Haoqi Zhang and Darren Gergle, 2022. Please do not distribute further.

HCC: Small: Human-AI tools for conceptualizing, aligning, and accounting for difference in expressing human situations and contexts to machines

1 Introduction

How can we imbue machines with a deeper understanding of human situations and contexts, and the interactions they afford? Despite significant advances in machine sensing and machine learning technologies—and the availability of rich APIs for creating AI-powered applications—it remains difficult for designers to express their concept of a human situation (e.g., places to hold a private conversation; good places to take a bike ride with young children) to machines so that applications can be aware and responsive to the situation across a variety of distributed contexts. Failing to do so could result in costly errors, violate safety and privacy, or lead to inequitable access to AI supported experiences. To fill this gap, this project advances new programming environments and tools that support designers using their understanding of human situations to construct machine representations using available context features. Specifically, this proposal seeks to help designers (1) **form rich and accurate conceptions of human situations to encode into machines systems**; (2) **align concepts to machine representations and execution on actual use cases**; and (3) **identify and account for differences across contexts of use**.

Over the last two decades, research and development in machine learning (ML) and context-aware computing have significantly expanded the ability of computing systems to understand rich human context across the dimensions of location, identity, activity, and time [1]. This work led to numerous component detectors for various facets of context through better data, algorithms, and sensors (e.g., [60]). Today, many *context features* are widely available through public APIs (e.g., the Google Awareness and Vision APIs) that implement a range of component detectors including time, location, place, activity, object, and weather. But while existing APIs provide developers with access to a large set of low-level context features, they provide few supports for expressing and encoding higher-level concepts of a situation. As a consequence, existing applications often respond to simplistic queries matching specific context features (e.g., "parks," "restaurants"), but fail to understand conceptually-rich situations that may be several levels of abstraction removed (e.g., "starting a new adventure with young children" or "finding a place to discuss a difficult topic").

With AI projected to deliver $13 trillion in added global economic activity by 2030 [10], and the promise that AI will produce transformative applications that advance every industry and support every aspect of human life [62, 33], the lack of high-level understanding of human situations and contexts is fatal to the vision of AI systems broadly supporting human activities. A core issue is meaning-making: without effective ways to capture the cultural and social aspects of human situations [23, 52], there are fundamental limits to the extent to which AI systems can understand and act on situations for human interaction. For AI systems to support societal concerns such as family cohesion and workplace readiness, and to support diverse application modalities such as personal assistants, augmented reality, social apps, and workplace tools, we need effective ways of encoding human concepts associated with these aims into AI systems (e.g., what is needed to build strong relationships; what struggling at work looks like). Unfortunately, it is prohibitively costly and impractical to train ML systems to learn accurate models of high-level concepts of human situations [57]. As a consequence, we are left with many low-level machine detectors that are easier to train, but that by themselves cannot capture and reflect the wealth of human situations, experience, and activity.

To address this challenge, our recent HCI work [49] and other recent work in ML [57, 63] are advocating for tools that support humans directly expressing their conceptions of a situation to machines. Broadly, we are interested in developing principles and tools that will make it easier to create *concept expressions* that translate a high-level human concept of a situation into a low-level machine representation that can be acted upon computationally (akin to *labeling functions* in ML). Instead of focusing on hand labeling data to be used for training, we focus instead on helping designers flesh out and expand their conceptions of how a situation affords engaging in a human activity or experience into a language that machines can understand. By improving our ability to "bridge" between the two, our proposed methods and tools can enable new applications and technologies that recognize and act on rich, human situations and push the boundaries of how computing systems can support our activities in the world.

However, a human's own conception of situations and the concept expressions they create can be riddled with problems that compromise the application of AI systems. In our preliminary work, we have observed that designers' concept expressions (1) can be vague and incomplete; (2) can fail to match machine representations and execution; and (3) can be insensitive to differences across settings and regions. In other words, encoding the human's conceptual understanding into a machine system can never be effective if the human's own conception and expression thereof is inaccurate or impoverished. This is analogous to issues in data coverage and bias in machine learning, but at the level of expressing concepts from the human to the machine. Therefore, to extend the benefits of this approach, we need explicit cognitive support for humans creating concept expressions, through which they enrich their own conception, reflect on their (mis)understanding and positionality, and on the machine's understanding. Specifically, this project will produce four deliverables that tackle general failures in expressing human concepts of a situation to an AI or context-aware system:

1. Human-AI tools for concept development and expression (Section 3)
2. Human-AI tools for aligning human concepts to machine representations and execution (Section 4)
3. Human-AI tools for accounting for differences across contexts of use (Section 5)
4. A public platform that integrates our learnings and tools across Sections 3-5 (Section 6)

To ensure that our tools are widely applicable and to maximize the intellectual merit and broader impacts of our approach, we will collaborate with leading human-AI and responsible AI industry researchers at Google People + AI Research (Carrie Cai); the Allen Institute for AI (Doug Downey); and Microsoft Research (Ece Kamar) (see accompanying letters of collaboration).

Intellectual Merit: Our work will contribute principles, models, and tools for bridging from human understanding to machine understanding of human situations to enable machines to effectively support varied human activities. Our core contribution is creating a deeper interfacing modality between human and machine systems at the *conceptual level*, by explicitly linking human concepts to the machine's representation and execution. This is in contrast to the typical ML approach of training on manually labelled data, which is insufficient for creating the kinds of rich, human-centered AI systems that we are envisioning [3]. Moreover, while tooling is abundant for training labelling-based ML systems (e.g., [4, 13, 44]), few human-centered AI tools exist for directly encoding human situations into AI systems. To advance the design and operation of human-AI systems, we propose to create a human-AI toolkit that (1) addresses failures in not only data bias and coverage but conceptual bias and coverage [64]; (2) promotes ongoing back-and-forth between the human and machine system to avoid and address misalignment issues in early stages of development; and (3) incorporates diversity and representation in design and development by accounting for differences across contexts of use. We will achieve this by advancing how AI models address deficits in human designers, and countering machine deficits with human skill and reflection. By supporting an integrated human-machine understanding of a human situation, our approach broadly advances socio-technical approaches that reflect the entirety of the human-machine system, and not just the technical and algorithmic components. In doing so, we develop general human-AI principles and techniques that expand our capacity to democratize rich, human-centered AI to be broadly applicable across fields and areas of application.

2 Background and Our Approach

The core premise of the proposed research is to advance bridges from human understanding to machine understanding by developing principles, models, and tools for directly encoding experiential aspects of human situations into AI and context-aware systems. While existing APIs provide developers with access to a large set of low-level context features, they provide few supports for encoding higher-level concepts of a human situation or experience into a machine-understandable representation. Recent excitement and interest in deep learning notwithstanding, it remains prohibitively costly and impractical to train ML models to accurately learn high-level concepts of human situations [57], as they typically require large sets of hand labeled training data. For instance, a concept such as "places to have an intimate conversation" is difficult for machines to understand, as it relies on understanding how private conversations can take place in both public and private spaces, but require different affordances (e.g., while a quiet private space works well, a public space can also work when there some background noise to mask the conversation and retain confidentiality).

As a consequence, AI-powered applications today have very little high-level understanding of human situations on which *they can act reliably and autonomously*, as would be required in most context-aware applications. To illustrate this point, while recent advances in AI models (e.g., knowledge models (KMs) [37], large language models (LLMs) [8]) have led to numerous applications that use AI to generate ideas and solutions for humans to consider [56, 32, 54, 53, 42], none of these systems can be entrusted to act without human oversight [25]. A core issue is that while model responses are sometimes useful and sensible, they are not reliably so [15] and can be inaccurate and non-sensical just as well. As we advance from training low-level detectors to representing higher-level concepts of human situations that are more nuanced and abstract and imbued with cultural and social meaning, there is an increasing need for humans to be in control of the process [34], and to reflect on their own and the machine's position and understanding [11].

To address this need, our research argues for creating an *expression layer* with support for directly expressing human situations and concepts to machines. Broadly, we are interested in developing human-AI principles and tools for constructing *concept expressions* that translate a high-level human concept of a situation into a low-level machine representation that can be acted upon computationally. Constructing concept expressions can be viewed as a form of *machine teaching* [63], in which humans express their understanding through constructing a rich representation thereof (a wide communication channel) than through labeling data (a very narrow channel). But unlike prior machine teaching tools which largely focus on data-driven exploration, labeling, and insight (e.g., [13, 12, 66, 43]), we focus instead on supporting concept development and expression, and on bridging from human conception to the available context features. In doing so, our work will provide new interactive tools for integrating human understanding into machine systems, and advance principles for combining and integrating modular AI capabilities [33].

