not meet the inclusion criteria. However, English was their primary language at their job as an engineer in a large technology company.

their English reading a 4 and writing a 3. All participants who used English in the task rated their reading and writing proficiency in English a 5 out of 5. 22 participants were men, 13 were women, and 1 was non-binary. Participants self-reported their race or ethnicity as White (11), Asian (6), Middle Eastern (5), Black or African-American (4), Hispanic or Latinx (4), Iranian (3), Black or African-American and Hispanic or Latinx (1), and South Asian (1); 3 did not report. The median age was 36.5 years, with a range of 22-60.[6] All participants were familiar with machine translation prior to the study, and most were infrequent, casual users. The 10 English-Farsi participants were all full-time employees at a technology company, and 2 of the remaining 28 participants worked in the technology industry. We refer to participants by the study session they participated in (e.g., I1 through I4 for the Igbo-English sessions), as well as the language they wrote and received messages in.

We selected language pairs across a range of low to high-resource languages in terms of NLP training data as well as diverse geographic origins. First, we selected English-Spanish as a high-resource language pair because it is highly relevant to real world use cases in the United States. Next, we curated a list of 32 languages that were supported by Google Translate and were spoken at home by at least 100,000 people in the United States,[7] excluding Western European languages, Chinese, Japanese, and Arabic, which receive relatively high MT research attention. From this list we selected Tagalog, Farsi, and Igbo based on geographic diversity and participant availability. According to Joshi et al.'s six-point (0-5) scale of language resources in NLP, Igbo is classified as a 1 (very little labelled training data available), Tagalog a 3, Farsi a 4, and Spanish a 5 (massive investment in data collection and model development)[8][235].

Study procedure

Each study session involved two participants communicating via Google chat. We randomly assigned each pair to one of three role play scenarios developed to reflect realistic use cases for machine translation based on Liebling et al. [290]: a tenant-landlord discussion about building repairs; a cleaner and client discussing workplace safety; and a parent interviewing a prospective nanny (see supplementary material). To further reflect real-world social dynamics, we consistently assigned the role with relatively less social status in the United States (tenant, cleaner, or nanny) to the non-English language, and the role with relatively more social status (landlord, client, or parent) to English. We were conscious of the risk that this choice would reinforce stereotypical associations between immigrants in the U.S. and low wage care professions. However, given our goal to evaluate MT challenges in realistic settings, we decided that it was important to simulate situations in which people who are marginalized on the basis of their English proficiency may face discrimination. Before the conversation began, participants gave informed consent to participate, and we reminded them that they could

[6] See supplementary material for details on demographic data collection.

[7] https://www.census.gov/data/tables/2013/demo/2009-2013-lang-tables.html

[8] https://microsoft.github.io/linguisticdiversity/assets/lang2tax.txt

end the task at any time. One author was virtually co-present with each of the participants throughout the session, and continuously monitored the chat.

Each scenario required participants to resolve a disagreement and schedule a time to meet. The conversation was mediated by a custom Google Apps Script bot that translated messages using the public Google Translate API.[9] In all but four sessions, the participant using English did not have proficiency in the non-English language, and the other participant had written and reading proficiency in both languages. In four of the Farsi-English sessions both participants knew English and Farsi. Participants could not see or communicate with each other apart from the chat interface, and they saw only their sent messages and the translations of their partner's messages (*not* the untranslated source messages) (Figure 7.1). Participants were asked to use any emoji to mark messages from their partner that were unclear or confusing. These emojis were not visible to the other person and served only as flags for the debrief interviews and data analysis. The conversation task was complete when the participants agreed on a time, or after approximately 20 minutes. The conversations ranged from 9 to 34 minutes (median 22), and 10 of the 19 pairs completed the task.

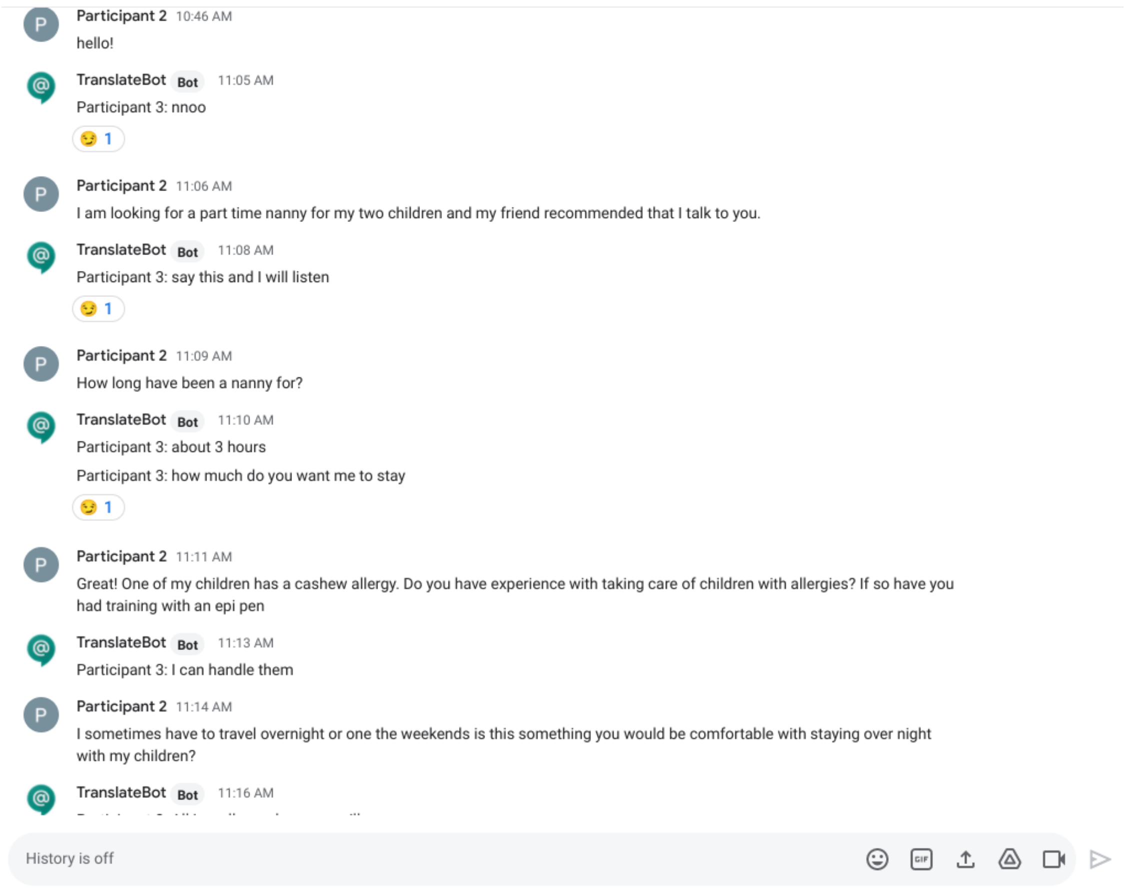
After the chat task was complete, we conducted a semi-structured debrief interview with each participant one-on-one over video call. These interviews were conducted in English and lasted 38 minutes on average (range: 16-62 minutes). We asked the participant to reflect on their conversation, discuss anything they found challenging or surprising, and whether they altered anything about the way they read or wrote messages because of the MT. Next, we showed participants a transcript of their conversation with the source text and machine translation of every turn. For each turn they received, we asked bilingual participants whether their understanding of the message had changed after seeing the source message. We also asked them to correct the translations where relevant.[10] For participants who wrote in English, we focused on messages they marked confusing and asked about their strategies to make sense of them. Interviews were recorded with consent and transcribed for analysis.

Data Analysis

Our dataset contained 19 conversation transcripts with confusion annotations and post-edits, and 38 debrief interview transcripts. First, we conducted inductive qualitative analysis on the interview transcripts [318]. The authors each independently conducted line-by-line open coding [102] on one interview transcript from each language pair. We then compared and discussed our codes. Next, we repeated this process on another set of one interview per language pair. At this point, we converged on a tentative code book, containing seven high-level codes including "error attribution," "uncaught mistranslation," "criteria for assessing quality," "error types," and "confusion strategies," each with up to seven subcodes describing

[9] We accessed the API via the Language service for Apps Script (`https://developers.google.com/apps-script/reference/language/language-app`) between July 30, 2021 and September 30, 2021.

[10] This process is called post-editing in the machine translation literature and is often a part of professional translation workflows [172]. One way to evaluate MT is to compare an MT-generated translation to a post-edited version using a string distance metric, e.g. Translation Error Rate (TER) [45].



Participant 2 10:46 AM
hello!

TranslateBot Bot 11:05 AM
Participant 3: nnoo
😏 1

Participant 2 11:06 AM
I am looking for a part time nanny for my two children and my friend recommended that I talk to you.

TranslateBot Bot 11:08 AM
Participant 3: say this and I will listen
😏 1

Participant 2 11:09 AM
How long have been a nanny for?

TranslateBot Bot 11:10 AM
Participant 3: about 3 hours
Participant 3: how much do you want me to stay
😏 1

Participant 2 11:11 AM
Great! One of my children has a cashew allergy. Do you have experience with taking care of children with allergies? If so have you had training with an epi pen

TranslateBot Bot 11:13 AM
Participant 3: I can handle them

Participant 2 11:14 AM
I sometimes have to travel overnight or one the weekends is this something you would be comfortable with staying over night with my children?

TranslateBot Bot 11:16 AM

Figure 7.1: An extract of the user study interface from the perspective of I4, English.

specific examples (e.g., “confusion strategies: ignore confusing part” and “criteria for assessing quality: effort.”) We then split the remaining interview transcripts and each continued this coding process on half of the data. We were in regular communication throughout this process, adding codes as necessary and resolving instances where we were unsure about our coding.

We then identified two phenomena of interest. First, we noticed that there were several instances where a participant thought they knew what their partner was trying to say, but had in fact misinterpreted an incorrect translation (“uncaught mistranslation”). Second, we noted several strategies that participants used when they were not sure about the meaning of a translation. We conducted deductive coding of the conversation transcripts. We read each conversation transcript in parallel with the associated interview transcripts and coded each conversation turn for: (a) recipient’s understanding of the intended meaning (total / partial / none / bilingual), where “partial” referred to turns that the participant indicated they understood a portion but not all of a message, and “bilingual” referred to turns where the person explicitly relied on their knowledge of the other language to understand a literal translation; (b) response action (ignore / repeat / simplify / add detail / generalize / clarify / guess); and (c) uncaught mistranslation (yes / no). This coding process was non-exhaustive, because we relied heavily on what users verbalized in the debrief interview. We used these codes to connect our analysis across the interview transcripts and the conversation transcripts, and report counts of these codes as *rough* estimates of the frequency of different phenomena in the data. Although the method is non-exhaustive, triangulating participant behavior using both interview and chat transcripts allowed us to make sense of participant responses without interrupting the flow of live conversation.

The 19 conversations featured 628 turns and 938 sentences. In total, 228 turns were either marked confusing (in situ with an emoji) or post-edited. Including turns that we coded as "none" or "partial" understanding, or "uncaught mistranslation," there were a total of 236 turns that caused miscommunication or were post-edited. Throughout the paper, when referring to or reporting the intended meaning of messages in Spanish, Farsi, Tagalog, or Igbo, we use participants' post-edits, i.e., their own translations of their messages to English.

7.4 Results

In this section we present our findings regarding how users identified errors, when they were unable to confidently identify issues, and how these challenges shaped conversations. In the following section we discuss the implications of these findings for the design of MT systems.

Fluency and dialogue flow are used as a (misleading) proxy for adequacy

Participants used the fluency of the translations they received and the logical flow of the dialogue as a proxy to judge translation quality. While fluency, dialogue flow, and adequacy were often correlated, this was not always true, leading to unidentified mistranslations.

Participants often referred to the fluency or flow of a conversation as evidence that the translations must have been accurate. Participants described conversations as *"smooth, easy"* (S2, Spanish), *"flowing"* (I4, Igbo), and *"pretty straightforward,"* (S1, English). Several participants felt very confident that the translations were accurate, even without verifying that the other person felt similarly or receiving any additional information about the quality of the translations. As S5, English put it, *"I would say like a hundred percent of the time. I was able to understand everything that the person was trying to say."*

However, this perception did not always align between conversation partners. The Igbo speaker in I1 indicated the conversation was *"smooth, there was no confusion,"* while his partner, who was using English, complained that she *"would be super frustrated"* if it had been a real life conversation. As shown in Figure 7.2, there was asymmetry in how often participants were confused by translations between the two directions of a language pair. This sometimes hindered repair if the participant receiving higher quality translations was

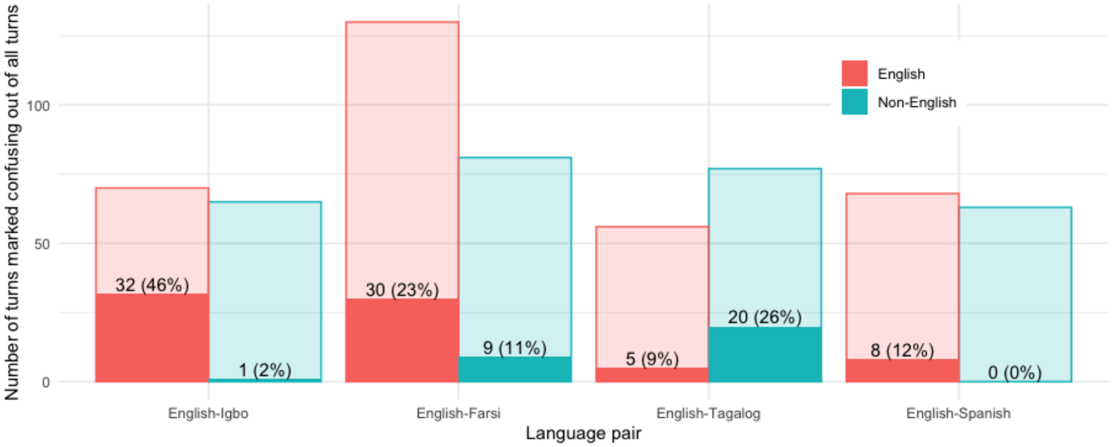


Figure 7.2: Participants using English were confused more than their partners when speaking with an Igbo, Farsi, or Spanish speaker, with Igbo to English translations annotated for confusion at the highest rate. In English-Tagalog sessions, Tagalog-speakers marked a higher proportion of messages they received as confusing.

confused by clarifying questions, not realizing how poorly their own messages were translated. This may reflect not only underlying asymmetries in quality, but also asymmetries in users' tolerance for errors. In the debrief interviews, Igbo and Farsi speakers, in particular, showed a higher tolerance for errors, informed by negative past MT experiences in their languages.

In some circumstances, participants were misled by messages that were both fluent and seemed to make sense in context, despite conveying the wrong meaning. In I4, The parent-participant explained to the nanny-participant that: *"One of my children has a cashew allergy. Do you have experience with taking care of children with allergies? If so have you had training with an epi pen,"* but the Igbo speaker interpreted the translation of this message to mean that the child was stubborn, and replied, *"ewerem ike ijikwa ha* [I can handle them]" The parent-participant took this as confirmation that the nanny-participant could handle a severe allergy and moved forward with the conversation. At the start of the debrief interview, the nanny-participant said, *"I understood everything,"* (I4, Igbo) but after seeing the conversation transcript realized he had not.

In the real world, a parent might make more effort to be certain that a prospective caretaker has fully understood their child's medical needs. Nevertheless, this is a powerful example of how harm could arise from translation errors that users are not able to identify. In another example T1, English received an untranslated sentence, *"Sige po* [Okay]," but the rest of the message was translated fluently, leading him to believe that it might refer to some kind of generational slang he was unfamiliar with. Our data indicates that fluent but inadequate translations are a particularly risky type of mistranslation, offering additional qualitative evidence in support of Martindale and Carpuat's findings that fluency has a greater impact than adequacy on people's perceptions of translation quality [312].

Difficulty attributing errors has social consequences

Participants were more skeptical of messages that seemed disfluent or out of place, but still lacked information to identify whether poor quality machine translation was to blame or if the other participant had said something they perceived to be odd or inappropriate. This uncertainty made it more difficult for participants to decide how best to resolve confusion and risks negative social consequences.

Issues with tone and formality were frequent across language pairs. One participant described messages he received as *"abrupt," "rude," "demanding,"* and even *"flirtatious"* (S6, Spanish). Other participants noticed instances where the other person seemed to be *"blaming"* (T5, Tagalog), or *"not respectful"* (T1, Tagalog). Recipients of these messages seemed to struggle to separate their judgment of the translation from their judgment of the other person.

The English-speaker in one Igbo-English session easily attributed errors to the machine translation when words weren't translated or when the English did not make any sense to him (e.g., when *"onwe ihemcho igwagi* [I have something to tell you]" (I3, Igbo) was translated to *"self-interest to talk"*).

> *"If [...] the words were English, but there were a few non-English words, then I assumed that the other person typed a legible, totally legit [message] and that the translator had for some reason not worked on a few of those words."* (I3, English)

However, when more nuanced translation errors occurred later in the conversation, the same participant assumed the message reflected the other person's intent. The Igbo-speaker, playing a tenant, wrote *"biko ke mbe i ga kpota mmadu idozie uko ulo a* [Please, when can you get someone to fix this ceiling?]," but it translated to, *"Please hurry up and get someone to fix the ceiling."* In the debrief interview, I3, English expressed surprise by this wording, *"The please hurry up was, that was like, I don't expect that as a landlord, I guess."*

T1, English also found it difficult to distinguish MT errors. In reference to one translation he initially said, *"see, that's confusing to me, but I'm chalking that up to the person writing it rather than the translator,"* but later admitted that *"I'm not sure if that was the translator or the person writing it."* In I1, a straightforward clarifying question was completely mistranslated and was interpreted as rude. *"I felt like the person was like, you know, this person was tired of talking with me they just wanted me to go away"* (I1, Igbo).

These difficulties influence the social dynamic both by shaping people's perceptions of others, but also by shaping how people communicate themselves. People tend to mirror the language of the person (or agent) they are interacting with [69, 67, 185], and we saw that this remained the case even if participants weren't sure whether they were mirroring their partner or the MT.

> *"Now that I'm looking at it in English, it looks like it [the other person's messages] would just be as if I was chatting with a friend, but I guess a translator makes*

> *it formal, so I was responding more formal based on how [the messages] were translating."* (S2, Spanish)

Ultimately, participants struggled to distinguish MT errors from genuine interpersonal miscommunications, with potentially negative consequences for the conversation and interpersonal dynamic.

Guessing and ignoring errors can widen the understanding gap

When users were able to identify errors, they then had to determine how to move forward. Consistent with prior work [522, 524, 187], the most common strategy was to ignore parts they couldn't understand (Table 7.1). Guessing the meaning was also common. When participants ignored errors, they either responded to the parts they did understand, crafted a more general response, or changed the subject altogether. While these strategies frequently kept the conversation going, they sometimes created a false perception of mutual understanding.

If a participant believed they understood enough of a message to formulate an appropriate response, they often chose to ignore the part they did not understand.

> *"Maybe half of the sentences that translations were not correct, but because [the other person] for a few times said a few sentences and next to each other, I was able to understand what he means by understanding at least one or two of them."* (F1, English)

Participants also ignored mistranslations when they felt that the information was not critical to completing the task at hand, seeking *enough* understanding rather than perfect understanding.

> *"I had to [ignore an untranslated part] because I looked at the bigger picture and I said, okay. Well, whatever [the other participant] said there was kind of not relevant to what I was trying to solve."* (T1, English)

When messages seemed slightly off, participants were able to (often subconsciously) make meaning by guessing related terms that would allow a clearer interpretation of the meaning.

> *"When I go back and look at them, I think the fact that it was basically one word, that made me think that it was either typo or a translation thing, like the one where they said, "I don't have a CPR, but if you can pay for the training, you can take a class." Well, I just swapped out "you" for "I," meaning them, and that's what I figured was happening there."* (S3, English)

After interpreting a confusing message, participants generally responded as best they could to continue the conversation. Some were able to formulate a relevant response even with very little understanding of what their partner had said.

Strategy	N turns
Ignore	60
Guess	21
Clarify	20
Generalize	3
Repeat	3
Simplify	2

Table 7.1: Ignoring confusing parts of a message was the most popular strategy, followed by guessing the meaning or asking for clarification.

> *"[The translation] basically didn't make any sense. [. . .] That's why I just kind of went and answered with something in general. Like, "When can we start this?" [. . .] because I wasn't sure like this, it's telling me that their roofs are not going to be able to be used on Sundays and I'm like, I don't I don't know what to do with this information."* (T5, Tagalog)

If a response was sufficiently relevant, the sender of the original confusing message often accepted this as evidence they had been understood [113], not realizing that the other person may have made an incorrect guess or even ignored part of their message altogether. For example, I2, Igbo sent a message that said *"enwere m oge abalị* [Have a good time]," but was translated to *"I have a night time."* The recipient interpreted the message incorrectly, explaining, *"He's okay with nights. That's what that means to me. I don't think that's terribly off,"* (I2, English) and moved on. This misunderstanding was never caught and the Igbo speaker's well wishes were never received.

An accumulation of partially understood messages and vague responses made it difficult to have a specific conversation. For example, two participants playing the landlord role (I3, English and F2, English) understood that the tenant-participant had a leak, but found it difficult to ascertain its seriousness.

Avoiding errors is difficult even with conscious effort

While many participants had theories about how to produce the clearest translations, it was difficult to control translation quality in practice. First, it was difficult for participants to know which strategies worked. Moreover, even strategies that work in one case may fail in another, and can be difficult to maintain throughout a conversation.

The most common beliefs about MT among the participants were that it performs poorly on long sentences and complex sentence structures, that it is sensitive to spelling and grammar, and that it often translates idioms and metaphorical language literally, failing to convey their meaning. While these theories were largely consistent with the limitations of MT models, putting them into practice proved difficult. For example, two participants mentioned a trade-off between keeping messages simple and including sufficient detail.

> *"I was like, should I just like, you know, just give one word [response]? Like "no?" [. . . O]r should I say "no experience?" [. . .] I don't really know how I'll be able to respond to it for the person to understand."* (I2, Igbo)

More broadly, the MT model's limitations were at odds with realistic features of casual language. In addition to occasional typos, users rarely used complete punctuation, and the system frequently failed to translate messages at all when users omitted diacritics.[11] Participants also frequently used idiomatic and metaphorical language, even when they had intentionally tried to avoid it, leading to strange literal translations, e.g., "standing water" was translated into Farsi using the word to describe a person standing up.

An issue with attempts to simplify language is that it risks shaping and constraining human communication around the limitations of existing machine translation tools. For example, the English participant in F1 tried rephrasing a message several times in an effort to improve the translation. However, this process changed the tone of his message.

> *"So I repeated the same sentence a few times. [. . .] And every time I made it simpler and simpler because he wasn't understanding that I need the building cleaned by tonight. So for example, "I need this building cleaned by tomorrow," and very direct way of saying things, which I usually don't say. For example, if you were a real cleaner, I would say, "can you please clean the building for me," but when I was doing [the task], I told him, "I want this building cleaned by tomorrow." So I gave him very direct orders."* (F1, English)

Although the translation may have eventually conveyed the core meaning of the source text, the MT is indirectly shaping the interpersonal dynamic in an undesired way.

7.5 Discussion

Two key principles of good user interface design are to prevent errors where possible, and when errors do occur, help users quickly recognize, diagnose, and recover from them [340]. Our findings identify important challenges that users face in identifying, recovering from, and preventing miscommunication due to translation errors in MT-mediated chat. We contextualize these findings in existing theories of computer-mediated communication and human-AI interaction to identify next steps for MT model development and user interface design that could improve the user experience in each of these areas. We end with a discussion of how systems could adapt support to different contexts to provide relevant and useful information without becoming intrusive.

[11] This was particularly a problem for Igbo-English because several of the Igbo-speaking participants did not know how to access diacritics on their computer's keyboard.

Identifying and recovering from errors

It is difficult for MT users to identify translation errors without knowing both the source and target languages. Our findings echo concerns that this is especially difficult with current state of the art neural machine translation systems, which can produce very fluent translations that are not necessarily adequate [312, 43]. One risk of language models that produce seemingly fluent and coherent output is that people are inherently driven to make meaning from such outputs, regardless of how they were produced or whether they reflect any meaning or intent [43]. In this study, participants had confidence in their interpretation of apparently fluent and coherent translations, even when their interpretation did not match their partner's intent. Thus, there is a need for novel approaches that interrupt this process and help users identify and recover from translation errors.

Make MT more visible

Participants found it difficult to identify MT errors and frequently attributed system errors to their conversation partner. In real world scenarios, translations that are erroneously offensive or rude translations, or that fail to convey well wishes could alter how users are able to present themselves and jeopardize interpersonal relationships, with potentially serious consequences in cases where users are seeking employment or assistance [185]. MT-mediated communication has historically been designed to feel seamless and as close as possible to a chat with someone speaking the same language [445]. However, this seamlessness may actually make it more difficult for users to identify and attribute errors, and easier for them to forget that MT is in use. A 2014 study by Gao et al. found that users attributed errors *less* to their conversation partner when they believed the conversation was mediated by MT, compared to when they believed they were speaking to someone for whom English was a second language [156]. As MT becomes more fluent, it becomes less salient to users, possibly making them more likely to attribute errors to their partner even when they are initially aware of MT.

Future work could investigate how designs that make MT more visible (see, e.g., seamful design [222, 100]) or adopt alternative metaphors for MT, such as that of an agent or interpreter [445], could heighten users' awareness of MT, help users identify errors, and reduce their tendency to attribute MT errors to their conversation partner. This is aligned with approaches to explainable AI that seek to encourage more deliberate and critical thinking about model predictions before making a decision about whether to rely on them [75]. At the same time, increased visibility may not always be appropriate or desired. One challenge will be designing tools that help users rely on predictions appropriately without adding frustration. Making MT more visible may limit users who want to retain control over how and whether they share aspects of their identity, including their language abilities, which may be associated with stigmatized social categories [39] and which can be the basis for linguistic discrimination (e.g., [181]).

Warn users when errors occur

Helping users identify incorrect predictions is a challenge across machine learning domains [293, 141]. In MT, there has been sustained effort to develop quality estimation (QE) models, which predict the quality of a translation without comparison to a reference translation and could thus be used to warn users of low quality translations in real time [56, 460, 96, 459]. This prediction task has proved difficult, and it is not clear what kind of quality indicators (prediction targets) would be both feasible to predict and helpful and actionable for end-users [460]. One study by Miyabe and Yoshino suggests that it is difficult for users to apply numeric quality indicators to repair translation errors, particularly if those quality indicators could, themselves, be inaccurate [322]. One direction for future work is to focus QE and other translation-level information interventions on specific kinds of errors that are particularly difficult for users to identify. For example, our findings suggest that supporting users to identify fluent but inadequate translations, as well as errors that change the tone of a message should be a high priority for conversational MT. The dominant approach to QE has been supervised learning, which requires expensive labelled training data, favoring high-resource languages. Given that people using MT with low-resource languages are those most in need of support, QE methods that are effective for low-resource languages will be especially critical.

Support collaborative repair

Theories of repair in communication suggest that people prefer to identify and correct errors in their own messages before sending them, avoiding the need to expend collaborative effort on repair [422]. Prior research has proposed interfaces that encourage self-repair, for example, by showing users the back-translation of their message [446, 526, 325], or suggesting changes to improve the translation [323, 445, 324], but even with support this process is challenging for users who do not speak the target language. Repair costs are shaped by the medium of communication [112]; in MT-mediated communication, users' preference for self-repair may be much weaker because they are forced to guess if their self-repair attempt is likely to be successful. Future work could examine lower cost mechanisms for engaging in collaborative repair. For example, Hu et al. developed a system that allows two monolingual people who each know a different language to collaboratively produce high quality translations from one language to the other using MT [215]. Another possibility is to develop interactions that support repair without relying on MT. In this study, participants annotated messages with emojis to indicate to the research team that they found a message confusing, but the other participant could not see those annotations. One participant often received more confusing translations than the other (Fig. 7.2), but it was difficult to communicate that to the other person. Offering a specific annotation that both participants can see to indicate that an entire translation, or a portion of it, is unclear could enable lower cost repair activities. Encouraging collaborative repair could avoid disproportionately burdening one person in a conversation with identifying and resolving misunderstandings. Another possibility is to offer standard clarification utterances that have been professionally translated. Such phrases could ease the difficulty of communicating specific issues such as pointing out untranslated words or asking for a statement to be reworded.

Preventing errors

It is important to be able to identify and recover from errors when they happen, but it is even better when users can prevent those errors from happening in the first place. Existing MT systems offer little insight into model performance, despite widely varying performance across language pairs, and even directions within a language pair [337, 532]. Although MT developers are aware of systematic weaknesses (e.g., [227, 46, 483, 434]), this information is not conveyed to end-users. Instead, users must develop their own theories about MT's strengths and weaknesses through interacting with these systems over time. Theories based on interactions with MT in a particular language pair at a particular time may be misleading when applied to a different language pair or after updates to the model. Moreover, not acknowledging disparities in performance between languages with large investments and those with less support reinforces an expectation that speakers of lower-resource languages should accept poorer performance. Greater transparency into model performance across language pairs and on specific types of language could help users better adjust their expectations, calibrate their trust in the system, and learn how to minimize the risk of translation errors.

One path forward is to develop onboarding materials for new users to teach them about the system's capabilities and known failure modes [88, 17, 510]. While lengthy instructions may not be feasible across all use cases, visual and contextual indicators or warnings could be a first step toward onboarding nudges. Further research is needed to identify what would help users understand what the system can and cannot do, and then apply that understanding to avoid harmful translation errors. This engagement must be ongoing; when the MT model is updated and improved, users should be kept up to date with specific guidance about how to update their strategies for reliable use [17].

Systems could also use interactive teaching strategies [510] and provide reminders when a user tries to use the system in a way that is not supported. For example, human communication is rarely fluent and free of errors [68], but MT systems perform poorly on text with typos, abbreviations, grammatical errors, and other normal features of casual language. While telling users about this limitation upfront would be useful, even users who are explicitly aware of these limitations struggle to abide by them consistently, especially in text messaging where casual language is broadly accepted and expected. As MT models are integrated into messaging apps and social media, a priority should be to ensure they are robust to casual language. A complementary approach could be to interactively assist users to write in a way that is suited to current MT capabilities, from interventions that are straightforward with existing technology like spelling and grammar correction, to more sophisticated interventions like detecting and suggesting alternatives for idiomatic and metaphorical language. Certainly, such an approach would constrain how people are able to communicate when using MT. Over time, this could lead to changes in language use driven by the arbitrary constraints of MT models, especially if inputs to MT systems are then used as training data for future systems.

However, with careful attention to this dynamic we can help users work within the limitations of existing MT systems, while simultaneously expanding system capabilities to reduce those constraints in the future.

Adapting support to the context of use

A consistent finding across study sessions was that people are tolerant of translation errors and can have successful conversations without perfect MT. Because this study involved a role play, participants may have been more accepting of partial understanding than they would be in real life. This is consistent with the idea in grounding theory that people's *grounding criterion*, or how much evidence a person needs that the other person has understood them before they move on with the conversation, changes not only with the medium of communication, but also with the purpose [112]. In different situations, we would expect users to hold MT to a different standard and adopt different strategies for assessing translation quality.

Accordingly, when designing and evaluating MT models and user interfaces, we should be accounting for the purpose of communication and evaluating how well the system serves that purpose. Our findings and the next steps we have proposed above, for example, are specific to MT-mediated text chat with a clear task or goal. Translation systems that adequately serve this purpose may be less effective for conversations with open-ended or creative goals, such as story-telling or getting to know someone, or communicating information that needs to be understood verbatim. Translation systems that use other modalities, such as speech-to-speech translation, also introduce different challenges. Hara and Iqbal [187] found that people using MT over video call also face challenges identifying and recovering from errors, and that users employ similar strategies to recover, like simplifying their language. However, the most useful interventions to improve communication may differ. For instance, visual and audio cues may make it easier for a user to identify misunderstanding, while text may be more conducive to identifying and correcting specific errors [187].

Users' purposes for MT can also shift over time, or even within a conversation. For example, a conversation between a parent and a prospective caretaker could easily shift between friendly chat and building rapport, to sharing critical information about a child's health condition. In high-stakes discussions, such as discussing allergies, people's tolerance for errors may be very low. In fact, studies of MT use in healthcare have found that patients and healthcare providers prefer phrase-based translation tools over open-ended MT because they are more reliable [457, 490, 358]. One path forward could be considering ways to smoothly integrate different types of translation support to match users' relative need for accuracy and flexibility in different contexts. Ideally, MT systems would be designed for flexible use, offering more or less intrusive support based on the context and stakes.

On the other hand, given that users' criteria for assessing translation quality shift according to the context, users' perceptions of translation quality may be an inconsistent proxy for their actual understanding. Several prior studies that introduce new interface designs for MT rely on users' *perceptions* of clarity to evaluate the new system, but do not compare those perceptions against other measurements of translation quality (e.g., professional translators' evaluations) to determine whether users actually understood the intended meaning. This makes it difficult to know whether these designs improved users' *actual* understanding and quality of communication, or whether they only improve *perceived* understanding and quality. By engaging bilingual participants in debrief interviews with the full conversation transcript and translations, we were able to compare users' perceptions of quality in situ to their understanding of their partner's intended meaning. The fact that we saw several instances of uncaught mistranslations, where a participant thought they understood a message until they saw the original source text, suggests a need to consider this gap more explicitly in future evaluations of systems designed to improve understanding in MT-mediated communication.

Limitations & Opportunities for future work

We faced several trade-offs in designing the study to resemble real-world high-stakes communication while remaining feasible. Here we identify drawbacks of our approach and discuss how they could be addressed in future work.

We recruited bilingual participants for two reasons: first, bilingual participants were able to compare the source messages and translations in the debrief interview, offering us insight into the difference between in situ *perceived* quality and *actual* quality of MT-mediated communication; second, it allowed us to conduct recruitment and debrief interviews in English. However, users in the real world are unlikely to be using MT to translate between languages they are fluent in, making our set-up less realistic. Further, bilingual participants were sometimes able to infer meaning from poor quality translations that would be difficult for someone who does not speak the source language to understand. For example, idiomatic or metaphorical language translated literally may be intelligible to a bilingual person because of an ability to backtranslate. We partially addressed this limitation in the Spanish, Igbo, and Tagalog sessions by having only one bilingual participant in each pair. Future work could improve on this further by recruiting only participants who have limited or no knowledge of their target language and hiring professional translators to assess translation quality.

We also faced issues with the limitations of the input devices that participants had available. Particularly in the Igbo sessions, some users could not access certain diacritics on their laptop. It is possible that this reflects realistic real-world use, but this is not something that we investigated. Future work could identify what kinds of input devices users might typically have access to when using MT with a specific language and replicate this in user studies.

Finally, the participants knew that they were role playing, so our study only partially replicates realistic high-stakes scenarios and power dynamics. Our insights could be further understood by observing MT-mediated interactions in the real world, for instance, drawing on ethnographic methods [274] or contextual inquiry [50]. Our choice of task prompts, and the choice to put an English speaker in the position of relative social power reflect our context as U.S.-based researchers, and would be complemented by future work in other cultural and linguistic contexts.

7.6 Conclusion

In this chapter we conducted a user study to explore how users evaluate translation quality and recover from translation errors in MT-mediated text conversations. 19 participant pairs engaged in an MT-mediated role-play conversation modeled after real-world, high-stakes scenarios in English and one of Spanish, Persian/Farsi, Igbo, or Filipino/Tagalog. Through analysis of debrief interviews, chat transcripts, and annotations of confusion provided by participants in situ, we demonstrate that users have difficulty identifying translation errors and validating their own understanding, particularly when translations are fluent, but inadequate. Often these difficulties were asymmetric within conversation pairs and participants were not always aware of their partner's difficulties, at times leading them to attribute MT errors to their partner. We build on existing scholarship in explainable AI (XAI), FAccT, and HCI to identify directions for interdisciplinary research and design to support users in identifying and recovering from MT errors. These directions include handling typos and non-standard grammar, making MT in interfaces more visible to help users evaluate errors, supporting collaborative repair of conversation breakdowns, and communicating model strengths and weaknesses to users.

Chapter 8

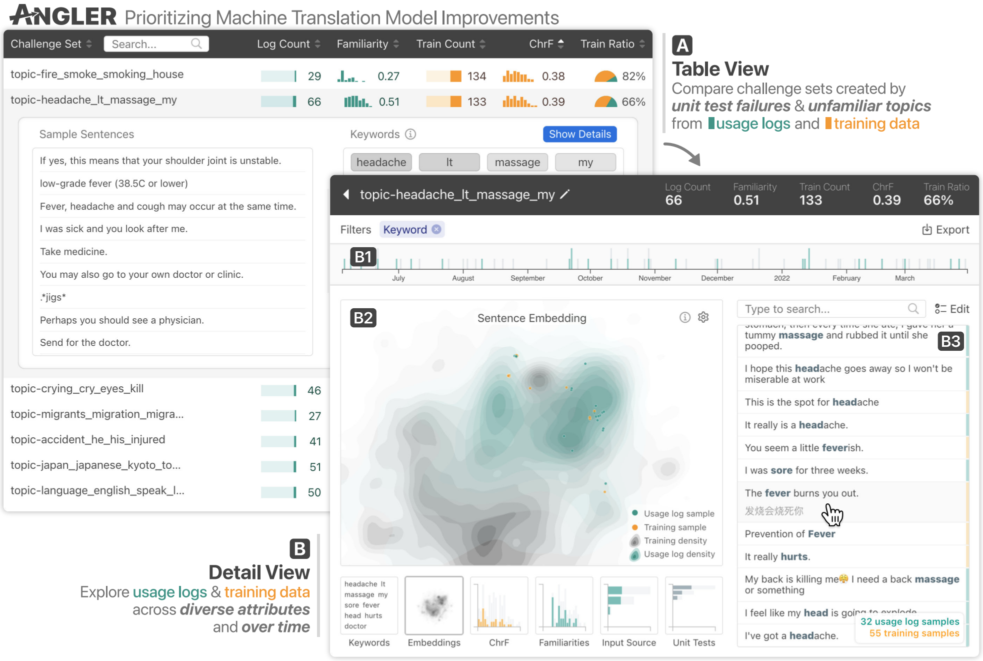
Angler: Helping Machine Translation Practitioners Prioritize Model Improvements

Machine learning (ML) models can fail in unexpected ways in the real world, but not all model failures are equal.[1] With finite time and resources, ML practitioners are forced to prioritize their model debugging and improvement efforts. Through interviews with 13 ML practitioners at Apple, we found that practitioners construct small targeted test sets to estimate an error's nature, scope, and impact on users. We built on this insight in a case study with machine translation models, and developed ANGLER, an interactive visual analytics tool to help practitioners prioritize model improvements. In a user study with 7 machine translation experts, we used ANGLER to understand prioritization practices when the input space is infinite, and obtaining reliable signals of model quality is expensive. Our study revealed that participants could form more interesting and user-focused hypotheses for prioritization by analyzing quantitative summary statistics and qualitatively assessing data by reading sentences.

8.1 Introduction

In machine learning (ML), out-of-sample evaluation metrics are used to approximate how well a model will perform in the real world. However, numerous high-profile failures have demonstrated that aggregate performance metrics only estimate how a model will perform *most of the time* and obscure harmful failure modes [e.g. 263, 125, 59, 349]. In response, researchers have explored how to anticipate model failures before they impact end users. For example, disaggregated error analysis has helped identify errors that impact people with marginalized identities. Prominent examples include facial recognition models failing

[1] This chapter was written in collaboration with Zijie J. Wang, Dominik Moritz, Mary Beth Kery, and Fred Hohman and published at ACM CHI 2023: `https://doi.org/10.1145/3544548.3580790`.



Challenge Set	Log Count	Familiarity	Train Count	ChrF	Train Ratio
topic-fire_smoke_smoking_house	29	0.27	134	0.38	82%
topic-headache_lt_massage_my	66	0.51	133	0.39	66%
topic-crying_cry_eyes_kill	46				
topic-migrants_migration_migra...	27				
topic-accident_he_his_injured	41				
topic-japan_japanese_kyoto_to...	51				
topic-language_english_speak_l...	50				

Sample Sentences

If yes, this means that your shoulder joint is unstable.

low-grade fever (38.5C or lower)

Fever, headache and cough may occur at the same time.

I was sick and you look after me.

Take medicine.

You may also go to your own doctor or clinic.

.*jigs*

Perhaps you should see a physician.

Send for the doctor.

Keywords: headache, lt, massage, my

topic-headache_lt_massage_my — Log Count 66, Familiarity 0.51, Train Count 133, ChrF 0.39, Train Ratio 66%

Keywords: headache lt massage my sore fever head hurts doctor

tummy **massage** and rubbed it until she pooped.

I hope this **head**ache goes away so I won't be miserable at work

This is the spot for **head**ache

It really is a **head**ache.

You seem a little **fever**ish.

I was **sore** for three weeks.

The **fever** burns you out.
发烧会烧死你

Prevention of **Fever**

It really **hurts**.

My back is killing me😤 I need a back **massage** or something

I feel like my **head** is going to explode.

I've got a **head**ache.

32 usage log samples
55 training samples

A **Table View**
Compare challenge sets created by ***unit test failures*** & ***unfamiliar topics*** from 🟩usage logs and 🟧training data

B **Detail View**
Explore usage logs & training data across ***diverse attributes*** and ***over time***

Figure 8.1: Angler enables ML developers to easily explore and curate challenge sets for machine translation. **(A) The *Table View*** lists all challenge sets, allowing users to compare them by metrics such as sample count, model performance, and familiarity score. After selecting a set, **(B) the *Detail View*** allows users to further explore samples in this set across various dimensions. **(B1) The *Timeline*** enables users to query data samples by time. **(B2) The *Spotlight*** presents visualizations with linking and brushing to help users characterize the set from different angles. **(B3) The *Sentence List*** shows all selected data samples and allows users to further fine-tune before exporting this challenge set for downstream tasks.

to recognize women with dark skin tones [76] or translation models perpetuating gender stereotypes [463]. However, subgroups where a model fails can be highly contextual, specific, and may not match any social category (i.e., "men wearing thin framed glasses" [87] or "busy/cluttered workspace" [121]). It remains an open challenge for ML practitioners to detect *which* specific use case scenarios are likely to fail out of a possibly infinite space of model inputs —and prioritize *which* failures have the greatest potential for harm [207, 35].

With finite time and resources, where should machine learning practitioners spend their model improvement efforts? **In this chapter, we aim to help practitioners detect and prioritize under-performing subgroups where failures are most likely to impact users**. Towards this goal, we contribute the following research:

- **A formative interview study with 13 ML practitioners at Apple** to understand their process for prioritizing and diagnosing potentially under-performing subgroups (Section 8.3).

Practitioners rely on the model type, usage context, and their own values and experiences to judge error importance. To test *suspected* issues, practitioners collect similar data to form **challenge sets**. Using a challenge set, practitioners rely on a combination of signals from model performance and usage patterns to gauge the prevalence and severity of a failure case. The most common fix for an under-performing subgroup is dataset augmentation to increase the model's **coverage** for that subgroup.

- **Angler (Figure 8.1), an open-source[2] interactive visualization tool for supporting error prioritization for machine translation** (MT) (Section 8.4). Since our research centers on the issue of *prioritization* (rather than specific error identification) we chose an ML domain where practitioners cannot directly observe model errors. MT developers do not speak all the languages that their translation models support. They rely on proxy metrics like BLEU [360] to estimate model performance but ultimately depend on human expert translators to obtain any ground-truth. Since gathering feedback from human translators is expensive and time-consuming, careful allocation of annotation resources is crucial. To help MT practitioners prioritize suspected error cases that most align with user needs, we worked with an industry MT product team to develop Angler. By adapting familiar visualization techniques such as *overview + detail*, *brushing + linking*, and *animations*, Angler allows MT practitioners to explore and prioritize potential model performance issues by combining multiple model metrics and usage signals.
- **A user study of 7 MT practitioners using Angler** to assess the relative importance of potentially under-performing subgroups (Section 8.5). MT practitioners completed a realistic exercise to allocate a hypothetical budget for human translators. Observing MT practitioners using Angler revealed how they use their intuition, values, and expertise to prioritize model improvements. Direct inspection of data showed the potential to encourage more efficient allocation of annotation resources than would have been possible by solely relying on quantitative metrics. While rule-based error analysis allowed participants to more successfully find specific model failure patterns, exploring data grouped by topic encouraged practitioners to think about how to improve support for specific use cases. The study also prompted discussion for future data collection and helped practitioners imagine new features for translation user experiences.

No model is perfect, and large production models have a daunting space of potential error cases. Prioritization of subgroup analysis is a practical challenge that impacts model end users. By exploring prioritization in the context of MT, where there are no reliable quality signals for previously unseen model inputs, we highlight the value of flexible visual analytics systems for guiding choices and trade-offs. Our findings support the potential for mixed-initiative approaches: where automatic visualizations & challenge sets help reveal areas of model uncertainty, and human ML practitioners use their judgment to decide where to spend time and money on deeper investigation.

[2] Angler code: `https://github.com/apple/ml-translate-vis`.

8.2 Related Work

This research builds on substantial prior work across general ML evaluation practices, visualization tooling for ML error analysis, and a broad body of work from our target domain, machine translation.

ML Evaluation and Error Analysis

First, we review standard evaluation practices in ML, and discuss how visualization tools can support ML error discovery.

How do Practitioners Evaluate ML Models?

Standard practice in machine learning is to evaluate models by computing aggregate performance metrics on held-out test sets before using them in the real world (offline evaluation) [19, 303]. The goal of using held-out test sets, i.e., data that was not used during model development, is to estimate how well the model will generalize to real world use cases. However, offline evaluations are limited. For example, held-out datasets can be very different from real usage data [365, 394], as data in the wild is often noisy [250] and the real world is ever-changing [264]. Held-out datasets tend to contain the same biases as the training data so they cannot detect potentially harmful behaviors of the model [390, 164]. While summarizing a model's performance in aggregate metrics is undeniably useful, it is insufficient for ensuring model quality.

To overcome these limitations, researchers have proposed additional approaches to help discover model weaknesses [e.g., 81, 160, 208]. For example, practitioners can apply subgroup analysis to discover fairness issues [139], use perturbed adversarial examples to evaluate a model's robustness to noise [51, 520, 390], create rule-based unit tests to detect errors [401, 412], and conduct interactive error analysis to expand known failure cases [334, 519, 398]. ML practitioners also continuously monitor a deployed model's performance and distribution shifts over time [35].

We build on this work by focusing on the question of *prioritization*: how ML practitioners judge where to spend their time and resources among many possible model failure cases. This understanding can help inform the design of future tooling and techniques for surfacing model issues that are more attuned to urgency or severity.

Visualization Tools for Supporting Error Discovery

Interactive visualization is a powerful method for helping ML developers explore and interpret their models [206, 40]. While many visualizations have been built to help practitioners evaluate models over time, one area of recent work has focused on designing and developing analytic tools for ML error discovery [e.g. 511, 296, 287, 352, 110, 533]. For example, FairVis [85] uses visualizations to help ML developers discover model bias by investigating known subgroups and exploring similar groups in the tabular data. Similarly, Visual Auditor [332] automatically surfaces underperforming subgroups and leverages graph visualizations to help practitioners interpret the relationships between subgroups and discover new errors. For image data, explAIner [461] combines interactive visualization and post-hoc ML explanation techniques [e.g., 400, 302] to help practitioners diagnose problems with image classifiers. For text data, Seq2seq-Vis [469] helps practitioners debug sequence-to-sequence models by visualizing the model's internal mechanisms throughout each of its inference stages.

The success of these recent visual ML diagnosis systems highlights the outstanding potential of applying visualization techniques to help ML developers detect errors. Instead of visualizing a model's internals [e.g., 469, 209], we treat ML models as black-box systems and focus on probing their behaviors on different data subsets. While prior systems have focused on ML applications like image captioning where errors are directly observable [85, 87], we designed Angler for the more challenging modeling domain where practitioners cannot always spot-check errors and must rely on proxy metrics to estimate the likelihood of an error.

Evaluating Machine Translation Models

Evaluating translation quality is extremely nuanced and difficult [184, 260, 503, 415]. Language can mean different things and be written in different ways by different people at different times. There are also often multiple "correct" translations of the same input [253].

The gold standard for machine translation evaluation is to have professional translators directly judge the quality of model outputs, for instance, by rating translation fluency and adequacy [95]. There are also automatic metrics for machine translation—such as BLEU [360], ChrF [375] and METEOR [31]—which measure the similarity between a candidate text translation (model output) and one or more reference translations. Intuitively, these metrics apply different heuristics to measure token overlap between two sentences. While these metrics are less reliable and nuanced than human judgment [377, 94, 292, 397, 314, 467, 415], they are intended to correlate as much as possible with human judgments and are widely used for comparing the aggregate performance of different MT models.

An overarching challenge in MT evaluation is that it is especially resource intensive. Both human and automatic evaluation depends on the expertise of human translators, to either directly judge translation quality, or generate reference translations. Since translators have high levels of expertise and are often difficult to find for rare language pairs [151, 329], it is expensive to evaluate translation quality if the input data does not already have a reference translation (e.g., users' requests to a model). In addition, it is difficult to maintain consistent quality within and across human evaluators [374, 417, 223]. Since evaluating the quality of translations from real world use cases requires human annotation, online monitoring and debugging MT models presents a resource allocation problem. In this chapter, we explore how interactive visualization of online model usage might help MT practitioners prioritize data for human evaluation.

