Disability-first Dataset Creation: Lessons from Constructing a Dataset for Teachable Object Recognition with Blind and Low Vision Data Collectors

Lida Theodorou
City, University of London
London, UK
leda.theodorou@gmail.com

Daniela Massiceti
Microsoft Research
Cambridge, UK
dmassiceti@microsoft.com

Luisa Zintgraf
University of Oxford
Oxford, UK
lmzintgraf@gmail.com

Simone Stumpf
City, University of London
London, UK
Simone.Stumpf.1@city.ac.uk

Cecily Morrison
Microsoft Research
Cambridge, UK
cecilym@microsoft.com

Edward Cutrell
Microsoft Research
Redmond,WA, USA
cutrell@microsoft.com

Matthew Tobias Harris
City, University of London
London, UK
toby@tobyz.net

Katja Hofmann
Microsoft Research
London, UK
Katja.Hofmann@microsoft.com

ABSTRACT

Artificial Intelligence (AI) for accessibility is a rapidly growing area, requiring datasets that are inclusive of the disabled users that assistive technology aims to serve. We offer insights from a multi-disciplinary project that constructed a dataset for teachable object recognition with people who are blind or low vision. Teachable object recognition enables users to teach a model objects that are of interest to them, e.g., their white cane or own sunglasses, by providing example images or videos of objects. In this paper, we make the following contributions: 1) a disability-first procedure to support blind and low vision data collectors to produce good quality data, using video rather than images; 2) a validation and evolution of this procedure through a series of data collection phases and 3) a set of questions to orient researchers involved in creating datasets toward reflecting on the needs of their participant community.

CCS CONCEPTS

• **Human-centered computing** → **Accessibility**; *accessibility systems and tools*; accessibility technologies; • **Computing methodologies** → Machine learning.

KEYWORDS

AI, accessibility, datasets, teachable object recognition, blind and low vision users

ASSETS '21, October 18–22, 2021, Virtual Event, USA

ACM ISBN 978-1-4503-8306-6/21/10…$15.00
https://doi.org/10.1145/3441852.3471225

ACM Reference Format:
Lida Theodorou, Daniela Massiceti, Luisa Zintgraf, Simone Stumpf, Cecily Morrison, Edward Cutrell, Matthew Tobias Harris, and Katja Hofmann. 2021. Disability-first Dataset Creation: Lessons from Constructing a Dataset for Teachable Object Recognition with Blind and Low Vision Data Collectors. In *The 23rd International ACM SIGACCESS Conference on Computers and Accessibility (ASSETS '21), October 18–22, 2021, Virtual Event, USA.* ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3441852.3471225

1 INTRODUCTION

Artificial Intelligence (AI) is opening up new ways for people with disabilities to access the world [12, 46]. Object recognition has been one of the early uses of AI by people who are blind or low vision, in apps such as Seeing AI [16] (Figure 1). However, generic AI object recognition is currently limited to finding common items, such as chairs or the door of a room. Hence, attention has shifted to teachable object recognisers [4, 37], which can help users identify objects that are: 1) not covered by common generic categories (e.g., white canes); and 2) specific instances of an object, such as a friend's car or their favourite mug. In this case, users can "teach" an object recogniser by providing a few, say 5-10, example images or videos of these objects, enabling users to extend current object recognition algorithms to meet their own needs.

Enabling this application requires new machine learning (ML) techniques that work well when a small number of examples are available, termed "few-shot" learning [15, 51], and, crucially, appropriate user-centric datasets to drive ML innovation in this space. Unfortunately, current datasets either are not structured in a way to support user-centric few-shot learning [33], are too small [28], or do not include data from blind and low vision people. Other approaches that try to build applications from data collected from sighted people [37], or simulate data as if collected from people with disabilities [53] risk de-valuing contributions from the beneficiary community. A large dataset that reflects the data of the eventual end users, the blind and low vision community, is desperately needed to create feasible teachable object recognisers in this space. But how