Recent ML research on *labeling functions* [57] also seeks to integrate human intelligence into machine systems, by providing tools for programming with data and context features (versus labeling data directly). But this work largely neglects the need to support designers overcoming the cognitive challenges associated with encoding one's understanding of a human situation, and creating effective machine representations. This task often requires designers to iteratively enrich and develop their own conceptions of the human situation, their understanding of conceptual boundaries, and their understanding of how the machine's representation may (mis)match their own. We call these *bridging challenges* [49], and they are central to our proposed work on human-AI tools that use AI models and computational techniques to help designers interactively overcome such challenges. Moreover, as rich human situations and contexts may be several layers of intelligibility above the base-level detectors that machine systems make available, with concept expressions we seek to encode intermediate, human understandable representations. This helps to capture more of the intentionality behind the human expression [63], and allows for an on-going human machine dialogue for refining the human and the machine's understanding of the situation that the designer is trying to express (in contrast to labeling functions, which are simply fed into machine learning systems for use). As we will illustrate, this *reflective and reflexive model building* process [11] is important because human conceptions

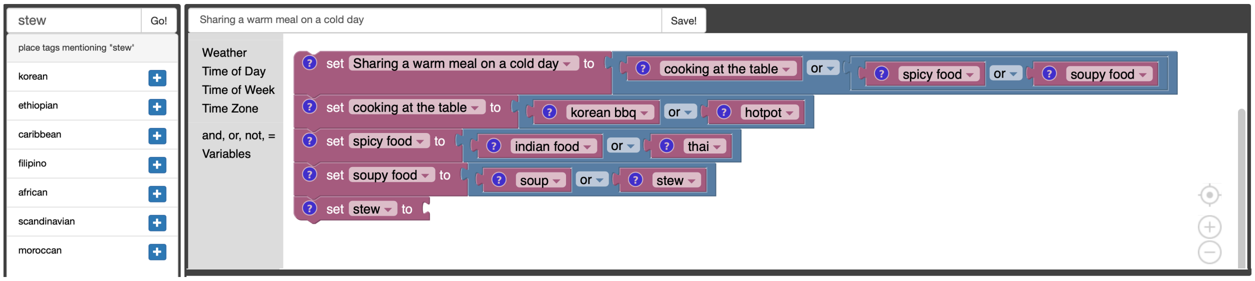

Figure 1: Our prior work on Affinder [49] led to a block-based programming environment for translating human conceptions of a situation into a machine representation using available context-features.

of human situations can itself be error-prone, and result in mismatches between the human's expression and the machine's representation. To truly widen the channel through which humans can express their understanding to machines, we need to consider ways in which human-machine feedback loops can help to refine human conceptions, while also improving machine representations.

2.1 Overview of the Proposed Research

Our early work on constructing concept expressions led to our CHI 2022 paper on Affinder [49], a block-based programming environment for translating human conceptions of a situation into a machine representation using available context features; see Figure 1. Affinder provides a visual workspace in which developers can (a) declare concept variables to represent intermediate concepts that serve as links between an abstract concept and available context features; (b) forage for context features by browsing and searching through categories of features; and (c) compose representations using logical operators. As examples, using Affinder, we may construct a concept expression for the high-level concept "a place for tossing a frisbee around" using simpler concepts and detectors such as open recreational areas (disc golf, parks, playgrounds, beaches), weather is not disruptive (not windy), and while there is daylight. Another construction may express "sharing a meal on a cold day" as cooking at the table (Korean BBQ; hotpot), spicy food (Indian and Thai food), or soupy food (restaurants serving soups and stews).

While our early work on Affinder provides a proof-of-concept for constructing concept expressions, it does not support designers (1) forming a broad and accurate conception of a human situation; (2) identifying and resolving generalizable causes for mismatches between the human's conception and the machine's representation and execution; and (3) accounting for differences across settings and contexts of use. To fully realize the potential for directly expressing human concepts to machines, our proposed work seeks to address three core bridging challenges in encoding a human's conceptual understanding into a machine system; see Figure 2. First, we will address **failures in conceptual coverage and precision** that arise when designers fixate on early solutions and generally struggle to comprehensively conceptualize the various facets of a situation [49, 38, 14]. In our early work [49], we found that even with *reflect and expand* prompts for considering alternative ways of realizing a situation (e.g., why are empty offices good for a private conversation? Where else would be good?), the ideas people generate tend to cluster around the same approach. Second, we will address **failures in aligning human concepts to machine representations and execution** that arise when concept expressions do not execute as the designer expects. In particular, we focus on two key issues: (1) identifying generalizable causes for mismatches between the human's conception and the machine's representation; and (2) reasoning about the availability of interactional resources in execution, so that created concept expressions take into account whether people will encounter the contexts that are defined. Third, we will address **failures to account for differences across settings and contexts of use** that arise due to insensitivity to differences across populations, settings, and regions.

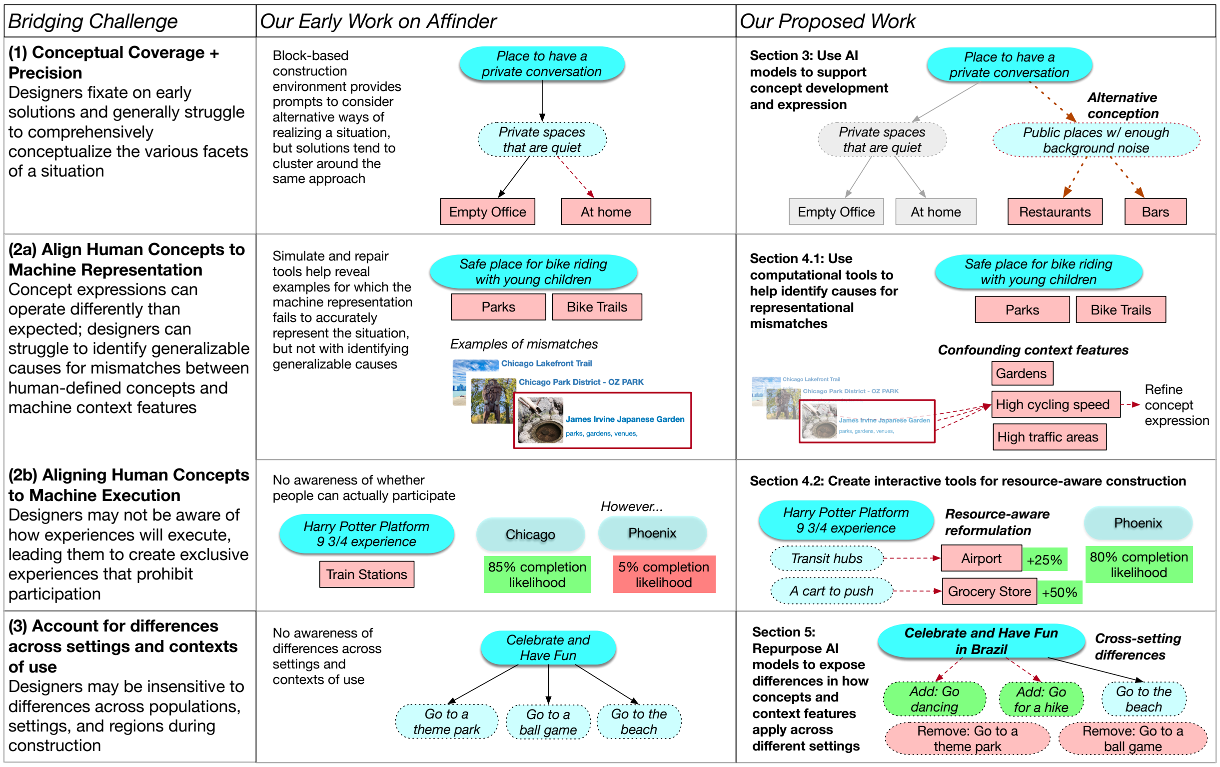

Figure 2: Our proposed work will address three general classes of bridging challenges by providing a variety of human-AI tools for expressing human situations and contexts to machines.

We propose to overcome these failures by creating new human-AI tools enabled by AI models and computational techniques that support conceptualizing, aligning, and accounting for difference in creating concept expressions; see Figure 2. Specifically, we propose to: (1) **use AI models to support designers creating concept expressions** to improve conceptual coverage and precision; (2) **use computational tools to identify generalizable causes** for representational mismatches between the human and the machine, and **to support a resource-aware construction process** that is aware of how created experiences will execute and provides insight to designers; and (3) **repurpose AI models to expose differences in how concepts and context features apply across different settings**, with respect to various reference systems (e.g., across geographies, industries, age, rural/urban strata, socio-economic, etc). We hypothesize that these innovations can (1) address failures in not only data bias and coverage but conceptual coverage and bias; (2) promote on-going back and forth between human and machine to avoid and address misalignment issues in early development; and (3) incorporate diversity and representation in design and development by accounting for differences across contexts of use.