Subgroups in Machine Translation

More recently, researchers have explored how to systematically identify specific kinds of errors in MT models [463, 227, 376]. Many of these are language-dependent challenge sets to probe the syntactic competence of MT models [77, 28, 304]. For example, Isabelle, Cherry, and Foster introduces a dataset of difficult sentences designed to test linguistic divergence phenomena between English and French. Stanovsky, Smith, and Zettlemoyer analyzed sentences with known stereotypical and non-stereotypical gender-role assignments in MT, which falls in a broader body of work on detecting gender bias in MT [535, 414, 498, 487].

While these approaches deepen our understanding of specific model failure modes, it is unclear how different errors impact end users of MT models. As a recent survey suggests [376], most of these challenge test sets are created either manually by MT experts or automatically with linguistic rule-based heuristics [e.g., 108, 384, 82]. An alternative approach has been to examine the performance of MT models in specific domains like law [255, 528, 379] or healthcare [249, 122, 475]. These domain-specific challenge sets are deeply informed by knowledge of a particular use case, but are limited in scope. It is difficult for researchers to develop broader challenge sets guided by real users' needs because we lack a clear understanding of how people use MT models, and how they can be impacted by errors. In this chapter we strive to narrow this gap by working with an industry MT team to understand how practitioners might prioritize model improvements based on their users' needs.

Visualization Tools for Evaluation in Machine Translation

There is a growing body of research focused on designing visual analytics tools for MT researchers and engineers [e.g., 276, 364]. For example, with the Chinese Room system MT practitioners can interactively decode and correct translations by visualizing the source and translation tokens side by side [13]. Similarly, NMTVis [333], SoftAlignments [402], and NeuralMonkey [200] use interactive visualization techniques such as parallel coordinate plots, node-link diagrams, and heatmaps to help MT practitioners analyze attention weights and verify translation results. MT researchers also use visual analytics tools [e.g., 257, 338, 507] to better understand MT evaluation metrics such as BLEU, ChrF, and METEOR scores. For example, iBLEU [306] allows researchers to visually compare BLEU scores between two MT models at both corpus and sentence levels. VEMV [466] uses an interactive table view and bar charts to help researchers compare MT models across various metric scores.

In contrast, our work focuses on evaluation based on *usage* of a deployed model. We use interactive visualization as a probe to understand how practitioners prioritize model evaluation when reliable quality signals are expensive to obtain. Our findings provide insight into what kind of information practitioners need to assess potential model failures with respect to their impact on users.

8.3 Interview Study: Prioritizing Model Improvements

This work began in close collaboration with an industry machine translation team, with the goal of helping them prioritize model debugging and improvement resources on problems that had the greatest potential to impact end users. From initial conversations with team members, we learned that their existing process for identifying and addressing problems was largely driven by specific errors (e.g., bug reports, or biases surfaced in academic research), or based on random sampling of online requests. Further, this process was limited to team members with the technical expertise to conduct one-off data analyses. To gain insight into a broader range of existing approaches to prioritization, we turned to practitioners across other ML domains.

We conducted a semi-structured interview study with 13 ML practitioners at Apple. In this section, we describe how practitioners identify and solve specific issues with ML models that impact the user experience. First, we discuss how practitioners navigate a large space of possible failure cases (Section 8.3). Next, we describe how they build challenge sets to assess the cause, scope and severity of an issue, which then informs which issues they address and how they fix them (Section 8.3). At each stage, we highlight how practitioners bring a range of approaches and perspectives to the task of prioritization. We synthesize these findings into four design implications for tooling to support prioritization in model debugging (Section 8.3).

Data Collection and Analysis

We recruited practitioners from both internal mailing lists related to ML and snowball sampling at Apple. Each interview lasted between 30 to 45 minutes. We recorded the interviews when participants gave permission, and otherwise took detailed notes. The study was approved by our internal IRB. We recruited practitioners who have worked on developing and/or evaluating models that are embedded in user-facing tools and products. Incorrect, offensive, or misleading model predictions are detrimental to users' experiences with these models. Therefore, engineers and data scientists that are working on evaluating and improving user-facing ML models are more likely to consider how their models shape users' experiences than other kinds of ML practitioners. Indeed, we found in our interviews that participants often considered how different kinds of model failures may impact end users. An overview of the participants' primary ML application is shown in Table 8.1.

Two authors conducted inductive qualitative analysis of the interview data. One author conducted three rounds of open coding, synthesizing and combining codes each round [318]. Next, a second author took the code book in development and independently coded two interviews, adding new codes where relevant and noting disagreements. These two authors then discussed these transcripts and codes to ensure mutual understanding and shared interpretation of the codes, and converged on a final code book. Lastly, they used this code book to code half of the transcripts each.

Table 8.1: Primary type of machine learning application that each interview participant works on.

Participant	**ML Application**	**Role**
P1	Business forecasting	Data Scientist
P2	Multiple NLP tasks	Data Scientist
P3	Image segmentation	ML Engineer
P4	ML Tooling	ML Engineer
P5	Image classification	Data Scientist
P6	Image classification	Research Scientist
P7	Various CV tasks	ML Manager
P8	Image classification	ML Engineer
P9	Resource use forecasting	ML Engineer
P10	Image segmentation	Research Scientist
P11	Recommender systems	ML Manager
P12	Image captioning	Robustness Analyst
P13	Gesture recognition	Research Scientist

Sourcing Potential Issues

Out of the many ways an ML model could fail, we found that practitioners want to prioritize those that are most consequential for end users. Participants discussed three approaches to find such issues: (1) analyzing errors reported by users; (2) brainstorming potential errors in collaboration with domain experts; and (3) comparing usage patterns against model training data to find areas where the distributions of these two data sources differ.

User Testing and User-Reported Errors

Six participants discussed identifying potential issues through direct feedback from users or from user testing [P3, P4, P5, P7, P10, P13]. Even *ad hoc* testing with a small sample of people can reveal issues that are not surfaced in standard offline evaluations. For example, P5 once *"just showed the [model-driven] app to other people"* and found a *"weird edge case"* where the model always (and often erroneously) classified images containing hands as a certain output class. This was an error that was not surfaced in offline model evaluations because the test data was drawn from the same distribution as the training data, and both contained a spurious correlation between images with hands and this particular output class. *"That's why you do your user testing"* (P5).

User feedback outside of user testing can be difficult to source. In some settings, failures are detectable from signals in usage data, e.g., whether a user accepts or rejects a suggested word from a predictive text model [P5]. More often, real users need to take additional steps to report errors, which they are unlikely to do for every error they encounter: *"I think it* *takes a lot of* [effort] *and willingness to go and file these things"* [P4]. In the most difficult case, users are not *able* to assess prediction quality themselves (e.g., if someone is using MT to translate to or from a language they do not know). In such contexts, direct user feedback is particularly rare.

Brainstorming with Domain Experts

Another approach is to brainstorm potential failure modes with people who hold specialized knowledge of what may be both *important* to users and *challenging* for the model [P1, P4, P9, P12, P13].

> *"Sometimes we involve partners, like other teams or providers who are specialized in the area, to attack the model make the model fail."* — P4

P13 works on gesture recognition models and followed this approach of brainstorming potential errors. P13 had a deep understanding of how their model worked, and thus what kinds of inputs might be difficult to classify accurately. They then collaborated with designers and accessibility experts, who have a deep understanding of users' needs, to identify how the model's weaknesses lined up with *realistic* and important use cases.

Sometimes ML practitioners have built this kind of expertise themselves over years of working with a similar model type [P1], or through precedent with prior reported model failures [P7]. P4 and P12 pointed to academic research (e.g., work published at venues focused on fairness and ethics in AI), the press, and social media as additional helpful sources of potential failure cases. These sources, while not necessarily directly related to any specific model they are working on, can help practitioners understand patterns in ML failures more systematically, and anticipate high-stakes failures.

Brainstorming is particularly useful for identifying *types* of failures that could impact users [401, 398]. However, it is difficult to translate these types into actual failure cases and keep them up to date [398, 51].

> *"We can go with things like what's known as the OVS list—offensive, vulgar, slur. Those are quite obvious, but things can be more subtly offensive... Frankly, there are ways to be offensive that we just simply probably haven't anticipated and language evolves and slang becomes apparent, and even the global situation changes and things that weren't offensive before could become offensive."* — P7

Identifying Usage Patterns with Low Training Data Coverage

A third approach is to identify suspected areas of weakness for the model by looking for differences between how users are interacting with a model and the data with which that model was trained. We use the term **coverage** to describe how well a model's training data "covers" the space of inputs that the model receives after deployment (i.e., use cases). The coverage of a particular use case is a measure of how much the model "knows" about those kinds of inputs, and can be used as a proxy for model performance when other quality signals are not available.

> *"We know that the* [training] *datasets that we have, however large they are, they don't cover the entire space. So wherever we don't have coverage we don't expect the* [model] *performance to be that great."* — P3

To detect coverage issues, practitioners monitor online data to see how people are using a model, and compare that against their training data:

> *"If it is a classification task, you were expecting to have a very balanced dataset, but online [you see] that almost 90% of the traffic is coming for 1 class. That means your offline [data] was not representative of what is going to happen in an online setting. So, by monitoring and looking at all data distributions, you will get a sense of those discrepancies."* — P2

Considering coverage allows practitioners to move beyond the kinds of failures that they already know of or suspect based on past experience and identify new failure modes that they were not previously aware of [P7].

Creating Challenge Sets to Validate and Evaluate Issues

When practitioners identify reported or suspected failures, they still need to determine whether this is a systematic problem with the model and, if so, assess the scope and severity of the error. Participants first wanted to understand if a potential failure is a one-off error (as can be expected given ML is probabilistic) or a more systematic problem [P4, P7, P12]. We found that practitioners shared a common general approach of:

> *"Collecting more similar data and testing the model behavior and seeing if it's systemically failing."* — P7

Practitioners in our sample referred to these curated data subsets as *aggressor sets*, *golden test sets*, and *stress tests*. In the remainder of this paper, we refer to these kinds of datasets as **challenge sets**.

Challenge sets differ from standard test sets in ML because they are designed to target a specific failure case, and are thus often more reflective of how people really use a model in practice. As P6 described, *"that kind of test while we call it stress test is probably closer to what happens in reality than when you do random sampling for testing."* Creating these sets can be challenging. In particular, it might not be immediately clear what kind of "similar" data will replicate a failure mode, and the axis of similarity that matters might not be annotated explicitly in the data.

> *"The length of the beard seemed to play a role [in a failure mode]. It [the dataset] was just annotated as has beard or not, and not so much the length."* — P6

Once practitioners have built a challenge set and determined that a suspected failure case is indeed a systematic problem with a model, they can then conduct quantitative and qualitative analyses of the challenge set to deeply understand the cause of the issue and how it might impact users. This understanding is critical to prioritizing issues to solve and informs the choice of solution.

Assessing the Cause of the Problem

Practitioners look for patterns in challenge sets to understand the potential cause of a problem. A first step is usually to compare the challenge set to the model's training data to identify coverage issues or other data problems, e.g., spurious correlations. This analysis could be a simple process of *"manually going through* [the challenge set] *and looking for any general trends"* [P5], although models with larger output spaces or high dimensional data may require more sophisticated techniques like embedding space visualization and dimensionality reduction techniques [P7].

Assessing the Impact of the Problem on Users

Practitioners also want to assess the impact of the problem on users to judge its urgency. Model failures might be prioritized if they impact many users, happen frequently [P1], or if they produce a negative user experience [P7, P11, P12]. In this way, prioritization is *"not just a pure data science question"* [P1], but involves considering different and possibly conflicting perspectives and values.

For example, in P7's work, *"certain mispredictions could be more offensive than others,"* so when gathering feedback from quality annotators, they ask annotators to *"exercise some judgment,"* and specifically flag anything they feel are *"potentially offensive or egregious mistakes."*

Practitioners might also prioritize improvements to ensure no subpopulation of users is experiencing particularly poor performance compared to others:

> *"What we want to do is, reduce the length of the tail end of users that have poor experience and talk more about, how can we bring these people up and what is it about* [their use context] *that causes the models to perform poorly."* — P11

Assessing Potential Solutions

The choice of appropriate solution depends on understanding the scope and nature of the problem, and discussing these with reference to how the issues impact end users. Often, problems are the result of poor coverage and can be addressed by increasing training data in a specific area and retraining the model [P9]. For some participants, this was the default approach: *"the answer is pretty much always going to be more representative data across all classes."* [P5]. However, our findings highlight a wider range of approaches that practitioners can take when they deeply understand the nature and stakes of a problem.

For example, three participants talked about strategies to augment the model's output space. This could mean adding or removing classes from a classification taxonomy [P4], preventing specific outputs using a block list or an additional classification model, or hard-coding outputs for certain inputs using a lookup table [P4, P7, P12]. Other approaches included improving annotation quality [P7, P10], removing problematic data from the training set [P5], changing the user interaction with the model to control the input environment in production [P8], adjusting the model architecture or loss function [P12], or adding additional data pre-processing steps [P5].

These approaches differ in complexity, cost, and effectiveness. The choice of solution is not solely based on technical and resource constraints, but could involve negotiating trade-offs, considering conflicting values, and accounting for the urgency of the error. For example, practitioners might select a fix that is faster to implement if an error impacts many users or is particularly offensive. Such decisions require input from stakeholders with a broad range of expertise. Therefore, ML practitioners must be able to discuss problems with reference to business metrics and user experience and in terms that are accessible to stakeholders without ML expertise [P1].

Design Implications

Our findings demonstrate how challenge sets allow practitioners to develop a deep understanding of a problem's cause, scope, and impact on users. This understanding is necessary to effectively prioritize resources on the most egregious and urgent model failures. Existing tooling for model evaluation and debugging have largely focused on *identifying* model weaknesses rather than *prioritizing resources* on weaknesses with the greatest potential impact on users. Based on the practices uncovered through our interview study, we developed four design implications for tooling to support prioritization:

D1. Compare usage patterns to training data to support exploration of suspected model weaknesses in addition to known errors.

D2. Build collections of similar data (challenge sets) to assess and prioritize problems, and allow users to compare challenge sets.

D3. Provide information about model performance and usage patterns to surface issues that matter most to users.

D4. Since prioritization is not solely a technical question and does not have a singular solution, account for prioritization subjectivity, and make the tools easy to use for stakeholders with diverse backgrounds.

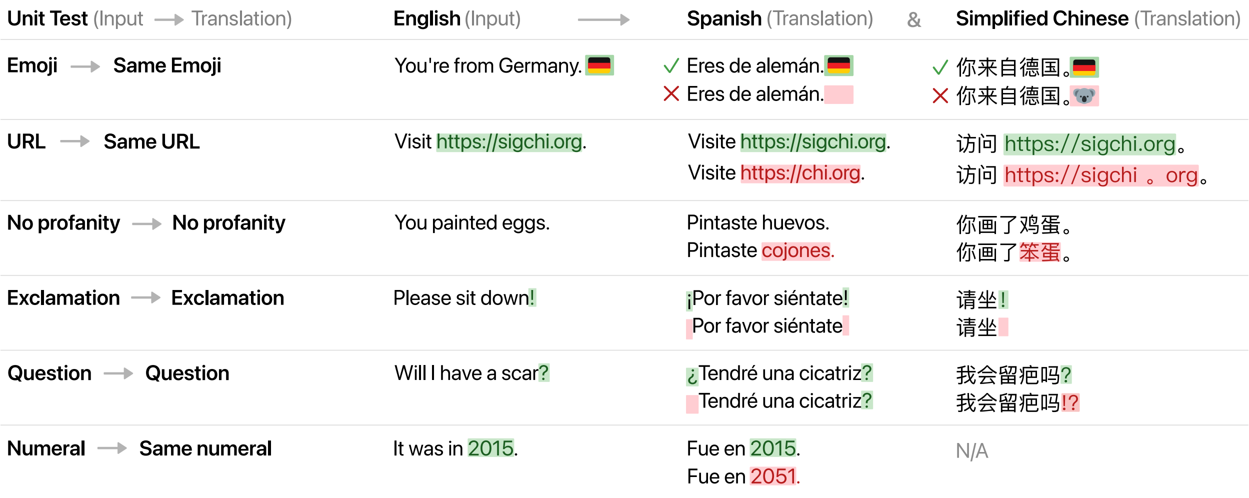
Interactive visualizations have successfully helped ML practitioners discover ML errors (Section 8.2) and understand model behaviors (Section 8.2). Interactive visualization techniques are especially useful for exploring data to support hypothesis generation and serendipitous discoveries (D1), comparing and contrasting slices of data (D2), analyzing data from multiple perspectives (D3), and supporting collaborative interpretation of data among stakeholders with diverse skill sets (D4). For these reasons, visual analytics is a promising choice for supporting prioritization.

The remainder of this paper focuses on Angler (Figure 8.1), a visual analytics tool to support prioritization in the context of machine translation (MT). While our interview study revealed common practices across ML domains, prioritization depends on measures of prediction quality and insight into usage patterns, both of which are extremely specific to a particular model. Therefore, to understand these practices more deeply and begin to explore what tooling support for prioritization might look like, it is useful to choose a specific ML task as a case study. We chose MT because it poses unique challenges that make prioritization both especially important and especially difficult: it is difficult and expensive to attain reliable measures of prediction quality [93]; MT models accept open-ended input from users, opening a vast space of possible failures; and we know relatively little about how people use MT models in the real world [**liebling2020unmet**].

8.4 Designing Angler: Exploring Machine Translation Usage with Challenge Sets

Given the design implications (D1–D4) described in Section 8.3, we present Angler (Figure 8.1), an interactive visualization tool that helps MT developers prioritize model improvements by exploring and curating challenge sets. Angler leverages both usage logs and training data to help users discover model weaknesses (D1, Section 8.4). Angler introduces two novel techniques to automatically surface challenge sets and expand challenge sets with similar data (D2, Section 8.4). Angler uses the *overview + detail* design pattern [114] to tightly integrate two major components: the *Table View* that summarizes challenge sets as table rows (Section 8.4) and the *Detail View* that enables users to explore one challenge set in depth with different attributes over time (D3, Section 8.4). Finally, to lower the barrier for different stakeholders to easily prioritize model improvements (D4), Angler allows users to conduct quantitative and qualitative analyses without needing to write custom code and manipulate complex data and model pipelines. We develop Angler with modern web technologies so that anyone can access it without installation overhead, and we open-source our implementation (Section 8.4).

We designed and developed Angler in conversation with an industry MT product team. To contextualize our design in the team's practices and gain iterative feedback, we used one of the team's MT models (English → Spanish), with a sample of their training data and usage logs. Usage logs are only available from users who have opted-in. For privacy and security reasons, members of our research team required special permissions and security protocols to access this data. Therefore, we cannot show Angler with the original model or dataset from our design process. Moreover, to demonstrate how Angler can support many different MT models and language pairs, we instead describe the Angler interface in this section using a public MT model (English → Simplified Chinese) and public datasets (Section 8.4).



Unit Test (Input → Translation)	**English** (Input) →	**Spanish** (Translation) &	**Simplified Chinese** (Translation)
Emoji → Same Emoji	You're from Germany.🇩🇪	✓ Eres de alemán.🇩🇪 ✗ Eres de alemán.	✓ 你来自德国。🇩🇪 ✗ 你来自德国。🐨
URL → Same URL	Visit https://sigchi.org.	Visite https://sigchi.org. Visite https://chi.org.	访问 https://sigchi.org。 访问 https://sigchi 。org。
No profanity → No profanity	You painted eggs.	Pintaste huevos. Pintaste cojones.	你画了鸡蛋。 你画了笨蛋。
Exclamation → Exclamation	Please sit down!	¡Por favor siéntate! Por favor siéntate	请坐! 请坐
Question → Question	Will I have a scar?	¿Tendré una cicatriz? Tendré una cicatriz?	我会留疤吗? 我会留疤吗!?
Numeral → Same numeral	It was in 2015.	Fue en 2015. Fue en 2051.	N/A

Figure 8.2: We create a suite of regex-based unit tests to detect translation errors without the need for ground-truth translation. For example, some tests check if the source and translation contain the same special words (e.g., Emoji, URLs, and roman numerals). Some tests check punctuation (e.g., question marks and exclamation points). Another test validates that a translation does not contain profane words if its source does not contain any profane words, by matching language-specific offensive, vulgar, slur word lists. In the translation columns, each row lists two examples that pass (top) and fail (bottom) the unit test.

Subgroup Analysis through Challenge Sets

In Angler, we introduce two novel techniques to surface interesting subsets from the usage logs and training data. We automatically extract *challenge sets* by sampling data that either fails our model performance unit tests (Section 8.4) or involves topics the model is less familiar with (Section 8.4).

Unit Test Failures

The current state-of-the-art approach to building challenge sets for machine translation is to build rule-based unit tests (8.2). In line with this practice, the first type of challenge sets that we include in Angler extends the team's existing suite of unit tests to identify unexpected model behavior (Figure 8.2). These unit tests use regex search to find patterns in a source-translation pair and verify that each match meets some pre-defined rules. For example, when a source includes an emoji, we expect the translation to have the same emoji. Similarly, when a source does not contain offensive, vulgar, or slur (OVS) words, we expect the translation not to include OVS words either. Some unit tests are language-specific: consider translating English to Spanish; if a source is a question, we expect the translation output to have both ``¿'' and ``?'' characters. For simplified Chinese, however, we would expect the translation output to end with the “？ ” unicode instead. For our English → Chinese demo in this section, we apply

5 unit tests to both usage logs and training data (D1), and collect data samples that fail any unit test into challenge sets (Figure 8.2). Each unit test corresponds to one challenge set.

Unfamiliar Topics

While unit tests can reveal some straightforward errors, they do not offer insight into issues of **coverage**, which our interview participants highlighted as critical to identifying failure modes that are highly consequential for end users (Section 8.3). In the context of MT, coverage refers to how much a translation model "knows" from its training data about a particular topic or way of speaking. For example, coverage is a major concern with *domain-specific language*: e.g., doctors use domain-specific phrases to talk about medicine, while video game players use language specific to their game. Coverage can be improved by collecting more training data to give the model more exposure to that particular language pattern.

Extending existing techniques for building challenge sets in MT, we sought to help MT practitioners prioritize *which* domains may need better coverage based on what their users are requesting. To identify topics that are not well represented in the training data, we first use a sentence transformer [396] to extract high-dimensional latent representations of sentences in the training data. This latent representation is trained to cluster sentences with similar meaning close together in high-dimensional space. We then apply UMAP [315], a dimensionality reduction method, to project the latent representation into 2D. We choose to use UMAP instead of other dimensionality reduction techniques, such as PCA [367] and t-SNE [497], because UMAP is faster and has better preservation of the data's global structure [315]. We use the cosine similarity to measure the distance between two samples in the high-dimensional space, as previous works have shown that cosine distance provides better and more consistent results than the Euclidean distance [501]. Following the suggested practice in applying UMAP [115], we fine-tune UMAP's hyperparameters `n_neighbors` and `min_dist` through a grid search of about 200 parameter combinations; we choose hyperparameters that spread out the training samples in the 2D space while maintaining local clusters. With about 47,000 training sentences and a latent representation size of 768, it takes about 50 seconds on average to fit a UMAP model with one parameter combination on a MacBook Pro laptop.

We use Kernel Density Estimation (KDE) [426] to estimate the training data's distribution. For the KDE, we choose a standard multivariate Gaussian kernel with a Silverman bandwidth $\mathbf{H}$ [450]. It only takes about 1 second to fit a KDE model on 47,000 training sentences' 2D representations. Then, we use this trained KDE model to compute the *familiarity score* (FA) [210] for each sentence from the usage logs. We define the familiarity score (Equation 8.1) of a sentence from the usage log as the log-likelihood of observing that sentence's UMAP 2D coordinate (x, y) under the training data's UMAP distribution $[(x_1, y_1), (x_2, y_2), \ldots, (x_i, y_i)]$. This concept of familiarity can be generalized to other data types and ML domains, and has shown to be a powerful tool for debugging data [210].

$$\mathrm{FA}(x, y) = \log\left(\frac{1}{n}\sum_{i=1}^{n}\frac{\exp\left(-\frac{1}{2}\begin{bmatrix}x - x_i & y - y_i\end{bmatrix}\mathbf{H}^{-1}\begin{bmatrix}x - x_i \\ y - y_i\end{bmatrix}\right)}{2\pi\sqrt{|\ \mathbf{H}\ |}}\right) \tag{8.1}$$

Computing FA is slow when the training data is large (i.e., n is large in Equation 8.1), because the algorithm needs to iterate through all n points in the training data for each sentence from the usage logs. Therefore, to accelerate FA computation, we apply a 2D binning approximation approach. We first pre-compute the log-likelihoods over a 2D grid of training data's UMAP 2D space $\mathbf{F} \in \mathbb{R}^{200\times 200}$, constrained by the range of the training data's UMAP coordinates. Then, to approximate the FA of a sentence, we only need to (1) locate the cell $\mathbf{F}_{i,j}$ in the grid that the sentence falls into, and (2) look up the pre-computed log-likelihood associated with that cell $\mathbf{F}_{i,j}$. If a sentence falls out of the 2D grid, we extrapolate its FA by using the log-likelihood associated with the closest grid cell. Note that one can choose a different grid density other than 200×200; we tune the grid density ($d = 200$) to balance the computation time and the approximation accuracy. Our binning approximation is scalable to large usage logs (m) and training data (n), as it decreases the FA computation's time complexity from a quadratic time $\mathcal{O}(kmn)$ to a linear time $\mathcal{O}(m + d^2kn)$, where k is the dimension of the UMAP space ($k = 2$ in our case), and d is the grid density ($d = 200$). In addition, we use the KDE implementation from *Scikit-Learn* [368], which leverages KD Tree [47] for more efficient distance computation. With a tree-based KDE, our FA computation method has a logarithmic time complexity $\mathcal{O}(m + d^2k\log(n))$ on average and a linear time complexity $\mathcal{O}(m + d^2kn)$ in the worst case.

After estimating the FAs for all sentences in the usage logs, we use BERTopic [176] to build a topic model on a sample of 50,000 sentences from the usage logs with the lowest FA and select the 100 largest topics from this model. To estimate the model's performance on these topics, we need labeled training data. Therefore, we extend each extracted topic set with a sample of training sentences that are close to the topic set in the high-dimensional space (D2). To reduce the computational cost of this search, we randomly sample 15 "seed sentences" from each topic and add any sentences from the training data that are close to at least one of the "seed sentences" in the high-dimensional space (threshold $\ell_2 < 0.6$ selected through manual inspection). We have tuned the number of "seed sentences" to balance the computational cost and the number of close training sentences that we can find. Finally, we have controlled random seeds for random sampling, UMAP computation, and BERTopic, so that our topic results are reproducible.

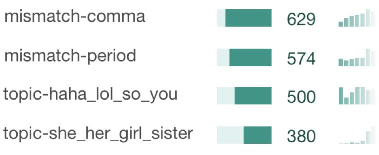
Limitations

Identifying new model failure modes and collecting examples to replicate the failure is extremely challenging. Developing automatic, expert-driven, and crowd-sourcing methods for identifying failures is an active area of research in machine learning and human-computer interaction [533, 110, 127, 443, 511, 85, 332]. Compared to prior research, it is especially difficult to automatically identify MT model failures because there are no explicit, interpretable features or metadata on which to slice data into subgroups, and automatic evaluation metrics are very noisy. Further, prior work largely focuses on identifying failure modes by comparing predictions to ground-truth labels [e.g. 533, 110], which does not give practitioners insight into failure modes that impact end-users but are not yet represented in offline, labeled datasets.

Our goal in this chapter is to understand how practitioners prioritize their resources across many potential failure modes, and what information they need to do so. We generate example challenge sets to guide this exploration using pattern-matching rules (the current state-of-the-art in MT) and topic modeling on areas of low coverage. However, further research is needed to evaluate and extend these methods. While we did not conduct a formal evaluation of our challenge sets in this work, both kinds of sets are certainly imperfect in terms of error identification – there are perfect translations included in the challenge sets, and there are translation errors in the larger data that are not included in any challenge set. Given the large space of possible inputs, and probabilistic nature of machine learning, we cannot expect to ever have methods to identify all possible failures with perfect accuracy. Thus, there is a need for interactive visualization tools that support practitioners to explore and make sense of *potential* failure modes and prioritize development and annotation resources under uncertainty.

Table View

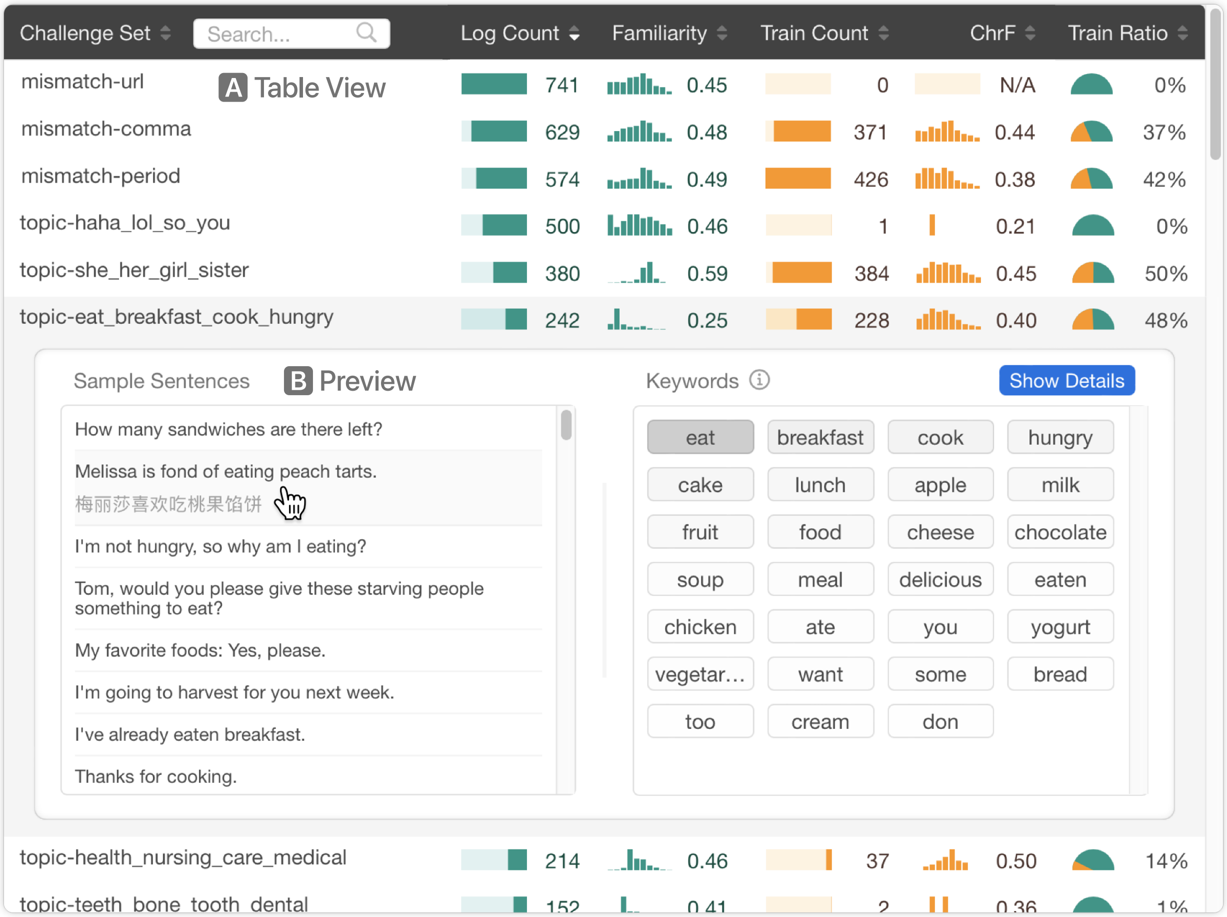
When users launch Angler, they first see the *Table View* listing all pre-computed challenge sets in a table (Figure 8.3A). Each challenge set can contain samples from the training data and usage logs, color coded as orange and green respectively throughout Angler. We name challenge sets based on their construction methods (Figure 8.4). For challenge sets created by unit test failures, we name them "mismatch-[*unit test name*]." For challenge sets created by unfamiliar topics, we name them "topic-[*top-4 keywords*]." These keywords are the same as keywords shown in the *Set Preview* (Figure 8.3-B). In addition to the names of challenge sets, the *Table View* view provides five metrics associated with each set:



mismatch-comma	629
mismatch-period	574
topic-haha_lol_so_you	500
topic-she_her_girl_sister	380

Figure 8.4: Angler distinguishes challenge sets created by unit test failures and unfamiliar topics by their names.

- *Train Count* and *Log Count*: the number of training and usage log samples in the set.
- *ChrF*: a measure of the model's performance on the training samples in the set. ChrF is the F-score based on character n-gram overlap between the hypothesis translation produced by the model and a validated reference translation [375]. We use the open-source SacreBLEU implementation of ChrF [377].



Challenge Set	Log Count	Familiarity	Train Count	ChrF	Train Ratio
mismatch-url	741	0.45	0	N/A	0%
mismatch-comma	629	0.48	371	0.44	37%
mismatch-period	574	0.49	426	0.38	42%
topic-haha_lol_so_you	500	0.46	1	0.21	0%
topic-she_her_girl_sister	380	0.59	384	0.45	50%
topic-eat_breakfast_cook_hungry	242	0.25	228	0.40	48%
topic-health_nursing_care_medical	214	0.46	37	0.50	14%
topic-teeth_bone_tooth_dental	152	0.41	2	0.36	1%

Sample Sentences

How many sandwiches are there left?

Melissa is fond of eating peach tarts.
梅丽莎喜欢吃桃果馅饼

I'm not hungry, so why am I eating?

Tom, would you please give these starving people something to eat?

My favorite foods: Yes, please.

I'm going to harvest for you next week.

I've already eaten breakfast.

Thanks for cooking.

Keywords ⓘ

eat	breakfast	cook	hungry
cake	lunch	apple	milk
fruit	food	cheese	chocolate
soup	meal	delicious	eaten
chicken	ate	you	yogurt
vegetar…	want	some	bread
too	cream	don	

Figure 8.3: **(A)** ***The Table View*** summarizes all challenge sets in a table, where users can compare the model's performance and familiarity across these sets. To help users better understand aggregate metrics, the *Table View* also visualizes the distributions of metrics and sets' compositions as sparkline-like charts. **(B)** ***The Preview*** presents more details of a challenge set after a user clicks a row. The *Sample Sentences* (left) lists 100 randomly selected source sentences from this set with their translations. The *Keywords* (right) visualizes the most representative words in this set, where a darker background indicates higher representativeness.

- *Familiarity*: a measure of how familiar the usage log samples in the set are to the model, by reference to the training data distribution (Section 8.4).
- *Train Ratio*: the percentage of samples in the set that are training samples.

Users can sort challenge sets by any of these metrics by clicking the sorting button ⬍ in the table header. To help users quickly compare these metrics across challenge sets, the

Table View also provides sparkline-like visualizations [488] in each row. For each challenge set, the *Table View* visualizes its sample counts as in-line bar charts, ChrF and Familiarity distributions as histograms, and the training sample ratio as a semi-circle pie chart.

After identifying an interesting challenge set, users can click the row to open a *Set Preview* (Figure 8.3B) in the table to see a preview of that set. This view provides users with a quick summary of the set on demand. On the left, users can browse 100 randomly sampled sentences from this challenge set; users can also click on each sentence to see the model's output translation. The number of sentences in each challenge set varies from about 100 to 1000; challenge sets constructed from unit tests tend to have more sentences than ones constructed from unfamiliar topics. We choose the number 100 because it gives a fair coverage of all sentences in the set and users can have a smooth experience in quickly browsing sentences from different sets. If they are interested in one particular challenge set, they can view all sentences in that set's *Detail View* (Section 8.4). On the right, users can inspect the most representative keywords from this set. Keywords are extracted and sorted by their class-based TF-IDF scores [176]. Intuitively, these keywords are words that appear more frequently in this set than in all other sets. In Angler, we list all keywords returned from BERTopic; future researchers and developers can determine a class-based TF-IDF score threshold to only display more frequent keywords (keywords with a darker background).

Detail View

To help users further analyze individual challenge sets (D3), Angler presents the *Detail View* (Figure 8.5) when a user clicks the Show Details button under a challenge set in the *Table View* (Figure 8.3). In the header of the *Detail View* (Figure 8.5A), users can inspect the metrics associated with this challenge set and edit the set's name. To explore sentences in this set through different perspectives, users can use the *Timeline* (Figure 8.5C) and *Spotlight* (Figure 8.5E) to filter sentences by different attributes. The *Filter Bar* (Figure 8.5B) displays the currently applied filters, and the *Sentence List* (Figure 8.5F) only shows sentences that satisfy these filters. There are six visualization variations of the *Spotlight* (Figure 8.6). Users can switch between them to fit their exploration needs (D4) by clicking the corresponding *Thumbnails* (Figure 8.5D). Each *Thumbnail* is a simplified version of a *Spotlight* variation, where the visualization also updates in real-time when users add or remove filters.

Timeline

To help users investigate how usage logs change over time, the *Detail View* provides a *Timeline* (Figure 8.5C) panel on top of the window. The *Timeline* visualizes the number of user requests in this set over time as a histogram, where the x-axis represents the time and the y-axis represents the request count. Users can zoom and pan to inspect different periods. Users can also brush the histogram to filter usage logs that are from a particular time window.

Keyword Spotlight

Similar to the *Keyword* panel in the *Set Preview* (Section 8.4), the *Keyword Spotlight* (Figure 8.6-1) displays the most representative words in a challenge set. It sorts keywords by their representativeness, which is measured by the class-based TF-IDF scores [176]. This view uses the darkness of the background color to encode a word's representativeness. Users can click keywords to filter sentences that contain selected keywords.

Embedding Spotlight

To help users explore the semantic similarity of sentences in a challenge set, the *Embedding Spotlight* (Figure 8.6-2) visualizes a 2D projection of the sentences' high-dimensional representations (Section 8.4) in a scatter plot. Each dot in the scatter plot represents a sentence, and it is positioned by its UMAP coordinates. Furthermore, we visualize the KDE density distributions (Section 8.4) of all training data and all usage logs as contour plots. Augmenting the scatter plot with density distributions of overall training data and usage logs allows users to discover use cases that are not well supported by existing training data (D1).

ChrF Spotlight

The *ChrF Spotlight* (Figure 8.6-3) visualizes the model's ChrF score distribution on the training data in this set as a histogram, allowing users to gain more insights regarding the model's performance on a particular set. The x-axis encodes the ChrF scores, and the y-axis encodes the distribution frequency of training data in the set. In addition, users can brush to select bins in the histogram, which would filter sentences with a ChrF score in the specified range.

Familiarity Spotlight

The *Familiarity Spotlight* (Figure 8.6-4) is similar to the *ChrF Spotlight*. However, the x-axis here represents the model's familiarity scores on usage logs in the set. The familiarity score is determined by the log-likelihood of observing a user request under the distribution of all training data (Section 8.4). Users can brush the histogram to filter sentences with particular familiarity scores.

Source Spotlight

To allow users to compare usage logs by the sources of these user requests, the *Source Spotlight* (Figure 8.6-5) visualizes usage log count as a horizontal bar chart. The x-axis encodes the count of usage logs from one particular source, and the y-axis encodes the source category. To focus on logs from particular sources, users can click the source names to create filters. In our open-source demo, the *Source Spotlight* shows the source dataset from which each sentence was sampled. In the version of the tool developed for the MT team, the *Source Spotlight* shows the source application from which a request was made (available for a sample of usage logs).

Overlap Spotlight

The design of the *Overlap Spotlight* (Figure 8.6-6) is similar to *Source Spotlight*. Instead of encoding the source category, the y-axis here represents other challenge sets. For example, for a challenge set created by unit test failures, the y-axis in its *Overlap Spotlight* represents other challenge sets created by unfamiliar topics. As unfamiliar topics are strictly non-overlapping, this view only shows overlap with challenge sets of the other type. By cross-referencing two challenge set types (unit test failures and unfamiliar topics), this visualization can help users explore syntactic errors within semantic topics, and vice versa.

Sentence List

The *Sentence List* (Figure 8.5F) shows all sentences in the challenge set that satisfy the currently applied filters. Users can click a sentence to see the model's translation. To further fine-tune a challenge set, users can click the Edit button and remove unhelpful sentences from the set. Finally, users can click the Export button to export sentences shown in the list along with their translations and attributes; users can then easily share these sentences with colleagues and human annotators (D4).

Open-source Implementation

Angler is an open-source interactive visualization system built with *D3.js* [61]: users with diverse backgrounds (D4) can easily access our tool directly in their web browsers without installing any software or writing code. We use the standard NLP suite for data processing (e.g., *NLTK* [298], *Scikit-learn* [368]) and topic modeling (e.g., *BERTopic* [176], *UMAP* [315]). We first implemented Angler with an industry-scale MT model (English → Spanish) and real training data and usage logs. For demonstrations in this chapter, we use the public MT model *OPUS-MT* (English → Simplified Chinese) [481] and its training data [480]. To simulate usage logs we augment a sample of the model's test data [480] with publicly available sources to emulate realistic use cases that can be difficult for MT models: social media [57], conversation [147], and scientific articles [455].

Usage Scenario

We present a hypothetical usage scenario to illustrate how Angler can help Priya, an MT engineer, explore usage logs and guide new training data acquisition. The first part of this usage scenario, in which the user explores and selects challenge sets of interest, is informed by real user interactions that we observed in the user study (Section 8.5). The second phase of the scenario describes how we envision extending Angler in future work to help practitioners use the datasets they collect with Angler to improve model performance.

Priya works on improving an English–Chinese translation model, and she only speaks English. Priya first applies the challenge set extraction pipeline (Section 8.4) to the training data and usage logs from the past 6 months. The pipeline yields 100 challenge sets from unfamiliar topics and 6 challenge sets from unit test failures. After Priya opens Angler to load extracted challenge sets in a web browser, she sees the *Table View* (Figure 8.1A) summarize all 106 sets in a table with a variety of statistics. Priya wants to prioritize subsets of data where the model may not perform well, but which are important to the end-users of the model, e.g., because they occur frequently in the usage logs, or represent a high-stakes use case. To focus on data on which the current MT model may not perform well, Priya sorts challenge sets in ascending order by their mean ChrF scores by clicking the sort button. After inspecting the top rows and their *Set Previews*, the `topic-headache` set draws Priya's attention—the MT model performs poorly on this set (mean ChrF score is only 0.39), and this set involves high-stakes medical topics where the MT quality is critical (observed from the *Keywords* in the *Set Preview*).

To learn more about this challenge set, Priya clicks the Show Details button to open the *Detail View* (Figure 8.1B). Priya notices that the number of usage logs is consistent across the past nine months (from July 2021 to March 2022) in the *Timeline* (Figure 8.1-B1). She then clicks the *Embedding Thumbnail* (Figure 8.5D) to switch the *Spotlight* from the default *Keywords* view (Figure 8.6-1) to the *Embedding* view (Figure 8.6-2). Through zooming and hovering over the scatter plot, Priya finds that most sentences from this set form a cluster in the high-dimensional representation space, and all these sentences are about health issues. She is surprised to see that people are using the model to communicate about health concerns, and wonders whether the training data covers this use case. To explore this, Priya opens the *Familiarity Distribution Spotlight* (Figure 8.6-4) and brushes the histogram to select the region with low familiarity scores. The *Timeline*, charts in *Thumbnails*, and the *Sentence List* update in-real time to focus on the usage logs with a familiarity score in the selected range. Browsing the sentences in the *Sentence List*, Priya realizes that many of these unfamiliar sentences are about fever. She worries that wrong translations about fevers could pose a health risk to users. Therefore, Priya decides to prioritize improving her model's performance on this challenge set; she clicks the ⤓ Export button to save all sentences along with their translations from this challenge set.

The current Angler prototype was designed to explore what information practitioners need to prioritize subsets of data to send to annotators. Priya follows a similar process to identify and export a few other challenge sets of high importance and sends all of this data to human annotators to acquire additional training data. Human annotators can speak both English and Chinese. They write reference Chinese translations for given English sentences by directly editing the translations produced by the model. In the future, Angler could be extended to allow Priya to continue her analysis after the data has been annotated.

For example, Priya could check a few sentences from the health-related challenge set whose reference translations are significantly different from their translations produced by

Priya's model. At this point, Priya might find that her model has made several serious translation errors. For example, her model translates the input sentence "The fever burns you out" to "发烧会烧死你" in Chinese (Figure 8.1-B3), which means "fever will burn you to death." After retraining the MT model with the newly annotated data, Priya would hope to see that the model's ChrF scores and familiarity scores on the original challenge sets have significantly improved. In this case, Priya would schedule to deploy her new MT model in the next software release cycle, now with better support for a safety-critical use case.

8.5 User Study

We used Angler in a user study with seven people who contribute to machine translation development at Apple as ML engineers (E1–3), and in user experience-focused roles, such as product management, design, or analytics (UX1–4). Our goal in this study was to understand how users with different expertise would use Angler to explore and prioritize challenge sets. We were also interested in whether exploring challenge sets using Angler could help practitioners to uncover new insights about their models, and identify new ways to improve their models in line with users' needs. Our goal was not to measure whether Angler can support prioritization more effectively than another tool, e.g., by finding more translation errors, but rather to explore what kind of information is useful to practitioners, and how the process of exploring challenge sets could shape future evaluation practices.

The study was approved by our internal IRB. Each session was conducted over video chat and lasted between 45 minutes and one hour. With participants' consent, each session was recorded and transcribed for analysis. During each study session, we introduced Angler with a tutorial demonstrating each view (Figure 8.3–8.6). Angler showed training and usage data from the team's own translation model, for the language pair English → Spanish. Next, we sent the participant a one-time secure link that allowed them to access the tool in their own browser, and asked them to share their screen while they explored the tool. For the remainder of the session, we asked the participant to think aloud as they completed three tasks:

T1. First, we asked the participant to navigate to the *Detail View* of a unit test challenge set that targeted mismatches in numbers between source and output translations, and to discuss what they saw.

T2. Second, we asked the participant to choose a topic-based challenge set that was interesting to them, explain their choice, and again explore the *Detail View* to learn more.

T3. Finally, we gave the participant a hypothetical budget of 2,000 sentences that they could choose to get evaluated by expert human translators, and asked them to explain how they would allocate that budget. The evaluation by professional translators could involve rating the quality of model-produced translations and/or correcting the translations to create gold standard reference translations that could be used for future model training.

We analyzed the transcripts following a similar qualitative data analysis procedure to that of the formative interview study (Section 8.3). One author conducted two rounds of open coding, synthesizing and combining codes each round [318]. Next, a second author took the code book in development and independently coded all of the transcripts, adding new codes where relevant and noting disagreements. These two authors then discussed and resolved disagreements, and converged on final coding scheme.

We found that **T1** and **T2** mainly served as a way for participants to acclimate to Angler's interface, and understand the two types of challenge sets. Although participants confirmed that **T3** was a realistic task for the team, most participants did not do the task as we had originally planned. We report our findings regarding how participants picked which challenge sets they deemed *important* for model improvements, but we do not report on their fictional budget allocations because the majority of participants were resistant to allocating concrete (even completely hypothetical) numbers. We discuss this tension more in limitations Section 8.5. All three tasks required participants to prioritize among the available challenge sets, but our findings focus largely on participants' judgments during **T3**, where they spent the majority of the study time.

As discussed in Section 8.3, the team's existing approaches to model debugging and improvement were either one-off, focused analyses, which do not require prioritization between issues, or random samples of usage logs, which implicitly prioritize use cases based on frequency of requests. Participants informally compared what they could do with Angler to these existing practices. Our goal in this study was to explore the space of possibilities for visual analytics tools to support prioritization, rather than quantify the relative benefit of our specific prototype compared to existing practices.

Results: Prioritizing under Uncertainty

Participants had to rely on imperfect and incomplete metrics to estimate the quality of translations. All participants knew a little Spanish, but not enough to spot-check the quality of a translation in most cases. Though Angler shows sentences from usage data and model training data, only the training data had *reference translations* certified as correct by human translators. Sentences from the usage logs have no quality annotations (Section 8.4). Thus the ChrF quality estimation for any challenge set was based on the limited data with reference translations. This evaluation setup is far from ideal, yet realistic to what MT practitioners ordinarily encounter.

Participants knew that no one metric was a reliable source of quality information, so they weighed multiple signals and still knew that human annotation would generally be required to get a reliable measure of quality. Three participants discussed how they incorporated uncertainty when interpreting metrics like average ChrF to ensure they were getting meaningful estimates of quality [E1, E2, UX4]:

> *"You might get a low metric or a low familiarity score, but the smaller the sample is the more likely it is [that] there's gonna be some noise in there that's kind of moving the metric."* — UX4

Future iterations of the tool could use re-sampling methods to estimate confidence intervals for challenge set summary statistics to make this uncertainty more explicit.

Despite available metrics in Angler being uncertain proxies for model performance, participants nonetheless used metrics to judge the relative importance of a challenge set:

> *"It's better to have a statistical way, I mean,* [rather than] *just by what I'm thinking, right?"* — E2

All participants tended to rank the challenge sets by potential risk of model failure by combining low ChrF, low familiarity, and low train-ratio. Low ChrF indicates that the limited *training data* that falls within the challenge set *might be poorly translated* [375]. Low familiarity and low train-ratio are proxies for where the training data set *might lack coverage* of the usage data. Low familiarity suggests that the user request data that falls within the challenge set is semantically different from the overall training data. Low train-ratio indicates that this challenge set represents a subgroup that is much more represented in the user requests than in the training data. Familiarity and train-ratio were calculated in a way that correlates (Section 8.4).