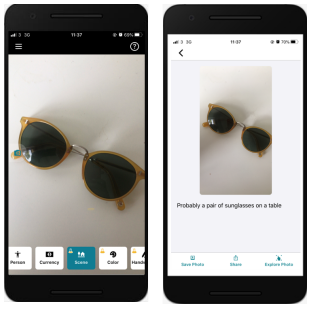

Figure 1: Seeing AI uses generic object recognition to identify a pair of sunglasses.

do we construct such a dataset that includes and supports people who are blind or low vision to contribute as data collectors while also ensuring that their contribution can successfully be integrated into the ML development pipeline?

Taking a disability-first approach to data collection, the multidisciplinary team in the ORBIT project (https://orbit.city.ac.uk/) collected a dataset[1] for teachable object recognisers from blind and low vision people. Disability-first suggests an approach that is used to serve a disability community first but then could be generalised to serve all people through the innovation that it enables. It stands in opposition to mainstream ML datasets and approaches which are later augmented or co-opted to address issues of importance to disabled communities. The team spent considerable time and iteration crafting a disability-first data collection procedure intended to strike a careful balance between the structure and fidelity of data required for machine learning innovation and the demands put on data collectors. In this paper, we describe the development of this procedure to support data collectors who are blind or low vision to contribute videos to our dataset. Using iterations of data collection, we reflect upon these design choices and how they balance between data collectors' needs and the requirements for a useful machine learning dataset. In doing so, we make the following contributions:

(1) An example disability-first procedure that enables data collectors who are blind or low vision to contribute to a dataset useful for developing teachable object recognisers.
(2) The validation and evolution of this procedure through a series of data collection phases, reflecting the tension between ML innovation and demands on the data collectors.
(3) Eight orienting questions to those creating datasets to encourage innovation in AI for accessibility, helping researchers reflect on the needs of their participant disability community as data collectors.

2 RELATED WORK

2.1 Disability-first ML Innovations

ML systems rely on data, both in their development of algorithms as well as in their application in the real world. Often data from disabled people differ in key ways from data from non-disabled people which leads to applications that do not perform as well for people with disabilities [20, 42]. Research by Lee et al. [37], for example, has used data collected from a sighted and a blind individual to explore teachable object recognisers for blind users. Others have used simulated data that mimics data from disabled users [48, 53]. These approaches lead to data that is used to develop ML algorithms being very different to data that is later encountered in application use; for example, pictures collected by blind and low vision people tend to be blurrier and might show an object partially or totally out-of-frame compared with sighted users.

In particular, this has been noted for computer vision systems for blind and low vision users, where data from blind and low vision people is scarce and has been treated, and arguably de-valued, as 'outliers' compared to sighted users [20, 41]. Hence, there have been calls to include disabled people more extensively in the creation of AI technologies, to design with and not for – and to be true to the credo of "nothing about us without us" [20, 41, 53, 56]. It is hence critical to include people with disabilities in data collection that underlies ML innovation.

The Incluset database [29], an online repository of datasets specifically curated for driving ML innovation in accessibility, has been released to address this challenge, bringing the broad diversity of datasets collected from persons with disabilities to researchers and practitioners. Datasets from people with disabilities however tend to be small, making it difficult to innovate in machine learning and hence, to engage the machine learning research community in the innovation process. This literature has informed our view that it is critical to tackle the creation of large disability-first datasets.

2.2 Datasets for Teachable Object Recognisers

Teachable object recognisers [30, 37] provide a way for people to "teach" an AI system about a new object that they may want to identify, by providing training examples themselves. This approach could potentially address many of the short-comings of generic object recognition for users as new objects can be added quickly and as needed. Achieving this vision, however, requires innovation in few-shot ML techniques [13, 18, 55, 61] that are optimised to work on small amounts of *real-world* data from people who are blind and low vision.