We propose a comprehensive evaluation strategy in which we make use of a range of pilot, laboratory, and deployment studies to empirically guide the development of our tools and systems as well as to assess their ability to support constructing effective concept expressions. We begin with a design-based research approach [7, 24] to advance each of our tools in concert with a series of small, targeted pilot studies that aim to inform and refine our approach (Year 1). We then progress to a series of controlled laboratory studies to inform how each tool supports expressing and refining concepts of situations (Year 2). These studies will help us advance toward our final goal of building and deploying a public platform that integrates our entire toolkit, which we will evaluate through *creator studies* that reveal how creators construct and refine concept expressions in less constrained settings (Year 3). By taking both a decompositional approach [31] (in the first two years) and advancing towards an integrative understanding of the entire system (in the final year), findings will provide new understanding of how ongoing engagement between humans and AI models can yield more robust machine representations and context-aware systems. We will measure conceptual coverage and bias, conceptual misalignments and resolutions, and variability in accuracy across cultures and settings. All developed tools will be made available to the public for research and use.

Throughout our research, we will consult with our advisory board of leading industry human-AI and responsible AI researchers who will (1) provide feedback on our proposed tools and our design approach; (2) ensure that we maintain and advance best practices in developing responsible human-AI systems; (3) connect us with the broader designer and developer community; and (4) help with coordinating the recruitment of designers of context-aware and AI applications for our user studies.

3 Human-AI tools for Concept Development and Expression

A core task in bridging from human concepts to machine representation is developing and enriching the designers' own conception of human situations so that they can better express them to the machine. Designers can fixate on early solutions, and generally struggle to comprehensively conceptualize the various facets of a situation [49, 38, 14]. For instance, to construct a concept expression for "places to have a private conversation," a designer might *fixate* on quiet places such as empty offices or quiet rooms at home, and miss entirely the *alternative conception* that private conversations can also happen in public spaces. They can also miss *hindrances*, for instance that a restaurant may be good public place for a private conversation, but not if it is too quiet, as others can easily overhear the conversation.

The cognitive challenges to overcoming fixation and recalling diverse solutions are well-documented in the design literature [38, 14, 47], but addressing them is challenging as it is difficult for designers to identify and move across varied conceptions when creating a concept expression. In our early work [49], we found that even when prompted to reflect on other ways of realizing a situation (e.g., why are empty offices good for private conversations? Where else would be good?), the ideas people generated tend to cluster around the same approach. To support designers reflecting on a situation at large, **we propose to use AI models** (knowledge models such as COMET-ATOMIC [37]; large language models (LLMs) such as GPT-3 [8]) **to support designers creating concept expressions** by surfacing diverse conceptions or examples that correspond to different conceptions. While such models are hardly perfect and contain inaccuracies and biases, they often capture a wider span of the various ways to conceptualize a situation, and of possible hinderances that a human designer can struggle to recall. This can help designers recognize relevant concepts to add to an expression rather than having to recall them, either as alternative conceptions (e.g., public spaces for having a private conversation) or as hindrances (e.g., heard by others) for realizing the situation.

We propose to operationalize this approach in three ways; see Figure 3. Augmenting the block-based programming tool for creating concept expressions from our prior work on Affinder [49], we will first surface relevant concepts provided by knowledge models and LLMs to support designers exploring and stretching their thinking. Second, we will use concepts surfaced by AI models that designers select as relevant (blue highlights and red text in Figure) to generate *smart reflection prompts*. This enables *reflective model building* [11] by promoting an iterative back-and-forth between human reflection and machine guidance. This also allows the designers to make use of ideas provided by the AI models that may not be exactly correct or directly usable, but upon further reflection, may help designers clarify their own understanding (e.g., recognizing that while not all bars and restaurants work for private conversation, that public places with enough background noise will work). Third, to support designers linking the concepts they identify to context features that are available, we will include terms selected from the AI model and other concepts the designer defines as search terms for query expansion when searching for context features. This gives

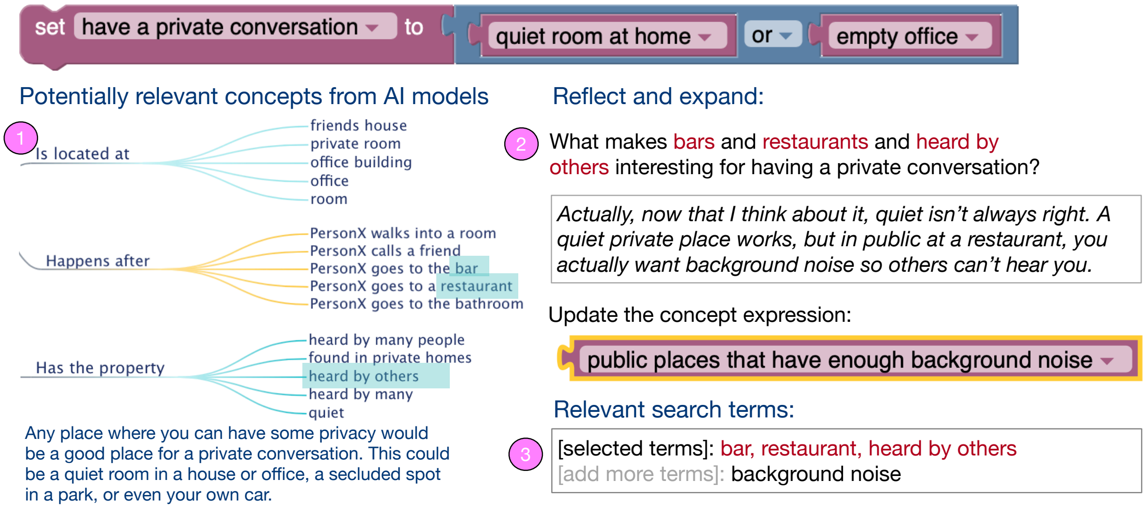

Figure 3: Mockup of human-AI tool for using AI models to support designers creating concept expressions, by (1) surfacing relevant concepts from AI models; (2) prompting for human reflection on selected concepts; and (3) using identified concepts to search for context features.

designers a richer vocabulary [27] with which to query into the metadata associated with context features (e.g., reviews within a Yelp category), making it easier to find relevant features and to increase coverage.

Evaluation Study: We will assess whether our proposed tools help designers develop and expand their conception of human situations, and translate it into a machine representation based on available context features. We will perform a between-subjects laboratory study that compares designers' use of Affinder with its basic features (concept variables, reflection prompts, and search) with the proposed AI-model enhancements (e.g., relevant concepts from AI models; smart reflection prompts; AI-assisted query expansion).

Participants: Early formative studies will draw on engineering design students from the Northwestern Segal Design Institute. Summative evaluations will enroll professional designers with 5+ years experience creating context-aware applications. We will use a stratified sampling approach to recruit a diverse population of designers across race, gender, socio-economic and urban/rural strata, and enroll approximately 50 creators total (25 in pilot studies, and 25 in the lab study). Final participant numbers will be based on preliminary testing, statistical power analysis, and sample size estimation.

Measures: To evaluate our tools, we will devise primary outcome measures that capture the breadth of concepts and ideas expressed, number of new concepts identified through interacting with the AI models, the accuracy of created concept expressions in identifying desired situations, affordances, and hindrances (i.e., precision), and the coverage of matching situations whereby users can engage in the defined context (i.e., recall). We will also collect process measures that capture how designers construct the context expressions and refine them. We will achieve this by analyzing changes in concept expressions, applying think-aloud protocols, analyzing screen recordings and log data, and via post-study interviews with designers. Our goal will be to understand not only the quality of a given concept construction, but also what the creators are trying to achieve and how their concept develops throughout the construction process. For instance, through interviews, we will assess a creator's ability to articulate how their concept expression allowed them to capture different ways of realizing their concept, and that work for a wider range of situations.