As a first pass, most participants used the sorting feature in the *Table View* (Figure 8.3) to rank which challenge sets scored the lowest on one or more of these three metrics [E1, E2, UX1, UX3, UX4]. Four of these five participants additionally considered set size, with a preference for larger sets. While participants could have stopped at this point, given the opportunity to explore further, none of the practitioners relied solely on the available metrics. Next, we discuss other ways that participants used Angler to explore the data beyond aggregate metrics and decide which sets to annotate.

Estimating Meaningful Use Cases

From the list of challenge sets in the *Table View* sorted by metrics, three participants chose to prioritize topics that appeared to represent a "meaningful," coherent use case [UX1, E2, UX2]. Partially, this is because the BERTopic [176] model tends to generate some "topics" with little meaning, e.g., the topic `topic-haha_lol_so_you` versus the more meaningful `topic-health_nursing_care_medical` in Figure 8.3.

Partially, participants needed to make judgments on the value of improving the model on various sentence types, since some challenge sets mostly contained data that appeared to be fraud, spam, or automated messages. Participants demonstrated the value of being able to directly read sentences in the *Detail View* (Figure 8.5) to make these judgments:

> *"Yeah, a lot of these are spam. [. . .] As I'm kind of going through them, it's like a lot of spam, a lot of porn and a lot of things that are like, automated messages.*

> *So I would use my discretion, of course, and wouldn't just use the numbers."* — UX1

E1 knew from prior experience that *"it's always good to look at what the data actually are [. . .] besides looking at the high level statistics."* They had seen in the past that even keyword summaries can be misleading and obscure complexity that is apparent when directly inspecting the data:

> *"From my past experience, sometimes we have seen some data contain some keywords and we imagine them to be, for example, articles, but looking at the actual example, they are kind of fraud messaging. [. . .] Combining them together as a single dedicated targeted test would not make too much sense for us to understand the performance on it."* — E1

Some of the dataset contained explicit sexual content and profanity, to which participants ascribed different value. UX1 argued for prioritizing model improvement resources towards use cases that aligned with their organization's values (such as supporting small business owners), *over* explicit content. UX4 was far more accepting of explicit content, arguing that if users were translating that content, there was no reason to treat it differently.

Estimating Impact on Users

Participants assessed how severe the consequences of specific model failures would be for end users. They considered the stakes of the interaction mediated by the model [UX1, UX2, UX3], whether a failure was especially sensitive, e.g., offensive outputs, and how likely a user was to be misled if they were to receive an erroneous translation of this nature, such as an incorrect date [UX2].

In task **T1** all participants looked at number mismatch translations (Figure 8.2). All participants skimmed the raw translations to focus on specific sub-cases of number translation, for instance how the model converts roman numerals, dates, or currency. Even though they could not read the Spanish translation, E2, UX2, and UX4 talked about wanting to find "obvious" errors where they could clearly see numbers changing from English to Spanish. e.g., 1,100 dollars to 100 dollars. An obvious error may not mislead a user, but could degrade their trust in the translation product. Participants dug into specific sub-cases through filtering and search in Angler to get a sense for the severity of an error:

> *"It's a really nice way of quickly getting into the patterns to see whether or not we're looking at something like a serious problem with translation or if it's just kind of surface level formatting issues."* — UX4

Estimating Complexity of the Error

Practitioners wanted to prioritize annotation resources on more complex kinds of failures, rather than those that could be solved internally without additional annotations [UX1, E1].

For example, issues with translating numbers could be identified within the team by using regular expressions to match source and translations. For pattern-based failures, such as translating numbers or translating automated messages, E1 proposed trying data augmentation techniques first. Data augmentation includes increasing the samples and variations on existing data, e.g., the sentence "I ate breakfast on Sunday" can be duplicated to create a sentence for all the weekdays "I ate breakfast on Monday":

> *"If it's a lack of data issue, it should be very easy to augment the data for this particular example."* — E1

While they used train-ratio and familiarity metrics to identify potential coverage issues, directly inspecting the data gave them insight into whether a problem was complex enough to warrant annotation. E3, E2, and UX2 used the *Embedding* view to develop more nuanced hypotheses about how customers' use of the model differs from the training data. Even within an apparently similar topic, participants used clusters of usage data with comparatively less training data to estimate subtopics that may need better coverage [UX1, E2, E3]. They used their experience to hypothesize why or why not an area of low-coverage might be difficult for the model. For instance a cluster with a lot of domain-specific language may be best improved by paying for additional annotations [E2].

While we had initially prompted practitioners to budget annotation resources between the challenge sets we gave them, more often we found that they wanted to prioritize subgroups *within those sets* to optimize annotation for the most complex and impactful issues.

We observed that practitioners were able to form more interesting and user-focused hypotheses for prioritization when they combined summary statistics with qualitative assessment of the data by reading sentences. Their use of Angler was promising evidence for the strength of a visual analytics approach in MT prioritization. At the same time, practitioners demanded more features for flexibly creating challenge sets and exploring more analytics lenses than the Angler prototype supported. We next discuss strengths and limitations of the tool to inform next steps.

Results: **Angler** Strengths and Usefulness

Develop Intuition for Model Behavior

Neural machine translation models are large language models that are not easily understandable even to those who have developed them [232]. We found that exploring data with Angler helped practitioners develop a deeper understanding of how their models work. UX4 said that *"a lot of what I like to do is just develop my like, mental model of how our translation models are working."* Exploring challenge sets by the quality metrics gave E2 *"some insights into the weaknesses of the model."* Given that practitioners' intuitions about model weaknesses guide their future debugging effort (Section 8.3), this can bring value beyond identifying specific failures in the moment.

Develop Intuition for Translation Usage

Participants used the topic-based challenge sets to improve their understanding of how customers use their translation products. The *Keyword Spotlight*, *Sentence List*, and *Source Spotlight* were especially useful for participants to develop hypotheses about how people use the model and the context where they are using it [UX1, UX2, UX3, E3]. UX1 works in a user experience focused role and said that, *"it helps us inform feature development when we understand the conversation topics that people are using* [the model for]*."* While browsing the unit test challenge set based on mismatches in numbers between source and output, E3 imagined potential future features that could give users greater control over how different number formats are handled in translation, e.g., when to use Roman numerals or convert date formats.

Develop a Shared Interdisciplinary Understanding

UX-focused participants expressed excitement for using a visual analytics tool like Angler to broadly explore the use cases surrounding potential failure-modes—as opposed to a purely metrics-driven report that does not allow them to develop their sense of the use context. UX1 and E3 pointed out that being able to describe specific use cases makes it easier to engage cross-functional teams in planning and prioritizing product improvements and developments. UX1 wanted to spend time using Angler to *"generate some insights about each of [the topics] in human understandable terms"* that they could then present to other team members.

> *"Our [team members] come from a variety of backgrounds [. . .] not all of them are engineers. So it's like, could I translate the high level findings here into something that they could understand in a brief?"* — UX1

Using Angler, UX2 and UX3 even learned new insights about where the model performs unexpectedly well on specific kinds of inputs: *"the date formatting changed. I didn't even know if that's something that we do"* — UX2.

Discussing improvements in terms of use cases and specific customer needs not only supports internal cross-functional collaboration, but also makes it easier to acquire new data targeting specific topics from external vendors [UX1].

Results: **Angler Limitations and Usability Issues**

As we mentioned in Section 8.5, most participants did not assign concrete numbers in the annotation budgeting task **T3**. Partially this was a limitation of the study setup, since we provided participants with a long list of pre-made challenge sets in Angler, and there was not enough time for a participant to closely examine all of them. However, in many cases participants also wanted more information and analytics features to drive their prioritization than Angler provides —or else wanted to refine challenge sets from those we pre-made. A major design takeaway for future work is that although Angler is *already* a relatively complex visualization tool, far more lenses and interactions were desired to complete the MT annotation prioritization task. We discuss these key areas for improvement next.

Provide More Context and Comparison

Participants wanted to contextualize the data they were seeing in each challenge set with reference to the overall distribution of data in the training data and usage logs. For example, UX3 and UX4 wanted to know what the average familiarity score was over all of the usage logs to interpret familiarity scores on specific challenge sets. Other participants suggested additional useful reference points, for instance, understanding overall ChrF score distribution by language pair [E2, UX3] or understanding the overall distribution of topics in the training data and usage logs [UX2, UX4]. E3 suggested that it would be helpful to be able to more easily compare challenge sets:

> *"It's not so easy for me to compare them [challenge sets]. So it would be great if I can somehow select a cluster and compare them side by side."* — E3

Support Authoring and Refining Challenge Sets

Our focus in this work was on how visual analytics tools could support the process of prioritizing specific areas for improving MT models. Therefore, we chose two plausible methods for constructing challenge sets to use as examples in the study. While participants found those sets interesting, there was a clear need for future tooling to support challenge set *creation* in addition to exploration.

Participants wanted to search all of the data by specific terms [UX3, E3] and refine the unit test logic to better capture specific types of errors [E1, E2, UX2, E3, UX4]. UX4 even asked if they could onboard their own datasets that they already had available to explore them with Angler's visualizations.

Offer Advanced Export Options for Custom Analysis

Two participants [UX1, E2] expressed a desire to conduct their own analyses on the data, e.g., conduct a custom analysis over time [UX1] or experiment with the underlying topic model [E2]. While Angler offers a simple export option, these options could be made more sophisticated to support users with advanced skills to build on the default visualizations.

Expand Filtering and Sorting Capabilities

Several participants found issues with the data filtering and sorting features that made it difficult to organize and prioritize data in the way they wished [UX1, E1, UX3, E3]. For example, two participants expected to be able to filter to all sentences containing *all* of the keywords selected, but the tool returned *any* such sentences [UX3, E3]. Two participants also struggled to keep track of which filters they had applied when navigating between views [E1, UX3], suggesting potentials to make the ability to view and remove filters more prominent in the interface. Two participants wanted to sort the *Table View* by multiple columns to be able to organize and prioritize sets by multiple metrics or factors, e.g., find large sets within the sets with the lowest average ChrF score [UX1, E1].

Validate and Extend Challenge Set Creation Methods

Grouping data using a topic model was a useful way for practitioners to explore data and better understand use cases for the model. However, it is not clear whether a group of sentences that are close together in a latent embedding space that was trained to group together sentences with similar meaning are also likely to be similarly difficult for an MT model.

There are certainly other dimensions by which to group data. As UX3 described,

> *"When I think about trying to determine where we're doing poorly, there are a lot of dimensions you can look at. Topic is one, right? But there could be other dimensions, like how long the sentence is."* — UX3

In this chapter, we found that exploring challenge sets has the potential to help practitioners prioritize their model evaluation and development resources on issues that are important to end-users. An important direction for future work is to validate that the challenge sets presented indeed represent areas where the translation model performance is relatively weaker. For instance, rather than asking practitioners to allocate hypothetical budgets, they could allocate real budgets and have professional translators evaluate the challenge sets selected to see whether they were able to identify new model failure cases using Angler.

In general, we designed Angler to be agnostic to the method of generating challenge sets. Thus, research developing and evaluating methods for generating challenge sets can proceed in parallel to efforts to improve visual analytics support for exploring, comparing, and prioritizing them.

8.6 Discussion and Future Work

To conclude, we discuss our findings in the broader context of tooling for ML evaluation and debugging and highlight directions for future work.

Trade-offs Between Automation and Human Curation

Angler encourages MT practitioners to inspect training data and usage logs, so that they can better understand how end-users use MT models and detect model failures. Practitioners can then annotate related data samples and retrain the model to address detected failures. Since exploring raw data is a manual and tedious process, we introduce an approach that uses unit tests and topic modeling to automatically surface interesting challenge sets (Section

8.4). Our approach yields many challenge sets, but it still takes time for MT practitioners to inspect and fine-tune these sets. One might argue that we should automate the whole pipeline and have human raters annotate all extracted challenge sets. However, annotating MT data is expensive [151, 329]. In our study, some challenge sets reveal MT errors that are trivial, where MT experts hesitate to spend the budget to annotate the challenge sets (Section 8.5). Besides data acquisition prioritization, our mixed-initiative approach can also help users interact with raw usage logs and gain insights into the real-life use cases of MT models. Future researchers can use Angler as a research instrument to further study the trade-offs between automation and human curation for challenge set creation. To further reduce human effort, researchers can surface challenge sets more precisely before presenting them to MT practitioners.

Generalization to Different Model Types

We situate Angler in the MT context, as it is particularly challenging to discover failures for MT models due to the scarcity of ground truth and the high cost of human annotators (Section 8.2). However, our method is generalizable to different model types. In our formative interview study, we find that it is a common practice to use challenge sets to test and monitor NLP, computer vision, and time-series models (Section 8.3). ML practitioners can adapt our *unit tests* and *topic modeling* (clustering) approach to surface challenge sets for other model types. Consider an image classification model; practitioners could define perturbation-based unit tests to detect model weaknesses. For example, we would expect a model to give the same prediction when the input image is rotated, resized, or with different lighting [393, 234]. Then, we can create challenge sets by collecting images where the model's prediction changes after adding image perturbations. Similar to topic modeling, practitioners could use embedding clustering [30, 91] and sub-group analysis [267, 283] to identify unfamiliar images from both usage logs and training data. For example, if an image classifier team receives a user's complaint about a misclassification, they can use embedding-based image search [383] to identify similar images and create a challenge set. Finally, researchers can adapt Angler's *overview+detail* design (Section 8.4) and open-source implementation (Section 8.4) to summarize extracted challenge sets and allow practitioners to explore and curate potentially error-prone images through different perspectives.

Unit Tests for Machine Learning

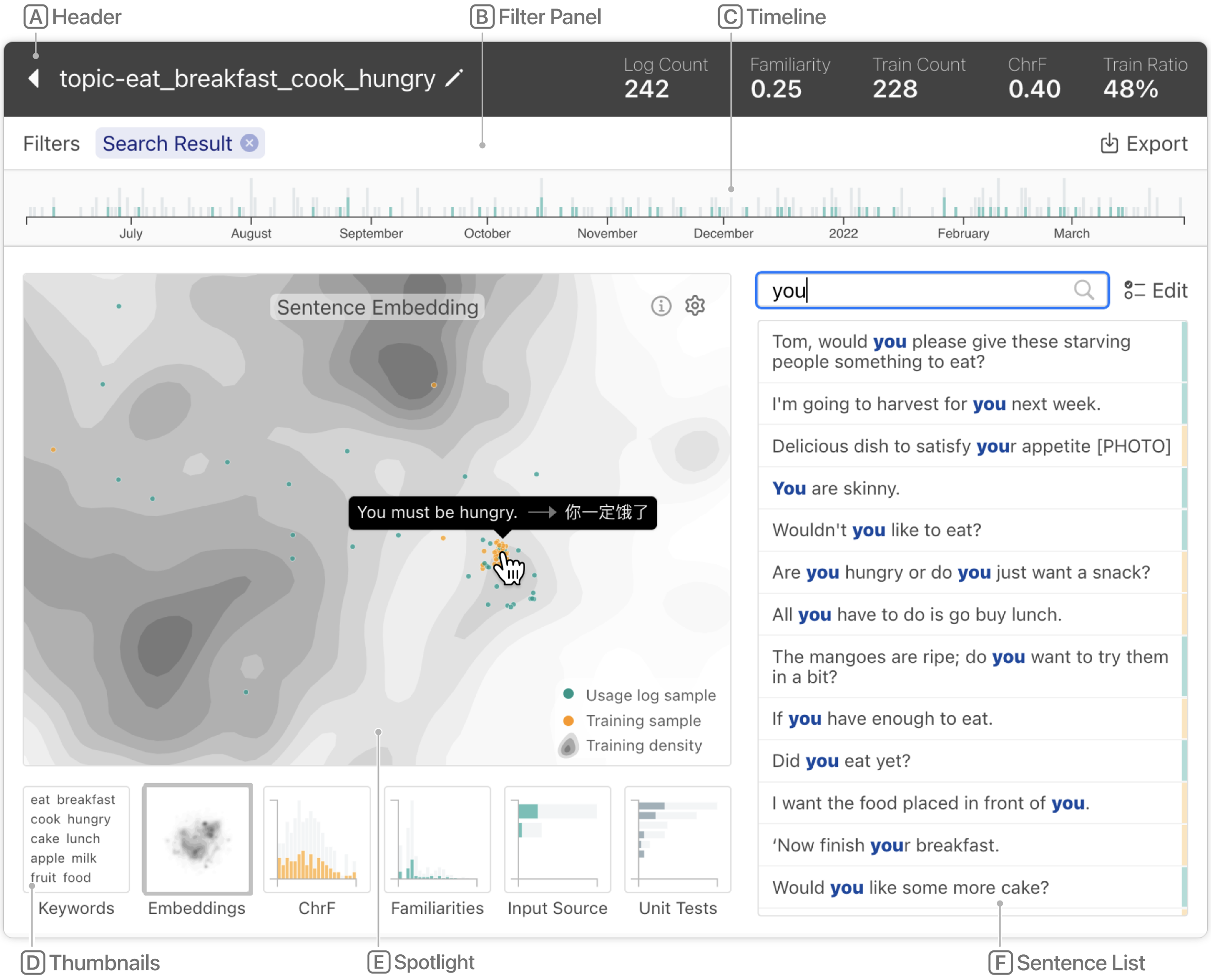
Researchers have argued that unit tests can help pay down the technical debt in ML systems [431, 430]. There are many different ways to apply unit tests to an ML system. For example, practitioners can write unit tests to validate the data quality [373, 423], verify a model's behavior [401, 299], and maintain ML operations (MLOps) [162, 220]. In Angler, we design simple rule-based unit tests, such as if the source does not contain offensive words, then the translation should not either. We then apply these tests to the training data and usage logs to surface challenge sets (Section 8.4). Since they were intended to be a proof of concept, our unit tests were blunt and imperfect. Still, MT practitioners in the evaluation study especially appreciated the unit tests, as they are powerful to detect glaring translation mistakes and yet are easy to adopt in the current model development workflow (Section 8.5). Therefore, we see rich research opportunities to study unit tests for ML systems. For example, future researchers could extend our unit tests to support MT data validation and MLOps in general. Researchers could also adapt Angler to design future interactive tools that allow ML practitioners to easily write, organize, and maintain unit tests for ML systems.

Broader Impact

To overcome the limitations of aggregate metrics on held-out test sets (Section 8.2, Section 8.2), Angler uses *real usage logs* to help MT practitioners gain a better understanding of how their models are used and prioritize model failures. Drawing on usage data raises privacy and security concerns. All of the authors have received training and license to use usage logs from their institution for this research. Researchers adapting Angler should carefully consider the ethical implications of their choice of data source [35]. Before collecting usage logs, researchers need to obtain consent from the users [389] and compensate them when applicable [25]. Usage logs must be de-identified before viewing them with Angler. Finally, we encourage researchers and developers to thoroughly document their process of adopting Angler with new models and datasets [199], including how they approach these ethical considerations.

8.7 Conclusion

This chapter presents Angler, an open-source interactive visualization system that empowers MT practitioners to prioritize model improvements by exploring and curating challenge sets. To inform the design of the system, we conducted a formative interview study with 13 ML practitioners to explore current practices in evaluating ML models and prioritizing evaluation resources. Through a user study with 7 MT stakeholders across engineering and user experience-focused roles, we revealed how practitioners prioritize their efforts based on an understanding of how problems could impact end users. We hope our work can inspire future researchers to design human-centered interactive tools that help ML practitioners improve their models in ways that enrich and improve the user experience.



You must be hungry. ⟶ 你一定饿了

Tom, would **you** please give these starving people something to eat?

I'm going to harvest for **you** next week.

Delicious dish to satisfy **you**r appetite [PHOTO]

You are skinny.

Wouldn't **you** like to eat?

Are **you** hungry or do **you** just want a snack?

All **you** have to do is go buy lunch.

The mangoes are ripe; do **you** want to try them in a bit?

If **you** have enough to eat.

Did **you** eat yet?

I want the food placed in front of **you**.

'Now finish **you**r breakfast.

Would **you** like some more cake?

Figure 8.5: After a user selects a challenge set from the *Table View*, ANGLER presents the *Detail View* to help the user further analyze this set from diverse perspectives. **(A) The Header** shows the name and statistics associated with this challenge set. **(B) The Filter Panel** helps users keep track of the currently active filters. **(C) The Timeline** visualizes the usage log count over time, allowing users to focus on traffic data from a particular time window. **(D) The Thumbnails** and **(E) Spotlight** visualize diverse representations of sentences in this set—users can click different *Thumbnails* to switch the *Spotlight*, on which users can further filter sentences with particular attributes. **(F) The Sentence List** displays all sentences that meet the active filters, where users can inspect translations, search words, and remove sentences from this set.

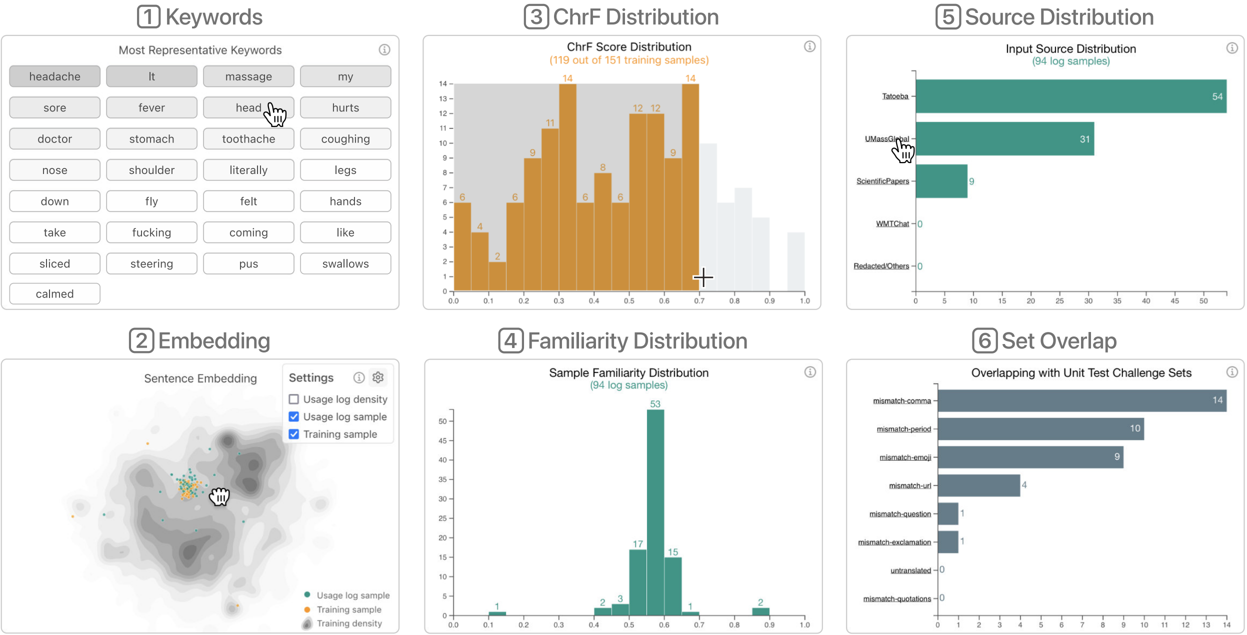


Figure 8.6: The *Detail View* presents six options for the *Spotlight* to help users explore a challenge set from diverse perspectives. **(1) The *Keywords*** shows the most representative words in a set. **(2) The *Embedding*** uses a zoomable scatter plot with contour backgrounds to help users explore the high-dimensional representations of sentences in a set. **(3) The *ChrF Distribution*** allows users to inspect and filter training sentences by their ChrF scores. **(4) The *Familiarity*** helps users filter usage logs by the models' familiarity scores. **(5) The *Source Distribution*** visualizes the usage log source as a horizontal histogram where users can filter usage logs from particular sources. **(6) The *Set Overlap*** allows users to see sentences that are also in other challenge sets.

Chapter 9

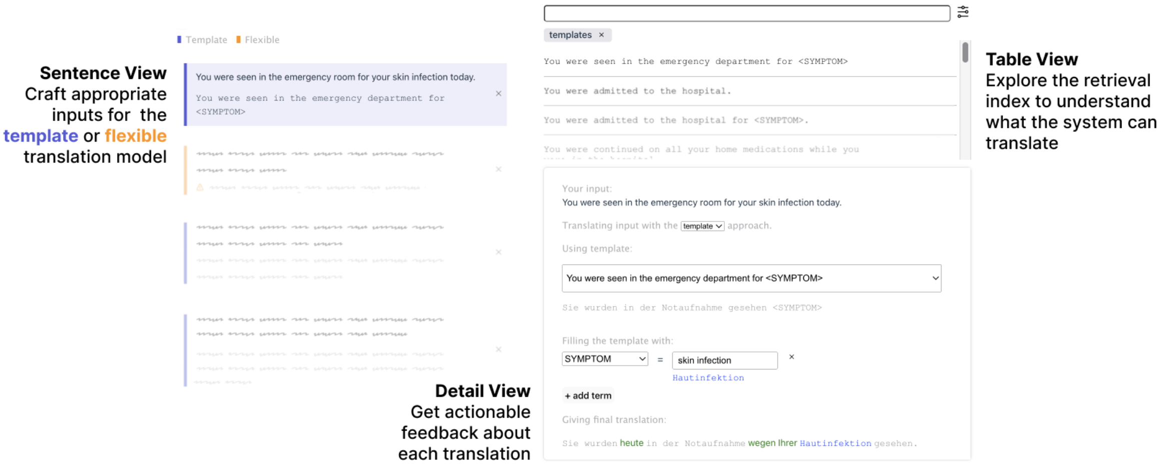
Cephalo: Leveraging Retrieval for Verifiable Machine Translation

Machine learning (ML) can be unsafe in high-stakes settings if users are unable to understand a model's output and identify errors.[1] In this chapter, I argue that one way to address this challenge is to help users understand what kinds of inputs are well supported by a system. I demonstrate this approach with Cephalo, a system that produces high-quality, verifiable translations of hospital discharge instructions. Cephalo uses two retrieval-augmented translation models to translate free-text instructions with reference to a database of professionally translated terms and phrases commonly found in discharge instructions. I developed Cephalo using over 3,000 emergency department discharge instructions in addition to publicly available medical domain data, and evaluated it in a user study with 18 participants. Compared to a typical neural machine translation system, Cephalo gives users a principled way to edit their inputs to a translation model to be make them more appropriate candidates for translation.

9.1 Introduction

Complex machine learning models fail in unpredictable ways. Many types of models produce outputs that are difficult for users to verify. For example, machine translation and code generation models take inputs in one language and produce outputs in a different (natural or computer) language, which the user may not understand. In these settings, it is difficult for a user to decide when they should rely on model output [259, 32, 482, 231]. Where possible, researchers have proposed providing users with explainable and interpretable models [499, 241, 242, 7, 289, 83] and setting appropriate expectations about what a model can and cannot do [259, 482]. However, such approaches are challenging to implement when the model can generate a wide range of outputs (e.g. text), and when error boundaries are highly stochastic

[1] Ana Milisavljevic, Elaine C. Khoong, Karan Bains, Katrin Jaradeh, Sylvie Venuto, and Niloufar Salehi contributed to the research presented in this chapter.



Sentence View
Craft appropriate inputs for the template or flexible translation model

You were seen in the emergency room for your skin infection today.

You were seen in the emergency department for <SYMPTOM>

Table View
Explore the retrieval index to understand what the system can translate

You were seen in the emergency department for <SYMPTOM>

You were admitted to the hospital.

You were admitted to the hospital for <SYMPTOM>.

You were continued on all your home medications while you

Detail View
Get actionable feedback about each translation

Your input:
You were seen in the emergency room for your skin infection today.

Translating input with the template approach.

Using template:
You were seen in the emergency department for <SYMPTOM>

Sie wurden in der Notaufnahme gesehen <SYMPTOM>

Filling the template with:
SYMPTOM = skin infection
Hautinfektion

Giving final translation:
Sie wurden heute in der Notaufnahme wegen Ihrer Hautinfektion gesehen.

Figure 9.1: Cephalo leverages retrieval-augmented machine translation models to help users craft appropriate inputs for translation and verify translation outputs. Cephalo integrates two translation models: a template-based approach, which matches input sentences to a professionally translated template and terms; and a flexible approach that combines professionally translated sentences to produce output translations. The table view displays the templates, terms, and sentences that these models use. Users can craft inputs closer to the data in the table view to improve translation reliability. The detail view allows users to refine translations through actionable feedback.

and unpredictable [32]. Without support to identify and recover from model failures, it is risky to rely on these models in high-stakes settings.

One example is when patients and clinicians use machine translation models to communicate in healthcare settings. Although clear and trustworthy patient-clinician communication is essential, there are often major barriers to reliable and timely language support when patients and clinicians do not share a common language [307]. Some researchers suggest that machine translation could be helpful for bridging language barriers when human interpreters are not available [249, 435, 327]. However, others argue that this is too risky because these tools are not specialized for medical settings, and it is difficult to ensure that patients are getting correct information from the translations [476, 123]. While pre-translated resources can offer some verified translations, their lack of flexibility limits their utility for broader clinical communication [490]. Therefore, there is great potential for tools that balance flexibility and verifiability.

In this chapter, I argue that retrieval-augmented machine translation offers opportunities to balance the flexibility of machine translation and the verifiability of pre-translated resources. Retrieval-augmented methods use a retrieval index of verified translations to generate new translations. These approaches can help a user generate more reliable translations by verifying that their inputs are similar to the data in the retrieval index. Because the retrieval index is bilingual, this verifiability does not require the user to speak the target language.

To demonstrate this idea, I designed Cephalo, a translation system for emergency department discharge instructions that relies on two different types of retrieval-augmented translation models. These models use a custom retrieval index that I developed using over 3,000 real emergency department discharge instructions, as well as three open sources of medical domain text data. I conducted a user study with 18 people to understand how Cephalo supports users to craft better inputs for machine translation.

I found that users were able to edit fictional discharge instructions to be closer to the retrieval index without changing the meaning. Their edits were specifically tailored to the strengths of the retrieval-augmented models. Participants preferred methods that gave clear and structured ways to verify how a translation was being generated, over more flexible methods that require a user to rely more heavily on their own judgment of input appropriateness.

I contextualize this work in the broader literature on interaction design for complex AI systems, specifically systems that produce outputs that users cannot directly verify. I argue that leveraging retrieval is one way to build a shared intent specification between the user and the system, that the user can then refine to ensure the system is achieving their goals. In some cases, developing such a specification requires constraining the scope of inputs that a system supports, but allows system designers to more precisely describe what a system can and cannot do. Ultimately, these new design patterns could increase users' agency to decide how to use a system, informed by a clear understanding of that system's capabilities and limitations.

9.2 Related work

This work builds on recent literature on human-centered AI, which seeks to design interaction patterns for AI systems that better meet the needs of users, developers, and other people who might be affected by a system [363, 259, 508, 169, 405, 17, 18, 5] . Here I review this literature and existing approaches to designing reliable machine translation systems. I also briefly review the literature on retrieval-augmented generation methods in natural language processing, which I used to build Cephalo. In the next section, I introduce the application domain of translation tools for hospital discharge instructions, and define the design goals for Cephalo.

Designing human-AI interaction

Recent advances in AI have resulted in a range of new possibilities and challenges for designing interactive user interfaces [17, 211, 198, 527]. These challenges stem from two main properties of AI systems: 1) uncertain capabilities: it is difficult to articulate what the system can and cannot do, and anticipate and mitigate unpredictable behaviors, and 2) output complexity:

the more possible outputs for a model, the more difficult it becomes to assess reliability and sketch effective interactions [527]. These challenges make it difficult for end-users to detect and recover from model errors [32, 74], especially in the context of natural language systems (e.g., conversational agents, machine translation, and code generation systems), which have large input and output spaces, and complex, stochastic error boundaries [436, 26, 406]. To this end, researchers in human-computer interaction and CSCW have proposed design guidelines and methodological frameworks for the design of interactive AI systems, including making clear what the system can do and how well it can do it, as well as supporting efficient correction when the system makes an error [17, 472, 536, 86]. Techniques for implementing these guidelines include breaking down the task into smaller subtasks (e.g. LLM chains [518]), grounding interaction with reference to specific user goals [26, 286], and setting expectations of model behavior, e.g. by showing performance metrics or examples of model behavior, or providing controls over performance trade-offs [259, 88, 89, 86].

The gulf of execution and the gulf of evaluation are helpful conceptual tools in the design and development of AI systems [345, 413]. The gulf of execution refers to the gap between the user's intentions and the actions that the system allows them to perform, while the gulf of evaluation refers to the difficulty of assessing the current state of the system and how well it supports the user's understanding of that state. To bridge the gulf of execution, systems must make their functions discoverable, allow users to do what they wish within safe constraints, and maintain consistency with other tools. To bridge the gulf of evaluation, systems must give constant feedback to the user that is immediate and matches the user's actions. Leveraging direct manipulation can also be helpful in providing immediate feedback to bridge the gulf of evaluation. One technique for bridging these gaps in AI systems is automatically inferring user intent based on their open-ended inputs [381, 509, 252], or showing the user a small number of possible intents and allowing them to pick one and change it [359, 26, 159]. In this chapter, I explore how retrieval from a set of verified input-output pairs can help bridge these gaps by more clearly specifying system capabilities and making explicit a system's interpretation of users' intent. Overall, designing AI systems that effectively bridge these gulfs is essential for creating systems that are usable, reliable, and verifiable for users.

Reliable machine translation

People often turn to machine translation (MT) systems because they need to communicate in a language they do not know, making it difficult for them to identify model failures [403]. The consequences of misunderstandings due to undetected MT errors can range from confusion, frustration, and embarrassment [**liebling2020unmet**, 404, 187], to serious safety and human rights concerns [249, 502, 123, 48, 484].

With existing systems, users have limited strategies available to detect and recover from errors. If users know the output language, they can assess translation fluency, but neural MT models can produce very fluent but incorrect translations [312, 43]. If users can understand the input but not the output, they might compare the back-translation to the input to look for discrepancies [446, 323, 526, 325]. This approach also has limitations, since the model can introduce new errors in the back-translation step. Researchers have found that showing users multiple translation hypotheses from one or different models [158, 522], or providing additional linguistic and contextual information like keyword highlighting and definitions [157, 291] can help improve MT-mediated communication. Quality estimation models predict the quality of a translation without access to reference translations [458]. It is not yet clear how to integrate these various, uncertain signals of quality in a way that is useful and reliable for end-users [325, 537, 322, 317].

MT systems also give users little guidance regarding how they can be used [406]. Research has found that users try to avoid slang, idioms, and metaphorical language, and use correct spelling and grammar to improve model performance [403, 187]. However, it is hard for people to follow those guidelines in real communication and to adapt their behavior as systems are updated [404]. One approach is to interactively guide users to craft inputs within a model's capabilities [320, 97], but these approaches have not been adapted to modern neural models. I build on this work by exploring how retrieval-augmented machine translation can give users insight into what kinds of inputs an MT system supports and offer better guarantees on output quality.

Retrieval-Augmented Generation in NLP

Retrieval-augmented language models have been developed in the field of natural language processing (NLP) for knowledge-intensive tasks (e.g., question answering, fact checking, and machine translation), where it is helpful for models to have access to verified information [285]. These approaches were motivated by language models' difficulties with producing factual information, as well as the desire to trace model output back to its sources, and keep models' implicit world knowledge up to date over time [285].

Given an input, retrieval-augmented language models retrieve examples from a retrieval index of verified input-output pairs (sometimes, but not necessarily, the model's training data), and use those examples to inform its output. Researchers have proposed a range of architectures, some specialized for specific tasks [104], and others that work across multiple tasks [285, 247]. Some of these approaches retrieve relevant documents at a sequence level [193, 285, 104], while others retrieve a different set of relevant documents for each output token [247, 285, 246]. Retrieval-augmented generation generally refers to models that flexibly draw on several retrieved documents, although these are closely related to retrieve-and-edit approaches [193, 309, 308, 310, 212], which retrieve a single similar input-output pair and then edit the retrieved example to produce a final output.

Cephalo uses a k-nearest neighbor machine translation model from Khandelwal et al. [247] and a template-based model that I developed, which is most similar to the retrieve-and-edit approach. I discuss each of these models in depth in Section 9.5. I build on the prior literature on retrieval-augmented models by exploring how retrieval can not only improve output quality, but also enable new kinds of user interactions that support verifiability.

9.3 Design goals for translating hospital discharge instructions

Cephalo is a tool for translating hospital discharge instructions. Here, I introduce this application domain and my design goals.

When a patient is discharged from the hospital, their care team will write instructions for them to take home. These patient discharge instructions typically describe why the patient came to the hospital, what tests and treatments they received while they were there, their diagnosis, and instructions for ongoing treatment, e.g. changes to their medications and follow-up appointments.

Written communication is a challenge when physicians and patients do not share a language. Even when a bilingual interpreter is available to discuss the instructions with the patient while they are still in the hospital, they may not have time to translate instructions for a patient to take home. This could be one factor contributing to systematically poorer post-discharge outcomes for patients in the U.S. with limited English proficiency [307].

One option is for clinicians to use machine translation to provide patients with their instructions in their preferred language [249]. Although using MT is not officially endorsed for use in clinical practice by any regulatory body, it is already widely unofficially used in written and verbal clinician-patient communication [316, 502, 489]. However, poor quality translations in this setting have the potential to cause serious harm to patients [249]. These risks are particularly high for patients who speak lower-resource languages [476]. Unfortunately, there is little to no guidance as to how users can mitigate these risks when using MT systems [476, 326, 391].

Phrase-based translation tools offer professional translations of a fixed set of statements. Clinicians and patients have expressed a preference for these tools over MT because communication is clearer and more reliable [490, 457, 358]. However, the usefulness and adoption of these tools is limited by their lack of flexibility. Some electronic health record (EHR) systems in the U.S. offer patient instructions that have been translated into several languages.[2] However, patients often need instructions that are customized to their specific situation, and it is very difficult (or impossible) for clinicians to edit these translated materials if they don't speak the target language.

Therefore, there exists a need for techniques that can balance the flexibility of open-ended MT tools with the reliability of pre-translated resources (Figure 9.2). Hospital discharge instructions offer a useful starting point to this work, since they are structured and repetitive but still require customization to individual patients. In fact, some hospital systems and public sources have developed templates for discharge instructions that offer providers an easy starting point that they can then customize. However, there is a lack of tooling to support customization of these multilingual resources in a way that is reliable and verifiable.

[2] E.g., `https://www.wolterskluwer.com/en/solutions/lexicomp/about/epic/integrated-patient-education`

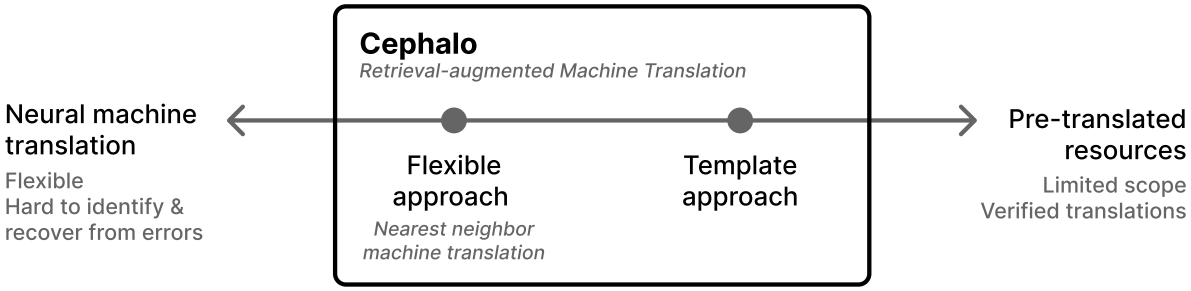

Cephalo
Retrieval-augmented Machine Translation
Neural machine translation
Flexible
Hard to identify & recover from errors
Flexible approach
Nearest neighbor machine translation
Template approach
Pre-translated resources
Limited scope
Verified translations

Figure 9.2: If there are no translators available, clinicians are faced with a choice between flexible but uncertain neural machine translation tools (left), and inflexible but verified pre-translated phrasebooks (right). I argue that retrieval-augmented translation methods, with appropriate user interface design, could provide a balance between flexibility and verifiability. I use two retrieval-augmented translation methods in this work: a flexible approach, based on Nearest Neighbor Machine Translation [247] (Section 9.5); and a template approach that I developed for discharge instructions (Section 9.5).

Based on prior research on translation needs in clinical settings and discussions within an inter-disciplinary research team of physicians and computer scientists, I determined **five design goals** for a translation system for hospital discharge instructions:

- **D1: Produce high quality translations.** Translations should be accurate, clear, and culturally appropriate.
- **D2: Allow users to verify translation quality without knowing the target language.** Users should be able to verify what information is communicated by a translation and avoid errors that could pose risk to a patient.
- **D3: Help users edit their input to improve translation quality.** It should be easy for users to understand system capabilities. Feedback should guide users towards well-supported inputs.
- **D4: Make efficient use of existing resources.** Flexibly re-using pre-translated materials can save resources and help prioritize the time and effort of human translators.
- **D5: Allow frequent and transparent updates to system capabilities.** Over time, the system should improve based on usage patterns and adapt to shifts in medical language. The costs of adding a new language should be relatively low across high- and low-resource languages.

I argue that *Retrieval-Augmented Machine Translation* models (Section 9.2) are one option for building systems that meet these design goals. Prior work has shown that retrieval can improve the performance of machine translation models, particularly for domain-specific translation tasks like medical translation (**D1**) [247]. These models draw on a retrieval index of translation pairs, meaning they make use of existing translated resources to improve translation quality (**D4**).

Cephalo integrates two different kinds of retrieval-augmented translation: a flexible approach, based on the k-nearest neighbor machine translation architecture from Khandelwal et al. [247], and a template approach, which uses a pre-defined set of templates, which can be flexibly filled with appropriate terms to generate full sentences (Figure 9.2). These two approaches do not require any augmentation of an underlying neural network, e.g. fine-tuning, making it very easy to add new translations to the retrieval index at any time (**D4**, **D5**).

My hypothesis in this work is that retrieval-augmented translation models make it easier to design actionable feedback that could help a user verify translation quality and craft inputs that are appropriate for the models (**D2**, **D3**). These models perform best on inputs similar to examples in the retrieval index. Since the index is bilingual, users can inspect the index to understand the system's capabilities (**D5**) and verify that their input can be translated well (**D2**). If they see that their input is not close to the index, they could edit their input accordingly (**D3**).

In the next section, I discuss how I built a custom retrieval index for translating hospital discharge instructions. In Section 9.5, I introduce the two retrieval-augmented translation models that I use in Cephalo. I then bring these two components together and introduce the Cephalo user interface in Section 9.6.

9.4 Building the Retrieval Index

My goal in this work is to show how retrieval-augmented MT can help users craft appropriate inputs for translation and verify translation quality. There are two main components to any retrieval-based translation model: a *retrieval index*, which is a set of verified translation pairs; and a *retrieval-augmented translation model*, which retrieves one or more examples from the index, and uses those to generate a final translation. In this section, I describe how I built a custom retrieval index for translating emergency department discharge instructions. In the next section, I introduce two models that use this index.

Ideally, we would want a large retrieval index that is specific to emergency department discharge instructions. One option would be to sample real discharge instructions from the hospital where the system will be used, de-identify them, and translate them into the target language. However, this approach is very costly and thus infeasible for early prototyping.

I reduce this cost by leveraging the structure and repetitiveness of the text to avoid collecting many extremely similar translations, mixing synthetic and real data sources, and using Google Translate to collect proxy reference translations. To reduce reliance on proxy reference translations, the largest data source is a parallel corpus of medical domain data that is used to train medical domain-specific translation models [262, 247]. This data is English-German, so I use German as the target language for prototyping Cephalo. Future work can easily extend this approach to other languages given a moderate translation budget.

The retrieval index contains four sources of synthetic and real medical domain data. First, I abstracted patterns from real discharge instructions, and use those patterns to generate realistic, synthetic sentences (Section 9.4). Second, I generated synthetic sentences from three widely used sources of discharge instruction templates (Section 9.4). I augmented this data with two sources that are closely related to discharge instructions: synthetic discharge summaries [270] (Section 9.4), and medication documentation from the European Medicines Agency [262, 9] (Section 9.4). In total, the retrieval index contains 8,706 sentence pairs: 641 from the custom synthetic discharge instructions, 912 from the online discharge instructions, 750 from the discharge summaries, and 7,044 from the medication documentation. I held out 2,083 pairs for model development and 2,022 as a test set for final model evaluation. The following sections describe each of these data sources in more detail. I conclude this section by describing how I translated the data and discussing the limitations of my method.

Real discharge instructions

To generate synthetic discharge instructions that reflect realistic language patters, I curated sentence templates from real discharge instructions, then filled the templates using GPT-4. To evaluate this approach, I measured how much of the meaning of a held out set of real instructions could be conveyed using only templates.

Data collection

I collected 3,089 emergency department discharge instructions from the UCSF emergency department, written between November, 2016 and September, 2021. Prior to us accessing the data, the instructions were run through an open source de-identification tool designed specifically for clinical notes to remove personal health information (PHI) [344]. I further filtered the data to include only free-text instructions that were written in English.

I randomly split the real instructions into a training set of 2,324 instructions for generating synthetic data, a development set of 447 instructions for iterative use during development, and a test set of 200 instructions for final evaluation. In this chapter, I only use 25 of the test set instructions for evaluation. I split each set of instructions into sentences using an open-source sentence tokenizer [53]. After data filtering, there was a total of 2,971 instructions split into 20,790 sentences across the three data splits. I manually corrected errors in sentence segmentation and PHI redaction on the test set.

Generating synthetic source sentences

I used a mix of quantitative and qualitative data analysis to capture patterns in the training data. To reduce redundancy and avoid directly using any patient data in the system, I defined templates to summarize common sentences, e.g., "You were seen in the emergency department for [SYMPTOM]," then used GPT-4 to generate synthetic examples from each template.

Real Sentence	**Type**	**Template**	**Synthetic Sentence**
Please take Seroquel 50mg once per day as needed for hallucinations.	Medication	Take [MEDICATION] [DOSAGE] for [SYMPTOM].	Take Ibuprofen 200 milligrams every 4 to 6 hours as needed for pain or fever.
Your pain may be caused by constipation or by the pregnancy.	Diagnosis	We think your [SYMPTOM] is due to [CONDITION].	We think your joint pain is due to rheumatoid arthritis.
We gave you some medication to increase your potassium.	Treatment	We gave you [MEDICATION] to increase your [SYMPTOM].	We gave you Albuterol to increase your lung function and reduce asthma symptoms.
You had an xray of your knee, which showed a fractured kneecap.	Tests	The [TEST] showed [CONDITION].	The X-ray showed a minor fracture in your left arm.

Table 9.1: Examples of the templates that we generated from clusters of real sentences, and synthetic sentences that we generated from each template. From 63 clusters of real sentences, we generated 163 templates of 13 types. From the templates we generated 1,172 synthetic sentences.

To define templates, I trained a BERTopic model[3] [176] to cluster the sentences in the training data into semantically similar groups. I analyzed the largest 63 topics to define templates that would cover the intent expressed in each cluster of sentences, stopping when she began to see more repetition in topics than new content.

A team of computer scientists and physicians iteratively refined and combined these templates to further reduce redundancy. In this stage, we went from 167 templates to 113 templates, grouped into 13 categories: “Greeting,” “Complaint,” “Consultation,” “Labs,” “Diagnosis,” “Treatment,” “Medication,” “At home instructions,” “Follow-up appointment,” “Discharge,” “Return instructions,” “Salutation,” and “Contact Information.” As I developed the translation models, we refined and added templates, ultimately converging on a final list of 163 templates that I evaluated on the test set.

Finally, for each template, I prompted GPT-4 to generate 10 sentences with the given template format with a zero-shot prompt[4]. 41 of the templates did not have any variables in them, so I did not generate synthetic data from those. I spot-checked the outputs at this stage to ensure they were realistic and true to the template. These 1,172 sentences formed the synthetic discharge instructions dataset. I used 600 synthetically filled templates plus the 41 templates without variables in the retrieval index. I held out 150 synthetically filled

[3] The default parameters were effective for my purposes.

[4] Full prompts in Appendix B.1.

sentences as the test set, and used the remaining 381 for development. Table 9.1 shows example sentences from the real discharge instructions, corresponding templates, and example synthetic sentences[5].

Evaluation

My goal with this method was to generate synthetic sentences that cover a wide range of topics that appear in discharge instructions. To evaluate my approach, I measured how much of the meaning in the 25 test set instructions can be conveyed using only templates.

Method I matched each unique sentence in the test set (N=203) to its best equivalent template(s), or "None of the above." I did two passes of this task to validate the selections. Two physicians on the team independently evaluated the matches and a third adjudicated disagreements (69 sentences, 34%). Disagreements between all three annotators were resolved through discussion (23 sentences). Each annotator rated the template's adequacy at conveying the meaning of the original sentence on a five-point scale, following common practice in translation evaluation [95]. For any matches that did not convey all of the meaning (adequacy < 5 out of 5), the annotators judged the potential for harm to a patient if they received only the information conveyed by the template using an established three-point rating scale: clinically non-significant, clinically significant, or life-threatening potential harm [335]. After resolving disagreements, none of the sentences were rated as having life-threatening potential for harm, so I report results using a binary significant/non-significant scale. The annotators rated the sentences in order, so that they had the full context for each set of instructions. Evaluating translation quality is a nuanced and difficult task, and we found that evaluating template matches was similarly difficult, reflected in relatively low inter-annotator agreement. Given a larger number of annotators, future work could explore using continuous scales to improve reliability [171].