Overall the studies will collect and analyze measures of performance, process and subjective experience. Outcome variables that involve subjective judgment will require expert raters for assessment. Quantitative data will be analyzed using appropriate statistical tests and supplemented with qualitative data. In some cases, standard statistical techniques will be applicable (e.g., ANOVA, Mann-Whitney). For repeated mea-

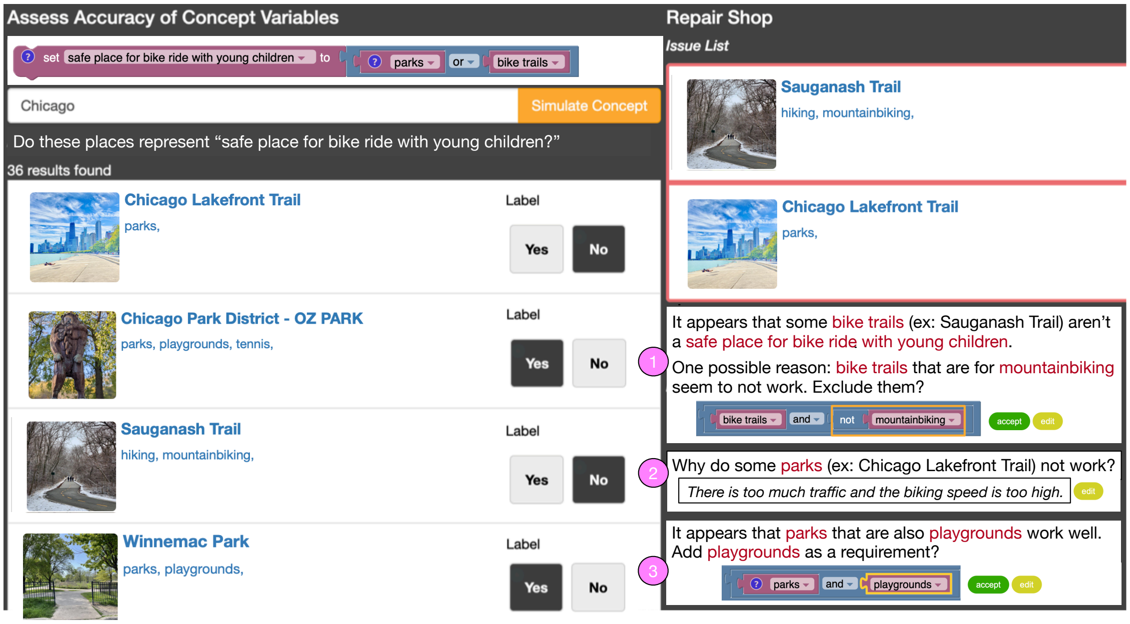

Figure 4: Mockup of an interactive tool for identifying and addressing causes for misalignments between the machine's representation and the expressed concept. After simulating the concept expression and labeling problematic cases as in our prior work, the proposed tools will (1) surface confounding context features that may need to be excluded; (2) prompt for reflection; and (3) suggest supporting context features to add.

sures, where the data are not independent, we will apply appropriate models such as Generalized Linear Mixed Models [9] or, for non-parametric repeated measures, Aligned Rank Transform techniques [67].

Expected Outcomes: Our work in this section will contribute an empirically-validated tool for improving concept development and expression. This work will advance human-AI principles for using AI models to support designers developing their own ideas, and linking them to available machine representations.

4 Human-AI Tools for Aligning to Machine Representations and Execution

Bridging is not only about conceptualizing and expressing a situation, but also how it is represented and executed by the machine. Machine context features may not accurately represent a designer's intentions, and target users may not encounter the situation defined in the concept expression during execution.

4.1 Addressing Failures in Machine Representation

Consider a designer constructing a concept expression to identify safe places to take a bike ride with young children. Initial context features may not operate as a designers intends; they may not precisely match the concept (e.g., mountain biking trails are bike trails, but not safe for young children), or they may reveal problems with the concept expression itself (e.g., not all parks are safe for riding with children). To support refining concept expressions, our prior work [49] designed affordances for simulating concept expressions to reveal cases in which the machine's representation fails to accurately represent specific concept variables. While this helps designers identify problematic cases, it remains difficult for designers to identify *generalizable causes* for such cases (e.g., specific confounding context features and their underlying machine learning models), based on which to refine and improve the concept expression.

To help with this, we propose new interactive tools that surface potential causes and prompt for reflection; see Figure 4. Taking a mixed-initiative approach [36, 26], upon labeling individual cases as in our prior work, our proposed tools will seek to identify generalizable causes for the issue by (1) surfacing *confounding context features* that may need to be excluded (e.g., 'mountain biking' as a confounding context feature of ‘bike trails’ that matched, but that are categorically not ‘safe places for bike ride with young children’), (2) prompting for human reflection (e.g., to recognize that high traffic or high bike speed areas are not safe); and (3) suggesting *supporting context features* to add (e.g., parks with playgrounds are safer). To do this, we will filter for context features that are not explicitly stated in the concept expression that match the problematic (or working) cases identified and that do not match the working (or problematic) cases identified. We expect designers using these tools can better understand why the machine may not understand a concept as intended, based on which they can make refinements on their own, or communicate the source of the issue to ML engineers on their team. Together, these affordances **support an ongoing human-machine feedback loop in which ML models and interactive tools help designers refine their concept expressions, and designers evaluate the usefulness of context features and the ML models they rely on.**

Evaluation Study: Following initial pilot studies as we develop our tools, we will conduct between-subjects lab studies to compare Affinder with the basic simulate and repair tools from our prior work to a version of Affinder with the proposed tools to assess whether designers are better able to find generalizable causes from problematic cases, based on which to refine concept expressions. We will curate a set of concept expressions that (1) contain machine features that don’t precisely match specific concept variables (feature issues); and (2) contain concept variables that do not precisely match the intended situation or experiences (concept issues). We will compare the number and types of confounding context features (e.g., hindrances) and supporting context features discovered, and changes in precision and recall from refining the concept expression. Participants will be similarly recruited as in the previous study.

Beyond these measures, we will study changes in designers’ understanding of the machine representation and its limitations, and in their ability to communicate gaps in human-machine understanding embedded in the concept expression. We will interview designers upon completing the task and design rubrics to measure their understanding of the machine representation, including: the depth of their description of the reasons for why an existing context feature fails to accurately represent a concept variable; unavailable context features that are needed to improve precision and recall; the embedded tradeoffs between precision and recall in the choice of context features; and any additional measures revealed through our pilot studies.

4.2 Addressing failures in execution

We turn now from issues of machine representational correctness to consider access issues in execution, that is, whether people encounter the contexts that are defined in the concept expression. Access issues can result from narrowly-defined concept expressions, but they can also occur due to a target population infrequently accessing contexts that match the concept expression (e.g., people rarely visit tourist sites in their own city); regional differences (i.e., there are many train stations in Chicago, few in Los Angeles and Phoenix); difference between the concept expression and people’s actual behavior (e.g., people perform actions that match the concept, but that differ from those tracked by the concept expression), and through the “conjunctions” of requirements across time, place, and other contextual factors (e.g., a late night commuting experience can be hard to access even though people commute daily).

Thinking about access demands that we consider what *interactional resources* might be available at the time of execution. Failure to do so can lead to context-aware experiences that are less inclusive than intended, or that otherwise prohibits participation. Instead of leaving the successful execution of context aware experiences to chance, we propose tools that **support designers reasoning about interactional resource availability** in the process of constructing concept expressions. Specifically, we propose to (1) build computational models of interactional resource availability; and to (2) operationalize them in interactive tools for building concept expressions so that the construction process is *interactional resource aware*. We hypothesize that this approach can help designers identify access issues early in the design process, and think

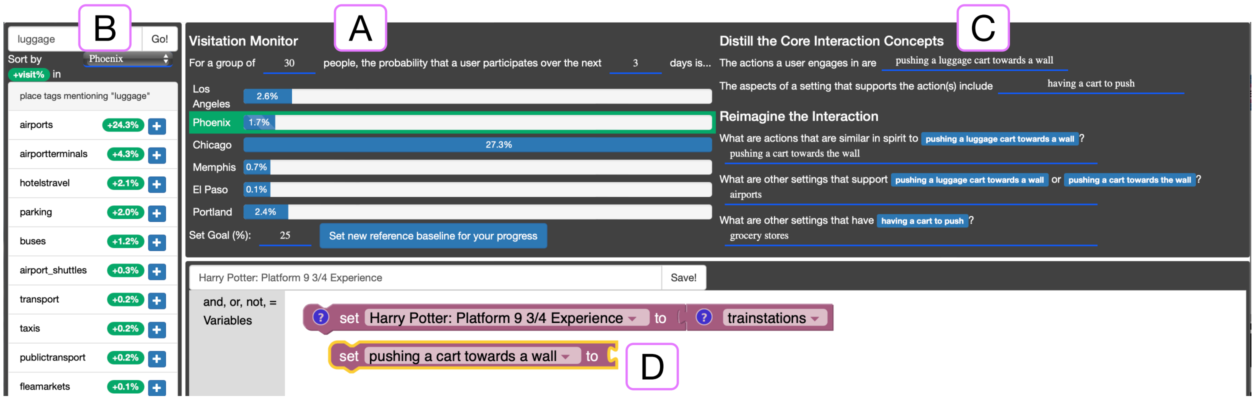

Figure 5: Mockup of interactive tools to support designers reasoning about interactional resource availability. In this example, a designer reformulates a concept expression for a Harry Potter social experience originally designed to take place at train stations by (A) monitoring the completion likelihood given the interactional resource availability across cities; (B) searching for context features in a resource-aware manner; (C) reflecting on core interactional requisites; and (D) updating the concept expression accordingly.

of alternative conceptions that will allow for more inclusive experiences that are more readily realizable in practice, based on predictions of what interactional resources will actually be available *during execution*.