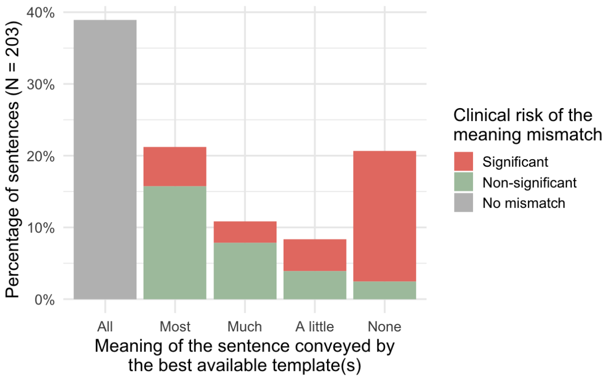


Figure 9.3: 60% of the sentences in a set of 25 discharge instructions could be matched to a template that conveyed all or most of the meaning. Overall, 31% of the best available matches had the potential for clinically significant harm if a patient received only the information in the template.

Results 60% of the sentences in the test set had matching template(s) that conveyed all or most of the meaning (adequacy $\geq$ 4 out of 5) (Figure 9.3). Overall, 31% of the best

[5] More examples in Appendix B.1.

available matches had the potential to cause clinically significant harm to a patient. The set of templates could be expanded to improve performance. 42 sentences (20%) did not have a match that conveyed even a little of the meaning (adequacy 1 out of 5). 23 of these sentences contained general information related to home activities (e.g. hand washing), or specific conditions (e.g. back pain), which could easily be added to the retrieval index and would likely apply to many patients.

In summary, the synthetic discharge instructions cover many but not all of the intents conveyed in the real emergency department discharge instructions. To improve the generalization of the retrieval index, I augment this data with three other data sources.

Online discharge instruction templates

I scraped 156 discharge instruction templates, which are commonly used as a starting point for emergency department discharge instructions, from three websites[6] in October 2023. I split the instructions into sentences using an open source sentence tokenizer [53], dropped the shortest 1% of sentences, and filtered data containing line breaks and special characters, as well as sentences containing only contact information for clinics (e.g. address, phone number, or opening hours). This process left 1,274 sentences, and 674 after dropping exact duplicates. Of these, there were 79 unique sentences that were templates (e.g. "You have been evaluated in the Emergency Department today for ***"). I filled these templates using GPT-4, as in Section 9.4, generating a total of 1,384 unique sentences. I split the data at the instruction level, holding out 15 instructions for development (168 sentences) and 15 instructions as a test set (291 sentences). The remaining 126 instructions (912 sentences) were added to the retrieval index.

Synthetic discharge summaries

I sampled 3,000 synthetic discharge summaries from the publicly available Asclepius dataset [270]. These notes were generated from publicly available case reports using GPT-3.5 Turbo and validated by clinicians. These notes differ from discharge instructions because they are intended for a clinical audience, not the patient, but they contain similar content. From these, I randomly sampled 2,000 notes to be a training set, and 500 each for development and test sets. I split each note into sentences on new lines and using an open-source sentence tokenizer [53], giving 72,028 sentences. Due to resource constraints, I further subsampled sentences from each of these sets. From the training data, I sampled 750 unique sentences to add to the retrieval index. From each of the development and test sets, I sampled 400 sentences for evaluation.

[6] `https://natedotphrase.com/portfolio/dc-instructions/`, `https://tydotphrase.wordpress.com/discharge/`, `http://brianemr.blogspot.com/p/discharge-reassessment.html`

Medication Documentation

Lastly, I sampled 12,000 out of 248,099 English-German sentence pairs from the medical domain data in the multi-domains dataset from Koehn and Knowles [262], and cleaned and re-split by Aharoni and Goldberg [9]. This data originates from PDF documents from the European Medicines Agency [9]. During model development, I found that there were some poor quality source-translation pairs that were severely impacting evaluation results, so I further filtered this data based on source-target similarity.

I computed sentence similarity between source and reference sentences using Language-Agnostic BERT Sentence Embeddings [148]. I filtered to pairs with cosine similarity greater than 0.8 and where the English and German were not identical. I also filtered out any sentences in the bottom or top 1% of sentences by sentence length, with the percentile cutoffs computed separately for English and German sentences. Finally, I removed document identifiers by removing sentences containing "EU/". This left 7,044 sentence pairs from the training data. I applied the same filtering procedure to the 2000 sentences in the development and test splits of this data, using the sentence length quantiles from the training set. This left 1,134 development set sentence pairs, and 1,181 test set pairs.

Translating the retrieval index

The real discharge instructions, the online discharge instructions, and the discharge summaries were all generated in English. For a real application, we would collect translations of all sentences in the retrieval index from professional medical translators and design rigorous quality assurance processes. To save resources during the iterative design process, I used Google Translate as a proxy for high-quality reference translations.[7] I discuss the limitations of this approach in the next section.

Limitations

Developing and translating high-quality standardized language resources requires collective, sustained effort from clinicians, linguists, medical translators, and technologists [213]. My goal in this work was to explore how the existence of such a resource can enable new interactions with translation systems that support more reliable and safe translation of hospital discharge instructions. The retrieval index offers a useful starting point, but systems deployed in hospital settings should undergo a thorough and rigorous development and quality assurance process.

By generating synthetic discharge instructions from the largest topics in the topic model, I prioritized content that occurs frequently in the data. However, criteria other than frequency may also be important to consider, for example, whether a sentence carries particularly

[7] I collected German translations of the real discharge instructions in June 2023, the online discharge instructions in October 2023, and the discharge summaries in September 2023 using the Google Translate API.

safety-critical information, or whether it contains information that could be particularly difficult for an MT model.

I generated synthetic sentences from templates without guidance as to how those templates should be filled, so it is possible that these sentences could contain misleading information. In a real application, any synthetically generated data should undergo more thorough quality assurance.

Finally, I relied on Google Translate for proxy reference translations. Given I am working with a high-resource language pair, it is likely that these translations are highly accurate, but they may still contain errors [521]. Prior work has already demonstrated that retrieval-augmented models improve machine translation performance [247]; my goal was to explore how we can help people use these models more reliably. My assumption is that if users can use the system to generate translations more consistent with proxy reference translations, then they would also be able to do the same with higher quality, verified references. I strongly advise against using proxy reference translations in any real system for clinical use.‘

9.5 Retrieval-Augmented Machine Translation

In the previous section, I described how I built a custom retrieval index for translating hospital discharge instructions. In this section, I introduce two models that use the retrieval index to generate translations. The first model is a nearest neighbor machine translation (kNN-MT) model, as introduced by Khandelwal et al. [247]. This model retrieves a new set of examples to predict each output token. This approach can improve translation quality on inputs close to the retrieval index, while still allowing flexibility. The second approach sacrifices some of this flexibility to further increase verifiability. This approach matches an input sentence to one of the templates I developed in the previous section (9.4), and fills the template with relevant terms to produce a final translation. For each of these models, I describe the implementation, evaluate the model's performance, and discuss how it supports verifiability.

Nearest Neighbor Machine Translation

First, I introduce the nearest neighbor machine translation model as described in Khandelwal et al. [247], then describe and evaluate my implementation. Finally, I propose two metrics for helping users verify outputs and improve their inputs to this model: utilization and relevance.

Background

First, I describe how a kNN-MT model works, so that I can define utilization and relevance in Section 9.5. I refer readers to Khandelwal et al. [247] for complete details. I have stayed consistent with their notation and terminology as much as possible.

Let $s = (s_1, ..., s_{M_1})$ be an input sequence that we would like to translate into a sequence in the target language, $\hat{t} = (\hat{t}_1, ..., \hat{t}_{M_2})$. A kNN-MT model is built on a pre-trained neural machine translation model, which we will refer to as the *baseline model*. The retrieval mechanism is added to the baseline model's decoder, and retrieves the k nearest neighbors at each generation step (each time it needs to predict a target token $\hat{t}_i$).

Recall that the retrieval index consists of source sentence-reference translation pairs. Given a source sequence s and a reference translation t, we refer to each token $t_i \in t$ as a *target token*, and the source sequence and preceding target tokens, $(s, t_{1:i-1})$, are the corresponding *translation context*. The *datastore* for a kNN-MT model is a map from translation context to target token, for each target token in each source-target pair in the retrieval index. The keys in this map are high-dimensional embeddings of the translation context, in this case as computed by the baseline model's decoder. Let f represent the decoder embedding step. Let $\mathcal{R} = (s, t)$ be the retrieval index of source-target pairs. The key-value pairs in the datastore $\mathcal{D}$ are then: $(f(s, t_{1:i-1}), t_i), \quad \forall t_i \in t, \forall (s, t) \in \mathcal{R}$.

Given an input sentence to translate, an autoregressive neural MT model generates a probability distribution over its vocubulary ($\mathcal{V}$) at each generation step, based on the input sequence and the previously generated tokens: $p_{\mathrm{MT}}(t_i|s, \hat{t}_{1:i-1}), \quad \forall t_i \in \mathcal{V}$. A kNN-MT model augments p_{MT} with a separate probability distribution over the vocabulary induced by the k nearest neighbors in the datastore. Given source sentence s and partial generation $\hat{t}_{1:i-1}$, we take the embedding of this translation context, $f(s, \hat{t}_{1:i-1})$, and retrieve the k-nearest neighbors ($\mathcal{N}$) in the datastore using a distance function d (I use squared-L^2 distance, following Khandelwal et al. [247]). The neighbors are then used to compute a distribution over the vocabulary, by taking the negative distances between the input translation context embedding and the neighbor embeddings, flattening the distribution using a softmax function with temperature T, and aggregating by target token:

$$p_{\mathrm{kNN}}(t_i|s, \hat{t}_{1:i-1}) \propto \sum_{(k_j, v_j) \in \mathcal{N}} \mathbb{1}_{t_i = v_j} \exp\left(\frac{-d\left(k_j, f\left(s, \hat{t}_{1:i-1}\right)\right)}{T}\right) \tag{9.1}$$

This distribution is then linearly interpolated with the baseline model:

$$p(t_i|s, \hat{t}_{1:i-1}) = \lambda p_{\mathrm{kNN}}(t_i|s, \hat{t}_{1:i-1}) + (1 - \lambda) p_{\mathrm{MT}}(t_i|s, \hat{t}_{1:i-1}) \tag{9.2}$$

The number of neighbors retrieved at each step, k, the softmax temperature parameter which flattens the p_{kNN} distribution, T, and the interpolation parameter, λ, are tunable hyperparameters.

Implementation

I implemented a nearest neighbor machine translation model on top of an open-source neural MT model for English to German translation[8]. I extended the HuggingFace transformers

[8] I use the `opus-mt-en-de` model, developed by the Language Technology Research Group at the University of Helsinki and trained on the OPUS training data `https://huggingface.co/Helsinki-NLP/opus-mt-en-de`

library[9] to support nearest neighbor decoding.

The retrieval index contains 8,706 sentence pairs. I computed 512-dimensional embeddings of each translation context in the retrieval index using the final hidden layer of the baseline model's decoder. This generated a datastore of 264,457 key-value pairs. Note that this is an order of magnitude smaller than the datastore used by Khandelwal et al. [247] for their medical domain-specific experiments, and 3-4 orders of magnitude smaller than their largest datastores. The size of the datastore was constrained by data availability, translation cost, and the need to reduce latency for interactive use. One of my contributions in this work is demonstrating the feasibility of these models in lower resource settings, in terms of data, storage, and/or compute resources.

Khandelwal et al. [247] use beam search to generate translations. I use greedy search for the sake of simplicity in implementation and interface design, which likely leads to lower translation performance. I leave extension to beam search to future work.

Evaluation

Khandelwal et al. [247] achieved a BLEU score of 54.35 on medical domain data with a datastore of 5.7 million key-value pairs, an improvement of 14.5 BLEU over their baseline model. Their findings clearly demonstrate that kNN-MT models have signficant potential to improve translation quality in specific domains. Here, I evaluate the model to validate my implementation and measure performance with a much smaller datastore.

I evaluated the model on the four data sources that are represented in the datastore: synthetic discharge instructions, online discharge instructions, synthetic discharge summaries, and medication documentation. I used the development sets to tune model hyperparameters and debug the retrieval index, and I report evaluation results on the test sets. I chose three different models to evaluate on the test set, all with $\lambda = 0.5$. The first model uses $k = 32$ neighbors at each step and softmax temperature $T = 1000$. The second model tests model performance with fewer neighbors, using $k = 10$ and $T = 1000$. The last model uses $k = 10$ and $T = 100$. I measured model performance using corpus BLEU [360] and ChrF [375] and used paired system-level resampling methods for significance testing [261].

On the two discharge instructions datasets, the 10-NN models outperform the baseline model by between 20-25 BLEU and 11-18 ChrF (Table 9.2). For the other datasets, the kNN models with $T = 1000$ performed comparably to the baseline model. This suggests that we can use a kNN model with a small number of neighbors and relatively small retrieval index on in-domain data with great advantage when the domain is very structured and repetitive, and without detriment to translation quality in more variable domains. The model with $k = 10$ and $T = 1000$ is the best performing model across domains, so I use this model as the default kNN model in the next section. Next, I explore how we can help users assess whether their inputs are likely to be translated well by the kNN-MT model.

[9] https://github.com/huggingface/transformers

	All	R-DCI	O-DCI	DCS	Meds
	(N=2,022)	(150)	(291)	(400)	(1,181)
Baseline	43.97	39.87	40.29	**41.71**	**45.71**
*k*NN-MT					
k=32, T=1000	**48.29**	56.41	**66.28**	**42.29**	***45.02***
k=10, T=1000	**48.34**	**60.30**	**65.76**	**42.34**	***44.91***
k=10, T=100	43.35	**58.64**	**66.19**	34.28	39.16

Table 9.2: Corpus BLEU score for the *k*NN-MT models and baseline translation model across the four domain-specific datasets in the retrieval index: real discharge instructions (R-DCI); online discharge instructions (O-DCI); synthetic discharge summaries (DCS); and medication documentation (Meds). The model with $k = 10$ and $T = 1000$ is most consistently high performing across datasets. Retrieval improves performance significantly for the two discharge instructions datasets, but does not improve performance over the baseline model for the other datasets. Bold: $p < 0.01$; bold italics: $0.01 < p < 0.05$.

Supporting Verification and Repair: Utilization and Relevance

My goal is to surface information to end users that would help them identify and recover from poor quality translations. In this section, I propose two sentence-level metrics that could help a user understand how the model is behaving on a specific input sentence: neighbor *utilization* and neighbor *relevance*. Intuitively, these metrics measure how much the retrieved neighbors are being used to predict the output tokens, and, when they are being used, how similar those neighbors are to the input sentence.

Using the same notation from the previous section, assume we have an input s and the model generates a translation $\hat{t} = (\hat{t}_1, ..., \hat{t}_M)$. Since we are using greedy search, each output token can be described by:

$$\hat{t}_i = \text{argmax } p(t_i | s, \hat{t}_{1:i-1}).$$

Recall that a nearest neighbor MT model generates output tokens $p(t_i|s, \hat{t}_{1:i-1})$ by interpolating between the baseline model's predictions $p_{\text{MT}}(t_i|s, \hat{t}_{1:i-1})$ and the nearest neighbor distribution $p_{\text{kNN}}(t_i|s, \hat{t}_{1:i-1})$. We define the "neighbor-voted" token $\hat{t}_{i,\text{kNN}}$ as:

$$\hat{t}_{i,\text{kNN}} = \text{argmax } p_{\text{kNN}}(t_i | s, \hat{t}_{1:i-1}).$$

For each token $\hat{t}_i$, let u_i be an indicator variable for whether the final output token agreed with the neighbor-voted token:

$$u_i = \mathbb{1}\left(\hat{t}_i = \hat{t}_{i,\text{kNN}}\right).$$

Then, we can define utilization (U) and relevance (R) for a source sentence, s, and translation hypothesis, $\hat{t} = (\hat{t}_1, ..., \hat{t}_M)$, based on $(u_1, ..., u_M)$ and the distance function d. Utilization is the proportion of output tokens that agreed with the neighbor-voted token:

$$\mathrm{U}(s, \hat{t}) = \frac{1}{M} \sum_{i=1}^{M} u_i. \tag{9.3}$$

Relevance is the average distance to the nearest utilized neighbor[10]:

$$\mathrm{R}(s, \hat{t}) = \frac{-\sum_{i=1}^{M} u_i \times \min_{(k_j, v_j) \in \mathcal{N}, \hat{t}_i = v_j} \left(d\left(k_j, f(s, \hat{t}_{1:i-1})\right)\right)}{\sum_{i=1}^{M} u_i}. \tag{9.4}$$

All of the information used to compute these metrics is generated as a byproduct of model inference.

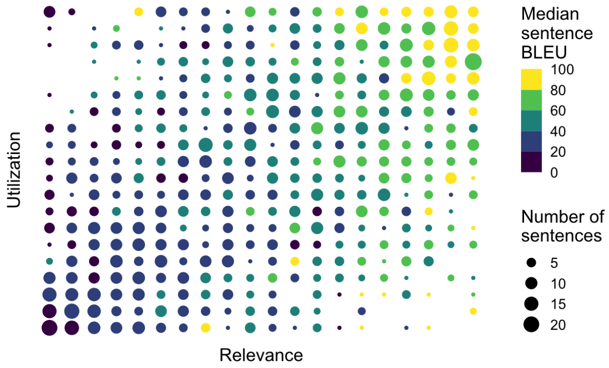


Figure 9.4: For translations generated using a k-Nearest Neighbor MT model [247], relevance (distance between an input sentence and the retrieval index) is correlated with translation quality, especially when utilization (how much the retrieval results are influencing the final translation) is high. Relevance and utilization are also correlated, because the model relies the retrieval results proportionally to their distance to the input.

Evaluation On the test set, I found that translation quality (measured using sentence BLEU and sentence ChrF) is correlated with relevance, especially when utilization is high (Figure 9.4). By definition, utilization and relevance are also highly correlated. The BLEU score for the kNN-MT model on sentences in the top 20% by relevance score (across all test sets) was 73.04 (compared to 48.34 over the entire test set). These sentences may be slightly easier than average, as the BLEU score for the baseline model was also higher (52.63 compared to 43.97), however the improvement is significantly larger with the kNN-MT model.

One way to interpret this finding is that when the model relies heavily on the neighbors to predict tokens (high utilization), the quality of the translation is heavily dependent on the relevance of those neighbors. When utilization is low, the relevance of the neighbors is less predictive of translation quality because they are not influencing the translation very much. Therefore, *we want a user to write sentences that are similar to the datastore*, which we can quantify as *high utilization*

[10] I chose this formulation to convey to a user how relevant the examples *that are being used* are. An alternative formulation could use the distances to the neighbors over all generation steps, which may be more informative when utilization is low.

and high relevance. Low utilization and relevance does not guarantee that a translation will be poor quality, but it is less likely that it will be consistent with the verified translations in the datastore.

In isolation, this approach differs little from other quality estimation methods for machine translation, which predict translation quality from source and hypothesis, and possibly model information (e.g. attention scores) [458]. However, the retrieval component of this approach allows us to provide more actionable information about how a user can change their input to increase the likelihood of a high quality translation (e.g., rephrase to be closer to the datastore). I discuss my approach to designing useful and actionable information using these metrics in Section 9.6.

Template approach

The second model matches input sentences directly to the discharge instruction templates that I developed in Section 9.4, and then fills the template with relevant terms in the target language. I hypothesized that this method would be less flexible than the *k*NN-MT approach, since it only works when there is a template that conveys the same meaning as the input sentence, but would provide better and more interpretable guarantees on output quality in those cases.

Given a source sentence, we generate a translation in three steps: first, we retrieve a template that conveys the same meaning as the source sentence; second, we extract entities from the source sentence to fill that template (e.g. symptoms, medications); third, we fill the template with the terms *in the target language* and edit it for fluency. In this section I describe and evaluate my approach to each of these steps.

Template retrieval

The first step is to retrieve a template that is similar to a source sentence. I started with a very simple retrieval model, ranking templates based on cosine similarity between high-dimensional sentence embeddings of the template and the source sentence. I used the `all-MiniLM-L6-v2` model from the HuggingFace `sentence-transformers` library[11] to compute 384-dimensional sentence embeddings. This model is specialized for tasks like semantic search and clustering.

Evaluation I evaluated the retrieval component on two data sources: the test set of 25 real discharge instructions that I used to evaluate the templates themselves in Section 9.4; and the synthetic discharge instructions test set of 150 sentences. I computed high-dimensional semantic sentence embeddings of the test set sentences and the 163 templates using the `all-MiniLM-L6-v2` model. For each sentence in the test set, I retrieved the top 10 templates, ranked by cosine similarity between sentence embeddings.

[11] `https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2`

In Section 9.4, I found that 60% of the sentences in the real discharge instructions test set have a matching template that conveys all or most of the meaning. For those sentences, the best available template was the retrieval model's top result 34% of the time, and in the top 10 70% of the time. The synthetic discharge instructions test set was constructed from the templates, and therefore every sentence in that set has a template that should convey close to all of the meaning. The top-1 template matched the original template 57% of the time (85 out of 150 sentences).

We cannot expect 163 templates to cover all of the things that a doctor might want to write in discharge instructions. Therefore, we would only want to suggest a template to a user when we think the top retrieval result is a good match for the input. To avoid suggesting irrelevant templates, I define a minimum threshold on cosine similarity for retrieval. With a higher cutoff, the model suggests fewer templates, but a higher proportion of the suggestions are good. A lower cutoff means that the model suggests templates more often, but a user would need to override more of those suggestions. With a threshold of 0.8, 10 sentences (5%) in the test set of 25 real discharge instructions have at least one similar template, and the top result is a good match for the input sentence for 7 of those 10. With a threshold of 0.6, 39% of sentences have a retrieval result and 44% of the top-1 retrievals are good. The choice of threshold in practice would depend on a user's tolerance for correcting model errors. I use a threshold of 0.6 in Cephalo.

Prior work on edit-based models have used a task-specific retrieval model, which has been specifically trained to retrieve inputs that are easiest for the editor model to edit to generate a good output [193]. In the case of machine translation, a simple and intuitive choice would be to use encoder hidden state representations from an NMT model, rather than a general purpose sentence embedding model. I leave exploration of optimal retrieval modules to future work.

Entity extraction

Once we have selected a template, the next step is to extract entities from the source sentence, which we will use to fill the template. I used GPT-3.5 Turbo to extract entities from the source sentence. The templates contain 13 different categories: "Medication," "Symptom," "Time," "N" (number), "Clinic," "Condition," "DEID" (deidentified information), "Test," "Units," "Frequency," "Treatment," "Office," and "Measurement." I found that attempting to extract all of these entity types from every sentence generally led to hallucination of entities that did not appear in the source sentence. Instead, I only prompted the model to extract the entities from the sentence that appear in the selected template.

The template filling step requires professional translations of all of these terms. For a real use case, I would integrate professionally translated knowledge bases, ontologies, and dictionaries, to provide a comprehensive termbase. For the sake of rapid prototyping, I developed an initial term base of 893 terms from the real discharge instructions dataset using a mix of Apache cTAKES™[12], regular expressions, and spaCy[13]. I pre-translated those terms using Google Translate. When a term is extracted in interactive use that is not in this termbase, we add it by calling the Google Translate API. In practice, we would need to warn a user that a specific term is not in the termbase.

Evaluation I extracted the relevant terms from each sentence in the synthetic discharge instructions test set, giving 303 terms. Of these, 138 were also extracted by the entity extraction module (46% accuracy). While this was sufficient for prototyping, future work could integrate custom clinical entity extraction modules to improve performance, for instance, a clinical text information extraction tool like Apache cTAKES™, or a clinical language model (e.g. [451, 270]).

Template filling

The final step is to take the selected template and extracted terms, each of which has a corresponding reference translation, and fill the template with the terms to generate a fluent output sentence in the target language. I used GPT-3.5 Turbo with a three-step prompt chain: first, we prompt the model to fill the template with the terms, given the input sentence (English), template (German) and terms (German); next, we ask the model whether the output from the first step is correct German grammar; finally, if the answer to the previous prompt is no, we prompt the model to correct the grammar[14]. I developed this prompting approach through iterative testing in several target languages (Spanish, French, Farsi, Chinese, and Malayalam) with speakers of those languages; this suggests my approach could likely generalize across target languages with further development.

Evaluation To evaluate the editor model, I use the synthetic discharge instructions generated from the templates, which by construction have a good match template. To isolate the performance of the edit step, I manually match these sentences to their corresponding template, and extract the correct terms. Then, I compared the output of the template filling prompt-chain on these templates and terms to the reference translation of each synthetic sentence using BLEU and ChrF.

The corpus BLEU score on the 150 test set sentences was 42.26 (compare to baseline model 39.87, and best kNN-MT model 66.28 (Table 9.2)) and ChrF was 68.48. In some cases, GPT-4 added additional content to the template during the synthetic sentence generation, and in other cases the reference translations for the synthetic sentences did not match the reference translation for the template. In either case, automatic metric are limited for measuring the performance of the template filling step. Excluding these left 120 sentences, on which the model had slightly better performance (BLEU: 44.46; ChrF: 70.13). Still, the model

[12] `https://ctakes.apache.org/`

[13] Full details and examples in Appendix B.1.

[14] Full prompts in Appendix B.2.

surprisingly underperforms compared to the *k*NN-MT models. I noticed that GPT-3.5 Turbo sometimes edited the translation quite significantly in the template filling process. While this does not necessarily mean the translations were *incorrect*, it does make it more difficult to support *verifiability*. Future work could build on a growing body of work on edit-based models [193, 309, 308, 310] to develop models that make the minimum possible edits to the template and term translations to produce fluent and adequate outputs.

Supporting Verification and Repair: Visualizing Template Filling

The template approach has three steps: we retrieve a relevant template, then extract terms to fill the template, then fill the template with the terms in the target language and edit for fluency. A major advantage of this approach is that the first two steps happen entirely in the source language, and are therefore directly verifiable and correctable by an end-user. A user can compare a template and extracted terms to the original input sentence, decide whether they convey the intended meaning, and if not, select alternative template(s) and term(s) if they exist. If the user finds that there are no relevant templates, they could switch to the more flexible *k*NN-MT model.

The design challenge for this approach is communicating to a user how well the edit model is performing to generate a consistent and fluent translation of the template, filled with the selected terms, in the target language. The system must be usable by someone with no target language fluency, so we cannot expect the user to be able to directly verify that the template filling has been done correctly. Pilot studies suggested that users try to assess how much the template has been edited to assess quality. To support this heuristic, I visualize the edits between the template translation and the filled translation using green for any added text that did not appear in the template and blue for the term translations.

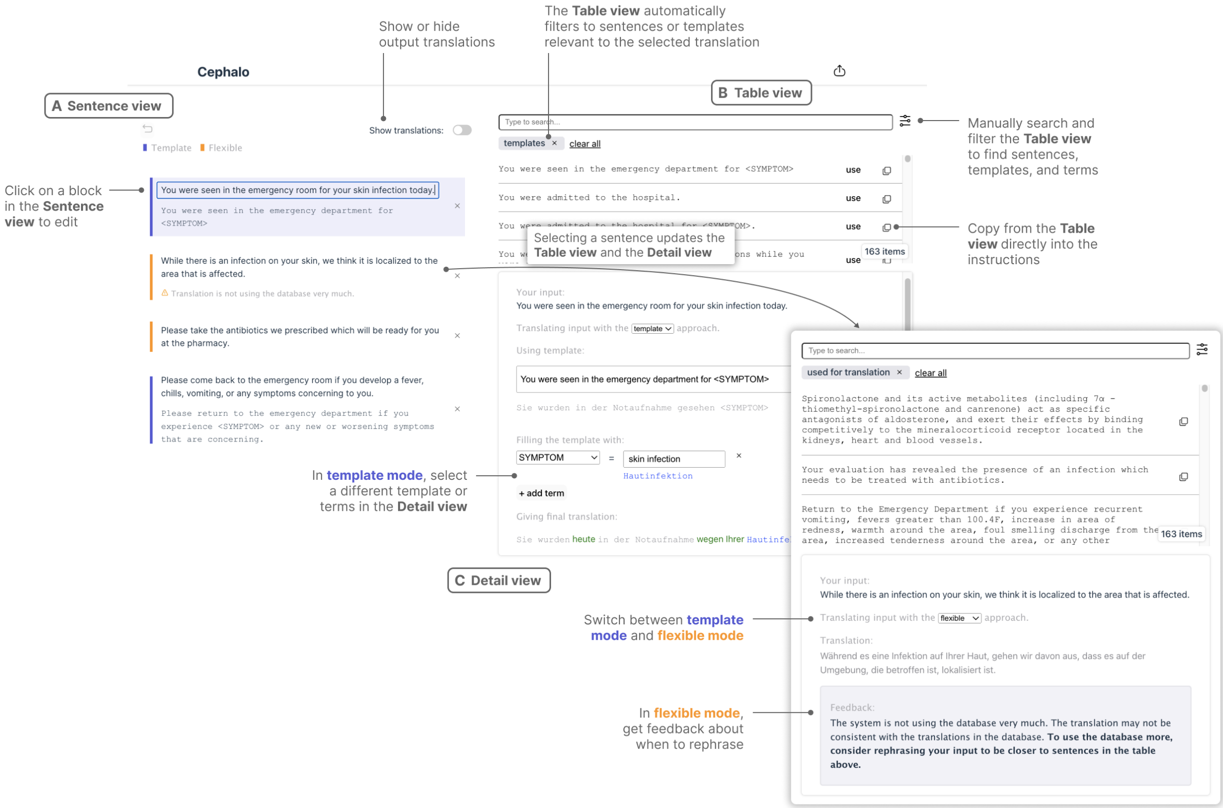
In the previous section, I explored how to help a user both identify failures and recover from failures with *k*NN-MT. With this model, I focus on identifying failures, with the assumption that a user would switch to *k*NN-MT as a recovery mechanism. Future work could look into more recoverable quality estimation approaches for edit-based models.

9.6 Cephalo

Cephalo is a system designed to help users craft appropriate inputs for MT. Cephalo does this by relying on the two retrieval-augmented translation models I introduced in the previous section: a nearest neighbor machine translation model [247], and the template approach. I describe Cephalo in this section, then evaluate it in a user study in the next section.

Cephalo Interface

Cephalo has three main interface components: the sentence view; the table view; and the detail view.



Show or hide output translations

The **Table view** automatically filters to sentences or templates relevant to the selected translation

Manually search and filter the **Table view** to find sentences, templates, and terms

Copy from the **Table view** directly into the instructions

Click on a block in the **Sentence view** to edit

Selecting a sentence updates the **Table view** and the **Detail view**

In **template mode**, select a different template or terms in the **Detail view**

Switch between **template mode** and **flexible mode**

In **flexible mode**, get feedback about when to rephrase

You were seen in the emergency room for your skin infection today.

You were seen in the emergency department for <SYMPTOM>

While there is an infection on your skin, we think it is localized to the area that is affected.

⚠ Translation is not using the database very much.

Please take the antibiotics we prescribed which will be ready for you at the pharmacy.

Please come back to the emergency room if you develop a fever, chills, vomiting, or any symptoms concerning to you.

Please return to the emergency department if you experience <SYMPTOM> or any new or worsening symptoms that are concerning.

```
You were seen in the emergency department for <SYMPTOM>
You were admitted to the hospital.
You were admitted to the hospital for <SYMPTOM>.
You we                                              ons while you
```

Your input:
You were seen in the emergency room for your skin infection today.

Translating input with the template approach.

Using template:
You were seen in the emergency department for <SYMPTOM>

Sie wurden in der Notaufnahme gesehen <SYMPTOM>

Filling the template with:
SYMPTOM = skin infection
Hautinfektion

Giving final translation:
Sie wurden heute in der Notaufnahme wegen Ihrer Hautinfe

```
Spironolactone and its active metabolites (including 7α -
thiomethyl-spironolactone and canrenone) act as specific
antagonists of aldosterone, and exert their effects by binding
competitively to the mineralocorticoid receptor located in the
kidneys, heart and blood vessels.

Your evaluation has revealed the presence of an infection which
needs to be treated with antibiotics.

Return to the Emergency Department if you experience recurrent
vomiting, fevers greater than 100.4F, increase in area of
redness, warmth around the area, foul smelling discharge from the
area, increased tenderness around the area, or any other
```

Your input:
While there is an infection on your skin, we think it is localized to the area that is affected.

Translating input with the flexible approach.

Translation:
Während es eine Infektion auf Ihrer Haut, gehen wir davon aus, dass es auf der Umgebung, die betroffen ist, lokalisiert ist.

Feedback:
The system is not using the database very much. The translation may not be consistent with the translations in the database. **To use the database more, consider rephrasing your input to be closer to sentences in the table above.**

Figure 9.5: CEPHALO has three main interface components: the sentence view (A); the table view (B); and the detail view (C). The sentence view shows each sentence in a document as an individual block, color-coded by the selected translation method. When a user clicks on a sentence block, they can edit the text, and view details about the translation in the table view and detail view. The table view lets users explore the retrieval index. The detail view shows details about the selected translation, and lets the user switch between translation methods. In template mode, the detail view lets a user select a template, add and remove terms, and view the final translation. In flexible mode, the detail view shows the final translation with actionable feedback about when to rephrase based on utilization and relevance scores.

Sentence view

The sentence view shows each sentence in a set of instructions as a block. The block is shown color coded based on which translation method is currently being used to translate that sentence. I call the kNN-MT approach the flexible approach in the interface, to reduce jargon and highlight its distinguishing advantage with respect to the template approach. Users can click on a sentence block to see detailed analysis of that sentence in the detail view, and templates or sentences used to translate that sentence in the table view. Users can edit sentence text and add sentence blocks at any position. A toggle above the sentences allows users to choose whether they would like to display the translation outputs; since users are *not* expected to read the target language, I make this optional and give any text in German low visual priority.

My goal is to guide users to improve their inputs. Since analysis is performed at a sentence level, I chose to design the sentence view in a block format to simplify implementation and focus users' attention on the sentence-level analysis. In a realistic use case, a more standard text editor would be a more appropriate input interface.

Table view

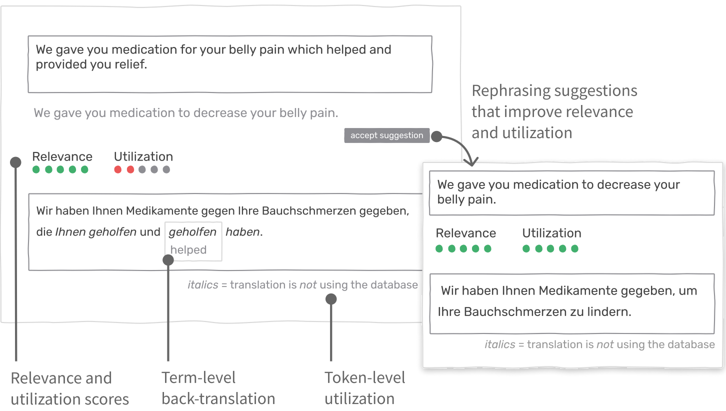
The table view allows users to explore the data in the retrieval index. Users can filter the data by type (sentence, term, or template), and a search bar supports full text search. Users can click on any row in the table to see the reference translation. A button in each row allows users to copy the English text from the table to the clipboard so they can add it to their input in the sentence view.

When a sentence is selected, the table shows data relevant to its translation. If the sentence is being translated with the template approach, the table shows all available templates, ranked by similarity to the selected source sentence. Users can click the "use" button next to any template in the table to override the suggested template. If the sentence is being translated with the flexible approach, the table shows all of the sentences retrieved as a nearest neighbor at any decoding step of the translation, ranked by how many times it was retrieved over the entire output translation. The "used for translation" filter is only available when a sentence is selected that is being translated with the flexible approach.

Detail view

The detail view shows more information about the translation of the selected sentence. When no sentence is selected, the detail view shows the number of sentences being translated with each method, and tells the user to select a sentence to see its analysis. When a sentence is selected, the detail view shows the source sentence, the hypothesis translation, and details specific to the method selected. Users can override the automatically suggested translation method using a dropdown in the detail view.

For sentences translated with a template, the detail view shows the source sentence, the selected template (and translation), the terms extracted (and translations), and the hypothesis translation (template filled with terms), color coded to show the results of the template filling step. The template is shown in a dropdown containing the top-10 templates, making it easy to override the defaul



We gave you medication for your belly pain which helped and provided you relief.

We gave you medication to decrease your belly pain.

Wir haben Ihnen Medikamente gegen Ihre Bauchschmerzen gegeben, die *Ihnen geholfen* und *geholfen* *haben*.

italics = translation is *not* using the database

We gave you medication to decrease your belly pain.

Wir haben Ihnen Medikamente gegeben, um Ihre Bauchschmerzen zu lindern.

italics = translation is *not* using the database

accept suggestion
Relevance
Utilization
helped
Relevance
Utilization
Rephrasing suggestions that improve relevance and utilization
Relevance and utilization scores
Term-level back-translation
Token-level utilization

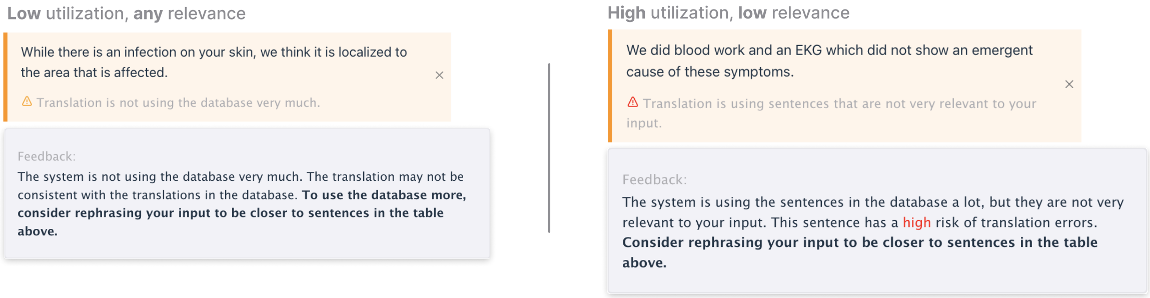
Figure 9.6: In early prototypes, I explored other information that we could show users in the flexible mode detail view. These included automatic rephrasing suggestions, sentence-level relevance and utilization scores, term-level back-translations, and token-level utilization scores (u_i).

For sentences translated with the flexible approach, the detail view shows the source sentence, the hypothesis translation, and actionable feedback based on the utilization and relevance scores for the translation. If utilization is low, then the system warns the user that the translation is not using the database very much, and suggests rephrasing (Figure 9.7). If utilization is high and relevance is low, the system warns the user that the translation is using sentences from the database that are not relevant to their input, and that the risk of translation errors is high. Again, the actionable suggestion is to rephrase the input sentence to be closer to the retrieval index. In each case, the user is given a short warning in the sentence view and a longer feedback message in the detail view. If relevance is moderate to high and utilization is moderate to high, then there is no warning in the sentence view, and the longer feedback tells the user: “the system is using the sentences in the database (a lot), and they are (very) relevant to your input. The translation is likely to be consistent with the translations in the database.” I defined low, moderate, and high utilization and relevance based on quantiles of these scores on the test set (low: 0-40th percentile; moderate: 40-60th percentile; high: 60-100th percentile).

The detail views are early prototypes designed to test what kinds of information could be useful for users to identify poor candidates for translation, and improve those inputs. During iterative prototyping, I tested showing quantitative sentence-level utilization and relevance scores, as well as token-level utilization (Figure 9.6), but found that it was difficult for people to interpret and act on this information. Future work could explore other kinds of information that could make the feedback more actionable or interpretable, e.g., automatic rephrasing suggestions or dictionary-based backtranslation (Figure 9.6).

Implementation

I built the front-end user interface using Svelte and hosted it as a static webpage on U.C. Berkeley enterprise GitHub pages. For the baseline neural MT model and the kNN-MT model, I ran model inference in Python 3 on Google Colab with a High-RAM T4 GPU. For



Low utilization, **any** relevance

While there is an infection on your skin, we think it is localized to the area that is affected.

⚠ Translation is not using the database very much.

Feedback:
The system is not using the database very much. The translation may not be consistent with the translations in the database. **To use the database more, consider rephrasing your input to be closer to sentences in the table above.**

High utilization, **low** relevance

We did blood work and an EKG which did not show an emergent cause of these symptoms.

⚠ Translation is using sentences that are not very relevant to your input.

Feedback:
The system is using the sentences in the database a lot, but they are not very relevant to your input. This sentence has a high risk of translation errors. **Consider rephrasing your input to be closer to sentences in the table above.**

Figure 9.7: Cephalo warns users when an input sentence has low utilization (high or low relevance), or high utilization and low relevance. A short warning appears in the sentence view (top) and actionable feedback is provided in the detail view (bottom).

the template approach, I queried GPT-3.5 Turbo through the OpenAI API for Python and translated terms that were not already in the termbase using the Google Cloud Translate API.

Before each user study session, I started a Flask server in the Colab notebook and forwarded requests from the front end user interface to the Flask server using an ngrok agent. The data and code is entirely open source,[15] allowing anyone to use the tool, given they have access to a Google Colab notebook with GPU[16], an ngrok account, an OpenAI account, and optionally a Google Cloud Project with the Cloud Translate API enabled.

9.7 User Study

Thus far, I have argued that retrieval-based methods can help users estimate translation quality and improve their inputs to a translation model because the retrieval index gives us a clear definition of a "good input" (inputs for which we can retrieve highly relevant examples). In this section, I evaluate these ideas in a user study with Cephalo.

Recruitment

I recruited 18 people through personal contacts and U.C. Berkeley Xlab. Participants were required to: be over the age of 18; be affiliated with U.C. Berkeley; have learned English as (one of) their first language(s); and not be able to read any German. In a pre-survey I asked people about their prior experience with machine translation, and their attitudes about the strengths and limitations of machine translation systems. In a post-survey, participants self-reported their age, gender, occupation, and area of study (if they were a student).

[15] Code will be open sourced upon publication.

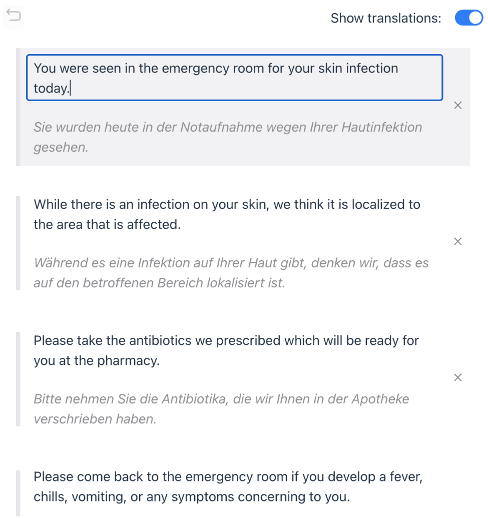
[16] GPUs are currently available on the Colab free tier, although without High-RAM and not guaranteed at all times.

Participants were between 18 and 27 years old (median: 22 years). 12 participants were female, 5 male or man, and 1 non-binary. 17 were students and 1 was a teacher. Six of the students were in a computer science or data science program.

The goal of this study is to gain insight into what kinds of information from retrieval-augmented translation models could be helpful for supporting translation verifiability. A limitation of this study is that participants are not clinicians, and the participant sample is heavily skewed towards undergraduate students at U.C. Berkeley. While I believe my findings still offer valuable insight to guide the design of verifiable translation tools, future work is necessary to apply my insights to design tools that could be used in clinical settings. I discuss these limitations and opportunities for future work in more detail in Section 9.8.

Study protocol

I conducted a within-subjects think-aloud user study to understand how people use Cephalo to craft inputs that are more appropriate for translation. Each participant used Cephalo to translate a set of fictional discharge instructions, written by physicians on the research team. Participants were asked to edit the instructions to improve the translations if and where they felt it was necessary.



You were seen in the emergency room for your skin infection today.

Sie wurden heute in der Notaufnahme wegen Ihrer Hautinfektion gesehen.

While there is an infection on your skin, we think it is localized to the area that is affected.

Während es eine Infektion auf Ihrer Haut gibt, denken wir, dass es auf den betroffenen Bereich lokalisiert ist.

Please take the antibiotics we prescribed which will be ready for you at the pharmacy.

Bitte nehmen Sie die Antibiotika, die wir Ihnen in der Apotheke verschrieben haben.

Please come back to the emergency room if you develop a fever, chills, vomiting, or any symptoms concerning to you.

Figure 9.8: In control mode, Cephalo shows only input sentences and output translations from the baseline MT model.

As a control, I showed a stripped back version of Cephalo, with only the sentence view (Figure 9.8). The translation model in the control condition was the baseline neural MT model. The control condition reflects the level of information provided by most MT products, input and translation only, while using the same interaction design as the full Cephalo interface.

Participants edited the same set of instructions twice: first in control mode, then using Cephalo. I randomized the set of instructions each participant saw to be one of three[17] options: one about a skin infection; one about fainting; and one about alcohol consumption. I did not randomize the order of the control and treatment conditions because I found during the pilots that learning effects were very strong after having seen Cephalo.

The first four sessions were pilot sessions to finalize the study design, and catch major bugs and usability issues. After the fourth session, I finalized the system design and study protocol. I include the pilot sessions in the qualitative analysis but not the quantitative

[17] Two physician authors of this paper wrote 20 fictional discharge instructions each. I ran each of these instructions through Cephalo and chose the three to use in the user study based on length (4-5 sentences because of time constraints), and a mix of sentences with and without template suggestions.

analyses. I stopped running sessions when I was no longer seeing new strategies for error identification and repair.

I began each session by introducing the system in the control mode, and explaining the task: edit the instructions to improve the translations, if necessary. When the participant said they had no more changes to make, I asked them which sentences they were most and least confident would be translated well and why.

Next, I introduced the interactive Cephalo interface, and asked them to repeat the task. I reset the instructions to the original text at this point. During the task, I helped participants if they needed help navigating the interface and understanding system output, but I did not give any guidance about what would improve the translations. When they finished, I again asked them which sentences they were most and least confident would be translated accurately and why. In a brief post-interview, I asked them how they approached these tasks overall, whether they had a preference between the template and flexible approaches, and whether the activity affected their perspective on automatic translation tools. Pilot sessions were approximately an hour, and the remaining sessions were 45 minutes.

Data Analysis

The study data included transcripts and usage logs from 18 study sessions. I analyzed the data using qualitative and quantitative methods to understand how people used Cephalo to identify and recover from errors in translations.

Qualitative data analysis

I recorded and transcribed each of the 18 sessions for qualitative analysis. During the task, I asked participants to think out loud and explain the edits they were making. I conducted two rounds of line-by-line coding on the transcripts, and memoing both between sessions and during analysis [318]. After the second round of coding, the code book contained eight high-level codes: factors (that make a sentence a good or bad candidate for translation), both generally and specific to the retrieval-augmented methods; verification (error identification strategies); revision (error repair strategies); source of beliefs (about translation difficulty); challenges; and give up/move on (without being totally confident in translation accuracy). Next, I combined second-level codes and drew connections between codes to organize findings around strategies for error identification and recovery in each condition.

Quantitative data analysis

The quantitative data includes time on task and participants' final edits, with corresponding model outputs. I only include quantitative data from P5-P18, as P1-P4 were pilot participants who each saw a slightly different version of the system and/or instructions.

I calculated time on task from the study transcripts. For each sentence in each condition, usage log data included the final edited source sentence, the selected translation method (template or flexible; treatment condition only), and the translation outputs for all three translation models (baseline, kNN-MT, and template). For sentences where users chose to use a template, I added their intended source sentence by filling their selected template with their selected terms in English.

My main hypothesis was that people would edit the instructions to be closer to the retrieval index when using Cephalo, but not when editing without guidance (control). I use the relevance score from the kNN-MT model to measure how close users' inputs were to the retrieval index. As a second outcome, I measured translation consistency with the translations in the retrieval index using the BLEU score relative to proxy reference translations from Google Translate[18] for the baseline and the kNN-MT models. I used resampling methods to estimate uncertainty in BLEU scores [261], and a linear mixed-effects model to estimate the effect of the treatment on similarity to index, with a random intercept and slope for each original source sentence that participants edited [72].[19] I also report descriptive statistics from the usage logs to quantify users' strategies when using Cephalo.

Results

Participants' goal in each condition was to improve the translation by editing the source and/or, in the treatment condition, choosing an appropriate translation method (template and terms, or flexible). Achieving this goal involves two main steps: first, the user has to assess whether and where there is potential for translation errors; then, once they have identified potential errors, they engage in repair to improve the likelihood of an accurate translation. Both of these steps involves uncertainty when users cannot read the translation to assess accuracy. My goal in developing Cephalo is to decrease the uncertainty in each of these steps.

In this section I describe participants' strategies in each of these statges: identifying errors; and recovering from errors. In the control condition, participants made edits according to theories about the strengths and limitations of machine translation that they had developed over past experiences with MT and translation in general. When using Cephalo, participants focused more on selecting appropriate templates and rephrasing their input to be closer to the data in the retrieval index, allowing them to generate translations that were more consistent with the proxy reference translations.

Identifying errors

Without any additional guidance, participants assessed translation quality either by reading the English and looking for features that may be difficult for a model to translate, and/or by looking at both the English input and the German translation. When using Cephalo, participants generally focused less on these strategies and more on assessing the appropriateness of suggested templates or the similarity of their input to sentences in the retrieval index.

[18] I computed reference translations via the Cloud Translate API in November 2023

[19] Also including a participant-level random effect led to singular fit.

Assessing input appropriateness for MT The most common strategy in the control condition was to read the English input sentences and identify features that might not be translated well by an MT model. Participants drew on their own theories about the strengths and limitations of MT models, built through their prior experiences with MT and language translation more broadly. I primed them to think about these theories in the pre-survey which they completed on their own at the start of the session.