We illustrate our proposed approach through the application of constructing concept expressions for supporting location-based, context-aware social experiences, such as in our prior work on designing opportunistic collective experiences [48]; see Figure 5. Such an experience structures an activity for people to engage in when they are in a situation (defined via a concept expression) that satisfies the interactional needs of the experience (e.g., engaging in a Harry Potter Platform 9 3/4's experience while at a train station). In this setting, we propose to build a model of people's interactional resource availability by modeling people's availability for accessing defined situations in their daily routine based on their visitation and participation likelihood, trained on their mobility data and their past participation histories (as we had done in our prior work [40]). Using this model, we propose to support designers actively monitoring for the likelihood of an experience completing given their current concept expression definition (Figure 5(A)). When the likelihood is low, we propose to help designers reformulate the concept expression by augmenting Affinder's search for context features with a *resource-aware search* that accounts for incremental changes in interactional resources when adding a context feature (Figure 5(B)). This helps designers reflect on the accessibility of their concept expression, and to build concept expressions that are accurate *and* accessible. Second, we propose to surface *reflection prompts* that ask designers to reflect on the core interactional requisites for their experience, so that they may consider removing reliance on surface features that hinder accessibility and finding alternative ways of realizing the core actions and qualities of the experience (Figure 5(C)). Similar to how, in the domain of design fixation, finding a functional schema can help with finding other relevant mechanisms for achieving a function or purpose [47], we hypothesize that focusing on core aspects of an experience should help a designer find useful alternative conceptions for reimagining the interactional needs of the experience, and reformulating the concept expression accordingly (Figure 5(D)).

Evaluation Study: We will conduct lab studies to understand how the proposed tools help designers reformulate opportunistic collective experiences (OCEs) [48] for which interactional needs are restrictive or hard to meet. An OCE defines a shared activity for a group of friends or family to engage in at distance when they are in situations that meet the interactional needs of the experience. We will specifically focus on understanding designers' processes for reformulation, and how our proposed tools help them broaden interactional needs so more people are able to participate, while mitigating the risk of changing an experience to the point where it loses its essential qualities and benefits. To do this, we will design initial concept expressions for social experiences that contain access issues (i.e., low interactional resource availability) due to: (1) infrequent visitations; (2) regional differences; (3) difference between user behavior and context tracked by the concept expression; and (4) "conjunctions" of context features. Prior to using any tools, we will first interview designers about their initial ideas about which and why certain interactional needs may be restrictive and how they would redesign the experience to fix this. Then, we will observe them using the reformulation tools on this task. Following each task, we will ask designers to describe how they reformulated the experience, and about which parts of the reformulation process they found effective or challenging, and how the tools helped with their making reasonable tradeoffs.

To further evaluate the effectiveness of reformulations for increasing interactional resource availability without negatively impacting the experience, we will conduct a deployment study to compare experiences pre- and post- reformulation. In a 2-week within-subjects study, we will measure the rate of participation and completion of experiences, and users' assessment of the experience for helping them connect with others, using similar social connectedness measures (e.g., [65, 5, 28, 68, 58, 45]) as we had used in our prior work [48]. We will recruit participants from university/college alumni groups and clubs, friend and family groups, and shared interest groups that only occasionally meet in person (e.g., hobbyists).

Expected Outcomes: Our work in this section will contribute empirically-validated (1) interactive tools for identifying and addressing causes for misalignments between the machine's representation and the expressed concept; and (2) a computational model and interactive tools for surfacing access issues in execution, and constructing more inclusive experiences. This work will also advance human-AI principles for identifying mismatches between human conceptions and machine representations and execution.

5 Human-AI Tools for Accounting for Differences Across Contexts of Use

A human experience that works in one context may not work in another. For example, while "cafes" in European cities may be good places to "linger over food," they are much less so in American cities. As another example, while some workplace challenges are common across occupations, many are job-specific or may manifest in very different ways. Being aware of such differences, and making refinements to how a concept expression should work *across settings*, is critical for creating robust human-AI and context-aware systems that are useful, equitable and accessible across diverse populations, settings, and regions.

5.1 Building differentiated models to account for differences across contexts of use

To support designers thinking about how a situation or experience may unfold differently across settings or regions, we propose techniques for evaluating, *differentially*, how well a concept or context feature matches a designer's intentions when considered across settings. Specifically, we propose to **repurpose AI models to expose differences in how concepts and context features apply** across different settings, by measuring the relative relevance of a context feature to a concept, as evaluated through building a *differentiated model* of how concepts connect to context features across settings and regions.

We illustrate this approach by applying it to evaluate (1) the relevance of a concept to a context-feature via a differentiated TF-IDF model, and (2) the relevance of concepts to an experience by querying large language models (LLMs) *differentially* across regions. In the first case, instead of computing the overall relevance of a context feature to a concept (e.g., a TF-IDF model in which the metadata associated with each context feature is a document), we propose to instead build a setting-specific TF-IDF model (where each document contains only metadata associated with a feature in a particular setting), and measure the difference in relevance across settings:

$$\text{TF-IDF}(d_{f,s1}) - \text{TF-IDF}(d_{f,s2})$$

where d_{fs} is the metadata associated with the context feature f in setting s. When this difference is large, it indicates that the context-feature f may be much more relevant in $s1$ than $s2$, suggesting that the designer may wish to find other ways to realize their concept in $s2$, for example by looking at which other context features achieve a high TF-IDF score in $s2$. The designer can then refine their concept expression accordingly, based on their refined understanding of such differences.

As another technique, we consider how LLMs that support conceptualization (see Section 3.1) can be repurposed to reveal cross-setting/cross-region differences by querying it with setting-specific prompts. We can then measure the relative relevance of a response (i.e., concepts associated with a situation) to a prompt (a human situation), when considered in one setting vs. another. We do this by constructing a region/setting-specific prompt (e.g., "what do people do for fun in *Brazil* (or, *Korea*)?," "what are challenges facing *tech workers* (or, *medical workers*)?"), and measuring the difference in perplexity [39] between the generic response R_g (when not conditioned on the region/setting) and the region specific response R_s:

$$\text{Perplexity}(R_g) - \text{Perplexity}(R_s)$$

When the difference is high, we can surface to designers the discrepancy in how well the concept is matching, and also show the best response from the model for that particular setting (R_s). This too can help designers conceptualize alternative ways of realizing the concept in this other setting.

5.2 Applying reference systems to differentiated models

To operationalize the differentiated models we propose to construct, we need *reference systems* on which they are to be built. Here a reference system refers to an *accountable perspective* of difference to consider in creating the concept expression (e.g., across geographies, industries, age, rural/urban strata, socio-economic, etc.) Each entity in a reference system refers to a particular setting in which the concept expression may be applied (e.g., a city or country in a geographic reference system; an occupation or field in a jobs reference system). The reference systems we choose to consider can be based on the goals of an application (e.g., creating context-aware experiences that work across geographies, or AI support systems that work across jobs and industries), or on known biases that can lead to costly errors or issues of exclusion if not accounted for (e.g., socio-economic differences; rural/urban divides, etc).

Applying a differentiated model to a reference system requires building and querying the models differentially across the entities (i.e., settings) defined in the reference system. For instance, to build a differentiated TF-IDF model across a geographic reference system, we can train the model by segmenting metadata associated with a place-based context feature (e.g., the reviews for each Yelp category) by city or country. Similarly, for LLMs, we can query an LLM (e.g., GPT-3) by replacing the generic terms in our query (e.g., workers) with a more specific entity in the reference system (e.g., tech workers, medical workers). We hypothesize that whenever the differentiated models accurately encode meaningful differences across entities, such differences can be surfaced to designers through our approach to support their reflecting and refining their concept expression to better account for differences across populations, regions, and settings.