Participants wanted sentences to be *"simple"* (12 participants), *"straightforward"* (7 participants), *"short"* (7 participants), *"clear"* (8 participants), and *"specific"* (5 participants). They expressed concern about words and phrases that they thought were colloquial, idiomatic, jargon, polysemous, complex, or uncommon.

These assessments were subjective, vague, and difficult to apply thoroughly and consistently. For example, many participants said they wanted to avoid idiomatic language, but it is easy for common metaphors and idioms to go unnoticed. For example, P12 suggested that the phrase, *"'feeling back to your normal self,' may not be translated well to different languages, because it's kind of idiomatic."* However, only one of the other four participants who saw this phrase chose to change it. Sometimes participants found it difficult to explain their edits, citing a *"gut feeling"* that something would be difficult to translate (P5).

Pattern matching inputs and translations 14 of the 18 participants also attempted to identify errors by inspecting the translation output, although only a few relied heavily on this approach. For example, participants looked across translations with a shared word in English to verify that the translations also shared a word in German. Another common approach was to look for expected patterns between the source and translation. For instance, participants could check that numbers in the input also appeared in the output, or that commas or other sentence structure looked similar. Many noticed cognates (e.g., infection and *infektion*, blood and *blut*), or knew a couple of German words: *"I know* bitte *is, please, that's all I know, and this* [Apotheke] *means pharmacy"* (P3). Some were guided by familiarity with other features of German, for example, P4 had *"heard in German that they have singular words for things that are really complex ideas. So I would expect this* [the German translation of emergency room] *to be multiple words,* [but] *sure, it could be one word."*

For high-stakes settings, these approaches may be risky. In many cases, participants would be comforted by their ability to verify one or two words, but this does not necessarily reflect the most important kinds of errors (e.g. negation errors or subject/object references). Participants were misled by apparent but not actual cognates (e.g. see and *sie* (she)), incorrect assumptions about word-level translations (e.g., that ER wasn't translated properly because there was no abbreviation in the output), or distracted by low priority issues, e.g. placement of punctuation.

Compared to standard MT tools, I lightly discouraged these practices by making it optional to display translations. Although participants were still inclined to look at the translations, some realized that they were unhelpful without knowing German:

> *"I guess I should be looking at the translation. Let's see. Not that I know German,*

> *but okay,* [seeing the German word *reaktion*] *a reaction in your body, I guess. Okay, no, this is not helpful to me."* (P4)

Future work could either further explore designs that discourage engagement with languages the user does not know, or consider interface elements that could support these processes more reliably (e.g., term-level back-translation or integrating quality estimation models). These strategies were likely more tempting for participants due to the similarity between English and German, and I expect these findings would have been very different if the languages had different writing systems.

Comparing inputs to the retrieval index To identify whether repair was necessary when using Cephalo, participants compared source sentences to the sentences and templates in the retrieval index. When there was a suggested template, participants compared the source sentence to the template and the extracted terms, and judged whether it was an appropriate alternative. When using the flexible approach, they looked at the nearest neighbors to decide whether they seemed similar to their input.

Engagement with the specific verification information I provided (utilization and relevance feedback in flexible mode, and color-coded filled template in template mode) was mixed and lower than expected. Six participants engaged with the utilization and relevance feedback to judge whether their inputs were good enough. Four participants engaged with the output of the template filling step, and tried to judge whether the changes made seemed appropriate.

While the retrieval component allows for some verification in English only, it still involves some uncertainty in the translation step. It was difficult for participants to know whether a kNN-MT translation or template filling was correct. For the template approach, they could verify that the terms appeared in the output (*"it translates stones in your gallbladder and just inserted that, so I'm pretty confident in that one,"* P2) and could judge whether the amount of added information seemed appropriate given the template and terms. However, they were still left uncertain about their assessment, particularly if there were multiple terms that needed to be added to the template:

> *"There's just a lot going on here. There are these 3 symptoms in addition to 'any new or worsening symptoms that are concerning.'* [There is] *a lot packed into this one phrase alone towards the end, which might be difficult to translate."* (P11)

Although there is no reason that the strategies used in the control condition could not also be applied when using Cephalo, even participants who very actively tried to inspect the translation for expected patterns and cognates in the control condition rarely applied the same strategies in the treatment condition.

Recovering from Errors

In the control condition, participants usually tried to rephrase in line with their theories about what would be translated well by an MT model. Using Cephalo, participants relied more on the retrieval index to guide rephrasing or rephrased by selecting a different template.

Crafting appropriate inputs for MT The only repair strategy available in the control condition was to change input sentences to be something the participants believed would be a better candidate for machine translation. Most participants only removed or added a few words, but they sometimes made more drastic changes like rephrasing the whole sentence, or splitting up a sentence into two or three. This process could be challenging: participants struggled to think of synonyms, and were generally anchored to the original sentence structure and phrasing.

Although the process of error identification and repair should be iterative in theory, it was difficult for participants to predict the impact of the changes they made, and time consuming to iterate. Ultimately, participants often gave up and moved on without being entirely confident that a translation would be correct. For example, when asked which sentence they were least confident would be translated well, P15 pointed to one where the wording seemed difficult to translate, but they were *"not really sure how else to edit."*

As expected, given that they had never seen the retrieval index, participants did not edit the instructions to be closer to the retrieval index in the control condition (Figure 9.10a). The kNN-MT model performed marginally better than the baseline model on the edited inputs in the control condition ($p = 0.038$[20]), but there was insignificant evidence that either model performed better on these edited inputs than on the original sentences[21] (Figure 9.10b).

Crafting inputs close to the retrieval index Participants' first strategy was usually to look for a template to match each input sentence. Participants almost universally preferred the template approach over the flexible approach (17 out of 18 participants). Participants found it easier to understand how the template approach worked and to *"know what should be right"* (P6). P15 said that the template approach, *"made me feel more secure, it's just a lot more structured."* 12 of the 13 sentences in the original instructions were translated using a template by at least one participant, and 6 were translated with the template approach by all of the participants. Overall, out of the 65 sentences that participants submitted in the treatment condition, 37 (57%) were translated using a template.

Of the 37 templates that were used in participants' final submissions, 27 were the top ranked (suggested) template. When participants did not like the suggested template, they usually looked for an alternative rather than immediately switching to the flexible method (Figure 9.9). Six of the other ten templates used were in the top 3 suggestions, two more were in the top 10, and two were ranked 21st and 28th. Participants were generally hesitant

[20] System-level comparison using paired resampling.

[21] Based on 95% confidence intervals from single-system resampling. We cannot conduct pairwise testing across conditions, because the sentences submitted in each condition are different.

to search for templates, meaning it is very important to ensure that the top suggestions are highly relevant. Sometimes, participants settled for templates they felt were sufficient but not ideal, even though they had the option of using the flexible approach which would not constrain their input: *"I'll just leave it like this because I don't think there's a better alternative"* (P17). Only one participant switched to flexible when the system originally suggested a template.

In flexible mode, if participants felt that the sentences in the table were dissimilar from their input, they would explore the table to find relevant sentences, then copy terms, phrases, or even entire sentences from the table into their input. Occasionally, participants looked for templates even when the system did not initially suggest one. Of the 33 sentences submitted for which the system originally suggested the flexible method, participants decided to use a template for 6 of them.

Similar to the control condition, few participants split complex sentences up to use multiple templates or sentences, despite this being an effective strategy since the templates tended to be short and simple. Six participants overall split any sentence into more than one. This happened more often in the control, with only three of them splitting any sentence into two or more in both the control and the treatment condition.

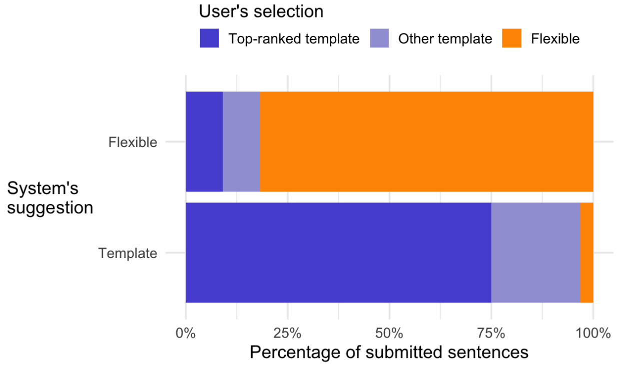


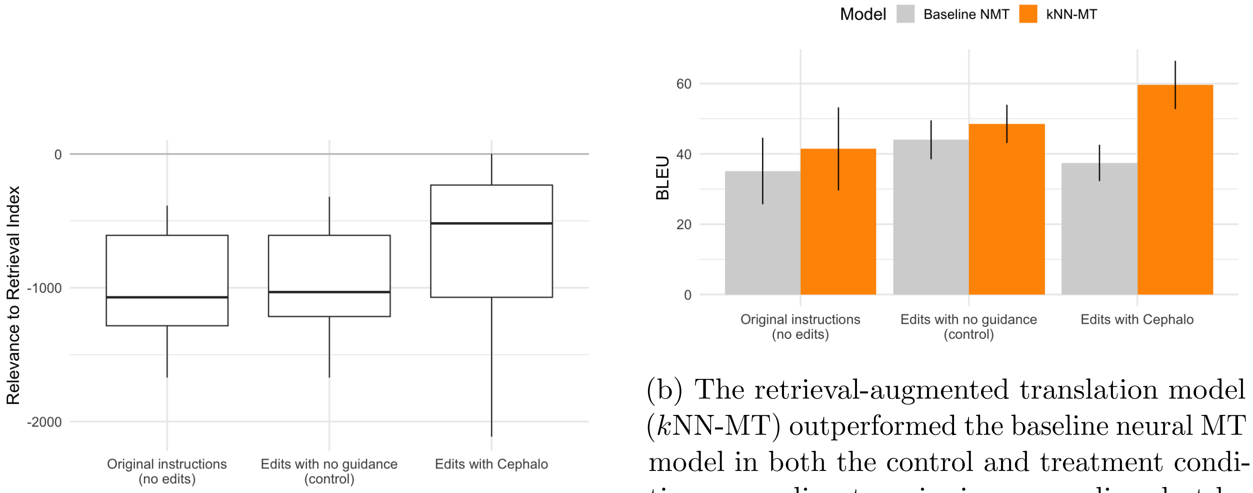
Figure 9.9: Participants usually used whatever translation method the system initially suggested. It was more common for users to use a template when the system did not suggest one, than for them to use the flexible approach when the system suggested a template. Users chose a different template than the original suggestion 27% of the time.

With these strategies, participants' edits when using CEPHALO were clearly informed by the retrieval index. On average, participants edited the instructions to be closer to the retrieval index when using CEPHALO ($p < 0.001$[22]) (Figure 9.10a). The kNN-MT model outperformed the baseline by a much larger margin in the treatment condition (22.17 BLEU score improvement; $p = 0.0080$[23]) than in the control (4.54 BLEU improvement; $p = 0.038$[24]) (Figure 9.10b). This suggests that CEPHALO helps users make edits that are specifically tailored to the strengths of the kNN-MT model. The BLEU score for the kNN-MT model was higher on the sentences edited using CEPHALO (59.60), than without guidance (48.54), but due to the small size of the corpora we cannot conclude this difference is significant at a 95% confidence level.

[22] A likelihood-ratio test indicated that a linear mixed-effects model including the experimental condition provided a better fit for the data than a model without it ($\chi^2(3) = 32.00, p < 0.001$).

[23] System-level comparison using paired resampling.

[24] System-level comparison using paired resampling.



(a) Editing the instructions with CEPHALO increased the relevance scores, compared to editing with no guidance.

(b) The retrieval-augmented translation model (kNN-MT) outperformed the baseline neural MT model in both the control and treatment condition, according to pairwise resampling, but by a significantly larger margin in the treatment condition. Error bars show 95% confidence intervals, computed using single system resampling.

Figure 9.10: CEPHALO helps users make edits that are specifically tailored to the strengths of the kNN-MT model.

It was cognitively intensive for participants to compare their source sentences against the large retrieval index. The average time on task was close to double when using CEPHALO (6.5 minutes control; 13.8 minutes treatment; p ¡ 0.001[25]). Participants spent most of their time in the table view, looking for sentences that were similar to the instructions. This was made more difficult by participants' lack of clinical domain knowledge: it was hard to judge what information was most critical, whether terms or phrases were equivalent, or to know what kinds of search terms might surface similar sentences.

Participants were unsure what they should do if there was nothing useful in the table. Although a reasonable approach would be to default to the editing strategies from the control condition, very few did this in practice. In a real application, this could be an opportunity to integrate a translator in the loop: if a user would like to say something that is not represented in the table, they could send that sentence to a professional translator and it could be added to the retrieval index for future use.

Design Implications

The user study surfaced several challenges that can inform future work designing user interfaces that leverage retrieval to help users craft good inputs for machine translation. Here, I discuss four design implications: efficient onboarding; high performing retrieval models; complementary editing strategies; and actionable feedback.

[25] Two-sided paired t-test for difference in means.

Provide efficient onboarding and prioritize learnability. When they were first introduced to Cephalo, participants needed help understanding exactly how each translation method worked, and keeping track of which translations were verified and which were from a model. Future work could improve this through better onboarding and clearer visual design, e.g., to distinguish professional translations from machine translations. While the retrieval index is large, I expect that if someone were to use the system consistently over time, they would begin to learn the contents and be able to write closer to the index without support. Future work should explore ways to support this kind of learnability.

Use high performing retrieval models and usable search interfaces. Participants spent most of their time in the treatment condition looking for relevant data to help them rephrase their inputs. Even with good onboarding and learnability, it will be important to make good suggestions and support efficient search. The retrieval model was very basic; improving this would likely significantly improve the user experience. Following established principles for search user interface design would also improve usability [197].

Make repair strategies consistent with users' existing mental models of machine translation. Sometimes participants' existing theories about machine translation conflicted with the best strategies for using Cephalo. For example, participants did not want to use a template that had complex sentence structure because they believed simplicity is important for accurate machine translation, even though the templates were pre-translated. Inevitably, the mental models that people have built through their experience with other kinds of MT systems will influence how they interpret any new system. Future work could explore ways to ensure that strategies are complementary, e.g., templates should be short and simple. This could also improve comprehensibility for patients beyond translation quality.

Ensure all feedback is actionable. In the flexible mode, the system suggested that users rephrase their input to be closer to the retrieval index. Participants tried to follow this guidance, but generally found it difficult to understand how their edits would affect the relevance score. Future work should make this feedback more easily actionable, e.g. by providing rephrasing suggestions. A useful direction for future exploration would be developing models and metrics that have an interpretable or intuitive definition of what makes two sentences similar, so that users can build simple mental models of what kinds of changes will influence the metrics.

9.8 Limitations and Opportunities for Future Work

The user study demonstrated that Cephalo can help users craft inputs that are specifically tailored to the strengths of retrieval-augmented machine translation. In this section, I discuss the limitations of my work and highlight promising directions for future research.

The user study is limited by the fact that the participants were not domain experts. My findings provide insight into the kinds of strategies users employ to improve translation quality and how retrieval can augment these strategies. However, I expect actual use to look different for experts. For example, clinicians would likely be able to make more informed decisions about what information is most important for a patient, and more quickly assess the appropriateness of a template or rephrasing. On the other hand, clinicians would have less time to spend verifying translations and editing inputs. Future work should focus on designing interactions that are reliable yet efficient for use under time-constraints.

Future evaluations should also include perspectives beyond healthcare providers, for example, patients and medical translators. It will be important to ensure that Cephalo fits into translators' workflow and respects their autonomy and expertise. Ultimately, the goal of improving translation quality in clinician-patient communication is to improve patients' clinical outcomes. Future work could conduct clinical trials to understand the extent to which improving translation quality can improve patient outcomes.

One benefit of my retrieval-based approach is that the retrieval index can be carefully designed to contain only content that is appropriate for the patient audience. For example, translations should be culturally appropriate, use plain language, and assume a basic level of health literacy. These benefits are not limited to cross-lingual communication. Researchers have been exploring ways to automatically give feedback to physicians to help them simplify their writing for patients [424]. Future work could extend my approach to this setting.

Finally, I recognize that my approach of increasing reliability through interface design could lead to dynamics where the responsibility for avoiding translation errors is unduly put on the individual user. My goal is to provide guidance that can help people use a system within its capabilities, rather than to offer a stopgap for unsafe systems.

9.9 Discussion

In this chapter, I have shown how retrieval can support new design patterns that increase verifiability for complex ML-driven systems when users cannot directly verify system outputs. I conclude by contextualizing this contribution in the broader space of designing for verifiability in ML-driven systems.

One important goal in the design of complex systems is to bridge the gulf of execution, i.e., make it easy for the user to understand what the system can do and how they can control its behavior, and the gulf of evaluation, i.e., make it easy for the user to evaluate whether the system did what they expected [217]. When the output of a system is difficult for a user to understand, it is extremely hard for them to evaluate outputs and verify that those outputs meet their needs (gulf of evaluation). ML-driven systems, particularly those with natural language interfaces, often provide little to no guidance as to how the system can be used, or what kinds of inputs are best supported (gulf of execution). This challenge is exacerbated when the output language is hard to understand, since users cannot build a mental model of the system's error boundaries over time through interaction with the model [32]. My work contributes one approach to building ML systems that help users understand exactly what a system can do, give them clear mechanisms to control it's behavior, and allow them to verify its outputs.

There are many other types of ML models that suffer similar challenges because their output language is difficult to understand and their error boundaries are stochastic. For example, code generation systems produce complex output that may look very close to a correct solution, but contain small bugs that can be difficult to find [504, 294]. Conversational agents built on top of large language models can produce confident sounding responses that are factually incorrect [73]. Automatic image captioning models provide descriptions of images for blind users, who cannot directly verify whether the caption is adequate [180]. In response, developers have warned useers to exercise caution and ensure human review when relying on model outputs in safety-critical settings [73]. In practice, exercising such caution is often time-consuming and expensive (e.g., fact-checking conversational agents [419], carefully debugging AI-generated code [504], or having professional translators post-edit machine translations [45]). In some cases expert human oversight is infeasible, for example, if a physician is using MT as a last-resort in a time-constrained setting [316]. This motivates alternative interaction techniques that support users to efficiently and independently verify model outputs.

At a high level, Cephalo addresses this problem by clearly scoping and describing the kinds of inputs that the system can support well. This allows users to directly verify whether their goals are supported by the system, and how they can express their intent within that supported scope of inputs. For instance, Cephalo users can explore the templates to find one that matches their intent.

This approach is generalizable to other domains where users have pragmatic goals and the space of inputs and outputs is relatively structured. For instance, researchers have used similar design patterns to improve natural language interfaces (NLIs) for data visualization. NLIs can make it easier and faster to generate visualizations, but it can be difficult for users to understand what these systems can do, and recover from misunderstandings [436, 444]. In response to these challenges, researchers have explored different ways of iteratively refining users' intent. A common pattern is to make a guess about a user's intent from a natural language query, then allow them to refine and correct that guess using direct manipulation [159, 252, 286, 509]. These approaches build a shared representation of the user's intent between the system and the user, that can be refined and then transformed deterministically into code (or another desired output language). Having a shared representation allows systems to show users alternative interpretations of their intent, giving them insight into system breakdowns and clear actions they can take towards repair [26].

In some contexts, systems must restrict their functionality to a fixed, but easily extensible, range of operations in order to build a shared specification. Cephalo constrains users' inputs more than an open-ended MT tool, but by doing this reduces uncertainty about the quality of the output translations. As another example, Promptiverse automatically generates scaffolding prompts for learners based on a knowledge graph representing content in a lecture video [284]. Compared to an open-ended conversational agent, Promptiverse guarantees that dialogue with learners will cover diverse topics and be grounded in pedagogical theory to improve learning outcomes. As in my work, there is a trade-off between the flexibility offered by the system, and the reliability and verifiability of its outputs. In high-stakes settings like healthcare and education, it may be better to produce no output than to produce an incorrect output that a user cannot recognize is incorrect. Ultimately, new design patterns focused on verifiability could enable end-users to have control over appropriate trade-offs [259].

9.10 Conclusion

In this chapter I introduced Cephalo, a tool for translating hospital discharge instructions that leverages retrieval to increase verifiability. Cephalo uses two approaches to translation: a template approach that matches open-ended input to a fixed set of pre-translated templates; and a flexible approach that translates any input but performs better on sentences close to a pre-translated retrieval index. In a user study with 18 participants, I found that users can verify whether their input is likely to be translated well without knowledge of the target language by directly inspecting the retrieval index, and use this insight to rephrase sentences that they think will not be well supported. My approach demonstrates a new way to use retrieval to bridge the gulfs of execution and evaluation in interfaces to ML systems.

Chapter 10

Discussion

This dissertation has explored design for reliability in the context of two different algorithmic systems: matching algorithms that assign students to schools; and machine translation models that translate between natural languages. Through these case studies, I have leveraged three different approaches to designing for reliability: teaching users what a system can and cannot do; aligning system evaluations with people's actual use cases, needs, and goals; and helping users recover from errors. I conclude by discussing three directives for designing reliable algorithmic systems, summarizing insights from my work and highlighting opportunities for future research.

Clearly define a system's purpose and capabilities

Several of the challenges I have studied, both in the student assignment context and the machine translation context, arise from a lack of clarity about what the system is supposed to achieve. In the student assignment context, school districts wanted to achieve collective-level goals, like diverse classrooms, but these goals conflicted with the objectives of the algorithms, which prioritized individual students' preferences. Machine translation systems have known strengths and limitations (e.g., some language pairs have much better performance than others, and general models can struggle with domain-specific translations), however, existing systems present users with a uniform and minimal interface that obscures these specifics.

My work has shown that designing for reliability is significantly more difficult if we must cater to an ambiguous, conflicting, or unspecified range of needs and goals. For instance, in Chapter 6, I argue that aggregating individual preferences is an insufficient mechanism to meet collective-level goals. Instead, collective deliberation of appropriate goals for the system should precede algorithm design. In Chapter 9, I suggest that building MT tools that have a more specific scope and range of capabilities could enable us to use these tools more reliably in high-stakes settings. This is in contrast to concurrent MT research and NLP research more broadly, where dominant trends are moving towards larger, multilingual, and even multi-task models for MT [477].

Throughout this dissertation I have argued that a deep, contextual understanding of how an algorithmic system will be used is critical to developing reliable approaches to designing those systems. This kind of deep understanding demands a specific, well-scoped task and usage context. While “general” purpose AI systems may offer useful new capabilities, achieving reliability will require carefully scoping specific applications of the technology.

Guide people towards well-supported use cases and inputs

Once we have clearly defined what a system is intended to do, we need to set people’s expectations accordingly. Users and those impacted by a system should understand what a system can do, and how well it can do it. While onboarding is one way to achieve this upfront with new users, my work has shown that this should be an ongoing and interactive process. Further, this understanding must be updated over time as systems change and improve.

Chapters 7 and 9 explore this idea in the most depth, in the context of machine translation. In Chapter 7 I describe some of the beliefs that MT users hold about the strengths and limitations of these systems. For instance, people say they would use simple language and avoid idioms and jargon when using MT. However, they struggle to abide by their own guidance when using MT for interpersonal communication. In Chapter 9 I show that this is still the case even when a user’s task is explicitly to edit text to make it a better candidate for machine translation. Therefore, I argue that users need interactive and real-time guidance to craft better inputs for MT. Google Translate already integrates spelling and grammar checking and recently added term disambiguation[1]. This could be extended to more model-specific interventions, e.g. to help users rephrase inputs to be closer to model training data. The core contribution of Chapter 9 is arguing that retrieval-augmented machine translation models are uniquely well-suited to this kind of guidance because they make it easier to clearly describe the range of supported inputs. I see promising future directions exploring other kinds of guidance that actively teaches people how to use an AI system as they are interacting with it. This kind of guidance may come at a cost to system flexibility, as it restricts or at least nudges a user to express themselves in a specific way. For example, one concern would be that people will shift how they communicate with others based on what an MT model is able to translate well. However, if a system has known strengths and limitations, teaching users how to use the system to its strengths arguably gives them more agency, by allowing them to make informed decisions about how to use the system, than allowing them to use the system however they would like without awareness of potential risks.

Develop intermediate intent specifications

In Chapter 9 I discussed how Cephalo bridges the gulfs of execution and evaluation by making it easier to describe what the system can do and verify that users’ inputs are within those capabilities [217]. Bridging these gulfs reduces the amount of effort required for a user to translate between how they conceive of their own goals and the system’s input and output

[1] `https://blog.google/products/search/google-search-generative-ai-international-expansion/`

formats. Often, algorithmic systems have seemingly “natural” or “obvious” input and output languages. For example, accepting natural language input and producing natural language output seems like an obvious choice for machine translation systems. Similarly, asking families for a ranked list of the available schools is a natural choice for student assignment. Part of my contribution has been questioning these assumptions, and demonstrating that rethinking input and output languages can be a flexible and powerful lever to increase reliability.

One way to think about more effective input and output languages for reliable algorithmic systems is in terms of developing a shared language for the system and the user to communicate and refine a shared understanding of the user’s goals [8]. When outputs are difficult for a user to directly verify (e.g., in a language they don’t know), this may require developing an intermediate intent specification that users can refine, and that a system can then transform into the desired output language with high certainty. For example, in Chapter 9 I use pre-translated templates as an intermediate intent specification for writing multilingual hospital discharge instructions. Since templates are available in both English and the target language, users can verify that the selected template conveys their intent by reading the English, then this template can be transformed to a fluent sentence in the output language with high-certainty using minimal edit-based models.

Developing an intermediate representation of user intent can also offer greater control over outputs for complex and nuanced tasks, since the intermediate representation can be made highly specific and granular. An example is multimodal search, especially in medicine, where users’ goals are highly domain-specific. For example, in contrast to familiar natural language interfaces for search, the SMILY system for similar image retrieval allows pathologists to define specific types of similarity that they are interested in, and refine search results accordingly [90]. Another system, Surch, enables users to search and compare surgical videos based on how individual steps are ordered and composed [251]. The system uses a graph-based representation to enable a shared understanding of video content between the system and the user. This breadth of applications demonstrates the promise of intent-matching approaches for reliable, verifiable, and controllable human-AI interaction.

10.1 Conclusion

This dissertation has explored how we can design more reliable interactions with complex, unpredictable algorithmic systems. I examine this task in two uniquely challenging domains: student assignment algorithms, where stakeholders have conflicting individual and collective goals; and machine translation models, where output quality is highly contextual and nuanced, and users often lack the language abilities to directly verify outputs. By making it easier for users to understand what a system can and cannot do, what kinds of inputs the system supports, and how they can verify that the system has understood their intent, we can increase their agency to appropriately rely on algorithmic systems in high-stakes settings.

Bibliography

[1] Atila Abdulkadiroglu et al. *Do Parents Value School Effectiveness?* Working Paper 23912. National Bureau of Economic Research, Oct. 2017. DOI: 10.3386/w23912. URL: http://www.nber.org/papers/w23912.

[2] Atila Abdulkadiroğlu and Tayfun Sönmez. “School Choice: A Mechanism Design Approach”. In: *American Economic Review* 93.3 (June 2003), pp. 729–747. DOI: 10.1257/000282803322157061. URL: https://www.aeaweb.org/articles?id=10.1257/000282803322157061.

[3] Rediet Abebe et al. “Adversarial Scrutiny of Evidentiary Statistical Software”. In: *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*. FAccT ’22. Seoul, Republic of Korea: Association for Computing Machinery, 2022, pp. 1733–1746. ISBN: 9781450393522. DOI: 10.1145/3531146.3533228. URL: https://doi.org/10.1145/3531146.3533228.

[4] Rediet Abebe et al. “Roles for Computing in Social Change”. In: *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. FAT* ’20. Barcelona, Spain: Association for Computing Machinery, 2020, pp. 252–260. ISBN: 9781450369367. DOI: 10.1145/3351095.3372871. URL: https://doi.org/10.1145/3351095.3372871.

[5] Hammaad Adam et al. “Mitigating the impact of biased artificial intelligence in emergency decision-making”. In: *Communications Medicine* 2.1 (2022), p. 149.

[6] African American Parent Advisory Council. *AAPAC Reflections on SFUSD’s Student Assignment Policy*. 2017. URL: https://archive.sfusd.edu/en/assets/sfusd-staff/enroll/files/AAPAC_Student_Assignment_Presentation_3.8.17.pdf?_ga=2.81786516.225077105.1586807928-1027271765.1579115407.

[7] Chirag Agarwal et al. “Openxai: Towards a transparent evaluation of model explanations”. In: *Advances in Neural Information Processing Systems* 35 (2022), pp. 15784–15799.

[8] Maneesh Agrawala. “Unpredictable Black Boxes are Terrible Interfaces”. In: (Mar. 2023). URL: https://magrawala.substack.com/p/unpredictable-black-boxes-are-terrible.

[9] Roee Aharoni and Yoav Goldberg. “Unsupervised Domain Clusters in Pretrained Language Models”. In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, July 2020, pp. 7747–7763. DOI: `10.18653/v1/2020.acl-main.692`. URL: `https://aclanthology.org/2020.acl-main.692`.

[10] Allison Woodruff et al. “A Qualitative Exploration of Perceptions of Algorithmic Fairness”. In: *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems - CHI ’18*. ACM Press, 2018.

[11] Haiyi Zhu et al. “Value-Sensitive Algorithm Design”. In: *Proceedings of the ACM on Human-Computer Interaction* 2.CSCW (Nov. 2018), pp. 1–23.

[12] Yannai A. Gonczarowski et al. “Matching for the Israeli ”Mechinot” Gap-Year Programs: Handling Rich Diversity Requirements”. In: *EC ’19*. ACM Press, 2019.

[13] Joshua Albrecht, Rebecca Hwa, and G. Elisabeta Marai. “The Chinese Room: Visualization and Interaction to Understand and Correct Ambiguous Machine Translation”. In: *Computer Graphics Forum* 28 (June 2009), pp. 1047–1054. DOI: `10.1111/j.1467-8659.2009.01443.x`.

[14] Muhammad Ali et al. “Discrimination through Optimization: How Facebook’s Ad Delivery Can Lead to Biased Outcomes”. In: *Proc. ACM Hum.-Comput. Interact.* 3.CSCW (Nov. 2019). DOI: `10.1145/3359301`. URL: `https://doi.org/10.1145/3359301`.

[15] Oscar Alvarado and Annika Waern. “Towards Algorithmic Experience: Initial Efforts for Social Media Contexts”. In: *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. New York, NY, USA: Association for Computing Machinery, 2018, pp. 1–12. ISBN: 9781450356206.

[16] Maureen L. Ambrose and Anke Arnaud. “Are procedural justice and distributive justice conceptually distinct?” In: *Handbook of Organizational Justice*. Ed. by Jerald Greenberg and Jason A. Colquitt. Lawrence Erlbaum Associates Publishers, 2005, pp. 59–84.

[17] Saleema Amershi et al. “Guidelines for Human-AI Interaction”. In: *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. CHI ’19. Glasgow, Scotland Uk: Association for Computing Machinery, 2019, pp. 1–13. ISBN: 9781450359702. DOI: `10.1145/3290605.3300233`. URL: `https://doi.org/10.1145/3290605.3300233`.

[18] Saleema Amershi et al. “Power to the people: The role of humans in interactive machine learning”. In: *Ai Magazine* 35.4 (2014), pp. 105–120.

[19] Saleema Amershi et al. “Software Engineering for Machine Learning: A Case Study”. In: *2019 IEEE/ACM 41st International Conference on Software Engineering: Software Engineering in Practice (ICSE-SEIP)*. Montreal, QC, Canada: IEEE, 2019, pp. 291–300. DOI: `10.1109/ICSE-SEIP.2019.00042`.

[20] Morgan G. Ames. "Learning consumption: Media, literacy, and the legacy of One Laptop per Child". In: *The Information Society* 32.2 (2016), pp. 85–97. DOI: 10.1080/01972243.2016.1130497. URL: https://doi.org/10.1080/01972243.2016.1130497.

[21] Tahir Andrabi, Jishnu Das, and Asim Ijaz Khwaja. "Report Cards: The Impact of Providing School and Child Test Scores on Educational Markets". In: *American Economic Review* 107.6 (June 2017), pp. 1535–63. DOI: 10.1257/aer.20140774. URL: https://www.aeaweb.org/articles?id=10.1257/aer.20140774.

[22] Julia Angwin et al. "Machine Bias". In: *ProPublica* (May 2016).

[23] Ben Casselman Anna Maria Barry-Jester and Dana Goldstein. "The New Science of Sentencing". In: (2015). URL: https://www.themarshallproject.org/2015/08/04/the-new-science-of-sentencing.

[24] Paul M. Aoki et al. "A Vehicle for Research: Using Street Sweepers to Explore the Landscape of Environmental Community Action". In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. CHI '09. Boston, MA, USA: Association for Computing Machinery, 2009, pp. 375–384. ISBN: 9781605582467. DOI: 10.1145/1518701.1518762. URL: https://doi.org/10.1145/1518701.1518762.

[25] Imanol Arrieta-Ibarra et al. "Should We Treat Data as Labor? Moving beyond "Free"". In: *AEA Papers and Proceedings* 108 (May 2018), pp. 38–42. DOI: 10.1257/pandp.20181003.

[26] Zahra Ashktorab et al. "Resilient Chatbots: Repair Strategy Preferences for Conversational Breakdowns". en. In: *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems - CHI '19*. Glasgow, Scotland Uk: ACM Press, 2019, pp. 1–12. ISBN: 978-1-4503-5970-2. DOI: 10.1145/3290605.3300484. URL: http://dl.acm.org/citation.cfm?doid=3290605.3300484 (visited on 10/20/2020).

[27] A. Avizienis et al. "Basic concepts and taxonomy of dependable and secure computing". In: *IEEE Transactions on Dependable and Secure Computing* 1.1 (2004), pp. 11–33. DOI: 10.1109/TDSC.2004.2.

[28] Eleftherios Avramidis et al. "Linguistic Evaluation of German-English Machine Translation Using a Test Suite". In: *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*. Florence, Italy: Association for Computational Linguistics, 2019, pp. 445–454. DOI: 10.18653/v1/w19-5351.

[29] Edmond Awad et al. "The Moral Machine experiment". In: *Nature* 563 (2018), pp. 59–64. DOI: 10.1038/s41586-018-0637-6.

[30] Mahsa Baktashmotlagh et al. "Distribution-Matching Embedding for Visual Domain Adaptation". In: *Journal of Machine Learning Research* 17.108 (2016), pp. 1–30. URL: http://jmlr.org/papers/v17/15-207.html.

[31] Satanjeev Banerjee and Alon Lavie. "METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments". In: *Proceedings of the Acl Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*. Ann Arbor, Michigan: Association for Computational Linguistics, 2005, pp. 65–72.

[32] Gagan Bansal et al. "Beyond accuracy: The role of mental models in human-AI team performance". In: *Proceedings of the AAAI conference on human computation and crowdsourcing*. Vol. 7. 2019, pp. 2–11.

[33] Shaowen Bardzell and Jeffrey Bardzell. "Towards a Feminist HCI Methodology: Social Science, Feminism, and HCI". In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. CHI '11. Vancouver, BC, Canada: Association for Computing Machinery, 2011, pp. 675–684. ISBN: 9781450302289. DOI: 10.1145/1978942.1979041. URL: https://doi.org/10.1145/1978942.1979041.

[34] Matt Barnum and Gabrielle LaMarr LeMee. *Looking for a home? You've seen GreatSchools ratings. Here's how they nudge families toward schools with fewer black and Hispanic students*. Dec. 2019. URL: https://www.chalkbeat.org/2019/12/5/21121858/looking-for-a-home-you-ve-seen-greatschools-ratings-here-s-how-they-nudge-families-toward-schools-wi.

[35] Solon Barocas et al. "Designing Disaggregated Evaluations of AI Systems: Choices, Considerations, and Tradeoffs". In: *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*. AIES '21. Virtual Event, USA: Association for Computing Machinery, 2021, pp. 368–378. ISBN: 9781450384735. DOI: 10.1145/3461702.3462610. URL: https://doi.org/10.1145/3461702.3462610.

[36] Jeanne Batalova and Jie Zong. *Language Diversity and English Proficiency in the United States*. Migration Policy Institute. Nov. 2016. URL: https://www.migrationpolicy.org/article/language-diversity-and-english-proficiency-united-states-2015.

[37] Eric P.S. Baumer and M. Six Silberman. "When the Implication is Not to Design (Technology)". In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. CHI '11. Vancouver, BC, Canada: Association for Computing Machinery, 2011, pp. 2271–2274. ISBN: 9781450302289. DOI: 10.1145/1978942.1979275. URL: https://doi.org/10.1145/1978942.1979275.

[38] Eric PS Baumer. "Toward human-centered algorithm design". In: *Big Data & Society* 4.2 (2017), pp. 1–12. DOI: 10.1177/2053951717718854. URL: https://doi.org/10.1177/2053951717718854.

[39] Nicole Baumgarten and Inke Du Bois. "Linguistic discrimination and cultural diversity in social spaces". In: *Journal of Language and Discrimination* 3.2 (2019), pp. 85–91.

[40] Emma Beauxis-Aussalet et al. “The Role of Interactive Visualization in Fostering Trust in AI”. In: *IEEE Computer Graphics and Applications* 41 (Nov. 2021), pp. 7–12. DOI: `10.1109/mcg.2021.3107875`.

[41] Emma Beede et al. “A Human-Centered Evaluation of a Deep Learning System Deployed in Clinics for the Detection of Diabetic Retinopathy”. In: *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. CHI ’20. Honolulu, HI, USA: Association for Computing Machinery, 2020, pp. 1–12. ISBN: 9781450367080. DOI: `10.1145/3313831.3376718`. URL: `https://doi.org/10.1145/3313831.3376718`.

[42] Gerdus Benade et al. “Preference Elicitation for Participatory Budgeting”. In: *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*. AAAI’17. San Francisco, California, USA: AAAI Press, 2017, pp. 376–382.

[43] Emily M. Bender et al. “On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?” In: *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. FAccT ’21. Virtual Event, Canada: Association for Computing Machinery, 2021, pp. 610–623. ISBN: 9781450383097. DOI: `10.1145/3442188.3445922`. URL: `https://doi.org/10.1145/3442188.3445922`.

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

[45] L. Bentivogli et al. “Machine Translation Human Evaluation: an investigation of evaluation based on Post-Editing and its relation with Direct Assessment”. In: *Proceedings of International Conference on Spoken Language Translation*. IWSLT ’18. 2018, pp. 62–69.

[46] Luisa Bentivogli et al. “Neural versus Phrase-Based Machine Translation Quality: a Case Study”. In: *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. Austin, Texas: Association for Computational Linguistics, Nov. 2016, pp. 257–267. DOI: `10.18653/v1/D16-1025`. URL: `https://aclanthology.org/D16-1025`.

[47] Jon Louis Bentley. “Multidimensional Binary Search Trees Used for Associative Searching”. In: *Communications of the ACM* 18 (1975). DOI: `10.1145/361002.361007`. URL: `https://dl.acm.org/doi/10.1145/361002.361007` (visited on 11/17/2022).

[48] Yotam Berger. “Israel Arrests Palestinian Because Facebook Translated ’Good Morning’ to ’Attack Them’”. In: *Haaretz* (Oct. 2017). URL: `https://www.haaretz.com/israel-news/palestinian-arrested-over-mistranslated-good-morning-facebook-post-1.5459427`.

[49] Peter Bergman, Eric Chan, and Adam Kapor. *Housing Search Frictions: Evidence from Detailed Search Data and a Field Experiment*. CESifo Working Paper No. 8080. 2020. URL: `https://ssrn.com/abstract=3535290`.

[50] Hugh Beyer and Karen Holtzblatt. *Contextual Design: Defining Customer-Centered Systems.* San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1997. ISBN: 9780080503042.

[51] Shaily Bhatt et al. "A Case Study of Efficacy and Challenges in Practical Human-in-Loop Evaluation of NLP Systems Using Checklist". In: *Workshop on Human Evaluation of NLP Systems.* Apr. 2021.

[52] Umang Bhatt et al. "Explainable Machine Learning in Deployment". In: *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency.* FAT* '20. Barcelona, Spain: Association for Computing Machinery, 2020, pp. 648–657. ISBN: 9781450369367. DOI: `10.1145/3351095.3375624`. URL: `https://doi.org/10.1145/3351095.3375624`.

[53] Steven Bird, Ewan Klein, and Edward Loper. *Natural Language Processing with Python.* O'Reilly Media Inc., 2009. URL: `https://www.nltk.org/book/`.

[54] Abeba Birhane et al. "Power to the People? Opportunities and Challenges for Participatory AI". In: *Equity and Access in Algorithms, Mechanisms, and Optimization.* EAAMO '22. Arlington, VA, USA: Association for Computing Machinery, 2022. ISBN: 9781450394772. DOI: `10.1145/3551624.3555290`. URL: `https://doi.org/10.1145/3551624.3555290`.

[55] Sandra E. Black. "Do Better Schools Matter? Parental Valuation of Elementary Education". In: *The Quarterly Journal of Economics* 114.2 (1999), pp. 577–599. ISSN: 00335533, 15314650. URL: `http://www.jstor.org/stable/2587017`.

[56] John Blatz et al. "Confidence Estimation for Machine Translation". In: *COLING 2004: Proceedings of the 20th International Conference on Computational Linguistics.* Geneva, Switzerland, Aug. 2004, pp. 315–321. URL: `https://www.aclweb.org/anthology/C04-1046` (visited on 01/17/2021).

[57] Su Lin Blodgett, Johnny Wei, and Brendan O'Connor. "A Dataset and Classifier for Recognizing Social Media English". In: *Proceedings of the 3rd Workshop on Noisy User-generated Text.* Copenhagen, Denmark: Association for Computational Linguistics, Sept. 2017, pp. 56–61. DOI: `10.18653/v1/W17-4408`. URL: `https://aclanthology.org/W17-4408`.

[58] Tolga Bolukbasi et al. "Man Is to Computer Programmer as Woman Is to Homemaker? Debiasing Word Embeddings". In: *NeurIPS.* Vol. 29. 2016.

[59] Tolga Bolukbasi et al. "Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings". In: *Advances in Neural Information Processing Systems 29.* Ed. by D. D. Lee et al. Curran Associates, Inc., 2016, pp. 4349–4357. URL: `http://papers.nips.cc/paper/6228-man-is-to-computer-programmer-as-woman-is-to-homemaker-debiasing-word-embeddings.pdf`.

[60] Alan Borning and Michael Muller. "Next Steps for Value Sensitive Design". In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. CHI '12. Austin, Texas, USA: Association for Computing Machinery, 2012, pp. 1125–1134. ISBN: 9781450310154. DOI: 10.1145/2207676.2208560. URL: https://doi.org/10.1145/2207676.2208560.

[61] M. Bostock, V. Ogievetsky, and J. Heer. "D^3 Data-Driven Documents". In: *IEEE TVCG* 17 (Dec. 2011). DOI: 10.1109/tvcg.2011.185.

[62] Richard Bowe, Sharon Gewirtz, and Stephen J. Ball. "Captured by the Discourse? Issues and concerns in researching 'parental choice'". In: *British Journal of Sociology of Education* 15.1 (1994), pp. 63–78. DOI: 10.1080/0142569940150104.

[63] G. C. Bowker et al. "Layers of Silence, Arenas of Voice: The Ecology of Visible and Invisible Work". In: *Boundary Objects and Beyond: Working with Leigh Star*. 2016, pp. 351–373.

[64] G. E. P. Box. "Robustness in the Strategy of Scientific Model Building". In: *Robustness in Statistics*. Ed. by Robert L. Launder and Graham N. Wilkinson. Cambridge, MA, USA: Academic Press, 1979, pp. 201–236. ISBN: 978-0-12-438150-6. DOI: 10.1016/B978-0-12-438150-6.50018-2. URL: http://www.sciencedirect.com/science/article/pii/B9780124381506500182.

[65] Robert J. Tibshirani Bradley Efron. *An Introduction to the Bootstrap*. Chapman & Hall/CRC, 1993.

[66] Felix Brandt et al. *Handbook of Computational Social Choice*. 1st. USA: Cambridge University Press, 2016. ISBN: 1107060435.

[67] Holly P. Branigan et al. "Linguistic alignment between people and computers". In: *Journal of Pragmatics* 42 (2010), pp. 2355–2368.

[68] Susan E. Brennan. "The Grounding Problem in Conversations With and Through Computers". In: *Social and cognitive psychological approaches to interpersonal communication*. Hillsdale, NJ: Lawrence Erlbaum, 1998, pp. 201–225.

[69] Susan E. Brennan and Justina O. Ohaeri. "Effects of Message Style on Users' Attributions toward Agents". In: *Conference Companion on Human Factors in Computing Systems*. CHI '94. Boston, Massachusetts, USA: Association for Computing Machinery, 1994, pp. 281–282. ISBN: 0897916514. DOI: 10.1145/259963.260492. URL: https://doi.org/10.1145/259963.260492.

[70] Fiona Brooks and Sally Kendall. "Making sense of assets: what can an assets based approach offer public health?" In: *Critical Public Health* 23.2 (2013), pp. 127–130. DOI: 10.1080/09581596.2013.783687. eprint: https://doi.org/10.1080/09581596.2013.783687. URL: https://doi.org/10.1080/09581596.2013.783687.

[71] Anna Brown et al. “Toward Algorithmic Accountability in Public Services: A Qualitative Study of Affected Community Perspectives on Algorithmic Decision-Making in Child Welfare Services”. In: *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. CHI ’19. Glasgow, Scotland Uk: Association for Computing Machinery, 2019. ISBN: 9781450359702. DOI: `10.1145/3290605.3300271`. URL: `https://doi.org/10.1145/3290605.3300271`.

[72] Violet A. Brown. “An Introduction to Linear Mixed-Effects Modeling in R”. In: *Advances in Methods and Practices in Psychological Science* 4.1 (2021). DOI: `10.1177/2515245920960351`.

[73] Sébastien Bubeck et al. *Sparks of Artificial General Intelligence: Early experiments with GPT-4*. 2023. arXiv: `2303.12712 [cs.CL]`.

[74] Zana Buçinca, Maja Barbara Malaya, and Krzysztof Z Gajos. “To trust or to think: cognitive forcing functions can reduce overreliance on AI in AI-assisted decision-making”. In: *Proceedings of the ACM on Human-Computer Interaction* 5.CSCW1 (2021), pp. 1–21.

[75] Zana Buçinca, Maja Barbara Malaya, and Krzysztof Z. Gajos. “To Trust or to Think: Cognitive Forcing Functions Can Reduce Overreliance on AI in AI-Assisted Decision-Making”. In: *Proc. ACM Hum.-Comput. Interact.* 5.CSCW1 (Apr. 2021). DOI: `10.1145/3449287`. URL: `https://doi.org/10.1145/3449287`.

[76] Joy Buolamwini and Timnit Gebru. “Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification”. In: *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*. Ed. by Sorelle A. Friedler and Christo Wilson. Vol. 81. Proceedings of Machine Learning Research. PMLR, Feb. 2018, pp. 77–91. URL: `https://proceedings.mlr.press/v81/buolamwini18a.html`.

[77] Aljoscha Burchardt et al. “A Linguistic Evaluation of Rule-Based, Phrase-Based, and Neural MT Engines”. In: *The Prague Bulletin of Mathematical Linguistics* 108 (2017). DOI: `10.1515/pralin-2017-0017`.

[78] Simon Burgess et al. “Parental choice of primary school in England: what ’type’ of school do parents choose?” In: *The Centre for Market and Public Organisation* (2009).

[79] Simon Burgess et al. “What Parents Want: School Preferences and School Choice”. In: *The Economic Journal* 125.587 (2015), pp. 1262–1289. DOI: `10.1111/ecoj.12153`. URL: `https://onlinelibrary.wiley.com/doi/abs/10.1111/ecoj.12153`.

[80] Robin Burke et al. “Algorithmic Fairness, Institutional Logics, and Social Choice”. In: *AI for Social Good Workshop* (2020).

[81] Franck Burlot and François Yvon. “Evaluating the Morphological Competence of Machine Translation Systems”. In: *Proceedings of the Second Conference on Machine Translation*. Sept. 2017. DOI: `10.18653/v1/w17-4705`.

[82] Franck Burlot et al. "The WMT'18 Morpheval Test Suites for English-Czech, English-German, English-Finnish and Turkish-English". In: *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*. Oct. 2018. DOI: `10.18653/v1/w18-6433`.

[83] Adrian Bussone, Simone Stumpf, and Dympna O'Sullivan. "The role of explanations on trust and reliance in clinical decision support systems". In: *2015 international conference on healthcare informatics*. IEEE. 2015, pp. 160–169.

[84] Yves Cabannes. "Participatory budgeting: a significant contribution to participatory democracy". In: *Environment and Urbanization* 16.1 (2004), pp. 27–46. DOI: `10.1177/095624780401600104`. eprint: `https://doi.org/10.1177/095624780401600104`. URL: `https://doi.org/10.1177/095624780401600104`.

[85] Angel Alexander Cabrera et al. "FAIRVIS: Visual Analytics for Discovering Intersectional Bias in Machine Learning". In: *2019 IEEE Conference on Visual Analytics Science and Technology (VAST)*. Oct. 2019. DOI: `10.1109/vast47406.2019.8986948`.

[86] Ángel Alexander Cabrera, Adam Perer, and Jason I. Hong. "Improving Human-AI Collaboration With Descriptions of AI Behavior". In: *Proc. ACM Hum.-Comput. Interact.* 7.CSCW1 (Apr. 2023). DOI: `10.1145/3579612`. URL: `https://doi.org/10.1145/3579612`.