5.3 Visualizing cross-setting differences

We will generate interactive visualizations to support designers reflecting on differences across settings by applying the *explicit spatialization* approach that we developed [35]. We describe how to overlay information from the differentiated models, and surface differences in interactional resources (from Section 4.2).

Specifically, given a visual representation of a reference system (e.g., a map for a geographic reference system) and a concept expression, we propose to generate an interactive thematic cartography layer (e.g.,

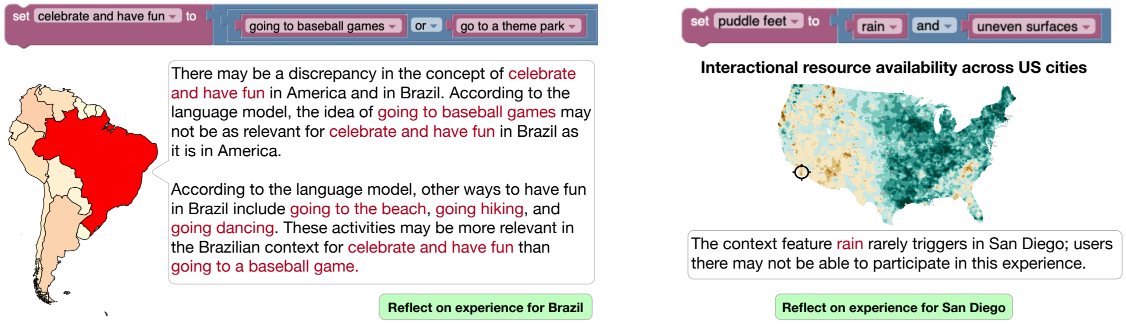

Figure 6: Mockup of interactive visualizations for analyzing differences across settings. The visualization and system-generated explanations explains how concept variables that work well in one region fail to match the high-level concept in another region (Left), and how interactional resource availability across settings considered in the reference system may vary in ways that can impact the experience (Right).

a heat map or choropleth) on top to visualize differences in (1) how well concept variables match context features, and how well concept variables match the high-level situation or experience; and (2) interactional resource availability across the entities in the reference system; see Figure 6. In addition, we will generate human-readable explanations and suggestions using our models. For example, upon clicking on a country for which a concept expression may not operate well as constructed, our system will use the differentiated models to highlight context-features and concept variables that may be poor matches for that country, and suggest alternative conceptions and context-features that may work better for the designer to reflect upon (Figure 6(left)). Similarly, we will help designers reflect on the cause for low interactional resources by surfacing which context features in the expression rarely match people's contexts in a particular setting (Figure 6(right)), and engage designers to consider resource-aware reformulations (as in Section 4.2).

Evaluation Study: We will assess (1) the accuracy of the differentiated models and (2) the usefulness of our interactive visualizations that use these models to support designers reflecting on differences across settings and reformulating concept expressions accordingly. To assess the accuracy of our models, we will create concept expressions that contain concept variables and context features that are likely to work better in one setting versus another (e.g., based on known differences across regions, industries, or populations), and also include concept expressions created in earlier studies. Running our differentiated models, we will measure *differential precision* based on the frequency that known differences were detected by our models; and *differential recall* based on coverage across the concepts and concept expressions we test. In addition to measuring performance with respect to known differences we expect to be see, we will also examine what the models uncover to better understand the capabilities and limitations of our approach.

To evaluate our interactive visualizations, we will examine the extent to which provided visualizations, human-readable explanations, and suggestions help designers reflect on and form a clearer and broader conception of differences across settings than what they are aware of and can recognize on their own. Here we are interested in not only understanding how designers are able to refine concept expressions to account for cross-setting differences and low interactional resources, but in how the process of exploring our visualizations with respect to different reference systems can help designers gain a deeper understanding of the design space and the various facets of difference that they may need to further account for. In this direction, we will additionally conduct exploratory studies that project differences onto additional reference systems on which the differentiated models are not trained. For instance, we can train a differentiated model to capture differences across occupations, but surface to designers a visualization that further groups occupations by industry, median annual income, size and growth of the sector, etc. We expect that this can help designers gain a broader perspective as they reflect on the underlying causes for the differences that are surfaced.

Expected Outcomes: Our work in this section will contribute general techniques for building (1) *differentiated models* that account for differences across settings, (2) *reference systems* to apply to these models; and (3) *interactive visualizations* that support reflecting on differences and refining concept expressions. This work broadly advances human-AI principles for accounting for differences across contexts of use.

6 Creator Studies

Beyond studying the proposed tools for conceptualizing, aligning, and accounting for difference in creating conception expressions, we plan to build and deploy a public Affinder platform that will integrate our learnings across the three areas of study in Year 3. Our goal is to use this platform to study how creators construct, refine, and use concept expression across all aspects of our system, and over longer-term deployments. With this approach, we expect to learn more about how our platform will be used by designers in the ways they intend, some of which will be unexpected and reveal new obstacles and opportunities for our research.

Our platform will integrate three new components into Affinder: (1) interactive tools that use AI models to support designers creating concept expressions (section 3); (2) interactive tools and computational models for identifying and acting on general mismatches between the machine's representation and the expressed concept, and access issues in execution due to low interactional resources (section 4); and (3) interactive visualizations of cross-setting differences powered by differentiated models (section 5). With this platform, we will conduct lab-based and deployment-based *creator studies*:

Lab-based creator studies will center on understanding the range of concept expressions and experiences created by designers with access to our entire toolkit, both in response to preset scenarios that we will curate and in completely open-ended tasks. Empirical analyses will focus on classifying the types of experiences and concept expressions created, and a reflection on the overall design space. In observed sessions, we will use think-alouds and post-interviews to analyze the process by which designers build concept expressions, particularly the ways that our tools support (1) their broadening and refining their conceptions; (2) addressing mismatches in machine representation and execution; and (3) accounting for cross-cultural and cross-setting differences. Using screen recordings, we will conduct followup interviews to investigate aspects of a designer's process that exhibit interesting uses of our tools, and major shifts in their conception.

Deployment-based creator studies will focus on evaluating how creators make use of our platform to create and refine experiences over time. Here we are interested in not only understanding how creators use Affinder, but also in how concept expressions designers create using Affinder actually support (or not) the experiences that they are making. To do this, we will deploy Affinder as a public platform and recruit users of our platform for monthly interviews to understand their intended goals for the context-aware experiences they are creating, their perceptions of their usage of Affinder, any refinements they make to concept expressions and the experience over the course of the deployment, and why.

Expected Outcomes: Our work on creator studies will contribute (1) a public Affinder platform that integrates all our learnings; (2) general principles for incorporating human concepts into AI-enable applications; and (3) an enhanced understanding of the limitations and opportunities for our proposed approach.

7 Broader Impacts

Our work seeks to provide a critical bridge between human conception and AI systems so that AI systems can more effectively support a variety of humanly important activities. We do this by promoting transparency in AI systems, by constructing and surfacing how they are used to represent human concepts. This largely advances a human-first approach to using AI technologies that mitigates the shortcomings of AI-first approaches that can lead to costly errors in safety and privacy, or lead to inequitable access to AI supported experiences. To further account for and counter the harmful effects of AI bias, our work explicitly promotes cross-cultural inclusion by promoting understanding of the differential impacts of AI technologies on different populations and across settings. Inline with computational social scientists and journalists who critically examine problems in AI systems, our work helps to expose challenges in using AI systems to support human activities and experiences by surfacing their limitations, biases, and exclusions. But unlike critiques offered post-deployment that are not always actionable [55], our proposed work provides system builders and experience designers with tools that make issues that arise immediately noticeable and actionable early in the design and development process. Moreover, by making AI systems more accessible to designers for supporting a wide range of human activities, the proposed work can expand the US workforce that is empowered by AI [62] by enabling more workers across sectors of our economy to leverage the powers of AI systems to enable new applications. This democratizes AI to a larger set of domain experts and individuals who can access and program AI systems beyond ML engineers [63].

To maximize our broader impacts, we will (1) release our developed tools as a public platform for constructing concept expressions across a wide range of applications and use cases; (2) actively recruit study participants from diverse backgrounds to construct concept expressions that match their lived experiences; and (3) advance research education through Design, Technology, and Research (DTR), a long-running research training program the PI directs [69, 61]. The DTR program has trained 100+ undergraduate and graduate students in independent research to-date, and we have created resources, tools, training programs, and a documentary film [61] to support 70+ faculty across institutions [2]. The proposed research will directly train at least 10 more undergraduate students in research. The PIs will make continued efforts to Broaden Participation in Computing (BPC) by actively recruiting underrepresented students quarterly through Women in Computing; National Society of Black Engineers; and Society of Hispanic Professional Engineers. Already, 40% of DTR undergraduates are female, far surpassing the national average of only 20% as reported in the latest CRA Taulbee Survey [70].

8 Results from Prior NSF Support

IIS-1618096 CHS: Small: Coordination of Opportunistic Actions to Produce Globally Effective Behaviors for Physical Crowdsourcing (PI: Zhang, $496,380, 07/1/2016-6/30/2021): **Intellectual Merit:** This work developed theory, methods and frameworks for physical crowdsourcing. The project produced (1) techniques for collecting physical data through people's daily routines [30, 46]; and (2) decision-theoretic frameworks for flexible coordination [40, 41]. **Broader Impact:** This project enables new socio-technical systems that empower people to solve local, communal problems. The project has trained 16 students (3 PhD, 13 undergrads), including 7 female students. **Evidence of Research Products:** This project produced five publications [30, 40, 46, 41, 48] in top venues (CHI, CSCW, UIST, HCOMP). Software products include ExperienceKit and the Physical Crowdsourcing API for creating physical crowdsourcing applications.

IIS-1901456 CHS: Medium: Next Generation Content Production Tools for People with Vision Impairments (PI: Gergle, $1,199,903, 10/01/2019-9/30/2023): **Intellectual Merit:** This work advances accessible content production tools by bridging the gulf between current tools and the accessibility needs of blind professionals. The work establishes accessibility practices [22, 59, 18], innovates content production tools for blind producers in both collaborative writing and physical "maker" contexts [6, 17], and evaluates their effectiveness for collaboration [21, 51, 16]. **Broader Impacts:** The project, thus far, has trained 5 PhDs 4 undergraduate students and established a research registry to support the development of collaborative systems for ability-diverse teams. **Evidence of Research Products:** The project has, to date, produced 12 publications [22, 6, 20, 19, 59, 18, 17, 51, 16, 29, 50] in top venues (CHI, CSCW, ASSETS, TOCHI) and includes 3 Best Paper/Honorable Mention Awards and 1st Place ACM SRC Grand Finals. Software products include the V11 Framework, a programming interface to support community-driven assistive technology.

References cited

[1] Gregory D Abowd and Elizabeth D Mynatt. Charting past, present, and future research in ubiquitous computing. *ACM Transactions on Computer-Human Interaction (TOCHI)*, 7(1):29–58, 2000.

[2] Agile Research University. http://agileresearch.io.

[3] Saleema Amershi, Maya Cakmak, William Bradley Knox, and Todd Kulesza. Power to the people: The role of humans in interactive machine learning. *Ai Magazine*, 35(4):105–120, 2014.

[4] Saleema Amershi, Max Chickering, Steven M Drucker, Bongshin Lee, Patrice Simard, and Jina Suh. Modeltracker: Redesigning performance analysis tools for machine learning. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, pages 337–346, 2015.

[5] D.T. Bel, van, K.C.H.J. Smolders, W.A. IJsselsteijn, and Y.A.W. Kort, de. Social connectedness : concept and measurement. In *International Conference on Intelligent Environments*, pages 67–74, Netherlands, 2009. IOS Press.

[6] Katya Borgos-Rodriguez, Maitraye Das, and Anne Marie Piper. Melodie: A design inquiry into accessible crafting through audio-enhanced weaving. *ACM Trans. Access. Comput.*, 14(1), March 2021.

[7] Ann L Brown. Design experiments: Theoretical and methodological challenges in creating complex interventions in classroom settings. *The journal of the learning sciences*, 2(2):141–178, 1992.

[8] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

[9] Joe Bruin. Introduction to linear mixed models, February 2011.

[10] Jacques Bughin, Jeongmin Seong, James Manyika, Michael Chui, and Raoul Joshi. Notes from the AI frontier: Modeling the impact of AI on the world economy. *McKinsey Global Institute*, 4, 2018.

[11] Scott Allen Cambo and Darren Gergle. Model positionality and computational reflexivity: Promoting reflexivity in data science. In *CHI Conference on Human Factors in Computing Systems*, pages 1–19, 2022.

[12] Joseph Chee Chang, Saleema Amershi, and Ece Kamar. Revolt: Collaborative crowdsourcing for labeling machine learning datasets. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, pages 2334–2346, 2017.

[13] Nan-Chen Chen, Jina Suh, Johan Verwey, Gonzalo Ramos, Steven Drucker, and Patrice Simard. Anchorviz: Facilitating classifier error discovery through interactive semantic data exploration. In *23rd International Conference on Intelligent User Interfaces*, pages 269–280, 2018.

[14] Nathan Crilly and Carlos Cardoso. Where next for research on fixation, inspiration and creativity in design? *Design Studies*, 50:1–38, 2017.

[15] Robert Dale. Gpt-3: What's it good for? *Natural Language Engineering*, 27(1):113–118, 2021.

[16] Maitraye Das. *Designing for Collaborative Content Creation for People with Vision Impairments*, page 105–110. Association for Computing Machinery, New York, NY, USA, 2020.

[17] Maitraye Das, Katya Borgos-Rodriguez, and Anne Marie Piper. Weaving by touch: A case analysis of accessible making. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, CHI '20, page 1–15, New York, NY, USA, 2020. Association for Computing Machinery.

[18] Maitraye Das, Darren Gergle, and Anne Marie Piper. "it doesn't win you friends": Understanding accessibility in collaborative writing for people with vision impairments. *Proc. ACM Hum.-Comput. Interact.*, 3(CSCW), November 2019.

[19] Maitraye Das, Thomas Barlow McHugh, Anne Marie Piper, and Darren Gergle. Co11ab: Augmenting accessibility in synchronous collaborative writing for people with vision impairments. CHI '22, New York, NY, USA, 2022. Association for Computing Machinery.

[20] Maitraye Das, Anne Marie Piper, and Darren Gergle. Design and evaluation of accessible collaborative writing techniques for people with vision impairments. *ACM Trans. Comput.-Hum. Interact.*, 29(2), jan 2022.

[21] Maitraye Das, Anne Marie Piper, and Darren Gergle. Design and evaluation of accessible collaborative writing techniques for people with vision impairments. *ACM Trans. on Computer-Human Interaction*, In Press.

[22] Maitraye Das, John Tang, Kathryn E. Ringland, and Anne Marie Piper. Towards accessible remote work: Understanding work-from-home practices of neurodivergent professionals. *Proc. ACM Hum.-Comput. Interact.*, 5(CSCW1), April 2021.

[23] Paul Dourish. Seeking a foundation for context-aware computing. *Human–Computer Interaction*, 16(2-4):229–241, 2001.

[24] Matt Easterday, Dan Rees Lewis, and Elizabeth Gerber. Design-based research process: Problems, phases, and applications. In *Proc. of International Conference of Learning Sciences*, volume 14, 2014.

[25] Luciano Floridi and Massimo Chiriatti. Gpt-3: Its nature, scope, limits, and consequences. *Minds and Machines*, 30(4):681–694, 2020.

[26] James Fogarty, Desney Tan, Ashish Kapoor, and Simon Winder. Cueflik: interactive concept learning in image search. In *Proceedings of the sigchi conference on human factors in computing systems*, pages 29–38, 2008.

[27] George W. Furnas, Thomas K. Landauer, Louis M. Gomez, and Susan T. Dumais. The vocabulary problem in human-system communication. *Communications of the ACM*, 30(11):964–971, 1987.

[28] Simon Gächter, Chris Starmer, and Fabio Tufano. Measuring the closeness of relationships: a comprehensive evaluation of the'inclusion of the other in the self'scale. *PloS one*, 10(6):e0129478, 2015.

[29] Hugo Flores Garcia, Aldo Aguilar, Ethan Manilow, Dmitry Vedenko, and Bryan Pardo. Deep learning tools for audacity: Helping researchers expand the artist's toolkit, 2021.

[30] Kapil Garg, Yongsung Kim, Darren Gergle, and Haoqi Zhang. 4x: A hybrid approach for scaffolding data collection and interest in low-effort participatory sensing. In *Proceedings of the 22nd ACM Conference on Computer Supported Cooperative Work and Social Computing (CSCW 19')*, 2019.

[31] Darren Gergle, Robert E Kraut, and Susan R Fussell. Using visual information for grounding and awareness in collaborative tasks. *Human–Computer Interaction*, 28(1):1–39, 2013.

[32] Katy Ilonka Gero, Vivian Liu, and Lydia Chilton. Sparks: Inspiration for science writing using language models. In *Designing Interactive Systems Conference*, pages 1002–1019, 2022.