[87] Ángel Alexander Cabrera et al. "Discovering and Validating AI Errors With Crowdsourced Failure Reports". In: *Proceedings of the ACM on Human-Computer Interaction* 5 (Oct. 2021). DOI: `10.1145/3479569`.

[88] Carrie J Cai et al. ""Hello AI": uncovering the onboarding needs of medical practitioners for human-AI collaborative decision-making". In: *Proceedings of the ACM on Human-computer Interaction* 3.CSCW (2019), pp. 1–24.

[89] Carrie J Cai et al. "Onboarding Materials as Cross-Functional Boundary Objects for Developing AI Assistants". In: *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems*. CHI EA '21. Yokohama, Japan: Association for Computing Machinery, 2021. ISBN: 9781450380959. DOI: `10.1145/3411763.3443435`. URL: `https://doi.org/10.1145/3411763.3443435`.

[90] Carrie J. Cai et al. "Human-Centered Tools for Coping with Imperfect Algorithms During Medical Decision-Making". In: *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. CHI '19. Glasgow, Scotland Uk: Association for Computing Machinery, 2019, pp. 1–14. ISBN: 9781450359702. DOI: `10.1145/3290605.3300234`. URL: `https://doi.org/10.1145/3290605.3300234`.

[91] Carrie J. Cai et al. "Human-Centered Tools for Coping with Imperfect Algorithms During Medical Decision-Making". In: *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. May 2019. DOI: `10.1145/3290605.3300234`.

[92] Fabio Calefato et al. "Assessing the impact of real-time machine translation on multilingual meetings in global software projects". en. In: *Empirical Software Engineering* 21.3 (June 2016), pp. 1002–1034. ISSN: 1382-3256, 1573-7616. DOI: 10.1007/s10664-015-9372-x. URL: http://link.springer.com/10.1007/s10664-015-9372-x (visited on 08/24/2020).

[93] Chris Callison-Burch. "Fast, Cheap, and Creative: Evaluating Translation Quality Using Amazon's Mechanical Turk". In: *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing.* Singapore: Association for Computational Linguistics, Aug. 2009, pp. 286–295. URL: https://aclanthology.org/D09-1030.

[94] Chris Callison-Burch, Miles Osborne, and Philipp Koehn. "Re-Evaluating the Role of Bleu in Machine Translation Research". In: *11th Conference of the European Chapter of the Association for Computational Linguistics.* Apr. 2006.

[95] Chris Callison-Burch et al. "(Meta-) Evaluation of Machine Translation". In: *Proceedings of the Second Workshop on Statistical Machine Translation.* Prague, Czech Republic: Association for Computational Linguistics, June 2007, pp. 136–158. URL: https://aclanthology.org/W07-0718.

[96] Chris Callison-Burch et al. "Findings of the 2012 Workshop on Statistical Machine Translation". In: *Proceedings of the Seventh Workshop on Statistical Machine Translation.* Montréal, Canada: Association for Computational Linguistics, June 2012, pp. 10–51. URL: https://www.aclweb.org/anthology/W12-3102.

[97] Sylviane Cardey, Peter Greenfield, and Xiahong Wu. "Designing a controlled language for the machine translation of medical protocols: the case of English to Chinese". In: *Proceedings of the 6th Conference of the Association for Machine Translation in the Americas: Technical Papers.* Washington, USA: Springer, Sept. 2004, pp. 37–47. URL: https://link.springer.com/chapter/10.1007/978-3-540-30194-3_5.

[98] John M. Carroll and Mary Beth Rosson. "Participatory design in community informatics". In: *Design Studies* 28.3 (2007), pp. 243–261. ISSN: 0142-694X. URL: http://www.sciencedirect.com/science/article/pii/S0142694X07000191.

[99] Jay Cassano. *NYC students take aim at segregation by hacking an algorithm.* Fast Company. 2019. URL: https://www.fastcompany.com/90331368/nyc-students-take-aim-at-segregation-by-hacking-an-algorithm.

[100] M. Chalmers, I. MacColl, and M. Bell. "Seamful design: showing the seams in wearable computing". In: *2003 IEE Eurowearable.* 2003, pp. 11–16. DOI: 10.1049/ic:20030140.

[101] Stevie Chancellor, Eric P. S. Baumer, and Munmun De Choudhury. "Who is the "Human" in Human-Centered Machine Learning: The Case of Predicting Mental Health from Social Media". In: *Proc. ACM Hum.-Comput. Interact.* 3.CSCW (Nov. 2019). DOI: 10.1145/3359249. URL: https://doi.org/10.1145/3359249.

[102] Kathy Charmaz. *Constructing grounded theory: A practical guide through qualitative research.* SAGE, 2006.

[103] Elfreda A. Chatman. “The impoverished life-world of outsiders”. In: *Journal of the American Society for Information Science* 47.3 (1996), pp. 193–206. DOI: https://doi.org/10.1002/(SICI)1097-4571(199603)47:3<193::AID-ASI3>3.0.CO;2-T.

[104] Danqi Chen et al. “Reading Wikipedia to Answer Open-Domain Questions”. In: *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Ed. by Regina Barzilay and Min-Yen Kan. Vancouver, Canada: Association for Computational Linguistics, July 2017, pp. 1870–1879. DOI: 10.18653/v1/P17-1171. URL: https://aclanthology.org/P17-1171.

[105] Yan Chen and Yinghua He. *Information Acquisition and Provision in School Choice : An Experimental Study*. Working Paper. 2020. URL: http://yanchen.people.si.umich.edu/papers/Chen_He_2020_09_Distribute.pdf.

[106] Hao-Fei Cheng et al. “Explaining Decision-Making Algorithms through UI: Strategies to Help Non-Expert Stakeholders”. In: *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. CHI ’19. Glasgow, Scotland Uk: Association for Computing Machinery, 2019, pp. 1–12. ISBN: 9781450359702. DOI: 10.1145/3290605.3300789. URL: https://doi.org/10.1145/3290605.3300789.

[107] Alexander Cho et al. “The ”Comadre” Project: An Asset-Based Design Approach to Connecting Low-Income Latinx Families to Out-of-School Learning Opportunities”. In: *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. CHI ’19. Glasgow, Scotland Uk: Association for Computing Machinery, 2019, pp. 1–14. ISBN: 9781450359702. DOI: 10.1145/3290605.3300837. URL: https://doi.org/10.1145/3290605.3300837.

[108] Leshem Choshen and Omri Abend. “Automatically Extracting Challenge Sets for Non-Local Phenomena in Neural Machine Translation”. In: *CoNLL*. Nov. 2019. DOI: 10.18653/v1/k19-1028.

[109] Alexandra Chouldechova. “Fair Prediction with Disparate Impact: A Study of Bias in Recidivism Prediction Instruments”. In: *Big Data* 5.2 (2017). PMID: 28632438, pp. 153–163. DOI: 10.1089/big.2016.0047. URL: https://doi.org/10.1089/big.2016.0047.

[110] Yeounoh Chung et al. “Automated Data Slicing for Model Validation: A Big Data - AI Integration Approach”. In: *IEEE Transactions on Knowledge and Data Engineering* 32.12 (2020), pp. 2284–2296. DOI: 10.1109/TKDE.2019.2916074.

[111] Marika Cifor et al. *Feminist data manifest-no*. 2019. URL: https://www.manifestno.com/.

[112] H. H. Clark and S. E. Brennan. “Grounding in communication”. In: *Perspectives on socially shared cognition*. Ed. by L. B. Resnick, J. M. Levine, and S. D. Teasley. American Psychological Association, 1991, pp. 127–149. URL: https://doi.org/10.1037/10096-006.

[113] Herbert H. Clark and Deanna Wilkes-Gibbs. “Referring as a collaborative process”. In: *Cognition* 22 (1 1986), pp. 1–39. URL: `https://doi.org/10.1016/0010-0277(86)90010-7`.

[114] Andy Cockburn, Amy Karlson, and Benjamin B. Bederson. “A Review of Overview + Detail, Zooming, and Focus+Context Interfaces”. In: *ACM Comput. Surv.* 41 (Jan. 2009). DOI: `10.1145/1456650.1456652`.

[115] Andy Coenen and Adam Pearce. *Understanding UMAP*. 2019. URL: `https://pair-code.github.io/understanding-umap/`.

[116] Eric Corbett and Emily Denton. “Interrogating the T in FAccT”. In: *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*. FAccT ’23. Chicago, IL, USA: Association for Computing Machinery, 2023, pp. 1624–1634. DOI: `10.1145/3593013.3594104`. URL: `https://doi.org/10.1145/3593013.3594104`.

[117] Sean P Corcoran et al. *Leveling the Playing Field for High School Choice: Results from a Field Experiment of Informational Interventions*. Working Paper 24471. National Bureau of Economic Research, Mar. 2018. DOI: `10.3386/w24471`. URL: `http://www.nber.org/papers/w24471`.

[118] Sasha Costanza-Chock. “Design Justice: Towards an Intersectional Feminist Framework for Design Theory and Practice”. In: *Proceedings of the Design Research Society*. London, United Kingdom: Design Research Society, June 2018. URL: `https://ssrn.com/abstract=3189696`.

[119] Kate Crawford. “Can an Algorithm be Agonistic? Ten Scenes from Life in Calculated Publics”. In: *Science, Technology, & Human Values* 41.1 (2016), pp. 77–92. DOI: `10.1177/0162243915589635`. URL: `https://doi.org/10.1177/0162243915589635`.

[120] Kimberlé W Crenshaw. *On intersectionality: Essential writings*. The New Press, 2017.

[121] Greg d’Eon et al. “The Spotlight: A General Method for Discovering Systematic Errors in Deep Learning Models”. In: *2022 ACM Conference on Fairness, Accountability, and Transparency*. 2022. DOI: `10.1145/3531146.3533240`.

[122] Prithwijit Das et al. “Dangers of Machine Translation: The Need for Professionally Translated Anticipatory Guidance Resources for Limited English Proficiency Caregivers”. In: *Clinical Pediatrics (Phila)* 58 (Feb. 2019).

[123] Prithwijit Das et al. “Dangers of Machine Translation: The Need for Professionally Translated Anticipatory Guidance Resources for Limited English Proficiency Caregivers”. In: *Clinical Pediatrics (Phila)* 58.2 (Feb. 2019), pp. 247–249. DOI: `10.1177/0009922818809494`.

[124] Janet Davis and Lisa P. Nathan. “Value Sensitive Design: Applications, Adaptations, and Critiques”. In: *Handbook of Ethics, Values, and Technological Design*. Ed. by Jeroen van den Hoven, Pieter E. Vermaas, and Ibo van de Poel. Dordrecht, Netherlands: Springer, 2015, pp. 11–40.

[125] Terrance de Vries et al. "Does Object Recognition Work for Everyone?" In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*. June 2019.

[126] Fernando Delgado et al. *Stakeholder Participation in AI: Beyond "Add Diverse Stakeholders and Stir"*. 2021. arXiv: `2111.01122 [cs.AI]`.

[127] Alicia DeVos et al. "Toward User-Driven Algorithm Auditing: Investigating Users' Strategies for Uncovering Harmful Algorithmic Behavior". In: *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. CHI '22. 2022. DOI: `10.1145/3491102.3517441`.

[128] Mark Díaz et al. "CrowdWorkSheets: Accounting for Individual and Collective Identities Underlying Crowdsourced Dataset Annotation". In: *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*. FAccT '22. Seoul, Republic of Korea: Association for Computing Machinery, 2022, pp. 2342–2351. ISBN: 9781450393522. DOI: `10.1145/3531146.3534647`. URL: `https://doi.org/10.1145/3531146.3534647`.

[129] Jessa Dickinson et al. ""The Cavalry Ain't Coming in to Save Us": Supporting Capacities and Relationships through Civic Tech". In: *Proc. ACM Hum.-Comput. Interact.* 3.CSCW (Nov. 2019). DOI: `10.1145/3359225`. URL: `https://doi.org/10.1145/3359225`.

[130] Thomas Dietz, Elinor Ostrom, and Paul C. Stern. "The Struggle to Govern the Commons". In: *Science* 302.5652 (2003), pp. 1907–1912. ISSN: 0036-8075. DOI: `10.1126/science.1091015`. URL: `https://science.sciencemag.org/content/302/5652/1907`.

[131] Tawanna R. Dillahunt and Alex Lu. "DreamGigs: Designing a Tool to Empower Low-Resource Job Seekers". In: *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. CHI '19. Glasgow, Scotland Uk: Association for Computing Machinery, 2019, pp. 1–14. ISBN: 9781450359702. DOI: `10.1145/3290605.3300808`. URL: `https://doi.org/10.1145/3290605.3300808`.

[132] Tawanna R. Dillahunt et al. "Designing Future Employment Applications for Underserved Job Seekers: A Speed Dating Study". In: *Proceedings of the 2018 Designing Interactive Systems Conference*. DIS '18. Hong Kong, China: Association for Computing Machinery, 2018, pp. 33–44. ISBN: 9781450351980. DOI: `10.1145/3196709.3196770`. URL: `https://doi.org/10.1145/3196709.3196770`.

[133] Tawanna R. Dillahunt et al. "Uncovering the Values and Constraints of Real-Time Ridesharing for Low-Resource Populations". In: *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*. CHI '17. Denver, Colorado, USA: Association for Computing Machinery, 2017, pp. 2757–2769. ISBN: 9781450346559. DOI: `10.1145/3025453.3025470`. URL: `https://doi.org/10.1145/3025453.3025470`.

[134] Tingting Ding and Andrew Schotter. “Learning and Mechanism Design: An Experimental Test of School Matching Mechanisms with Intergenerational Advice”. In: *The Economic Journal* 129.623 (May 2019), pp. 2779–2804.

[135] Tingting Ding and Andrew Schotter. “Matching and chatting: An experimental study of the impact of network communication on school-matching mechanisms”. In: *Games and Economic Behavior* 103 (2017). John Nash Memorial, pp. 94–115. ISSN: 0899-8256. DOI: `https://doi.org/10.1016/j.geb.2016.02.004`. URL: `http://www.sciencedirect.com/science/article/pii/S0899825616000294`.

[136] Betsy DiSalvo, Parisa Khanipour Roshan, and Briana Morrison. “Information Seeking Practices of Parents: Exploring Skills, Face Threats and Social Networks”. In: *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. CHI ’16. San Jose, California, USA: Association for Computing Machinery, 2016, pp. 623–634. ISBN: 9781450333627. DOI: `10.1145/2858036.2858586`. URL: `https://doi.org/10.1145/2858036.2858586`.

[137] Lynn Dombrowski, Ellie Harmon, and Sarah Fox. “Social Justice-Oriented Interaction Design: Outlining Key Design Strategies and Commitments”. In: *Proceedings of the 2016 ACM Conference on Designing Interactive Systems*. DIS ’16. Brisbane, QLD, Australia: Association for Computing Machinery, 2016, pp. 656–671. ISBN: 9781450340311. DOI: `10.1145/2901790.2901861`. URL: `https://doi.org/10.1145/2901790.2901861`.

[138] Joseph Donia and Jay Shaw. “Co-Design and Ethical Artificial Intelligence for Health: Myths and Misconceptions”. In: *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*. AIES ’21. Virtual Event, USA: Association for Computing Machinery, 2021, p. 77. ISBN: 9781450384735. DOI: `10.1145/3461702.3462537`. URL: `https://doi.org/10.1145/3461702.3462537`.

[139] Miro Dudík et al. “Fairlearn: A Toolkit for Assessing and Improving Fairness in AI”. In: (May 2020).

[140] Upol Ehsan and Mark O. Riedl. “Human-Centered Explainable AI: Towards a Reflective Sociotechnical Approach”. In: *HCI International 2020 - Late Breaking Papers: Multimodality and Intelligence*. Ed. by Constantine Stephanidis et al. Cham: Springer International Publishing, 2020, pp. 449–466. ISBN: 978-3-030-60117-1.

[141] Upol Ehsan et al. “Expanding Explainability: Towards Social Transparency in AI Systems”. In: *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. New York, NY, USA: Association for Computing Machinery, 2021. ISBN: 9781450380966. URL: `https://doi.org/10.1145/3411764.3445188`.

[142] Mica R. Endsley. “From Here to Autonomy: Lessons Learned From Human–Automation Research”. In: *Human Factors* 59.1 (2017), pp. 5–27. URL: `https://doi.org/10.1177/0018720816681350`.

[143] Sheena Erete and Jennifer O. Burrell. “Empowered Participation: How Citizens Use Technology in Local Governance”. In: *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*. CHI ’17. Denver, Colorado, USA: Association for Computing Machinery, 2017, pp. 2307–2319. ISBN: 9781450346559. DOI: 10.1145/3025453.3025996. URL: https://doi.org/10.1145/3025453.3025996.

[144] Florian Evequoz et al. “Diverse Representation via Computational Participatory Elections - Lessons from a Case Study”. In: *Equity and Access in Algorithms, Mechanisms, and Optimization*. EAAMO ’22. Arlington, VA, USA: Association for Computing Machinery, 2022. ISBN: 9781450394772. DOI: 10.1145/3551624.3555297. URL: https://doi.org/10.1145/3551624.3555297.

[145] Eve L. Ewing. *Ghosts in the Schoolyard: Racism and School Closings on Chicago’s South Side*. University of Chicago Press, 2018.

[146] Gregory Ewing and Ibrahim Demir. “An ethical decision-making framework with serious gaming: a smart water case study on flooding”. In: *Journal of Hydroinformatics* 23.3 (May 2021), pp. 466–482. DOI: 10.2166/hydro.2021.097.

[147] Ana C. Farinha et al. *WMT22 Chat Task*. 2022. URL: https://wmt-chat-task.github.io/.

[148] Fangxiaoyu Feng et al. “Language-agnostic BERT Sentence Embedding”. In: *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Dublin, Ireland: Association for Computational Linguistics, May 2022, pp. 878–891. DOI: 10.18653/v1/2022.acl-long.62. URL: https://aclanthology.org/2022.acl-long.62.

[149] Mary Flanagan, Daniel C. Howe, and Helen Nissenbaum. “Values at Play: Design Tradeoffs in Socially-Oriented Game Design”. In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. CHI ’05. Portland, Oregon, USA: Association for Computing Machinery, 2005, pp. 751–760. ISBN: 1581139985. DOI: 10.1145/1054972.1055076. URL: https://doi.org/10.1145/1054972.1055076.

[150] Rachel Freedman et al. “Adapting a Kidney Exchange Algorithm to Align with Human Values”. In: *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*. AIES ’18. New Orleans, LA, USA: Association for Computing Machinery, 2018, p. 115. ISBN: 9781450360128. DOI: 10.1145/3278721.3278727. URL: https://doi.org/10.1145/3278721.3278727.

[151] Markus Freitag et al. “Experts, Errors, and Context: A Large-Scale Study of Human Evaluation for Machine Translation”. In: *Transactions of the Association for Computational Linguistics* 9 (Dec. 2021). DOI: 10.1162/tacl_a_00437.

[152] B Friedman, PH Jr Kahn, and A Borning. “Value Sensitive Design and Information Systems”. In: *Human-Computer Interaction in Management Information Systems: Foundations*. Ed. by Ben Shneiderman, Ping Zhang, and Dennis Galletta. Armonk, NY, USA: M. E. Sharpe, Inc., 2006, pp. 348–372.

[153] Batya Friedman, David G. Hendry, and Alan Borning. “A Survey of Value Sensitive Design Methods”. In: *Foundations and Trends in Human–Computer Interaction* 11.2 (2017), pp. 63–125. ISSN: 1551-3955. DOI: 10.1561/1100000015. URL: http://dx.doi.org/10.1561/1100000015.

[154] Batya Friedman and Peter H. Kahn Jr. “Human Values, Ethics, and Design”. In: *The Human-Computer Interaction Handbook: Fundamentals, Evolving Technologies, and Emerging Applications.* Ed. by Jacko JA Sears A. USA: L. Erlbaum Associates Inc., 2003, pp. 1177–1201.

[155] D. Gale and L. S. Shapley. “College Admissions and the Stability of Marriage”. In: *The American Mathematical Monthly* 69.2 (Jan. 1962), pp. 9–15.

[156] Ge Gao et al. “How Beliefs about the Presence of Machine Translation Impact Multilingual Collaborations”. In: *Proceedings of the 17th ACM Conference on Computer Supported Cooperative Work & Social Computing.* CSCW ’14. Baltimore, Maryland, USA: Association for Computing Machinery, 2014, pp. 1549–1560. ISBN: 9781450325400. DOI: 10.1145/2531602.2531702. URL: https://doi.org/10.1145/2531602.2531702.

[157] Ge Gao et al. “Same translation but different experience: the effects of highlighting on machine-translated conversations”. In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems.* CHI ’13. Paris, France: Association for Computing Machinery, Apr. 2013, pp. 449–458. ISBN: 978-1-4503-1899-0. DOI: 10.1145/2470654.2470719. URL: https://doi.org/10.1145/2470654.2470719 (visited on 06/03/2020).

[158] Ge Gao et al. “Two is Better Than One: Improving Multilingual Collaboration by Giving Two Machine Translation Outputs”. en. In: *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing - CSCW ’15.* Vancouver, BC, Canada: ACM Press, 2015, pp. 852–863. ISBN: 978-1-4503-2922-4. DOI: 10.1145/2675133.2675197. URL: http://dl.acm.org/citation.cfm?doid=2675133.2675197 (visited on 08/08/2020).

[159] Tong Gao et al. “DataTone: Managing Ambiguity in Natural Language Interfaces for Data Visualization”. In: *Proceedings of the 28th Annual ACM Symposium on User Interface Software &amp; Technology.* UIST ’15. Charlotte, NC, USA: Association for Computing Machinery, 2015, pp. 489–500. ISBN: 9781450337793. DOI: 10.1145/2807442.2807478. URL: https://doi.org/10.1145/2807442.2807478.

[160] Matt Gardner et al. “Evaluating Models’ Local Decision Boundaries via Contrast Sets”. In: *Findings of the Association for Computational Linguistics: EMNLP 2020.* Nov. 2020. DOI: 10.18653/v1/2020.findings-emnlp.117.

[161] Nikhil Garg et al. “Who is in Your Top Three? Optimizing Learing in Elections with Many Candidates”. In: *AAAI Conference on Human Computation and Crowdsourcing (HCOMP ’19).* 2019.

[162] Satvik Garg et al. "On Continuous Integration / Continuous Delivery for Automated Deployment of Machine Learning Models Using MLOps". In: *2021 IEEE Fourth International Conference on Artificial Intelligence and Knowledge Engineering (AIKE)*. 2021. DOI: `10.1109/aike52691.2021.00010`.

[163] Timnit Gebru et al. *Datasheets for Datasets*. Mar. 2018. URL: `https://www.microsoft.com/en-us/research/publication/datasheets-for-datasets/`.

[164] Mor Geva, Yoav Goldberg, and Jonathan Berant. "Are We Modeling the Task or the Annotator? An Investigation of Annotator Bias in Natural Language Understanding Datasets". In: *EMNLP-IJCNLP*. Nov. 2019. DOI: `10.18653/v1/d19-1107`.

[165] Marzyeh Ghassemi et al. "A Review of Challenges and Opportunities in Machine Learning for Health." In: *arXiv preprint* (2019).

[166] James L. Gibson. "Understandings of Justice: Institutional Legitimacy, Procedural Justice, and Political Tolerance". In: *Law & Society Review* 23.3 (1989), pp. 469–496. ISSN: 00239216, 15405893. URL: `http://www.jstor.org/stable/3053830`.

[167] Steven Glazerman and Dallas Dotter. "Market Signals: Evidence on the Determinants and Consequences of School Choice From a Citywide Lottery". In: *Educational Evaluation and Policy Analysis* 39.4 (2017), pp. 593–619. DOI: `10.3102/0162373717702964`. URL: `https://doi.org/10.3102/0162373717702964`.

[168] Edward Golding, Laurie Goodman, and Sarah Strochak. *Is Limited English Proficiency a Barrier to Homeownership?* Urban Institute. Mar. 2018. URL: `https://www.urban.org/sites/default/files/publication/97436/is_limited_english_proficiency_a_barrier_to_homeownership.pdf`.

[169] Mitchell L. Gordon et al. "Jury Learning: Integrating Dissenting Voices into Machine Learning Models". In: *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. CHI '22. New Orleans, LA, USA: Association for Computing Machinery, 2022. ISBN: 9781450391573. DOI: `10.1145/3491102.3502004`. URL: `https://doi.org/10.1145/3491102.3502004`.

[170] Jeremy Gormley. *The problem Oakland schools live with*. July 2020. URL: `https://www.integrateoaklandschools.org/post/problem-oakland-schools-live-with`.

[171] Yvette Graham et al. "Continuous Measurement Scales in Human Evaluation of Machine Translation". In: *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*. Sofia, Bulgaria: Association for Computational Linguistics, Aug. 2013, pp. 33–41. URL: `https://aclanthology.org/W13-2305`.

[172] Spence Green et al. "Predictive translation memory: a mixed-initiative system for human language translation". en. In: *Proceedings of the 27th annual ACM symposium on User interface software and technology - UIST '14*. Honolulu, Hawaii, USA: ACM Press, 2014, pp. 177–187. ISBN: 978-1-4503-3069-5. DOI: `10.1145/2642918.2647408`. URL: `http://dl.acm.org/citation.cfm?doid=2642918.2647408` (visited on 10/11/2020).

[173] Mark Winston Griffith and Max Freedman. *Episode 5: The Disappearing District*. Oct. 2019. URL: https://www.schoolcolorspodcast.com/episodes/episode-5-the-disappearing-district.

[174] Mark Winston Griffith and Max Freedman. *School Colors (Episode 7: New Kids on the Block)*. Nov. 2019. URL: https://www.schoolcolorspodcast.com/episodes/episode-7-new-kids-on-the-block.

[175] Andrea Grimes et al. "EatWell: Sharing Nutrition-Related Memories in a Low-Income Community". In: *Proceedings of the 2008 ACM Conference on Computer Supported Cooperative Work*. CSCW '08. San Diego, CA, USA: Association for Computing Machinery, 2008, pp. 87–96. ISBN: 9781605580074. DOI: 10.1145/1460563.1460579. URL: https://doi.org/10.1145/1460563.1460579.

[176] Maarten Grootendorst. "BERTopic: Neural Topic Modeling with a Class-Based TF-IDF Procedure". In: *arXiv preprint arXiv:2203.05794* (2022).

[177] Pablo Guillen and Rustamdjan Hakimov. "Not quite the best response: truth-telling, strategy-proof matching, and the manipulation of others". In: *Experimental Economics* 20.3 (2017), pp. 670–686.

[178] Pablo Guillen and Rustamdjan Hakimov. "The effectiveness of top-down advice in strategy-proof mechanisms: A field experiment". In: *European Economic Review* 101.C (2018), pp. 505–511.

[179] Pablo Guillen and Alexander Hing. "Lying through their teeth: Third party advice and truth telling in a strategy proof mechanism". In: *European Economic Review* 70 (2014), pp. 178–185. ISSN: 0014-2921. DOI: https://doi.org/10.1016/j.euroecorev.2014.05.002. URL: http://www.sciencedirect.com/science/article/pii/S0014292114000737.

[180] Danna Gurari et al. "VizWiz Grand Challenge: Answering Visual Questions from Blind People". In: *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2018, pp. 3608–3617. DOI: 10.1109/CVPR.2018.00380.

[181] Craig Hadley and Crystal Patil. "Perceived discrimination among three groups of refugees resettled in the USA: associations with language, time in the USA, and continent of origin". In: *Journal of Immigrant and Minority Health* 11.6 (2009), pp. 505–512.

[182] Isa E. Hafalir, M. Bumin Yenmez, and Muhammed A. Yildirim. "Effective affirmative action in school choice". In: *Theoretical Economics* 8.2 (2013), pp. 325–363. DOI: 10.3982/TE1135. URL: https://onlinelibrary.wiley.com/doi/abs/10.3982/TE1135.

[183] Aaron Halfaker and R. Stuart Geiger. "ORES: Lowering Barriers with Participatory Machine Learning in Wikipedia". In: *ArXiv preprint* (2020). URL: https://arxiv.org/pdf/1909.05189.pdf.

[184] Lifeng Han. "Machine Translation Evaluation Resources and Methods: A Survey". In: *arxiv:1605.04515* (Sept. 2018).

[185] Jeffrey T Hancock, Mor Naaman, and Karen Levy. "AI-Mediated Communication: Definition, Research Agenda, and Ethical Considerations". In: *Journal of Computer-Mediated Communication* 25.1 (Jan. 2020), pp. 89–100. ISSN: 1083-6101. DOI: 10.1093/jcmc/zmz022. URL: https://doi.org/10.1093/jcmc/zmz022.

[186] Matt Haney, Steven Cook, and Rachel Norton. *Developing a Community Based Student Assignment System for SFUSD*. 2018.

[187] Kotaro Hara and Shamsi T. Iqbal. "Effect of Machine Translation in Interlingual Conversation: Lessons from a Formative Study". In: *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*. CHI '15. Seoul, Republic of Korea: Association for Computing Machinery, Apr. 2015, pp. 3473–3482. ISBN: 978-1-4503-3145-6. DOI: 10.1145/2702123.2702407. URL: https://doi.org/10.1145/2702123.2702407 (visited on 06/03/2020).

[188] Mike Harding et al. "HCI, Civic Engagement & Trust". In: *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*. CHI '15. Seoul, Republic of Korea: Association for Computing Machinery, 2015, pp. 2833–2842. ISBN: 9781450331456. DOI: 10.1145/2702123.2702255. URL: https://doi.org/10.1145/2702123.2702255.

[189] Shaun R. Harper. "An anti-deficit achievement framework for research on students of color in STEM". In: *New Directions for Institutional Research* 2010.148 (2010), pp. 63–74. DOI: https://doi.org/10.1002/ir.362. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/ir.362. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/ir.362.

[190] Christina Harrington, Sheena Erete, and Anne Marie Piper. "Deconstructing Community-Based Collaborative Design: Towards More Equitable Participatory Design Engagements". In: *Proc. ACM Hum.-Comput. Interact.* 3.CSCW (Nov. 2019). DOI: 10.1145/3359318. URL: https://doi.org/10.1145/3359318.

[191] Christina N. Harrington, Katya Borgos-Rodriguez, and Anne Marie Piper. "Engaging Low-Income African American Older Adults in Health Discussions through Community-Based Design Workshops". In: *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. CHI '19. Glasgow, Scotland Uk: Association for Computing Machinery, 2019, pp. 1–15. ISBN: 9781450359702. DOI: 10.1145/3290605.3300823. URL: https://doi.org/10.1145/3290605.3300823.

[192] Christina N. Harrington et al. "Designing Health and Fitness Apps with Older Adults: Examining the Value of Experience-Based Co-Design". In: *Proceedings of the 12th EAI International Conference on Pervasive Computing Technologies for Healthcare*. PervasiveHealth '18. New York, NY, USA: Association for Computing Machinery, 2018, pp. 15–24. ISBN: 9781450364508. DOI: 10.1145/3240925.3240929. URL: https://doi.org/10.1145/3240925.3240929.

[193] Tatsunori B. Hashimoto et al. “A Retrieve-and-Edit Framework for Predicting Structured Outputs”. In: *32nd Conference on Neural Information Processing Systems*. NeurIPS 2018. Montréal, Canada, 2018.

[194] Avinatan Hassidim, Assaf Romm, and Ran I. Shorrer. ““Strategic” Behavior in a Strategy-Proof Environment”. In: *Proceedings of the 2016 ACM Conference on Economics and Computation*. EC ’16. Maastricht, The Netherlands: Association for Computing Machinery, 2016, pp. 763–764. ISBN: 9781450339360. DOI: `10.1145/2940716.2940751`. URL: `https://doi.org/10.1145/2940716.2940751`.

[195] Justine S Hastings and Jeffrey M Weinstein. *Information, School Choice, and Academic Achievement: Evidence from Two Experiments*. Working Paper 13623. National Bureau of Economic Research, Nov. 2007. DOI: `10.3386/w13623`. URL: `http://www.nber.org/papers/w13623`.

[196] Justine S. Hastings, Thomas J. Kane, and Douglas O. Staiger. “Heterogeneous Preferences and the Efficacy of Public School Choice”. In: *NBER Working Paper* (2009).

[197] Marti A. Hearst. “The Design of Search User Interfaces”. In: *Search User Interfaces*. Cambridge University Press, 2009.

[198] Jeffrey Heer. “Agency plus automation: Designing artificial intelligence into interactive systems”. In: *Proceedings of the National Academy of Sciences* 116.6 (2019), pp. 1844–1850.

[199] Amy K. Heger et al. “Understanding Machine Learning Practitioners’ Data Documentation Perceptions, Needs, Challenges, and Desiderata”. In: *arxiv:2206.02923* (Aug. 2022).

[200] Jindřich Helcl and Jindřich Libovický. “Neural Monkey: An Open-Source Tool for Sequence Learning”. In: *The Prague Bulletin of Mathematical Linguistics* (2017). DOI: `10.1515/pralin-2017-0001`.

[201] Yoan Hermstrüwer. *Transparency and Fairness in School Choice Mechanisms*. Tech. rep. Max Planck Institute for Research on Collective Goods, 2019.

[202] Helen Hershkofft and Adam S. Cohen. “School Choice and the Lessons of Choctaw County”. In: *Yale Law & Policy Review* 10.1 (1992). URL: `https://digitalcommons.law.yale.edu/ylpr/vol10/iss1/2/`.

[203] Paul T. Hill. “NCLB School Choice and Children in Poverty”. In: *Standards-Based Reform and the Poverty Gap: Lessons for “No Child Left Behind”*. Ed. by Adam Gamoran. 2007. Chap. 8.

[204] Zoë Hitzig. “The normative gap: mechanism design and ideal theories of justice”. In: *Economics and Philosophy* (2019), pp. 1–28. DOI: `10.1017/S0266267119000270`.

[205] Anna Lauren Hoffmann. "Terms of inclusion: Data, discourse, violence". In: *New Media & Society* 23.12 (2021), pp. 3539–3556. DOI: 10.1177/1461444820958725. eprint: https://doi.org/10.1177/1461444820958725. URL: https://doi.org/10.1177/1461444820958725.

[206] Fred Hohman et al. "Visual Analytics in Deep Learning: An Interrogative Survey for the Next Frontiers". In: *IEEE Transactions on Visualization and Computer Graphics* 25 (Aug. 2019). DOI: 10.1109/tvcg.2018.2843369.

[207] Kenneth Holstein et al. "Improving Fairness in Machine Learning Systems: What Do Industry Practitioners Need?" In: *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. 2019. DOI: 10.1145/3290605.3300830.

[208] Sara Hooker et al. "Characterising Bias in Compressed Models". In: *arxiv:2010.03058* (Dec. 2020).

[209] Benjamin Hoover, Hendrik Strobelt, and Sebastian Gehrmann. "exBERT: A Visual Analysis Tool to Explore Learned Representations in Transformer Models". In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*. July 2020. DOI: 10.18653/v1/2020.acl-demos.22.

[210] Aspen Hopkins et al. "Designing data: Proactive data collection and iteration for machine learning". In: *arXiv preprint arXiv:2301.10319* (2023).

[211] Eric Horvitz. "Principles of Mixed-Initiative User Interfaces". In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. CHI '99. Pittsburgh, Pennsylvania, USA: Association for Computing Machinery, 1999, pp. 159–166. ISBN: 0201485591. DOI: 10.1145/302979.303030. URL: https://doi.org/10.1145/302979.303030.

[212] Nabil Hossain, Marjan Ghazvininejad, and Luke Zettlemoyer. "Simple and Effective Retrieve-Edit-Rerank Text Generation". In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Ed. by Dan Jurafsky et al. Online: Association for Computational Linguistics, July 2020, pp. 2532–2538. DOI: 10.18653/v1/2020.acl-main.228. URL: https://aclanthology.org/2020.acl-main.228.

[213] Asta Høy. *Guidelines for Translation of SNOMED CT®*. International Health Terminology Standards Development Organization. 2012. URL: https://www.snomed.org/SNOMED/media/SNOMED/documents/IHTSDO_Translation_Guidelines_v2_02_20121211-(1).pdf.

[214] Joey Chiao-Yin Hsiao and Tawanna R. Dillahunt. "Technology to Support Immigrant Access to Social Capital and Adaptation to a New Country". In: *Proc. ACM Hum.-Comput. Interact.* 2.CSCW (Nov. 2018). DOI: 10.1145/3274339. URL: https://doi.org/10.1145/3274339.

[215] Chang Hu et al. "MonoTrans2: a new human computation system to support monolingual translation". In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. CHI '11. Vancouver, BC, Canada: Association for Computing Machinery, May 2011, pp. 1133–1136. ISBN: 978-1-4503-0228-9. DOI: `10.1145/1978942.1979111`. URL: `https://doi.org/10.1145/1978942.1979111` (visited on 06/03/2020).

[216] Yuen J. Huo. "Procedural Justice and Social Regulation Across Group Boundaries: Does Subgroup Identity Undermine Relationship-Based Governance?" In: *Personality and Social Psychology Bulletin* 29.3 (2003). PMID: 15273011, pp. 336–348. DOI: `10.1177/0146167202250222`. eprint: `https://doi.org/10.1177/0146167202250222`. URL: `https://doi.org/10.1177/0146167202250222`.

[217] Edwin L Hutchins, James D Hollan, and Donald A Norman. "Direct manipulation interfaces". In: *Human–computer interaction* 1.4 (1985), pp. 311–338.

[218] Ben Hutchinson et al. "Evaluation Gaps in Machine Learning Practice". In: *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*. FAccT '22. Seoul, Republic of Korea: Association for Computing Machinery, 2022, pp. 1859–1876. ISBN: 9781450393522. DOI: `10.1145/3531146.3533233`. URL: `https://doi.org/10.1145/3531146.3533233`.

[219] Ben Hutchinson et al. "Towards Accountability for Machine Learning Datasets: Practices from Software Engineering and Infrastructure". In: *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. FAccT '21. Virtual Event, Canada: Association for Computing Machinery, 2021, pp. 560–575. ISBN: 9781450383097. DOI: `10.1145/3442188.3445918`. URL: `https://doi.org/10.1145/3442188.3445918`.

[220] Chip Huyen. *Designing Machine Learning Systems: An Iterative Process for Production-Ready Applications*. First edition. 2022.

[221] "IEEE Standard for Phasor Data Concentrators for Power Systems". In: *IEEE Std C37.247-2019* (2019), pp. 1–44. DOI: `10.1109/IEEESTD.2019.8830511`.

[222] Sarah Inman and David Ribes. ""Beautiful Seams": Strategic Revelations and Concealments". In: *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. New York, NY, USA: Association for Computing Machinery, 2019, pp. 1–14. ISBN: 9781450359702. URL: `https://doi.org/10.1145/3290605.3300508`.

[223] Panagiotis G. Ipeirotis, Foster Provost, and Jing Wang. "Quality Management on Amazon Mechanical Turk". In: *Proceedings of the ACM SIGKDD Workshop on Human Computation - HCOMP '10*. 2010. DOI: `10.1145/1837885.1837906`.

[224] Lilly Irani. "The Hidden Faces of Automation". In: *XRDS: Crossroads, The ACM Magazine for Students* 23.2 (2016), pp. 34–47. DOI: `10.1145/3014390`.

[225] Lilly Irani et al. "Postcolonial Computing: A Lens on Design and Development". In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. CHI '10. Atlanta, Georgia, USA: Association for Computing Machinery, 2010, pp. 1311–1320. ISBN: 9781605589299. DOI: 10.1145/1753326.1753522. URL: https://doi.org/10.1145/1753326.1753522.

[226] Lilly C. Irani and M. Six Silberman. "Turkopticon: Interrupting Worker Invisibility in Amazon Mechanical Turk". In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. CHI '13. Paris, France: Association for Computing Machinery, 2013, pp. 611–620. ISBN: 9781450318990. DOI: 10.1145/2470654.2470742. URL: https://doi.org/10.1145/2470654.2470742.

[227] Pierre Isabelle, Colin Cherry, and George Foster. "A Challenge Set Approach to Evaluating Machine Translation". In: *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. Sept. 2017. DOI: 10.18653/v1/d17-1263.

[228] Aarti Israni, Nicole B. Ellison, and Tawanna R. Dillahunt. "'A Library of People': Online Resource-Seeking in Low-Income Communities". In: *Proc. ACM Hum.-Comput. Interact.* 5.CSCW1 (Nov. 2021). DOI: 10.1145/3449226. URL: https://doi.org/10.1145/3449226.

[229] Alon Jacovi et al. "Formalizing trust in artificial intelligence: Prerequisites, causes and goals of human trust in ai". In: *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. 2021, pp. 624–635.

[230] Sarah Jaffe. *The Chicago Teachers Strike Was a Lesson in 21st-Century Organizing*. The Nation. Nov. 2019. URL: https://www.thenation.com/article/archive/chicago-ctu-strike-win.

[231] Heinrich Jiang et al. "To trust or not to trust a classifier". In: *Advances in neural information processing systems* 31 (2018).

[232] Melvin Johnson et al. "Google's Multilingual Neural Machine Translation System: Enabling Zero-Shot Translation". In: *Transactions of the Association for Computational Linguistics* (2017), pp. 339–351. DOI: 10.1162/tacl_a_00065.

[233] Caroline M. Johnston, Simon Blessenohl, and Phebe Vayanos. "Preference Elicitation and Aggregation to Aid with Patient Triage during the COVID-19 Pandemic". In: *Workshop on Participatory Approaches to Machine Learning at International Conference on Machine Learning (ICML)* (2020). URL: https://participatoryml.github.io/papers/2020/33.pdf.

[234] Aparna R Joshi et al. "Fair SA: Sensitivity Analysis for Fairness in Face Recognition". In: *Algorithmic Fairness through the Lens of Causality and Robustness Workshop*. PMLR. 2022.

[235] Pratik Joshi et al. "The State and Fate of Linguistic Diversity and Inclusion in the NLP World". In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, July 2020, pp. 6282–6293. DOI: `10.18653/v1/2020.acl-main.560`. URL: `https://aclanthology.org/2020.acl-main.560`.

[236] John T. Jost, Mahzarin R. Banaji, and Brian A. Nosek. "A Decade of System Justification Theory: Accumulated Evidence of Conscious and Unconscious Bolstering of the Status Quo". In: *Political Psychology* 25.6 (2004), pp. 881–919. DOI: `https://doi.org/10.1111/j.1467-9221.2004.00402.x`. eprint: `https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1467-9221.2004.00402.x`. URL: `https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1467-9221.2004.00402.x`.

[237] Anson Kahng et al. "Statistical Foundations of Virtual Democracy". In: *Proceedings of the 36th International Conference on Machine Learning*. Ed. by Kamalika Chaudhuri and Ruslan Salakhutdinov. Vol. 97. Proceedings of Machine Learning Research. Long Beach, California, USA: PMLR, June 2019, pp. 3173–3182. URL: `http://proceedings.mlr.press/v97/kahng19a.html`.

[238] Adam J. Kapor, Christopher A. Neilson, and Seth D. Zimmerman. "Heterogeneous Beliefs and School Choice Mechanisms". In: *American Economic Review* 110.5 (May 2020), pp. 1274–1315. DOI: `10.1257/aer.20170129`. URL: `https://www.aeaweb.org/articles?id=10.1257/aer.20170129`.

[239] Amir-Hossein Karimi et al. "A Survey of Algorithmic Recourse: Contrastive Explanations and Consequential Recommendations". In: *ACM Comput. Surv.* 55.5 (Dec. 2022). ISSN: 0360-0300. DOI: `10.1145/3527848`. URL: `https://doi.org/10.1145/3527848`.

[240] Matt Kasman and Jon Valant. "The opportunities and risks of K-12 student placement algorithms". In: *The Brookings Institute* (2019).

[241] Harmanpreet Kaur et al. "Interpreting Interpretability: Understanding Data Scientists' Use of Interpretability Tools for Machine Learning". In: *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. New York, NY, USA: Association for Computing Machinery, 2020, pp. 1–14. ISBN: 9781450367080. URL: `https://doi.org/10.1145/3313831.3376219`.

[242] Harmanpreet Kaur et al. "Sensible AI: Re-imagining interpretability and explainability using sensemaking theory". In: *2022 ACM Conference on Fairness, Accountability, and Transparency*. 2022, pp. 702–714.

[243] Anna Kawakami et al. ""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*. 2022, pp. 454–470.

[244] J. Keller and K. McDade. "Attitudes of low-income parents toward seeking help with parenting: implications for practice". In: *Child Welfare* 79.3 (2000), pp. 285–312.

[245] Kotaro Suzumura Kenneth J. Arrow Amartya Sen. *Handbook of Social Choice and Welfare*. Elsevier, 2010.

[246] Urvashi Khandelwal et al. "Generalization through Memorization: Nearest Neighbor Language Models". In: *International Conference on Learning Representations (ICLR)*. 2020.

[247] Urvashi Khandelwal et al. "Nearest Neighbor Machine Translation". In: *International Conference on Learning Representations*. 2021. URL: https://openreview.net/forum?id=7wCBOfJ8hJM.

[248] Parisa Khanipour Roshan et al. "Exploring How Parents in Economically Depressed Communities Access Learning Resources". In: *Proceedings of the 18th International Conference on Supporting Group Work*. GROUP '14. Sanibel Island, Florida, USA: Association for Computing Machinery, 2014, pp. 131–141. ISBN: 9781450330435. DOI: 10.1145/2660398.2660415. URL: https://doi.org/10.1145/2660398.2660415.

[249] Elaine C. Khoong et al. "Assessing the Use of Google Translate for Spanish and Chinese Translations of Emergency Department Discharge Instructions". In: *JAMA Internal Medicine* 179.4 (Apr. 2019), pp. 580–582. DOI: 10.1001/jamainternmed.2018.7653. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6450297/.

[250] Douwe Kiela et al. "Dynabench: Rethinking Benchmarking in NLP". In: *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. 2021. DOI: 10.18653/v1/2021.naacl-main.324.

[251] Jeongyeon Kim et al. "Surch: Enabling Structural Search and Comparison for Surgical Videos". In: *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. CHI '23. Hamburg, Germany: Association for Computing Machinery, 2023. ISBN: 9781450394215. DOI: 10.1145/3544548.3580772. URL: https://doi.org/10.1145/3544548.3580772.

[252] Tae Soo Kim et al. "Stylette: Styling the Web with Natural Language". In: *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. CHI '22. New Orleans, LA, USA: Association for Computing Machinery, 2022. ISBN: 9781450391573. DOI: 10.1145/3491102.3501931. URL: https://doi.org/10.1145/3491102.3501931.

[253] Margaret King and Bente Maeggard. "Issues in Natural Language Systems Evaluation." In: *LREC*. 1998.

[254] David L. Kirp. "Race, Schooling, and Interest Politics: The Oakland Story". In: *The School Review* 87.4 (1979), pp. 355–397. ISSN: 00366773. URL: http://www.jstor.org/stable/1084730.

[255] Chunyu Kit and Tak Ming Wong. "Comparative Evaluation of Online Machine Translation Systems with Legal Texts". In: *Law Library Journal* 100 (2008).

[256] René F Kizilcec. "How much information? Effects of transparency on trust in an algorithmic interface". In: *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. 2016, pp. 2390–2395.

[257] Ondrej Klejch et al. "MT-ComparEval: Graphical Evaluation Interface for Machine Translation Development." In: *Prague Bull. Math. Linguistics* 104 (2015). DOI: `10.1515/pralin-2015-0014`.

[258] Amy J. Ko. "Seeking Information". In: *Foundations of Information*. 2021. URL: `https://faculty.washington.edu/ajko/books/foundations-of-information/`.

[259] 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*. 2019, pp. 1–14.

[260] Philipp Koehn. "EuroMatrix – Machine Translation for All European Languages". In: *Proceedings of Machine Translation Summit XI: Invited Papers*. 2007.

[261] Philipp Koehn. "Statistical Significance Tests for Machine Translation Evaluation". In: *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*. Barcelona, Spain: Association for Computational Linguistics, July 2004, pp. 388–395. URL: `https://aclanthology.org/W04-3250`.

[262] Philipp Koehn and Rebecca Knowles. "Six Challenges for Neural Machine Translation". In: *Proceedings of the First Workshop on Neural Machine Translation*. Vancouver: Association for Computational Linguistics, Aug. 2017, pp. 28–39. DOI: `10.18653/v1/W17-3204`. URL: `https://aclanthology.org/W17-3204`.

[263] Allison Koenecke et al. "Racial Disparities in Automated Speech Recognition". In: *Proceedings of the National Academy of Sciences* 117 (Apr. 2020). DOI: `10.1073/pnas.1915768117`.

[264] Pang Wei Koh et al. "WILDS: A Benchmark of in-the-Wild Distribution Shifts". In: *ICML*. Vol. 139. Proceedings of Machine Learning Research. July 2021.

[265] James Rufus Koren. "Some lenders are judging you on much more than finances". In: (2015). URL: `https://www.latimes.com/business/la-fi-new-credit-score-20151220-story.html`.

[266] John Kretzmann and John P. McKnight. "Assets-based community development". In: *National Civic Review* 85.4 (1996), pp. 23–29. DOI: `https://doi.org/10.1002/ncr.4100850405`. eprint: `https://onlinelibrary.wiley.com/doi/pdf/10.1002/ncr.4100850405`. URL: `https://onlinelibrary.wiley.com/doi/abs/10.1002/ncr.4100850405`.

[267] Arvindkumar Krishnakumar et al. "UDIS: Unsupervised Discovery of Bias in Deep Visual Recognition Models". In: *arxiv:2110.15499* (Oct. 2021).