[33] Yolanda Gil, Bart Selman, Marie Desjardins, Ken Forbus, Kathy Mckeown, Dan Weld, and Ann Drobnis. *A 20-year community roadmap for Artificial Intelligence Research in the US*. Computing Community Consortium (CCC) and Association for the Advancement of Artificial Intelligence (AAAI), 2019.

[34] Ben Green and Yiling Chen. Disparate interactions: An algorithm-in-the-loop analysis of fairness in risk assessments. In *Proceedings of the conference on fairness, accountability, and transparency*, pages 90–99, 2019.

[35] Brent Hecht, Samuel H Carton, Mahmood Quaderi, Johannes Schöning, Martin Raubal, Darren Gergle, and Doug Downey. Explanatory semantic relatedness and explicit spatialization for exploratory search. In *Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval*, pages 415–424, 2012.

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

[37] Jena D Hwang, Chandra Bhagavatula, Ronan Le Bras, Jeff Da, Keisuke Sakaguchi, Antoine Bosselut, and Yejin Choi. (comet-)atomic 2020: On symbolic and neural commonsense knowledge graphs. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 6384–6392, 2021.

[38] David G Jansson and Steven M Smith. Design fixation. *Design studies*, 12(1):3–11, 1991.

[39] Daniel Jurafsky and James H Martin. N-gram language models. *Speech and language processing*, 23, 2018.

[40] Yongsung Kim, Darren Gergle, and Haoqi Zhang. Hit-or-wait: Coordinating opportunistic low-effort contributions to achieve global outcomes in on-the-go crowdsourcing. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, page 96. ACM, 2018.

[41] Yongsung Kim, Emily Harburg, Shana Azria, Aaron Shaw, Elizabeth Gerber, Darren Gergle, and Haoqi Zhang. Studying the effects of task notification policies on participation and outcomes in on-the-go crowdsourcing. *AAAI HCOMP*, 2016.

[42] Hyung-Kwon Ko, Gwanmo Park, Hyeon Jeon, Jaemin Jo, Juho Kim, and Jinwook Seo. Large-scale text-to-image generation models for visual artists' creative works. *arXiv preprint arXiv:2210.08477*, 2022.

[43] Todd Kulesza, Saleema Amershi, Rich Caruana, Danyel Fisher, and Denis Charles. Structured labeling for facilitating concept evolution in machine learning. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 3075–3084, 2014.

[44] Todd Kulesza, Margaret Burnett, Weng-Keen Wong, and Simone Stumpf. Principles of explanatory debugging to personalize interactive machine learning. In *Proceedings of the 20th international conference on intelligent user interfaces*, pages 126–137, 2015.

[45] Daniël Lakens and Mariëlle Stel. If they move in sync, they must feel in sync: Movement synchrony leads to attributions of rapport and entitativity. *Social Cognition*, 29(1):1–14, 2011.

[46] Katherine Lin, Henry Spindell, Scott Cambo, Yongsung Kim, and Haoqi Zhang. Habitsourcing: Sensing the environment through immersive, habit-building experiences. In *Proceedings of the 29th Annual Symposium on User Interface Software and Technology*, pages 639–650. ACM, 2016.

[47] J. S. Linsey, A. B. Markman, and K. L. Wood. Design by Analogy: A Study of the WordTree Method for Problem Re-Representation. *Journal of Mechanical Design*, 134(4), 04 2012. 041009.

[48] Ryan Louie, Kapil Garg, Jennie Werner, Allison Sun, Darren Gergle, and Haoqi Zhang. Opportunistic collective experiences: Identifying shared situations and structuring shared activities at distance. *Proceedings of the ACM on Human-Computer Interaction*, 4(CSCW3):1–32, 2021.

[49] Ryan Louie, Darren Gergle, and Haoqi Zhang. Affinder: Expressing concepts of situations that afford activities using context-detectors. In *CHI Conference on Human Factors in Computing Systems*, pages 1–18, 2022.

[50] Thomas B. McHugh. *Constructing Agency and Usability Through Community-Driven Assistive Technology Design*. Association for Computing Machinery, New York, NY, USA, 2021.

[51] Thomas B. McHugh and Cooper Barth. Assistive technology design as a computer science learning experience. In *The 22nd International ACM SIGACCESS Conference on Computers and Accessibility*, ASSETS '20, New York, NY, USA, 2020. Association for Computing Machinery.

[52] Melanie Mitchell. Artificial intelligence hits the barrier of meaning. *Information*, 10(2):51, 2019.

[53] Joon Sung Park, Lindsay Popowski, Carrie Cai, Meredith Ringel Morris, Percy Liang, and Michael S Bernstein. Social simulacra: Creating populated prototypes for social computing systems. In *Proceedings of the 35th Annual ACM Symposium on User Interface Software and Technology*, pages 1–18, 2022.

[54] Han Qiao, Vivian Liu, and Lydia Chilton. Initial images: Using image prompts to improve subject representation in multimodal ai generated art. In *Creativity and Cognition*, pages 15–28, 2022.

[55] Inioluwa Deborah Raji, Andrew Smart, Rebecca N White, Margaret Mitchell, Timnit Gebru, Ben Hutchinson, Jamila Smith-Loud, Daniel Theron, and Parker Barnes. 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*, pages 33–44, 2020.

[56] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 2022.

[57] Alexander Ratner, Stephen H Bach, Henry Ehrenberg, Jason Fries, Sen Wu, and Christopher Ré. Snorkel: Rapid training data creation with weak supervision. *The VLDB Journal*, 29(2):709–730, 2020.

[58] Paul Reddish, Ronald Fischer, and Joseph Bulbulia. Let's dance together: synchrony, shared intentionality and cooperation. *PloS one*, 8(8):e71182, 2013.

[59] Abir Saha and Anne Marie Piper. Understanding audio production practices of people with vision impairments. In *The 22nd International ACM SIGACCESS Conference on Computers and Accessibility*, ASSETS '20, New York, NY, USA, 2020. Association for Computing Machinery.

[60] Daniel Salber, Anind K Dey, and Gregory D Abowd. The context toolkit: aiding the development of context-enabled applications. In *Proceedings of the SIGCHI conference on Human Factors in Computing Systems*, pages 434–441. ACM, 1999.

[61] Sergio Salgado, Sarah Hanson, and Haoqi Zhang. *Forward: A Story about Learning and Growth*. FurnaceFPS, 2020.

[62] National Science and Technology Council (US). Select Committee on Artificial Intelligence. *The national artificial intelligence research and development strategic plan: 2019 update*. National Science and Technology Council (US), Select Committee on Artificial Intelligence., 2019.

[63] Patrice Y. Simard, Saleema Amershi, David Maxwell Chickering, Alicia Edelman Pelton, Soroush Ghorashi, Christopher Meek, Gonzalo A. Ramos, Jina Suh, Johan Verwey, Mo Wang, and John Robert Wernsing. Machine teaching: A new paradigm for building machine learning systems. *ArXiv*, abs/1707.06742, 2017.

[64] Harini Suresh and John Guttag. A framework for understanding sources of harm throughout the machine learning life cycle. In *Equity and Access in Algorithms, Mechanisms, and Optimization*, EAAMO '21, New York, NY, USA, 2021. Association for Computing Machinery.

[65] Daniel T van Bel, KCHJ Smolders, Wijnand A IJsselsteijn, and Yvonne de Kort. Social connectedness: concept and measurement. *Intelligent Environments*, 2:67–74, 2009.

[66] Emily Wall, Soroush Ghorashi, and Gonzalo Ramos. Using expert patterns in assisted interactive machine learning: A study in machine teaching. In *IFIP Conference on Human-Computer Interaction*, pages 578–599. Springer, 2019.

[67] Jacob O Wobbrock, Leah Findlater, Darren Gergle, and James J Higgins. The aligned rank transform for nonparametric factorial analyses using only anova procedures. In *Proceedings of the SIGCHI conference on human factors in computing systems*, pages 143–146. ACM, 2011.

[68] Svetlana Yarosh, Stephen Cuzzort, Hendrik Müller, and Gregory D. Abowd. Developing a Media Space for Remote Synchronous Parent-child Interaction. In *Proceedings of the 8th International Conference on Interaction Design and Children*, IDC '09, pages 97–105, New York, NY, USA, 2009. ACM.

[69] Haoqi Zhang, Matthew W Easterday, Elizabeth M Gerber, Daniel Rees Lewis, and Leesha Maliakal. Agile research studios: Orchestrating communities of practice to advance research training. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*, pages 220–232. ACM, 2017.

[70] Stuart Zweben and Betsy Bizot. 2020 CRA Taulbee survey. *Computing Research News*, 33(5), 2021.