[268] Amanda Kube et al. "Just Resource Allocation? How Algorithmic Predictions and Human Notions of Justice Interact". In: *Proceedings of the 23rd ACM Conference on Economics and Computation*. EC '22. Boulder, CO, USA: Association for Computing Machinery, 2022, pp. 1184–1242. ISBN: 9781450391504. DOI: `10.1145/3490486.3538305`. URL: `https://doi.org/10.1145/3490486.3538305`.

[269] T. Kulesza et al. "Too much, too little, or just right? Ways explanations impact end users' mental models". In: *2013 IEEE Symposium on Visual Languages and Human Centric Computing*. 2013, pp. 3–10.

[270] Sunjun Kweon et al. *Publicly Shareable Clinical Large Language Model Built on Synthetic Clinical Notes*. 2023. arXiv: `2309.00237 [cs.CL]`.

[271] Nina Lakhani. *America's year of hunger: how children and people of color suffered most*. The Guardian. 2020. URL: `https://www.theguardian.com/environment/2021/apr/14/americas-year-of-hunger-how-children-and-people-of-color-suffered-most`.

[272] Mariana Laverde et al. *Distance to Schools and Equal Access in School Choice Systems*. Tech. rep. 2022.

[273] Mariana Laverde. "Unequal Assignments to Public Schools and the Limits of School Choice". In: 2020.

[274] Jonathan Lazar, Jinjuan Heidi Feng, and Harry Hochheiser. "Chapter 9 - Ethnography". In: *Research Methods in Human Computer Interaction (Second Edition)*. Ed. by Jonathan Lazar, Jinjuan Heidi Feng, and Harry Hochheiser. Second Edition. Boston: Morgan Kaufmann, 2017, pp. 229–261. ISBN: 978-0-12-805390-4. DOI: `https://doi.org/10.1016/B978-0-12-805390-4.00009-1`. URL: `https://www.sciencedirect.com/science/article/pii/B9780128053904000091`.

[275] Christopher A. Le Dantec, Erika Shehan Poole, and Susan P. Wyche. "Values as Lived Experience: Evolving Value Sensitive Design in Support of Value Discovery". In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. CHI '09. Boston, MA, USA: Association for Computing Machinery, 2009, pp. 1141–1150. ISBN: 9781605582467. URL: `https://doi.org/10.1145/1518701.1518875`.

[276] Jaesong Lee, Joong-Hwi Shin, and Jun-Seok Kim. "Interactive Visualization and Manipulation of Attention-based Neural Machine Translation". In: *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. 2017.

[277] Min Kyung Lee and Su Baykal. "Algorithmic Mediation in Group Decisions: Fairness Perceptions of Algorithmically Mediated vs. Discussion-Based Social Division". In: *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*. CSCW '17. Portland, Oregon, USA: Association for Computing Machinery, 2017, pp. 1035–1048. ISBN: 9781450343350. DOI: `10.1145/2998181.2998230`. URL: `https://doi.org/10.1145/2998181.2998230`.

[278] Min Kyung Lee, Ji Tae Kim, and Leah Lizarondo. “A Human-Centered Approach to Algorithmic Services: Considerations for Fair and Motivating Smart Community Service Management That Allocates Donations to Non-Profit Organizations”. In: *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*. CHI ’17. Denver, Colorado, USA: Association for Computing Machinery, 2017, pp. 3365–3376. ISBN: 9781450346559. DOI: `10.1145/3025453.3025884`. URL: `https://doi.org/10.1145/3025453.3025884`.

[279] Min Kyung Lee, Ji Tae Kim, and Leah Lizarondo. “A Human-Centered Approach to Algorithmic Services: Considerations for Fair and Motivating Smart Community Service Management That Allocates Donations to Non-Profit Organizations”. In: *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*. CHI ’17. Denver, Colorado, USA: Association for Computing Machinery, 2017, pp. 3365–3376. ISBN: 9781450346559. DOI: `10.1145/3025453.3025884`. URL: `https://doi.org/10.1145/3025453.3025884`.

[280] Min Kyung Lee et al. “Participatory Algorithmic Management: Elicitation Methods for Worker Well-Being Models”. In: *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*. AIES ’21. Virtual Event, USA: Association for Computing Machinery, 2021, pp. 715–726. ISBN: 9781450384735. DOI: `10.1145/3461702.3462628`. URL: `https://doi.org/10.1145/3461702.3462628`.

[281] Min Kyung Lee et al. “Procedural Justice in Algorithmic Fairness: Leveraging Transparency and Outcome Control for Fair Algorithmic Mediation”. In: *Proc. ACM Hum.-Comput. Interact.* 3.CSCW (Nov. 2019). DOI: `10.1145/3359284`. URL: `https://doi.org/10.1145/3359284`.

[282] Min Kyung Lee et al. “WeBuildAI: Participatory Framework for Algorithmic Governance”. In: *Proc. ACM Hum.-Comput. Interact.* 3.CSCW (Nov. 2019). DOI: `10.1145/3359283`. URL: `https://doi.org/10.1145/3359283`.

[283] Seongmin Lee et al. “VisCUIT: Visual Auditor for Bias in CNN Image Classifier”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. June 2022.

[284] Yoonjoo Lee et al. “Promptiverse: Scalable Generation of Scaffolding Prompts Through Human-AI Hybrid Knowledge Graph Annotation”. In: *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. CHI ’22. New Orleans, LA, USA: Association for Computing Machinery, 2022. ISBN: 9781450391573. DOI: `10.1145/3491102.3502087`. URL: `https://doi.org/10.1145/3491102.3502087`.

[285] Patrick S. H. Lewis et al. “Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks”. In: *34th Conference on Neural Information Processing Systems*. NeurIPS 2020. Vancouver, Canada, 2020.

[286] Toby Jia-Jun Li et al. "Multi-Modal Repairs of Conversational Breakdowns in Task-Oriented Dialogs". In: *Proceedings of the 33rd Annual ACM Symposium on User Interface Software and Technology*. UIST '20. Virtual Event, USA: Association for Computing Machinery, 2020, pp. 1094–1107. ISBN: 9781450375146. DOI: 10.1145/3379337.3415820. URL: https://doi.org/10.1145/3379337.3415820.

[287] Zhen Li et al. "A Unified Understanding of Deep NLP Models for Text Classification". In: *IEEE Transactions on Visualization and Computer Graphics* (2022). DOI: 10.1109/tvcg.2022.3184186.

[288] Calvin A. Liang, Sean A. Munson, and Julie A. Kientz. "Embracing Four Tensions in Human-Computer Interaction Research with Marginalized People". In: *ACM Trans. Comput.-Hum. Interact.* 28.2 (Apr. 2021). ISSN: 1073-0516. DOI: 10.1145/3443686. URL: https://doi.org/10.1145/3443686.

[289] 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*. CHI '20. Honolulu, HI, USA: Association for Computing Machinery, 2020, pp. 1–15. ISBN: 9781450367080. DOI: 10.1145/3313831.3376590. URL: https://doi.org/10.1145/3313831.3376590.

[290] Daniel J. Liebling et al. "Unmet Needs and Opportunities for Mobile Translation AI". In: *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. New York, NY, USA: Association for Computing Machinery, 2020, pp. 1–13. ISBN: 9781450367080. URL: https://doi.org/10.1145/3313831.3376261.

[291] Hajin Lim, Dan Cosley, and Susan R. Fussell. "Beyond Translation: Design and Evaluation of an Emotional and Contextual Knowledge Interface for Foreign Language Social Media Posts". In: *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. CHI '18. Montreal QC, Canada: Association for Computing Machinery, Apr. 2018, pp. 1–12. ISBN: 978-1-4503-5620-6. DOI: 10.1145/3173574.3173791. URL: https://doi.org/10.1145/3173574.3173791 (visited on 06/03/2020).

[292] Chin-Yew Lin and Franz Josef Och. "ORANGE: A Method for Evaluating Automatic Evaluation Metrics for Machine Translation". In: *COLING 2004: Proceedings of the 20th International Conference on Computational Linguistics*. Aug. 2004. DOI: 10.3115/1220355.1220427.

[293] Zachary Chase Lipton. "The Mythos of Model Interpretability". In: *CoRR* (2016). arXiv: 1606.03490. URL: http://arxiv.org/abs/1606.03490.

[294] Geoffrey Litt. "Malleable software in the age of LLMs". In: (Mar. 2023). URL: https://www.geoffreylitt.com/2023/03/25/llm-end-user-programming.html.

[295] Lydia T. Liu et al. “Reimagining the machine learning life cycle to improve educational outcomes of students”. In: *Proceedings of the National Academy of Sciences* 120.9 (2023), e2204781120. DOI: `10.1073/pnas.2204781120`. eprint: `https://www.pnas.org/doi/pdf/10.1073/pnas.2204781120`. URL: `https://www.pnas.org/doi/abs/10.1073/pnas.2204781120`.

[296] Shixia Liu et al. “Visual Diagnosis of Tree Boosting Methods”. In: *IEEE TVCG* 24 (Jan. 2018). DOI: `10.1109/tvcg.2017.2744378`.

[297] Duri Long and Brian Magerko. “What is AI Literacy? Competencies and Design Considerations”. In: *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. CHI ’20. Honolulu, HI, USA: Association for Computing Machinery, 2020, pp. 1–16. ISBN: 9781450367080. DOI: `10.1145/3313831.3376727`. URL: `https://doi.org/10.1145/3313831.3376727`.

[298] Edward Loper and Steven Bird. “NLTK: The Natural Language Toolkit”. In: *Proceedings of the ACL-02 Workshop on Effective Tools and Methodologies for Teaching Natural Language Processing and Computational Linguistics* -. Vol. 1. 2002. DOI: `10.3115/1219044.1219075`.

[299] Charles Lovering and Ellie Pavlick. “Unit Testing for Concepts in Neural Networks”. In: *arxiv:2208.10244* (July 2022).

[300] Margaux Luflade. *The value of information in centralized school choice systems*. Working Paper. 2017. URL: `https://economics.sas.upenn.edu/sites/default/files/filevault/event_papers/LUFLADE_JobMarketPaper.pdf`.

[301] Niklas Luhmann. “Trust: A mechanism for the reduction of social complexity”. In: *Trust and power: Two works by Niklas Luhmann* (1979), pp. 1–103.

[302] Scott M. Lundberg and Su-In Lee. “A Unified Approach to Interpreting Model Predictions”. In: *Proceedings of the 31st International Conference on Neural Information Processing Systems*. NIPS’17. 2017.

[303] Hossin M and Sulaiman M.N. “A Review on Evaluation Metrics for Data Classification Evaluations”. In: *International Journal of Data Mining & Knowledge Management Process* 5 (Mar. 2015). DOI: `10.5121/ijdkp.2015.5201`.

[304] Vivien Macketanz et al. “Fine-Grained Evaluation of German-English Machine Translation Based on a Test Suite”. In: *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*. Oct. 2018. DOI: `10.18653/v1/w18-6436`.

[305] Michael A. Madaio et al. “”Everyone Brings Their Grain of Salt”: Designing for Low-Literate Parental Engagement with a Mobile Literacy Technology in CôTe d’Ivoire”. In: *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. CHI ’19. Glasgow, Scotland Uk: Association for Computing Machinery, 2019, pp. 1–15. ISBN: 9781450359702. DOI: `10.1145/3290605.3300695`. URL: `https://doi.org/10.1145/3290605.3300695`.

[306] Nitin Madnani. “iBLEU: Interactively Debugging and Scoring Statistical Machine Translation Systems”. In: *2011 IEEE Fifth International Conference on Semantic Computing*. Sept. 2011. DOI: 10.1109/icsc.2011.36.

[307] Lev Malevanchik et al. “Disparities After Discharge: The Association of Limited English Proficiency and Postdischarge Patient-Reported Issues”. In: *The Joint Commission Journal on Quality and Patient Safety* 47.12 (2021). DOI: 10.1016/j.jcjq.2021.08.013.

[308] Jonathan Mallinson et al. “EdiT5: Semi-Autoregressive Text Editing with T5 Warm-Start”. In: *Findings of the Association for Computational Linguistics: EMNLP 2022*. Ed. by Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang. Abu Dhabi, United Arab Emirates: Association for Computational Linguistics, Dec. 2022, pp. 2126–2138. DOI: 10.18653/v1/2022.findings-emnlp.156. URL: https://aclanthology.org/2022.findings-emnlp.156.

[309] Jonathan Mallinson et al. “FELIX: Flexible Text Editing Through Tagging and Insertion”. In: *Findings of the Association for Computational Linguistics: EMNLP 2020*. Online: Association for Computational Linguistics, Nov. 2020, pp. 1244–1255. DOI: 10.18653/v1/2020.findings-emnlp.111. URL: https://aclanthology.org/2020.findings-emnlp.111.

[310] Eric Malmi et al. “Encode, Tag, Realize: High-Precision Text Editing”. In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Ed. by Kentaro Inui et al. Hong Kong, China: Association for Computational Linguistics, Nov. 2019, pp. 5054–5065. DOI: 10.18653/v1/D19-1510. URL: https://aclanthology.org/D19-1510.

[311] Noëmi Manders-Huits. “What Values in Design? The Challenge of Incorporating Moral Values into Design”. en. In: *Science and Engineering Ethics* 17.2 (June 2011), pp. 271–287. ISSN: 1471-5546. DOI: 10.1007/s11948-010-9198-2. URL: https://doi.org/10.1007/s11948-010-9198-2 (visited on 09/04/2020).

[312] Marianna Martindale and Marine Carpuat. “Fluency Over Adequacy: A Pilot Study in Measuring User Trust in Imperfect MT”. In: *Proceedings of the 13th Conference of the Association for Machine Translation in the Americas (Volume 1: Research Papers)*. Boston, MA, USA: Association for Machine Translation in the Americas, Mar. 2018, pp. 13–25. URL: https://www.aclweb.org/anthology/W18-1803 (visited on 07/22/2020).

[313] Alison Mathie and Gord Cunningham. “From clients to citizens: Asset-based Community Development as a strategy for community-driven development”. In: *Development in Practice* 13.5 (2003), pp. 474–486. DOI: 10.1080/0961452032000125857. eprint: https://doi.org/10.1080/0961452032000125857. URL: https://doi.org/10.1080/0961452032000125857.

[314] Nitika Mathur, Timothy Baldwin, and Trevor Cohn. “Tangled up in BLEU: Reevaluating the Evaluation of Automatic Machine Translation Evaluation Metrics”. In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. 2020. DOI: `10.18653/v1/2020.acl-main.448`.

[315] L. McInnes, J. Healy, and J. Melville. “UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction”. In: *ArXiv e-prints* (Feb. 2018).

[316] Nikita Mehandru, Samantha Robertson, and Niloufar Salehi. “Reliable and Safe Use Machine Translation in Medical Settings”. In: *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*. FAccT ’22. Seoul, South Korea: Association for Computing Machinery, 2022.

[317] Nikita Mehandru et al. “Physician Detection of Clinical Harm in Machine Translation: Quality Estimation Aids in Reliance and Backtranslation Identifies Critical Errors”. In: *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. Ed. by Houda Bouamor, Juan Pino, and Kalika Bali. Singapore: Association for Computational Linguistics, Dec. 2023, pp. 11633–11647. URL: `https://aclanthology.org/2023.emnlp-main.712`.

[318] Sharan B Merriam and Associates. “Introduction to qualitative research”. In: *Qualitative research in practice: Examples for discussion and analysis*. Jossey-Bass, 2002, pp. 1–17.

[319] Tim Miller. “Explanation in artificial intelligence: Insights from the social sciences”. en. In: *Artificial Intelligence* 267 (Feb. 2019), pp. 1–38. ISSN: 0004-3702. DOI: `10.1016/j.artint.2018.07.007`. URL: `http://www.sciencedirect.com/science/article/pii/S0004370218305988` (visited on 11/09/2020).

[320] Teruko Mitamura and Eric H. Nyberg 3rd. “Controlled English for Knowledge-Based MT: Experience with the KANT System”. In: *Proceedings of the Sixth Conference on Theoretical and Methodological Issues in Machine Translation of Natural Languages*. Katholieke Universiteit, Leuven, July 1995. URL: `https://aclanthology.org/1995.tmi-1.12`.

[321] Margaret Mitchell et al. “Model Cards for Model Reporting”. In: *Proceedings of the Conference on Fairness, Accountability, and Transparency*. FAT* ’19. Atlanta, GA, USA: Association for Computing Machinery, 2019, pp. 220–229. ISBN: 9781450361255. DOI: `10.1145/3287560.3287596`. URL: `https://doi.org/10.1145/3287560.3287596`.

[322] Mai Miyabe and Takashi Yoshino. “Can Indicating Translation Accuracy Encourage People to Rectify Inaccurate Translations?” In: *Human-Computer Interaction. Interaction Techniques and Environments*. Ed. by Julie A. Jacko. Berlin, Heidelberg: Springer Berlin Heidelberg, 2011, pp. 368–377. ISBN: 978-3-642-21605-3.

[323] Mai Miyabe and Takashi Yoshino. "Influence of Detecting Inaccurate Messages in Real-Time Remote Text-Based Communication via Machine Translation". In: *Proceedings of the 3rd International Conference on Intercultural Collaboration*. ICIC '10. Copenhagen, Denmark: Association for Computing Machinery, 2010, pp. 59–68. ISBN: 9781450301084. DOI: `10.1145/1841853.1841863`. URL: `https://doi.org/10.1145/1841853.1841863`.

[324] Mai Miyabe, Takashi Yoshino, and Tomohiro Shigenobu. "Effects of Repair Support Agent for Accurate Multilingual Communication". In: *PRICAI 2008: Trends in Artificial Intelligence*. Ed. by Tu-Bao Ho and Zhi-Hua Zhou. Berlin, Heidelberg: Springer Berlin Heidelberg, 2008, pp. 1022–1027. ISBN: 978-3-540-89197-0.

[325] Mai Miyabe, Takashi Yoshino, and Tomohiro Shigenobu. "Effects of Undertaking Translation Repair Using Back Translation". In: *Proceedings of the 2009 International Workshop on Intercultural Collaboration*. IWIC '09. Palo Alto, California, USA: Association for Computing Machinery, 2009, pp. 33–40. ISBN: 9781605585024. DOI: `10.1145/1499224.1499232`. URL: `https://doi.org/10.1145/1499224.1499232`.

[326] Tom Moberly. "Doctors are cautioned against using Google Translate in consultations". In: *BMJ* 363 (2018). ISSN: 0959-8138. DOI: `10.1136/bmj.k4546`. eprint: `https://www.bmj.com/content/363/bmj.k4546.full.pdf`. URL: `https://www.bmj.com/content/363/bmj.k4546`.

[327] Tom Moberly. "Doctors choose Google Translate to communicate with patients because of easy access". In: *BMJ* 362 (2018). ISSN: 0959-8138. DOI: `10.1136/bmj.k3974`. eprint: `https://www.bmj.com/content/362/bmj.k3974.full.pdf`. URL: `https://www.bmj.com/content/362/bmj.k3974`.

[328] Luis C. Moll et al. "Funds of Knowledge for Teaching: Using a Qualitative Approach to Connect Homes and Classrooms". In: *Theory Into Practice* 31.2 (1992), pp. 132–141. ISSN: 00405841, 15430421. URL: `http://www.jstor.org/stable/1476399`.

[329] Joaquim Moré and Salvador Climent. "Machine Translationness: Machine-likeness in Machine Translation Evaluation". In: *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*. May 2014.

[330] Michael J Muller and Sarah Kuhn. "Participatory design". In: *Communications of the ACM* 36.6 (1993), pp. 24–28.

[331] Michael J. Muller and Allison Druin. "Participatory Design: The Third Space in HCI". In: *The Human-Computer Interaction Handbook: Fundamentals, Evolving Technologies and Emerging Applications*. 3rd ed. USA: L. Erlbaum Associates Inc., 2012, pp. 1125–1154.

[332] David Munechika et al. "Visual Auditor: Interactive Visualization for Detection and Summarization of Model Biases". In: *2022 IEEE Visualization Conference (VIS)*. 2022.

[333] Tanja Munz et al. "Visualization-Based Improvement of Neural Machine Translation". In: *Computers & Graphics* 103 (Apr. 2022). DOI: 10.1016/j.cag.2021.12.003.

[334] Aakanksha Naik et al. "Stress Test Evaluation for Natural Language Inference". In: *The 27th International Conference on Computational Linguistics (COLING)*. Aug. 2018.

[335] Anna M. Nápoles et al. "Inaccurate Language Interpretation and its Clinical Significance in the Medical Encounters of Spanish-speaking Latinos". In: *Medical care* 53.11 (Nov. 2015), pp. 940–947. ISSN: 0025-7079. DOI: 10.1097/MLR.0000000000000422. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4610127/ (visited on 10/04/2022).

[336] Yusuke Narita. *Match or Mismatch? Learning and Inertia in School Choice*. 2018. URL: https://ssrn.com/abstract=3198417.

[337] Wilhelmina Nekoto et al. "Participatory Research for Low-resourced Machine Translation: A Case Study in African Languages". In: *Findings of the Association for Computational Linguistics: EMNLP 2020*. Online: Association for Computational Linguistics, Nov. 2020, pp. 2144–2160. DOI: 10.18653/v1/2020.findings-emnlp.195. URL: https://www.aclweb.org/anthology/2020.findings-emnlp.195.

[338] Graham Neubig et al. "Compare-Mt: A Tool for Holistic Comparison of Language Generation Systems". In: *Proceedings of the 2019 Conference of the North*. 2019. DOI: 10.18653/v1/n19-4007.

[339] ThÃnh Nguyen and Rakesh Vohra. "Stable Matching with Proportionality Constraints". In: *Operations Research* 67.6 (2019), pp. 1503–1519. DOI: 10.1287/opre.2019.1909. URL: https://doi.org/10.1287/opre.2019.1909.

[340] Jakob Nielsen. *10 Usability Heuristics for User Interface Design*. Apr. 1994. URL: nngroup.com/articles/ten-usability-heuristics/.

[341] Helen Nissenbaum. "A Contextual Approach to Privacy Online". In: *Daedalus* 140.4 (2011), pp. 32–48.

[342] Safiya Umoja Noble. *Algorithms of Oppression: How Search Engines Reinforce Racism*. NYU Press, 2018.

[343] Ritesh Noothigattu et al. "A Voting-Based System for Ethical Decision Making". In: 2018. URL: https://www.aaai.org/ocs/index.php/AAAI/AAAI18/paper/view/17052/15857.

[344] Beau Norgeot et al. "Protected Health Information filter (Philter): accurately and securely de-identifying free-text clinical notes". In: *npj Digital Medicine* 3.57 (2020). URL: https://doi.org/10.1038/s41746-020-0258-y.

[345] Donald Norman. "The Design of Everyday Things: Revised and Expanded Edition". In: (2013).

[346] Chris Norval et al. "Disclosure by Design: Designing Information Disclosures to Support Meaningful Transparency and Accountability". In: *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency.* FAccT '22. Seoul, Republic of Korea: Association for Computing Machinery, 2022, pp. 679–690. ISBN: 9781450393522. DOI: 10.1145/3531146.3533133. URL: https://doi.org/10.1145/3531146.3533133.

[347] Cathy O'Neil. *Weapons of Math Destruction.* Crown Books, 2016.

[348] *Oakland Unified.* EdData. URL: http://www.ed-data.org/district/Alameda/Oakland-Unified.

[349] Ziad Obermeyer et al. "Dissecting Racial Bias in an Algorithm Used to Manage the Health of Populations". In: *Science* 366 (Oct. 2019). DOI: 10.1126/science.aax2342.

[350] Ihudiya Finda Ogbonnaya-Ogburu, Kentaro Toyama, and Tawanna R. Dillahunt. "Towards an Effective Digital Literacy Intervention to Assist Returning Citizens with Job Search". In: *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems.* CHI '19. Glasgow, Scotland Uk: Association for Computing Machinery, 2019, pp. 1–12. ISBN: 9781450359702. DOI: 10.1145/3290605.3300315. URL: https://doi.org/10.1145/3290605.3300315.

[351] Ihudiya Finda Ogbonnaya-Ogburu et al. "Critical Race Theory for HCI". In: *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems.* CHI '20. Honolulu, HI, USA: Association for Computing Machinery, 2020, pp. 1–16. ISBN: 9781450367080. DOI: 10.1145/3313831.3376392. URL: https://doi.org/10.1145/3313831.3376392.

[352] Jorge Piazentin Ono et al. "PipelineProfiler: A Visual Analytics Tool for the Exploration of AutoML Pipelines". In: *IEEE Transactions on Visualization and Computer Graphics* 27 (Feb. 2021).

[353] Hessel Oosterbeek, Sándor Sóvágó, and Bas Klaauw. "Why are Schools Segregated? Evidence from the Secondary-School Match in Amsterdam". In: *CEPR Discussion Paper No. DP13462* (2019). URL: https://ssrn.com/abstract=3319783.

[354] Gary Orfield. *Schools more separate: Consequences of a decade of resegregation.* Cambridge, MA: The Civil Rights Project, Harvard University, 2001.

[355] Google PAIR. "Explainability + Trust". In: *People + AI Guidebook.* 2019. URL: https://pair.withgoogle.com/chapter/explainability-trust/.

[356] Google PAIR. *People + AI Guidebook.* 2019. URL: https://pair.withgoogle.com/guidebook/.

[357] Joana Pais and Agnes Pintér. "School choice and information: An experimental study on matching mechanisms". In: *Games and Economic Behavior* 64.1 (2008), pp. 303–328.

[358] Anita Panayiotou et al. “The perceptions of translation apps for everyday health care in healthcare workers and older people: A multi-method study”. In: *Journal of Clinical Nursing* 29.17-18 (Sept. 2020), pp. 3516–3526. DOI: `10.1111/jocn.15390`.

[359] A. Pandey, A. Srinivasan, and V. Setlur. “MEDLEY: Intent-based Recommendations to Support Dashboard Composition”. In: *IEEE Transactions on Visualization & Computer Graphics* 29.01 (Jan. 2023), pp. 1135–1145. ISSN: 1941-0506. DOI: `10.1109/TVCG.2022.3209421`.

[360] Kishore Papineni et al. “Bleu: A Method for Automatic Evaluation of Machine Translation”. In: *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*. 2002.

[361] Hyanghee Park et al. “Designing Fair AI in Human Resource Management: Understanding Tensions Surrounding Algorithmic Evaluation and Envisioning Stakeholder-Centered Solutions”. In: *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. CHI ’22. New Orleans, LA, USA: Association for Computing Machinery, 2022. ISBN: 9781450391573. DOI: `10.1145/3491102.3517672`. URL: `https://doi.org/10.1145/3491102.3517672`.

[362] Joon Sung Park et al. “Power Dynamics and Value Conflicts in Designing and Maintaining Socio-Technical Algorithmic Processes”. In: *Proc. ACM Hum.-Comput. Interact.* 6.CSCW1 (Apr. 2022). DOI: `10.1145/3512957`. URL: `https://doi.org/10.1145/3512957`.

[363] Joon Sung Park et al. “Social Simulacra: Creating Populated Prototypes for Social Computing Systems”. In: *Proceedings of the 35th Annual ACM Symposium on User Interface Software and Technology*. 2022, pp. 1–18.

[364] Sebeom Park et al. “VANT: A Visual Analytics System for Refining Parallel Corpora in Neural Machine Translation”. In: *2022 IEEE 15th Pacific Visualization Symposium (PacificVis)*. Apr. 2022. DOI: `10.1109/pacificvis53943.2022.00029`.

[365] Kayur Patel et al. “Investigating Statistical Machine Learning as a Tool for Software Development”. In: *Proceeding of the Twenty-Sixth Annual CHI Conference on Human Factors in Computing Systems - CHI ’08*. 2008. DOI: `10.1145/1357054.1357160`.

[366] Parag A. Pathak. “What Really Matters in Designing School Choice Mechanisms”. In: *Advances in Economics and Econometrics: Eleventh World Congress*. Ed. by Bo Honoré et al. Vol. 1. Econometric Society Monographs. Cambridge University Press, 2017, pp. 176–214. DOI: `10.1017/9781108227162.006`.

[367] Karl Pearson. “On Lines and Planes of Closest Fit to Systems of Points in Space”. In: *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science* 2 (1901). DOI: `10.1080/14786440109462720`. URL: `https://www.tandfonline.com/doi/full/10.1080/14786440109462720` (visited on 11/16/2022).

[368] Fabian Pedregosa et al. “Scikit-Learn: Machine Learning in Python”. In: *the Journal of machine Learning research* 12 (2011).

[369] Lucy Pei and Bonnie Nardi. "We Did It Right, But It Was Still Wrong: Toward Assets-Based Design". In: *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems*. CHI EA '19. Glasgow, Scotland Uk: Association for Computing Machinery, 2019, pp. 1–11. ISBN: 9781450359719. DOI: 10.1145/3290607.3310434. URL: https://doi.org/10.1145/3290607.3310434.

[370] Lisa Pickoff-White. *S.F.'s Kindergarten Lottery: Do Parents' Tricks Work?* KQED. 2018. URL: https://www.kqed.org/news/11641019/s-f-s-kindergarten-lottery-do-parents-tricks-work.

[371] Nichole Pinkard et al. "Digital Youth Divas: Exploring Narrative-Driven Curriculum to Spark Middle School Girls' Interest in Computational Activities". In: *Journal of the Learning Sciences* 26.3 (2017), pp. 477–516. DOI: 10.1080/10508406.2017.1307199. eprint: https://doi.org/10.1080/10508406.2017.1307199. URL: https://doi.org/10.1080/10508406.2017.1307199.

[372] Lindsay Poirier. "Accountable Data: The Politics and Pragmatics of Disclosure Datasets". In: *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*. FAccT '22. Seoul, Republic of Korea: Association for Computing Machinery, 2022, pp. 1446–1456. ISBN: 9781450393522. DOI: 10.1145/3531146.3533201. URL: https://doi.org/10.1145/3531146.3533201.

[373] Neoklis Polyzotis et al. "Data Validation for Machine Learning". In: *Proceedings of Machine Learning and Systems*. Vol. 1. 2019.

[374] Maja Popović. "Agree to Disagree: Analysis of Inter-Annotator Disagreements in Human Evaluation of Machine Translation Output". In: *Proceedings of the 25th Conference on Computational Natural Language Learning*. Nov. 2021. DOI: 10.18653/v1/2021.conll-1.18.

[375] Maja Popović. "chrF: character n-gram F-score for automatic MT evaluation". In: *Proceedings of the Tenth Workshop on Statistical Machine Translation*. Lisbon, Portugal: Association for Computational Linguistics, Sept. 2015, pp. 392–395. DOI: 10.18653/v1/W15-3049. URL: https://aclanthology.org/W15-3049.

[376] Maja Popović and Sheila Castilho. "Challenge Test Sets for MT Evaluation". In: *Proceedings of Machine Translation Summit XVII: Tutorial Abstracts*. Aug. 2019.

[377] Matt Post. "A Call for Clarity in Reporting BLEU Scores". In: *Proceedings of the Third Conference on Machine Translation: Research Papers*. 2018. DOI: 10.18653/v1/w18-6319.

[378] Forough Poursabzi-Sangdeh et al. "Manipulating and Measuring Model Interpretability". In: *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. CHI '21. ¡conf-loc¿, ¡city¿Yokohama¡/city¿, ¡country¿Japan¡/country¿, ¡/conf-loc¿: Association for Computing Machinery, 2021. ISBN: 9781450380966. DOI: 10.1145/3411764.3445315. URL: https://doi.org/10.1145/3411764.3445315.

[379] Vinay Prabhu et al. "Did They Direct the Violence or Admonish It? A Cautionary Tale on Contronomy, Androcentrism and Back-Translation Foibles". In: *AfricaNLP Workshop at EACL*. 2021.

[380] Mahima Pushkarna, Andrew Zaldivar, and Oddur Kjartansson. "Data Cards: Purposeful and Transparent Dataset Documentation for Responsible AI". In: *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*. FAccT '22. Seoul, Republic of Korea: Association for Computing Machinery, 2022, pp. 1776–1826. ISBN: 9781450393522. DOI: `10.1145/3531146.3533231`. URL: `https://doi.org/10.1145/3531146.3533231`.

[381] Chen Qu et al. "Analyzing and characterizing user intent in information-seeking conversations". In: *The 41st international acm sigir conference on research & development in information retrieval*. 2018, pp. 989–992.

[382] Emilee Rader, Kelley Cotter, and Janghee Cho. "Explanations as Mechanisms for Supporting Algorithmic Transparency". In: *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. CHI '18. Montreal QC, Canada: Association for Computing Machinery, 2018, pp. 1–13. ISBN: 9781450356206. DOI: `10.1145/3173574.3173677`. URL: `https://doi.org/10.1145/3173574.3173677`.

[383] Alec Radford et al. "Learning Transferable Visual Models from Natural Language Supervision". In: *Proceedings of the 38th International Conference on Machine Learning*. Vol. 139. Proceedings of Machine Learning Research. 2021. URL: `https://proceedings.mlr.press/v139/radford21a.html`.

[384] Alessandro Raganato, Yves Scherrer, and Jörg Tiedemann. "The MuCoW Test Suite at WMT 2019: Automatically Harvested Multilingual Contrastive Word Sense Disambiguation Test Sets for Machine Translation". In: *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*. Aug. 2019. DOI: `10.18653/v1/w19-5354`.

[385] Manish Raghavan and Solon Barocas. "Challenges for mitigating bias in algorithmic hiring". In: (2019). URL: `https://www.brookings.edu/research/challenges-for-mitigating-bias-in-algorithmic-hiring/`.

[386] Deborah Raji et al. "AI and the Everything in the Whole Wide World Benchmark". In: *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*. Ed. by J. Vanschoren and S. Yeung. Vol. 1. Curran, 2021. URL: `https://datasets-benchmarks-proceedings.neurips.cc/paper_files/paper/2021/file/084b6fbb10729ed4da8c3d3f5a3ae7c9-Paper-round2.pdf`.

[387] Inioluwa Deborah Raji and Joy Buolamwini. "Actionable Auditing: Investigating the Impact of Publicly Naming Biased Performance Results of Commercial AI Products". In: *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*. AIES '19. Honolulu, HI, USA: Association for Computing Machinery, 2019, pp. 429–435.

ISBN: 9781450363242. DOI: 10.1145/3306618.3314244. URL: https://doi.org/10.1145/3306618.3314244.

[388] Inioluwa Deborah Raji et al. "Closing the AI Accountability Gap: Defining an End-to-End Framework for Internal Algorithmic Auditing". In: *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. FAT* '20. Barcelona, Spain: Association for Computing Machinery, 2020, pp. 33–44. ISBN: 9781450369367. DOI: 10.1145/3351095.3372873. URL: https://doi.org/10.1145/3351095.3372873.

[389] Inioluwa Deborah Raji et al. "Saving Face: Investigating the Ethical Concerns of Facial Recognition Auditing". In: *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*. AIES '20. 2020. DOI: 10.1145/3375627.3375820.

[390] Pranav Rajpurkar, Robin Jia, and Percy Liang. "Know What You Don't Know: Unanswerable Questions for SQuAD". In: *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. July 2018. DOI: 10.18653/v1/p18-2124.

[391] Gurdeeshpal Randhawa et al. "Using machine translation in clinical practice". In: *Canadian Family Physician* 59.4 (2013), pp. 382–383.

[392] *Real Integration*. IntegrateNYC. URL: https://integratenyc.org/mission.

[393] Sylvestre-Alvise Rebuffi et al. "Data Augmentation Can Improve Robustness". In: *Advances in Neural Information Processing Systems*. Vol. 34. 2021.

[394] Benjamin Recht et al. "Do ImageNet Classifiers Generalize to ImageNet?" In: *Proceedings of the 36th International Conference on Machine Learning*. Vol. 97. Proceedings of Machine Learning Research. June 2019.

[395] Alex Rees-Jones and Samuel Skowronek. "An experimental investigation of preference misrepresentation in the residency match". In: *Proceedings of the National Academy of Sciences* 115.45 (2018), pp. 11471–11476. ISSN: 0027-8424. DOI: 10.1073/pnas.1803212115. URL: https://www.pnas.org/content/115/45/11471.

[396] Nils Reimers and Iryna Gurevych. "Sentence-BERT: Sentence Embeddings Using Siamese BERT-Networks". In: *EMNLP*. Nov. 2019. DOI: 10.18653/v1/d19-1410.

[397] Ehud Reiter. "A Structured Review of the Validity of BLEU". In: *Computational Linguistics* 44 (Sept. 2018). DOI: 10.1162/coli_a_00322.

[398] Marco Tulio Ribeiro and Scott Lundberg. "Adaptive Testing and Debugging of NLP Models". In: *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. May 2022. DOI: 10.18653/v1/2022.acl-long.230.

[399] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. ""Why Should I Trust You?": Explaining the Predictions of Any Classifier". In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. KDD '16. San Francisco, California, USA: Association for Computing Machinery, 2016, pp. 1135–1144. ISBN: 9781450342322. DOI: `10.1145/2939672.2939778`. URL: `https://doi.org/10.1145/2939672.2939778`.

[400] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. ""Why Should I Trust You?": Explaining the Predictions of Any Classifier". In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. Aug. 2016. DOI: `10.1145/2939672.2939778`.

[401] Marco Tulio Ribeiro et al. "Beyond Accuracy: Behavioral Testing of NLP Models with CheckList". In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. 2020. DOI: `10.18653/v1/2020.acl-main.442`.

[402] Matīss Rikters, Mark Fishel, and Ondřej Bojar. "Visualizing Neural Machine Translation Attention and Confidence". In: *The Prague Bulletin of Mathematical Linguistics* 109 (Oct. 2017). DOI: `10.1515/pralin-2017-0037`.

[403] Samantha Robertson and Mark Díaz. "Understanding and Being Understood: User Strategies for Identifying and Recovering From Mistranslations in Machine Translation-Mediated Chat". In: *2022 ACM Conference on Fairness, Accountability, and Transparency*. FAccT '22. Seoul, Republic of Korea: Association for Computing Machinery, 2022, pp. 2223–2238. ISBN: 9781450393522. DOI: `10.1145/3531146.3534638`. URL: `https://doi.org/10.1145/3531146.3534638`.

[404] Samantha Robertson and Mark Díaz. "Understanding and Being Understood: User Strategies for Identifying and Recovering From Mistranslations in Machine Translation-Mediated Chat". In: *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*. FAccT '22. Seoul, South Korea: Association for Computing Machinery, 2022.

[405] Samantha Robertson, Tonya Nguyen, and Niloufar Salehi. "Modeling assumptions clash with the real world: Transparency, equity, and community challenges for student assignment algorithms". In: *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 2021, pp. 1–14.

[406] Samantha Robertson et al. "Three Directions for the Design of Human-Centered Machine Translation". In: *First Workshop on Bridging Human–Computer Interaction and Natural Language Processing at EACL 2021* (2021).

[407] Negar Rostamzadeh et al. "Healthsheet: Development of a Transparency Artifact for Health Datasets". In: *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*. FAccT '22. Seoul, Republic of Korea: Association for Computing Machinery, 2022, pp. 1943–1961. ISBN: 9781450393522. DOI: `10.1145/3531146.3533239`. URL: `https://doi.org/10.1145/3531146.3533239`.

[408] Alvin E. Roth. "Deferred acceptance algorithms: history, theory, practice, and open questions". In: *International Journal of Game Theory* 36.3-4 (Jan. 2008), pp. 537–569.

[409] Alvin E. Roth. *Who Gets What And Why: the hidden world of matchmaking and market design*. Harper Collins, 2015.

[410] Richard Rothstein. *The Color of Law: A Forgotten History of How Our Government Segregated America*. New York City, New York: Liveright, 2017.

[411] Richard Rothstein. "Why Our Schools Are Segregated". In: *Educational Leadership* 70.8 (2013), pp. 50–55.

[412] Paul Röttger et al. "HateCheck: Functional Tests for Hate Speech Detection Models". In: *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Aug. 2021. DOI: 10.18653/v1/2021.acl-long.4.

[413] Kaustav Roy. *Feedback cycle and the Gulfs of execution and evaluation*. May 2022. URL: https://bootcamp.uxdesign.cc/feedback-cycle-and-the-gulfs-of-execution-and-evaluation-84647b8028fe.

[414] Rachel Rudinger et al. "Gender Bias in Coreference Resolution". In: *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*. June 2018. DOI: 10.18653/v1/n18-2002.

[415] Ananya B. Sai, Akash Kumar Mohankumar, and Mitesh M. Khapra. "A Survey of Evaluation Metrics Used for NLG Systems". In: *Acm Computing Surveys* 55 (Jan. 2022). DOI: 10.1145/3485766.

[416] Herman Saksono et al. "Family Health Promotion in Low-SES Neighborhoods: A Two-Month Study of Wearable Activity Tracking". In: *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. CHI '18. Montreal QC, Canada: Association for Computing Machinery, 2018, pp. 1–13. ISBN: 9781450356206. DOI: 10.1145/3173574.3173883. URL: https://doi.org/10.1145/3173574.3173883.

[417] Belén Saldías Fuentes et al. "Toward More Effective Human Evaluation for Machine Translation". In: *Proceedings of the 2nd Workshop on Human Evaluation of NLP Systems (HumEval)*. May 2022. DOI: 10.18653/v1/2022.humeval-1.7.

[418] Niloufar Salehi et al. "We Are Dynamo: Overcoming Stalling and Friction in Collective Action for Crowd Workers". In: *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*. CHI '15. Seoul, Republic of Korea: Association for Computing Machinery, 2015, pp. 1621–1630. ISBN: 9781450331456. DOI: 10.1145/2702123.2702508. URL: https://doi.org/10.1145/2702123.2702508.

[419] Mia Sato and Emma Roth. "CNET found errors in more than half of its AI-written stories". In: *The Verge* (Jan. 2023). URL: https://www.theverge.com/2023/1/25/23571082/cnet-ai-written-stories-errors-corrections-red-ventures.

[420] Devansh Saxena and Shion Guha. “Conducting Participatory Design to Improve Algorithms in Public Services: Lessons and Challenges”. In: *Conference Companion Publication of the 2020 on Computer Supported Cooperative Work and Social Computing*. CSCW ’20 Companion. Virtual Event, USA: Association for Computing Machinery, 2020, pp. 383–388. ISBN: 9781450380591. DOI: 10.1145/3406865.3418331. URL: https://doi.org/10.1145/3406865.3418331.

[421] Devansh Saxena et al. “A Human-Centered Review of Algorithms Used within the U.S. Child Welfare System”. In: *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. CHI ’20. Honolulu, HI, USA: Association for Computing Machinery, 2020, pp. 1–15. ISBN: 9781450367080. DOI: 10.1145/3313831.3376229. URL: https://doi.org/10.1145/3313831.3376229.

[422] Emanuel A. Schegloff, Gail Jefferson, and Harvey Sacks. “The Preference for Self-Correction in the Organization of Repair in Conversation”. In: *Language* 53.2 (1977), pp. 361–382. ISSN: 00978507, 15350665. URL: http://www.jstor.org/stable/413107.

[423] Sebastian Schelter et al. “Unit Testing Data with Deequ”. In: *Proceedings of the 2019 International Conference on Management of Data*. SIGMOD ’19. 2019. DOI: 10.1145/3299869.3320210.

[424] Dean Schillinger et al. “Precision communication: Physicians’ linguistic adaptation to patients’ health literacy”. In: *Science Advances* 7.51 (2021), eabj2836.

[425] Mark Schneider and Jack Buckley. “What Do Parents Want from Schools? Evidence from the Internet”. In: *Educational Evaluation and Policy Analysis* 24.2 (2002), pp. 133–144. ISSN: 01623737, 19351062. URL: http://www.jstor.org/stable/3594140.

[426] David W. Scott. *Multivariate Density Estimation: Theory, Practice, and Visualization*. Wiley Series in Probability and Statistics. 2015. DOI: 10.1002/9781118575574.

[427] Janelle Scott. “School Choice and the Empowerment Imperative”. In: *Peabody Journal of Education* 88.1 (2013), pp. 60–73. DOI: 10.1080/0161956X.2013.752635. eprint: https://doi.org/10.1080/0161956X.2013.752635. URL: https://doi.org/10.1080/0161956X.2013.752635.

[428] Janelle T. Scott. “A Rosa Parks moment? School choice and the marketization of civil rights”. In: *Critical Studies in Education* 54.1 (2013), pp. 5–18. DOI: 10.1080/17508487.2013.739570. eprint: https://doi.org/10.1080/17508487.2013.739570. URL: https://doi.org/10.1080/17508487.2013.739570.

[429] Janelle T. Scott. “Market-Driven Education Reform and the Racial Politics of Advocacy”. In: *Peabody Journal of Education* 86.5 (2011), pp. 580–599. DOI: 10.1080/0161956X.2011.616445. URL: https://doi.org/10.1080/0161956X.2011.616445.

[430] D. Sculley. *A Data-Centric View of Technical Debt in AI*. 2022. URL: https://datacentricai.org/data-in-deployment/.

[431] D. Sculley et al. “Hidden Technical Debt in Machine Learning Systems”. In: *Advances in Neural Information Processing Systems*. Vol. 28. 2015.

[432] Andrew D. Selbst and Solon Barocas. “The Intuitive Appeal of Explainable Machines”. In: *87 Fordham Law Review 1085* (2018). URL: `http://dx.doi.org/10.2139/ssrn.3126971`.

[433] Andrew D. Selbst et al. “Fairness and Abstraction in Sociotechnical Systems”. In: *Proceedings of the Conference on Fairness, Accountability, and Transparency*. FAT* ’19. Atlanta, GA, USA: Association for Computing Machinery, 2019, pp. 59–68. ISBN: 9781450361255. DOI: `10.1145/3287560.3287598`. URL: `https://doi.org/10.1145/3287560.3287598`.

[434] Rico Sennrich. “How Grammatical is Character-level Neural Machine Translation? Assessing MT Quality with Contrastive Translation Pairs”. In: *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*. Valencia, Spain: Association for Computational Linguistics, Apr. 2017, pp. 376–382. URL: `https://aclanthology.org/E17-2060`.

[435] Emre Şentürk, Mukadder Orhan-Sungur, and Tülay Özkan-Seyhan. “Google Translate: Can It Be a Solution for Language Barrier in Neuraxial Anaesthesia?” In: *Turkish journal of anaesthesiology and reanimation* 49.2 (2021), pp. 181–182. URL: `https://doi.org/10.5152/TJAR.2021.101`.

[436] Vidya Setlur et al. “Eviza: A Natural Language Interface for Visual Analysis”. en. In: *Proceedings of the 29th Annual Symposium on User Interface Software and Technology*. Tokyo Japan: ACM, Oct. 2016, pp. 365–377. ISBN: 978-1-4503-4189-9. DOI: `10.1145/2984511.2984588`. URL: `https://dl.acm.org/doi/10.1145/2984511.2984588` (visited on 12/23/2021).

[437] SFUSD. “Changes to Student Assignment for Elementary Schools”. In: (2020). URL: `https://www.sfusd.edu/schools/enroll/student-assignment-policy/student-assignment-changes`.

[438] SFUSD. *Student Assignment: 4th Annual Report: 2014-15 School Year*. 2015. URL: `https://archive.sfusd.edu/en/assets/sfusd-staff/enroll/files/2015-16/4th-annual-report-april-8-2015.pdf`.

[439] SFUSD. “Why We’re Redesigning Student Assignment”. In: (2019). URL: `https://www.sfusd.edu/studentassignment/why-were-redesigning-student-assignment`.

[440] SFUSD Office of Education. *Board Policy 5101: Student Assignment*. 2010. URL: `https://go.boarddocs.com/ca/sfusd/Board.nsf/goto?open&id=B55QMC657423`.

[441] Lloyd Shapley and Herbert Scarf. “On cores and indivisibility”. In: *Journal of Mathematical Economics* 1.1 (1974), pp. 23–37. ISSN: 0304-4068. DOI: `https://doi.org/10.1016/0304-4068(74)90033-0`. URL: `http://www.sciencedirect.com/science/article/pii/0304406874900330`.

[442] Carla Shedd. *Unequal City: Race, Schools, and Perceptions of Injustice.* Russell Sage Foundation, 2015. ISBN: 9780871547965. URL: http://www.jstor.org/stable/10.7758/9781610448529.

[443] Hong Shen et al. "Everyday Algorithm Auditing: Understanding the Power of Everyday Users in Surfacing Harmful Algorithmic Behaviors". In: *Proc. ACM Hum.-Comput. Interact.* 5 (Oct. 2021). DOI: 10.1145/3479577.

[444] Leixian Shen et al. "Towards Natural Language Interfaces for Data Visualization: A Survey". In: *arXiv:2109.03506 [cs]* (Sept. 2021). arXiv: 2109.03506. URL: http://arxiv.org/abs/2109.03506 (visited on 12/23/2021).

[445] Chunqi Shi, Donghui Lin, and Toru Ishida. "Agent Metaphor for Machine Translation Mediated Communication". In: *Proceedings of the 2013 International Conference on Intelligent User Interfaces.* IUI '13. Santa Monica, California, USA: Association for Computing Machinery, 2013, pp. 67–74. ISBN: 9781450319652. DOI: 10.1145/2449396.2449407. URL: https://doi.org/10.1145/2449396.2449407.

[446] Tomohiro Shigenobu. "Evaluation and Usability of Back Translation for Intercultural Communication". In: *Usability and Internationalization. Global and Local User Interfaces.* Ed. by Nuray Aykin. Berlin, Heidelberg: Springer Berlin Heidelberg, 2007, pp. 259–265. ISBN: 978-3-540-73289-1.

[447] Katie Shilton. "Values Levers: Building Ethics into Design". In: *Science, Technology, & Human Values* 38.3 (2013), pp. 374–397. DOI: 10.1177/0162243912436985.

[448] JongHo Shin, Panayiotis G. Georgiou, and Shrikanth Narayanan. "Enabling effective design of multimodal interfaces for speech-to-speech translation system: An empirical study of longitudinal user behaviors over time and user strategies for coping with errors". In: *Computer Speech & Language* 27.2 (2013). Special Issue on Speech-speech translation, pp. 554–571. ISSN: 0885-2308. DOI: https://doi.org/10.1016/j.csl.2012.02.001. URL: https://www.sciencedirect.com/science/article/pii/S0885230812000113.

[449] Ben Shneiderman. "Human-Centered Artificial Intelligence: Reliable, Safe & Trustworthy". In: *International Journal of Human–Computer Interaction* 36.6 (2020), pp. 495–504. DOI: 10.1080/10447318.2020.1741118. eprint: https://doi.org/10.1080/10447318.2020.1741118. URL: https://doi.org/10.1080/10447318.2020.1741118.

[450] Bernard W Silverman. *Density Estimation for Statistics and Data Analysis.* 2018. DOI: 10.1201/9781315140919.

[451] Karan Singhal et al. "Large language models encode clinical knowledge". In: *Nature* 620.7972 (Aug. 2023), pp. 172–180. ISSN: 1476-4687. DOI: 10.1038/s41586-023-06291-2. URL: https://doi.org/10.1038/s41586-023-06291-2.

[452] Mona Sloane et al. *Participation is not a Design Fix for Machine Learning.* 2020. arXiv: 2007.02423 [cs.CY].

[453] C. Estelle Smith et al. “Keeping Community in the Loop: Understanding Wikipedia Stakeholder Values for Machine Learning-Based Systems”. In: *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. CHI ’20. Honolulu, HI, USA: Association for Computing Machinery, 2020, pp. 1–14. ISBN: 9781450367080. DOI: `10.1145/3313831.3376783`. URL: `https://doi.org/10.1145/3313831.3376783`.

[454] Jeremy Adam Smith. *As Parents Get More Choice, S.F. Schools Resegregate*. San Francisco Public Press. 2015.

[455] Felipe Soares, Viviane Moreira, and Karin Becker. “A Large Parallel Corpus of Full-Text Scientific Articles”. In: *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC-2018)*. Miyazaki, Japan: European Language Resource Association, 2018. URL: `http://aclweb.org/anthology/L18-1546`.

[456] Dean Spade. *Normal Life: Administrative Violence, Critical Trans Politics and the Limits of Law*. 2nd ed. Duke University Press, 2015.

[457] Hervé Spechbach et al. “Comparison of the quality of two speech translators in emergency settings : A case study with standardized Arabic speaking patients with abdominal pain”. In: *Proceedings of European Congress of Emergency Medicine*. EUSEM 2017. Athens, Greece, 2017. URL: `https://archive-ouverte.unige.ch/unige:100812`.

[458] Lucia Specia, Carolina Scarton, and Gustavo Henrique Paetzold. “Quality Estimation for Machine Translation”. In: *Synthesis Lectures on Human Language Technologies*. Ed. by Graeme Hirst. Morgan & Claypool, 2018. DOI: `10.2200/S00854ED1V01Y201805HLT039`.

[459] Lucia Specia et al. “Findings of the WMT 2018 Shared Task on Quality Estimation”. In: *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*. Belgium, Brussels: Association for Computational Linguistics, Oct. 2018, pp. 689–709. DOI: `10.18653/v1/W18-6451`. URL: `https://www.aclweb.org/anthology/W18-6451`.

[460] Lucia Specia et al. “Findings of the WMT 2020 Shared Task on Quality Estimation”. In: *Proceedings of the Fifth Conference on Machine Translation*. Online: Association for Computational Linguistics, Nov. 2020, pp. 743–764. URL: `https://www.aclweb.org/anthology/2020.wmt-1.79`.

[461] Thilo Spinner et al. “explAIner: A Visual Analytics Framework for Interactive and Explainable Machine Learning”. In: *IEEE Transactions on Visualization and Computer Graphics* (2019). DOI: `10.1109/tvcg.2019.2934629`.

[462] Aaron Springer and Steve Whittaker. "Progressive Disclosure: Empirically Motivated Approaches to Designing Effective Transparency". In: *Proceedings of the 24th International Conference on Intelligent User Interfaces*. IUI '19. Marina del Ray, California: Association for Computing Machinery, 2019, pp. 107–120. ISBN: 9781450362726. DOI: 10.1145/3301275.3302322. URL: https://doi.org/10.1145/3301275.3302322.

[463] Gabriel Stanovsky, Noah A. Smith, and Luke Zettlemoyer. "Evaluating Gender Bias in Machine Translation". In: *ACL*. 2019. DOI: 10.18653/v1/p19-1164.

[464] Logan Stapleton et al. "Imagining new futures beyond predictive systems in child welfare: A qualitative study with impacted stakeholders". In: *2022 ACM Conference on Fairness, Accountability, and Transparency*. 2022, pp. 1162–1177.

[465] Steve Stecklow. "Why Facebook is losing the war on hate speech in Myanmar". In: *Reuters* (Aug. 2018). URL: https://www.reuters.com/investigates/special-report/myanmar-facebook-hate/.

[466] David Steele and Lucia Specia. "Vis-Eval Metric Viewer: A Visualisation Tool for Inspecting and Evaluating Metric Scores of Machine Translation Output". In: *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations*. June 2018.

[467] Amanda Stent, Matthew Marge, and Mohit Singhai. "Evaluating Evaluation Methods for Generation in the Presence of Variation". In: *Proceedings of the 6th International Conference on Computational Linguistics and Intelligent Text Processing*. CICLing'05. 2005. DOI: 10.1007/978-3-540-30586-6_38.

[468] Jonathan Stray. "Aligning AI Optimization to Community Well-Being". In: *International Journal of Community Wellbeing* 3.4 (Nov. 2020), pp. 443–463. DOI: 10.1007/s42413-020-00086-3.

[469] Hendrik Strobelt et al. "S Eq 2s Eq-v Is: A Visual Debugging Tool for Sequence-to-Sequence Models". In: *IEEE transactions on visualization and computer graphics* 25 (2018). DOI: 10.1109/tvcg.2018.2865044.

[470] Angelika Strohmayer, Rob Comber, and Madeline Balaam. "Exploring Learning Ecologies among People Experiencing Homelessness". In: *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*. CHI '15. Seoul, Republic of Korea: Association for Computing Machinery, 2015, pp. 2275–2284. ISBN: 9781450331456. DOI: 10.1145/2702123.2702157. URL: https://doi.org/10.1145/2702123.2702157.

[471] Lucy Suchman. "Making Work Visible". In: *Commun. ACM* 38.9 (Sept. 1995), pp. 56–64. ISSN: 0001-0782. DOI: 10.1145/223248.223263. URL: https://doi.org/10.1145/223248.223263.

[472] Jennifer Sukis. *Ai design & practices guidelines (a review)*. 2019.

[473] Jason Sunshine and Tom R. Tyler. "The Role of Procedural Justice and Legitimacy in Shaping Public Support for Policing". In: *Law and Society Review* 37.3 (2003), pp. 513–548.

[474] Latanya Sweeney. "Discrimination in Online Ad Delivery". In: *SSRN Scholarly Paper* (2013).

[475] Breena R Taira et al. "A Pragmatic Assessment of Google Translate for Emergency Department Instructions". In: *Journal of Geneneral Internal Medicine* 36 (Nov. 2021). DOI: `10.1007/s11606-021-06666-z`.

[476] Breena R. Taira et al. "A Pragmatic Assessment of Google Translate for Emergency Department Instructions". In: *Journal of General Internal Medicine* 36 (2021), pp. 3361–3365. URL: `https://doi.org/10.1007/s11606-021-06666-z`.

[477] NLLB Team et al. *No Language Left Behind: Scaling Human-Centered Machine Translation*. 2022. arXiv: `2207.04672 [cs.CL]`.

[478] Jakita O. Thomas et al. "Exploring the Difficulties African-American Middle School Girls Face Enacting Computational Algorithmic Thinking over Three Years While Designing Games for Social Change". In: *Comput. Supported Coop. Work* 26.4–6 (Dec. 2017), pp. 389–421. ISSN: 0925-9724. DOI: `10.1007/s10606-017-9292-y`. URL: `https://doi.org/10.1007/s10606-017-9292-y`.

[479] Lauren Thornton, Bran Knowles, and Gordon Blair. "Fifty Shades of Grey: In Praise of a Nuanced Approach Towards Trustworthy Design". In: *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. 2021, pp. 64–76.

[480] Jörg Tiedemann. "The Tatoeba Translation Challenge – Realistic Data Sets for Low Resource and Multilingual MT". In: *Proceedings of the Fifth Conference on Machine Translation*. Online: Association for Computational Linguistics, Nov. 2020, pp. 1174–1182. URL: `https://www.aclweb.org/anthology/2020.wmt-1.139`.

[481] Jörg Tiedemann and Santhosh Thottingal. "OPUS-MT — Building Open Translation Services for the World". In: *Proceedings of the 22nd Annual Conferenec of the European Association for Machine Translation (EAMT)*. 2020.

[482] Richard Tomsett et al. "Rapid trust calibration through interpretable and uncertainty-aware AI". In: *Patterns* 1.4 (2020), p. 100049.

[483] Antonio Toral and Víctor M. Sánchez-Cartagena. "A Multifaceted Evaluation of Neural versus Phrase-Based Machine Translation for 9 Language Directions". In: *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*. Valencia, Spain: Association for Computational Linguistics, Apr. 2017, pp. 1063–1073. URL: `https://aclanthology.org/E17-1100`.

[484] Yeganeh Torbati. "Google Says Google Translate Can't Replace Human Translators. Immigration Officials Have Used It to Vet Refugees." In: *Pro Publica* (Sept. 2019). URL: `https://www.propublica.org/article/google-says-google-translate-cant-replace-human-translators-immigration-officials-have-used-it-to-vet-refugees`.

[485] Ehsan Toreini et al. "The relationship between trust in AI and trustworthy machine learning technologies". In: *Proceedings of the 2020 conference on fairness, accountability, and transparency.* 2020, pp. 272–283.

[486] Kentaro Toyama. *Geek Heresy: Rescuing Social Change from the Cult of Technology.* PublicAffairs, 2015.

[487] Jonas-Dario Troles and Ute Schmid. "Extending Challenge Sets to Uncover Gender Bias in Machine Translation: Impact of Stereotypical Verbs and Adjectives". In: *Proceedings of the Sixth Conference on Machine Translation.* Nov. 2021.

[488] Edward R. Tufte. *The Visual Display of Quantitative Information.* 2nd ed., 8th print. 2013.

[489] Anne M Turner, Hannah Mandel, and Daniel Capurro. "Local health department translation processes: potential of machine translation technologies to help meet needs". In: *AMIA annual symposium proceedings.* Vol. 2013. American Medical Informatics Association. 2013, p. 1378.

[490] Anne M Turner et al. "Evaluating the Usefulness of Translation Technologies for Emergency Response Communication: A Scenario-Based Study". In: *JMIR Public Health Surveill* 5.1 (Jan. 2019). DOI: `10.2196/11171`.

[491] Tom Tyler, Peter Degoey, and Heather Smith. "Understanding why the justice of group procedures matters: A test of the psychological dynamics of the group-value model". In: *Journal of Personality and Social Psychology* 70.5 (1996), pp. 913–930. URL: `https://doi.org/10.1037/0022-3514.70.5.913`.

[492] Tom R. Tyler. "Affirmative Action in an Institutional Context: The Antecedents of Policy Preferences and Political Support". In: *Social Justice Research* 17.1 (Mar. 2004), pp. 5–24. ISSN: 1573-6725. DOI: `10.1023/B:SORE.0000018090.84298.5b`. URL: `https://doi.org/10.1023/B:SORE.0000018090.84298.5b`.

[493] Tom R. Tyler. "Multiculturalism and the Willingness of Citizens to Defer to Law and to Legal Authorities". In: *Law & Social Inquiry* 25.4 (2000), pp. 983–1019. ISSN: 08976546, 17474469. URL: `http://www.jstor.org/stable/829122`.

[494] Tom R. Tyler. "What is Procedural Justice?: Criteria used by Citizens to Assess the Fairness of Legal Procedures". In: *Law & Society Review* 22.1 (1988), pp. 103–135. ISSN: 00239216, 15405893. URL: `http://www.jstor.org/stable/3053563`.

[495] Tom R. Tyler. *Why People Obey the Law.* Princeton University Press, 2006.

[496] Tom R. Tyler and Robert J. Bies. “Beyond formal procedures: The interpersonal context of procedural justice”. In: *Applied Social Psychology and Organizational Settings* (2015), pp. 77–98.

[497] Laurens van der Maaten and Geoffrey Hinton. “Visualizing Data Using T-SNE”. In: *Journal of Machine Learning Research* 9 (2008). URL: `http://jmlr.org/papers/v9/vandermaaten08a.html`.

[498] Eva Vanmassenhove and Johanna Monti. “gENder-IT: An Annotated English-Italian Parallel Challenge Set for Cross-Linguistic Natural Gender Phenomena”. In: *Proceedings of the 3rd Workshop on Gender Bias in Natural Language Processing.* Aug. 2021. DOI: `10.18653/v1/2021.gebnlp-1.1`.

[499] Helena Vasconcelos et al. “Explanations Can Reduce Overreliance on AI Systems During Decision-Making”. In: *arXiv preprint arXiv:2212.06823* (2022).

[500] Michael Veale, Max Van Kleek, and Reuben Binns. “Fairness and accountability design needs for algorithmic support in high-stakes public sector decision-making”. In: *Proceedings of the 2018 chi conference on human factors in computing systems.* 2018, pp. 1–14.

[501] Marc Vermeulen et al. “Application of Uniform Manifold Approximation and Projection (UMAP) in Spectral Imaging of Artworks”. In: *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy* 252 (2021). DOI: `10.1016/j.saa.2021.119547`. URL: `https://linkinghub.elsevier.com/retrieve/pii/S1386142521001232` (visited on 11/16/2022).

[502] Lucas Nunes Vieira, Minako O’Hagan, and Carol O’Sullivan. “Understanding the societal impacts of machine translation: a critical review of the literature on medical and legal use cases”. In: *Information, Communication & Society* 0.0 (2020), pp. 1–18. DOI: `10.1080/1369118X.2020.1776370`. eprint: `https://doi.org/10.1080/1369118X.2020.1776370`. URL: `https://doi.org/10.1080/1369118X.2020.1776370`.

[503] David Vilar et al. “Error Analysis of Statistical Machine Translation Output”. In: *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC’06).* May 2006.

[504] James Vincent. “AI-generated answers temporarily banned on coding Q&A site Stack Overflow”. In: *The Verge* (Dec. 2022). URL: `https://www.theverge.com/2022/12/5/23493932/chatgpt-ai-generated-answers-temporarily-banned-stack-overflow-llms-dangers?campaign_id=158&emc=edit_ot_20230329&instance_id=88922&nl=on-tech%5C%3A-a.i.&regi_id=89807153&segment_id=129057&te=1&user_id=543bed7fe8679619de6783f84310c37f`.

[505] John Vines et al. "Configuring participation: on how we involve people in design". en. In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems - CHI '13*. Paris, France: ACM Press, 2013, p. 429. ISBN: 978-1-4503-1899-0. DOI: 10.1145/2470654.2470716. URL: http://dl.acm.org/citation.cfm?doid=2470654.2470716 (visited on 09/27/2020).

[506] Amy Voida et al. "Shared Values/Conflicting Logics: Working around e-Government Systems". In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. CHI '14. Toronto, Ontario, Canada: Association for Computing Machinery, 2014, pp. 3583–3592. ISBN: 9781450324731. DOI: 10.1145/2556288.2556971. URL: https://doi.org/10.1145/2556288.2556971.

[507] Changhan Wang et al. "VizSeq: A Visual Analysis Toolkit for Text Generation Tasks". In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP): System Demonstrations*. 2019. DOI: 10.18653/v1/d19-3043.

[508] Qiaosi Wang et al. "Towards mutual theory of mind in human-ai interaction: How language reflects what students perceive about a virtual teaching assistant". In: *Proceedings of the 2021 CHI conference on human factors in computing systems*. 2021, pp. 1–14.

[509] Yun Wang et al. "Towards Natural Language-Based Visualization Authoring". In: *arXiv preprint arXiv:2208.10947v2* (2022).

[510] Justin D. Weisz et al. "BigBlueBot: Teaching Strategies for Successful Human-Agent Interactions". In: *Proceedings of the 24th International Conference on Intelligent User Interfaces*. IUI '19. Marina del Ray, California: Association for Computing Machinery, 2019, pp. 448–459. ISBN: 9781450362726. DOI: 10.1145/3301275.3302290. URL: https://doi.org/10.1145/3301275.3302290.

[511] James Wexler et al. "The What-If Tool: Interactive Probing of Machine Learning Models". In: *IEEE TVCG* 26 (2019). DOI: 10.1109/tvcg.2019.2934619.

[512] Grover J. "Russ" Whitehurst. *New evidence on school choice and racially segregated schools*. Dec. 2017. URL: https://www.brookings.edu/research/new-evidence-on-school-choice-and-racially-segregated-schools/.

[513] Elisabeth Wilson et al. "Effects of limited English proficiency and physician language on health care comprehension". In: *Journal of General Internal Medicine* 20.9 (2005), pp. 800–6. DOI: 10.1111/j.1525-1497.2005.0174.x.

[514] Marisol Wong-Villacres, Neha Kumar, and Betsy DiSalvo. "The Work of Bilingual Parent-Education Liaisons: Assembling Information Patchworks for Immigrant Parents". In: *Proc. ACM Hum.-Comput. Interact.* 3.CSCW (Nov. 2019). DOI: 10.1145/3359288. URL: https://doi.org/10.1145/3359288.

[515] Marisol Wong-Villacres et al. "Culture in Action: Unpacking Capacities to Inform Assets-Based Design". In: *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. CHI '20. Honolulu, HI, USA: Association for Computing Machinery, 2020, pp. 1–14. ISBN: 9781450367080. DOI: `10.1145/3313831.3376329`. URL: `https://doi.org/10.1145/3313831.3376329`.

[516] Marisol Wong-Villacres et al. "Design Guidelines for Parent-School Technologies to Support the Ecology of Parental Engagement". In: *Proceedings of the 2017 Conference on Interaction Design and Children*. IDC '17. Stanford, California, USA: Association for Computing Machinery, 2017, pp. 73–83. ISBN: 9781450349215. DOI: `10.1145/3078072.3079748`. URL: `https://doi.org/10.1145/3078072.3079748`.

[517] Daniel Wood. *As Pandemic Deaths Add Up, Racial Disparities Persist — And In Some Cases Worsen*. NPR. 2020. URL: `https://www.npr.org/sections/health-shots/2020/09/23/914427907/as-pandemic-deaths-add-up-racial-disparities-persist-and-in-some-cases-worsen`.

[518] Tongshuang Wu, Michael Terry, and Carrie Jun Cai. "Ai chains: Transparent and controllable human-ai interaction by chaining large language model prompts". In: *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. 2022, pp. 1–22.

[519] Tongshuang Wu et al. "Errudite: Scalable, Reproducible, and Testable Error Analysis". In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. 2019. DOI: `10.18653/v1/p19-1073`.

[520] Tongshuang Wu et al. "Polyjuice: Generating Counterfactuals for Explaining, Evaluating, and Improving Models". In: *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Aug. 2021. DOI: `10.18653/v1/2021.acl-long.523`.

[521] Yonghui Wu et al. "Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation". In: *CoRR* abs/1609.08144 (2016). URL: `http://arxiv.org/abs/1609.08144`.

[522] Bin Xu et al. "Improving Machine Translation by Showing Two Outputs". In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. CHI '14. Toronto, Ontario, Canada: Association for Computing Machinery, 2014, pp. 3743–3746. ISBN: 9781450324731. DOI: `10.1145/2556288.2557171`. URL: `https://doi.org/10.1145/2556288.2557171`.

[523] Mohammad Yaghini, Andreas Krause, and Hoda Heidari. "A Human-in-the-Loop Framework to Construct Context-Aware Mathematical Notions of Outcome Fairness". In: *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*. AIES '21. Virtual Event, USA: Association for Computing Machinery, 2021, pp. 1023–1033.

ISBN: 9781450384735. DOI: 10.1145/3461702.3462583. URL: https://doi.org/10.1145/3461702.3462583.

[524] Naomi Yamashita and Toru Ishida. "Automatic Prediction of Misconceptions in Multilingual Computer-Mediated Communication". In: *Proceedings of the 11th International Conference on Intelligent User Interfaces*. IUI '06. Sydney, Australia: Association for Computing Machinery, 2006, pp. 62–69. ISBN: 1595932879. DOI: 10.1145/1111449.1111469. URL: https://doi.org/10.1145/1111449.1111469.

[525] Naomi Yamashita and Toru Ishida. "Effects of machine translation on collaborative work". en. In: *Proceedings of the 2006 20th anniversary conference on Computer supported cooperative work - CSCW '06*. Banff, Alberta, Canada: ACM Press, 2006, p. 515. ISBN: 978-1-59593-249-5. DOI: 10.1145/1180875.1180955. URL: http://portal.acm.org/citation.cfm?doid=1180875.1180955 (visited on 08/08/2020).

[526] Naomi Yamashita et al. "Difficulties in Establishing Common Ground in Multiparty Groups Using Machine Translation". In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. CHI '09. Boston, MA, USA: Association for Computing Machinery, 2009, pp. 679–688. ISBN: 9781605582467. DOI: 10.1145/1518701.1518807. URL: https://doi.org/10.1145/1518701.1518807.

[527] Qian Yang et al. "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*. 2020, pp. 1–13.

[528] Sarah Yates. "Scaling the Tower of Babel Fish: An Analysis of the Machine Translation of Legal Information". In: *Law Library Journal* 98 (2006).

[529] Takuya Yokota and Yuri Nakao. "Toward a Decision Process of the Best Machine Learning Model for Multi-Stakeholders: A Crowdsourcing Survey Method". In: *Adjunct Proceedings of the 30th ACM Conference on User Modeling, Adaptation and Personalization*. UMAP '22 Adjunct. Barcelona, Spain: Association for Computing Machinery, 2022, pp. 245–254. ISBN: 9781450392327. DOI: 10.1145/3511047.3538033. URL: https://doi.org/10.1145/3511047.3538033.

[530] Tara J. Yosso. "Whose culture has capital? A critical race theory discussion of community cultural wealth". In: *Race Ethnicity and Education* 8.1 (2005), pp. 69–91. DOI: 10.1080/1361332052000341006. eprint: https://doi.org/10.1080/1361332052000341006. URL: https://doi.org/10.1080/1361332052000341006.

[531] Angie Zhang et al. "Algorithmic Management Reimagined For Workers and By Workers: Centering Worker Well-Being in Gig Work". In: *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. CHI '22. New Orleans, LA, USA: Association for Computing Machinery, 2022. ISBN: 9781450391573. DOI: 10.1145/3491102.3501866. URL: https://doi.org/10.1145/3491102.3501866.

[532] Biao Zhang et al. "Improving Massively Multilingual Neural Machine Translation and Zero-Shot Translation". In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, July 2020, pp. 1628–1639. DOI: `10.18653/v1/2020.acl-main.148`. URL: `https://aclanthology.org/2020.acl-main.148`.

[533] Xiaoyu Zhang et al. "SliceTeller : A Data Slice-Driven Approach for Machine Learning Model Validation". In: *IEEE Transactions on Visualization and Computer Graphics* (2022), pp. 1–11. DOI: `10.1109/TVCG.2022.3209465`.

[534] Yunfeng Zhang, Q Vera Liao, and Rachel KE Bellamy. "Effect of confidence and explanation on accuracy and trust calibration in AI-assisted decision making". In: *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. 2020, pp. 295–305.

[535] Jieyu Zhao et al. "Gender Bias in Coreference Resolution: Evaluation and Debiasing Methods". In: *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*. June 2018. DOI: `10.18653/v1/n18-2003`.

[536] Haiyi Zhu et al. "Value-sensitive algorithm design: Method, case study, and lessons". In: *Proceedings of the ACM on Human-Computer Interaction* 2.CSCW (2018), pp. 1–23.

[537] Vilém Zouhar et al. "Backtranslation Feedback Improves User Confidence in MT, Not Quality". In: *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Online: Association for Computational Linguistics, June 2021, pp. 151–161. DOI: `10.18653/v1/2021.naacl-main.14`. URL: `https://aclanthology.org/2021.naacl-main.14`.

Appendix A

Supplementary Material to Chapter 7

Screener survey used to recruit bilingual Spanish and English speakers and English speakers who do not know Spanish (dscout)

1. Are you comfortable having a written chat conversation in Spanish (i.e. writing and reading messages)?
   - Yes
   - No → Skip to Q7
2. What dialect or variety of Spanish do you speak? (e.g. Mexican, Chilean, ...)
   - Open ended, up to 140 characters
3. How did you learn Spanish?
   - Open ended, up to 140 characters
4. What is your experience with using Spanish?
   - Open ended, up to 140 characters
5. How well do you READ in Spanish? (1 = Not well at all; 5 = Very well)
   - Scale from 1 to 5
6. How well do you WRITE in Spanish? (1 = Not well at all; 5 = Very well)
   - Scale from 1 to 5
7. How well do you READ in English? (1 = Not well at all; 5 = Very well)
   - Scale from 1 to 5
8. How well do you WRITE in English? (1 = Not well at all; 5 = Very well)
   - Scale from 1 to 5

Screener survey used to recruit bilingual speakers of English and a low-resource language (dscout)

1. Are you comfortable having a written conversation over instant messaging (i.e. writing and reading messages) in one of these languages? If you know more than one of these language, please select the one you know the best or use most frequently.

    - I do not know any of these languages → Knocked out
    - Albanian
    - Amharic
    - Armenian
    - Bengali
    - Croatian
    - Gujarati
    - Haitian Creole
    - Hebrew
    - Hindi
    - Hmong
    - Igbo
    - Khmer
    - Korean
    - Lao
    - Malayalam
    - Persian
    - Punjabi
    - Romanian
    - Russian
    - Serbian
    - Swahili
    - Tagalog
    - Tamil
    - Telugu
    - Thai
    - Turkish

- Ukrainian
- Urdu
- Vietnamese
- Yiddish
- Yoruba
- Zulu

2. If this language features letters or characters not used in standard English, are you able to set up your computer so you can type in this language for an instant messaging chat?
    - Yes
    - No
    - This language doesn't feature letters or characters not used in standard English.
3. What is your experience with using this language?
    - Open ended, up to 140 chars
4. What dialect or variety of this language do you speak? (Leave blank if you're not sure)
    - Open ended, up to 140 chars
5. How well do you READ in this language? (1 = Not well at all; 5 = Very well)
    - Scale from 1 to 5
6. How well do you WRITE in this language? (1 = Not well at all; 5 = Very well)
    - Scale from 1 to 5
7. How well do you READ in English? (1 = Not well at all; 5 = Very well)
    - Scale from 1 to 5
8. How well do you WRITE in English? (1 = Not well at all; 5 = Very well)
    - Scale from 1 to 5

Screener survey used to recruit bilingual speakers of English and a low-resource language (shared internally at a large technology company)

Let us know what languages you speak (other than English).

You can fill this out for up to three languages, with the option to let us know if there are other languages you can read and write in.

During the study you will have a conversation over Google Chat, so please list languages you can TYPE in on one of your devices.

1. Language (and dialect if applicable)
    - Short answer text
2. How well do you READ this language?
    - 1 - Not well at all
    - 2
    - 3
    - 4
    - 5 - Very well
3. How well do you WRITE in this language?
    - 1 - Not well at all
    - 2
    - 3
    - 4
    - 5 - Very well
4. Do you have another language to add? *(Shown up to 2 times)*
    - Yes → Return to (1)
    - No → End.

Post-session survey used to collect participant demographics (for English-Farsi participants only)

Note: dscout provided demographic information for participants recruited through their platform (including age, gender, education, employment status, job title, race and ethnicity, household income, and industry).

1. How well do you READ in English?
    - 1 - Not well at all
    - 2
    - 3
    - 4
    - 5 - Very well
2. How well do you WRITE in English?
    - 1 - Not well at all
    - 2
    - 3
    - 4
    - 5 - Very well
3. Which of these best describes how often you use an automatic translation tool (e.g. Google Translate)?
    - Never
    - A few times a year
    - About once a month
    - Multiple times a month
    - Multiple times a week
    - Every day
4. What is your age?
    - 18-25
    - 26-30
    - 31-35
    - 36-40
    - 41-45
    - 46-50
    - 51-55
    - 56-60
    - 61-65

- 66-70
- 71-75
- 76-80
- 80+

5. What is your gender?
    - Short answer text

6. What is your race and/or ethnicity?
    - Short answer text

Screenshot of the study user interface

Figure A.1 replicates Figure 7.1 with the full text shown.

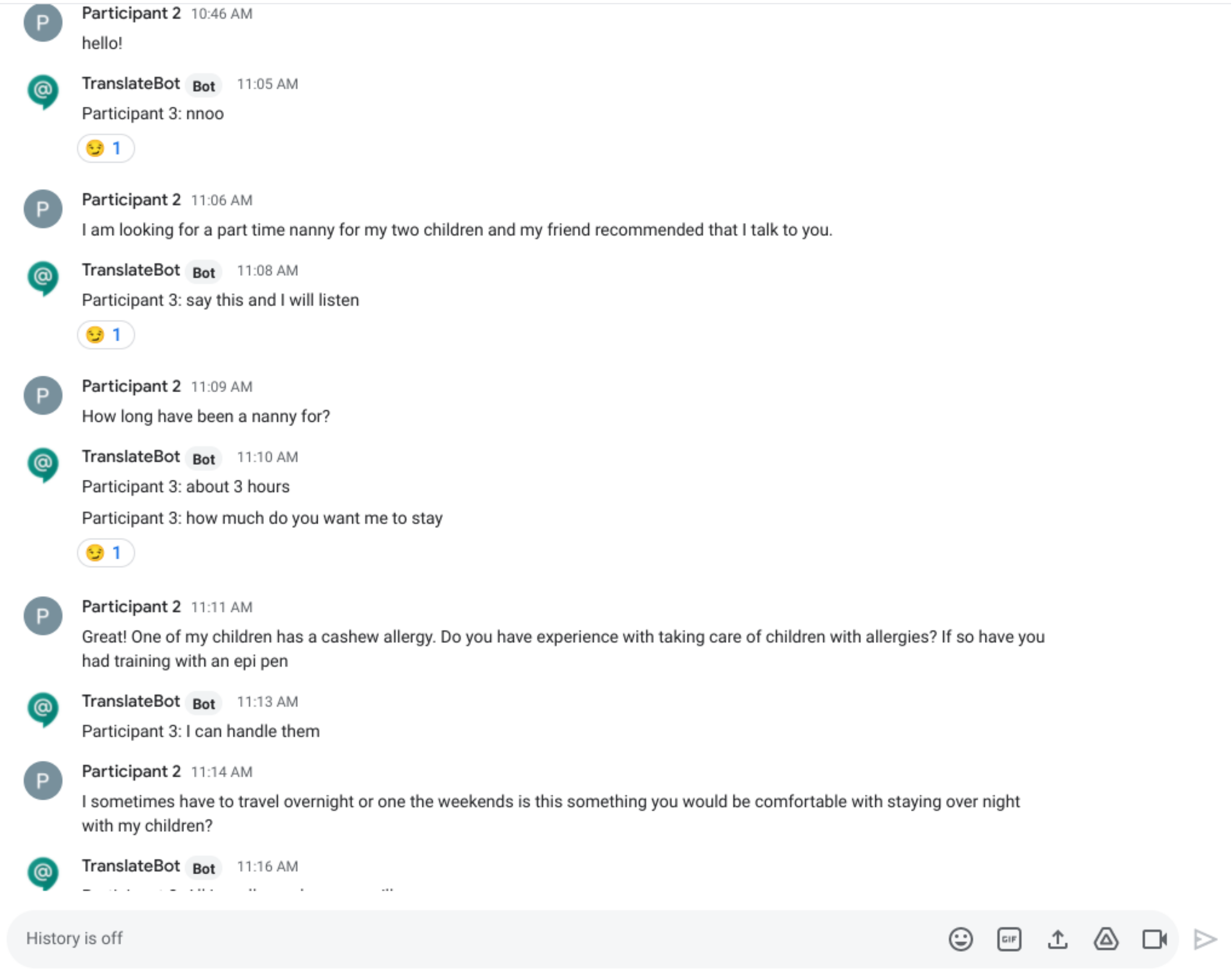
A.1 Instructions for participants

Instructions
Welcome to the study! We are so excited to have you participate today. Before we get started, please read these instructions and let us know if you have any questions.
This study has two parts:

1. Role play **conversation** over Google chat. (SEE YOUR ROLE ON THE NEXT PAGE)

2. One-on-one follow-up interview over video call. We will ask you to open and edit a Google doc during the interview.

IMPORTANT During the chat portion, **use emoji reactions (any emoji is fine) to mark messages that you receive whenever you are not sure whether you have understood what your partner is saying.**
TO GET STARTED:
Go to chat.google.com and log in with these credentials:
USERNAME: [*Participant gmail account*]
PASSWORD: [*Password, randomly generated and reset after each session.*]
You should start the conversation/ by sending a direct message to TranslateBot.
OR
Your partner will start the conversation and you will see their message (translated) in the chat with TranslateBot.
Key points:



Participant 2 10:46 AM

hello!

TranslateBot Bot 11:05 AM

Participant 3: nnoo

😏 1

Participant 2 11:06 AM

I am looking for a part time nanny for my two children and my friend recommended that I talk to you.

TranslateBot Bot 11:08 AM

Participant 3: say this and I will listen

😏 1

Participant 2 11:09 AM

How long have been a nanny for?

TranslateBot Bot 11:10 AM

Participant 3: about 3 hours

Participant 3: how much do you want me to stay

😏 1

Participant 2 11:11 AM

Great! One of my children has a cashew allergy. Do you have experience with taking care of children with allergies? If so have you had training with an epi pen

TranslateBot Bot 11:13 AM

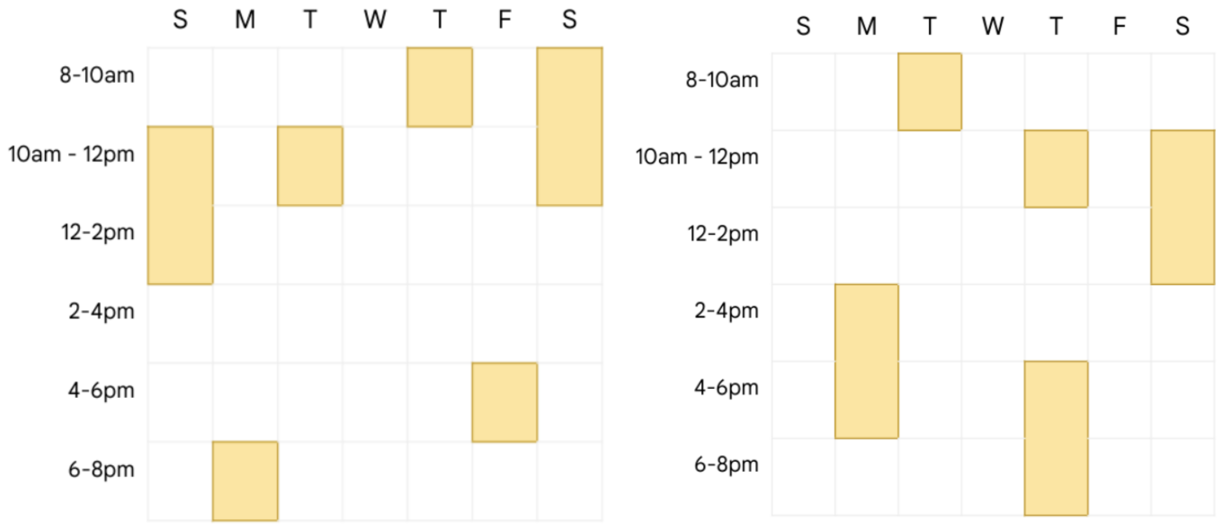
Participant 3: I can handle them

Participant 2 11:14 AM

I sometimes have to travel overnight or one the weekends is this something you would be comfortable with staying over night with my children?

TranslateBot Bot 11:16 AM

Figure A.1: The user study interface from the perspective of the participant using English in session I4.



	S	M	T	W	T	F	S
8-10am					🟨		🟨
10am - 12pm	🟨		🟨				🟨
12-2pm	🟨						
2-4pm							
4-6pm						🟨	
6-8pm		🟨					

	S	M	T	W	T	F	S
8-10am			🟨				
10am - 12pm					🟨		🟨
12-2pm							🟨
2-4pm		🟨					
4-6pm		🟨			🟨		
6-8pm					🟨		

Figure A.2: Each participant in a pair was shown one of these two calendars to indicate their availability. Each calendar has eight available two hour blocks, but only two overlap.

- Send messages in *[English/Spanish/Farsi/Tagalog/Igbo]* only.
- Stay in character. Do not share any personally identifiable information.
- Mark messages with an emoji reaction if you're not sure you understood what your partner is trying to communicate.
- Your partner's messages may come in slowly sometimes and you will not be able to see when they are typing, so please be patient when waiting for a response to your messages.
- The conversation is over when you schedule a time or after 30 minutes, whichever comes sooner. We will watch the conversation and confirm when you are done.
- If you have any questions, unmute and ask at any time.

YOUR ROLE IS ON THE NEXT PAGE! >>>
[page break]
You will be using: *[English/Spanish/Farsi/Tagalog/Igbo]*
Your role: *[Cleaner/Tenant/Nanny/Real estate agent/Landlord/Parent]*
Role description: *[Relevant description (see below)]*
Your availability: (yellow shows times when you are available) [*One of the two calendars in Figure A.2*]

Tenant

You live in a rental apartment that is old and poorly maintained. There's always something broken in your apartment, and it has started really interfering with your ability to work from home. You always let your landlord know when there are problems. They sometimes try to help, but recently they have been too slow to respond and you've had to fix things yourself. You've just noticed a drip from the ceiling in your bathroom. You've placed a bucket underneath but the paint is starting to sag and you're worried about flooding. **Text your landlord to let them know about the leak and ask for help.** You expect them to organize and pay for all the work; this looks like it could be a big job.
Find a time for someone to come and fix the leak. The calendar below shows your availability:

Landlord

You are a landlord and you manage a small apartment building. One of your tenants is always texting you complaining about problems in their apartment. The wifi is too slow, or the washing machine is too loud, or the neighbors are smoking. You try your best to be responsive but you feel like sometimes they are too demanding.
You get a text from your tenant - there's another issue. **Find out what is going on and how serious it is.** If they need help, determine who would be the appropriate person to call (e.g. plumber, electrician, roofer). Your task is over when you either: **agree that no help is needed, or arrange a time for someone to come fix it.** The calendar below shows when you or a plumber[1] is available to visit the apartment:

Nanny

You are a part-time nanny and you are looking for a new family to care for. A parent texts you wanting to find out more about your experience and availability. **Answer their questions and try to get the job.**
Some facts about you to help you answer the parent's questions:

- You have a flexible schedule, but you balance nannying with another job so you need advance notice before your shifts.
- You have a driver's license but no car.
- You don't have any specialized healthcare training (e.g. CPR), but if the parent pays for it you are happy to do a course.

(Note: you don't have to get all of this info across, just use it if you need help answering the parent's question.)

[1] Participants pointed out that this was an error in the instructions, as it gave away that a plumber was needed. Future work using this protocol should correct this.

If the parent asks any questions you don't know the answer to, feel free to get creative.
If the parent thinks you might be a good fit, they will ask you to come to their house for an interview. The calendar below shows your availability for the week:

Parent

You are looking to hire a part-time nanny. You have two children, a 6 month old and a 2 year old. A friend recommended someone and you would like to find out about their past nannying experience.
Here are a few of your constraints:

- One of the twins has a severe allergy to cashew nuts, so you want to make sure the nanny has training in how to care for kids with allergies and how to use an EpiPen.
- You have a large dog, so the nanny needs to be comfortable with dogs.
- You sometimes work late at night or travel on the weekends, so they need to be okay with staying over at your house occasionally.
- Your older child needs to be driven to and picked up from a kids play group twice a week.

Text the nanny to find out whether they meet these needs.
Once you have a sense of their experience, arrange a time for them to come to your house. The calendar below shows your availability for the week:

Cleaner

You work for a cleaning company and you've been assigned to a job. You normally do routine home cleaning, often for nice homes that are about to go on the market for sale.
You show up to find that the house is extremely run down. Worse, the walls are covered in mold in multiple rooms. You weren't warned about biohazards and you didn't bring any special equipment. You have severe asthma, and exposure to mold for long periods could make you very sick. You're nervous to confront the client because if they complain to your employer you could get in trouble and you can't afford to lose your job right now.
Text the client to explain the situation and let them know you cannot clean the house today. Negotiate with the client to find a solution that meets their needs and protects your health. Agree on a time by which the work can be finished. The calendar below shows your availability:

Real-estate agent

You are a real estate agent and you work for a large agency. You've been assigned to sell an old, run-down building for a very important client that needs to close the sale ASAP. You have a few potential buyers lined up for tomorrow and you've hired a professional cleaning company to come and clean up the place before they arrive.

Someone arrives to start cleaning, but they look concerned. You're frustrated because you're under a lot of stress with this sale and the last thing you need is a delay right now. **Discuss and resolve the situation with the cleaner.** Agree on a time by which the work can be finished. The calendar below shows your constraints:

Appendix B

Supplementary Material to Chapter 9

Table B.1 describes the data filtering and cleaning procedure I applied to the discharge instructions.

B.1 Retrieval index

Example templates and sentences

Table B.2 lists the types of templates in the retrieval index, the counts of each type, an example template of each type, and a corresponding free-text sentence that could match the example template.

Synthetic sentence generation

I filled templates to generate synthetic sentences using GPT-4.

System message: `You are writing discharge instructions for hospital patients`
User message: `Generate 10 sentences with the format:` {`template`}, where {`template`} is filled in with a template.
Hyperparameters:

- temperature = 0.7
- maximum length (tokens) = 256
- top p = 1
- frequency penalty = 0
- presence penalty = 0

Action	**Filter**	**Details**
Drop instructions	Written in Spanish	Filter out instructions written in Spanish with a regular expression containing the five most common words in Spanish (according to: `https://en.wikipedia.org/wiki/Most_common_words_in_Spanish`): de, la, que, el, and en (when not followed by route).
	Heavily repetitive or structured text	Regular expression specific to the data based on manual exploration. Goal was to exclude instructions that contained text copied-and-pasted from another source (e.g. patient education materials), and filter to only free-text discharge instructions written by the patient's providers.
	Very long	Drop instructions longer than 3400 characters. Threshold selected through manual exploration of the data to exclude instructions with long patient instructions that were very likely copied and pasted from another source.
	Heavily redacted	Drop instructions with two or more consecutive ?
	Patient eloped	Drop instructions containing the word "eloped." If the patient eloped, they did not receive discharge instructions.
	Duplicates	Drop duplicate instructions.
Remove text	Patient education	Remove patient education materials using regular expressions to capture patterns in those materials, based on manual exploration. Remove text after "Sincerely, *** Team."
	Extra text at the start of instructions	Remove any characters that appear at the start of the instructions before "Dear," or that are non-alphabetic.
	Trim whitespace	
Other	Split multiple instructions	Some rows of data contained multiple sets of instructions. If "Dear" appears once at the beginning and once in the middle of the instruction, split up at the second "Dear" into two rows.

Table B.1: Details of the data cleaning procedure. Some steps are very specific to the dataset, but here I generalize for reproducibility.

Type	**N**	**Example template**	**Example sentence**
Medication	49	Take [MEDICATION] [DOSAGE] for [SYMPTOM].	Please take Seroquel 50mg once per day as needed for hallucinations.
At home instructions	31	Stay hydrated.	Please continue to stay hydrated.
Follow-up appointments	22	[CLINIC] will call you to schedule an appointment.	The orthopedic foot/ankle clinic will call you for follow up.
Diagnosis	16	We think your [SYMPTOM] is due to [CONDITION].	Your pain may be caused by constipation or by the pregnancy.
Treatment	15	In the hospital we gave you [TREATMENT] for [SYMPTOM].	We gave you some medication to increase your potassium.
Tests	13	The [TEST] showed [CONDITION].	You had an xray of your knee, which showed a fractured kneecap.
Return instructions	4	Please return to the emergency department if you experience [SYMPTOM] or any new or worsening symptoms that are concerning.	Return to the ER for shortness of breath, chest pain, or worsening swelling or redness in your legs.
Complaint	3	You were seen in the emergency department for [SYMPTOM].	You were seen in the Emergency Department for shortness of breath.
Contact information	3	If you have questions, please contact [CLINIC].	If you have questions, please contact your primary care provider or hospital team.
Greeting	2	It was a pleasure taking care of you!	It was our pleasure caring for you.
Discharge	2	It is safe for you to go home at this time.	We think it is safe for you to be discharged at this time.
Signature	2	Sincerely, Dr. [DEID] - Intern Dr. [DEID] - Resident Dr. [DEID], MD - Attending	Sincerely, Dr. *** *** - Intern Dr. *** *** - Resident Dr. *** ***, MD - Attending
Consultation	1	You were evaluated by [CLINIC].	We consulted our psychiatry colleagues who evaluated you.

Table B.2: The retrieval index contains 163 templates across 13 types.

Entity extraction and example terms

Table B.3 lists the types of terms in the term base, the counts of each type, examples of each type, and details about how I extracted those entities. I first extracted medications, conditions, symptoms, treatments, doctors, tests, times and clinics using automatic methods (Apache cTAKES, spaCy named entity recognition, and regular expressions). Next, I manually filtered these terms to remove duplicates, group terms with the same meaning, and split terms into sub-categories (body parts and modifiers). I manually examined sentences containing dosages to add measurements and frequencies.

Term type	**N**	**Examples**	**Method**	**Details**
Medication	221	Advil, afinitor	cTAKES	250 most common of type “medication”
Condition	191	Anemia, asthma	cTAKES	250 most common of type “disease”
Symptom	104	Bleeding, chills	cTAKES	250 most common of type “symptom”
Treatment	68	Abltion, acupuncture	cTAKES	250 most common of type “procedure”
Doctor	62	Cancer doctor, cardiologist	Regex	Words that end with “gist,” bigrams that occurred more than once and end with “gist”, “ doctor”, or “ provider,” and trigrams that occur more than once, end with “ doctor” or “ provider,” excluding “date time provider” and “contact healthcare provider.”
Modifier	56	Abdominal, acute	By hand	Extracted from other term types by hand.
Test	44	Biopsy, blood glucose	cTAKES	250 most common of type “lab”
Time	39	1 day, daily	spaCy NER	50 most common of type “DATE” or “TIME”
Clinic	32	Acute care clinic, cardiology	Regex + by hand	Bigrams with second word “clinic,” excluding incomplete clinic names; Hand-selected trigrams and 4-grams ending with “unit” or “clinic.”
Body part	28	Ankle, bowel	By hand	Extracted from other term types by hand.
Measurement	6	mg, puffs	By hand	Hand constructed based on inspection of dosages.
Frequency	2	As needed, times a day	By hand	Hand constructed based on inspection of dosages.

Table B.3: The termbase contains 853 terms across 12 types.

B.2 Template Filling

I filled templates with terms using GPT-3.5-turbo. To develop prompts I selected four sentences that have a good match to a template, and extracted the terms from each sentence. I asked five bilingual volunteers to translate the template and terms from English into one of Spanish, French, Chinese, Russian, and Farsi. Next, we naively filled the template with the terms in the target language, and experimented with different prompts to GPT-3 and GPT-3.5-turbo that would generate a fluent sentence in the target language.

This approach was able to generate fluent output sentences across all four languages using zero-shot prompting. Text completion mode worked better than text editing mode, despite text editing being a more intuitive choice for this task.

The final prompts I used for Cephalo were:

Template filling:

System message: `You are a medical translation assistant that inputs terms into templates.`

User message: `Input: {input sentence}`
`Template: {template}`
`Terms: {term_string}`
`Template filled with terms:`

where `{input sentence}` is the original sentence in English, `{template}` is the template in German, and `{term_string}` has the format, "`TYPE = term`" with the term in German and a new line between terms.
Hyperparameters:

- temperature = 1
- maximum length (tokens) = 256
- top p = 1
- frequency penalty = 0
- presence penalty = 0

Grammar checking:

System message: `You are a medical translation assistant that checks German grammar.`

User message: `Is the grammar correct?`
`{translation}`.

Hyperparameters:

- temperature = 1
- maximum length (tokens) = 256
- top p = 1
- frequency penalty = 0
- presence penalty = 0

Grammar fixing:

System message: `You are a medical translation assistant that fixes German grammar.`

User message: `Fix the grammar in the German sentence.`
`Input: {translation}`
`Correct grammar:`

Hyperparameters:

- temperature = 1
- maximum length (tokens) = 256
- top p = 1
- frequency penalty = 0
- presence penalty = 0

In future work, I will explore methods that fill the template naively and then prompted the model to fix the grammar, rather than prompting the model to fill in the template and fix the grammar. Where there were more than one terms needed to fill a single hole in the template, performance was best when we indicated to the model how to join those terms (e.g. “We think your pain is due to constipation or the pregnancy” → “We think your [SYMPTOM=pain] is due to [CONDITION=constipation;pregnancy]” could be interpreted as constipation associated with the pregnancy (“pregnancy constipation”) rather than two separate causes.) In general, the model made very few modifications to the naively filled template. Minor grammatical corrections appear possible, but issues that require more significant restructuring of the sentence seem to be more difficult.