Existing few-shot learning datasets (and benchmarks) have led to valuable ML research insights; accuracy on existing benchmark datasets is often very high (e.g., >80% on MiniImagenet's 5-object classification task [55]). However, these benchmarks present overly simplistic tasks [35, 45, 55] or ones with low ecological validity [54], rendering trained models not appropriate for deployment in the real-world. Hence, further ML innovation is required to make these techniques useful to blind and low vision users, which in turn necessitates an appropriate dataset for further research.

A way forward would be to turn to existing datasets containing objects collected from blind and low vison people. VizWiz is the largest dataset containing images from people who are blind and low vision [8]. Containing 39,181 images, this dataset is derived from a visual question-answering app that crowdsourced answers to questions about photographs submitted by blind and low vision users. A subset of these images was curated and annotated to produce the VizWiz dataset [22]. However, VizWiz is not appropriate

[1]The ORBIT dataset is available for download at https://doi.org/10.25383/city.14294597.v2

for few-shot learning techniques which require a small number of examples per objects.

Standing in the way of collecting suitable datasets for teachable object recognisers involving the blind and low vision community are challenges around quality and privacy of the data. In terms of quality of the data, existing research developing teachable object recognisers with blind and low vision collectors [28] shows that one of the main reasons for performance degradation is the absence of the object of interest from the training examples due to challenges in photo-taking by people who are blind or have low vision [4, 37]. Researchers have described many of the challenges that blind people encounter while taking photos [1, 2, 6, 24, 27, 58]. Examining thousands of photos uploaded by blind and low vision users to the visual question-answering service VizWiz [10] and a Flicker group for blind users [1], researchers have identified problems in photo quality that were particular to these users, including blur, lighting, composition, framing and user's hand obscuring the object of interest. To manage some of these issues, researchers and blind photographers have adopted strategies including: 1) guiding blind and low vision users through audio feedback so that objects stay in frame [4, 37], 2) estimating the general location of the target, positioning the camera close to the target and backing up, and 3) taking multiple shots in the hope that some will be good [1, 27]. These are strategies drawn upon in the procedure presented in this paper.

Another significant challenge encountered by researchers are privacy issues. Many of the questions asked by users of the VizWiz app relate to intrinsically private information, e.g., a photo of a prescription bottle asking what the drug is, or a selfie asking about the colour of a shirt. Datasets of images taken by blind photographers are particularly at risk for privacy issues because they may not be aware of everything that is captured in a photo, such as the license plate on a car. A second dataset, VizWiz-Priv [21], was created for understanding the presence and purpose of private information in images taken by VizWiz users. This work has prompted us to think deeply about the approach to privacy in the procedure we developed.

To the best of our knowledge, no realistic dataset for teachable object recognisers has been collected by the blind and low vision community. This motivated the collection of the ORBIT dataset. We report on our experiences with collecting this dataset in this paper.

3 AN INITIAL PROCEDURE: INVOLVING BLIND AND LOW VISION PEOPLE AS DATA COLLECTORS

In taking a disability-first approach to collecting a dataset for developing teachable object recognisers, thought is needed in addressing the tension between inclusion and the work placed on people with disabilities to be included. The multi-disciplinary team covering HCI, accessibility, and machine learning aimed to craft a procedure for data collectors that sought a careful balance between the structure and fidelity of data required for machine learning innovation and the demands put on data collectors. We first articulate a set of key constraints followed by their exploration in a pilot study. We finish this section by detailing the initial procedure developed.

3.1 Key Constraints

3.1.1 Supporting Good Quality Data Capture. Machine learning datasets require good quality images to build models that achieve reasonable recognition accuracy. Such well-framed, well-lit, stable photographs can be difficult for a data collector who is blind or low vision to confidently provide without sighted assistance [1, 2, 6, 24, 27, 58]. Another approach is to use video. Video increases both the number of images that are collected as it is simply a series of images, and also the chance that the object will be in frame at some point. Superfluous frames can be discarded [31, 60]. Moreover, models can also learn from the temporal information in a video, e.g., an object's 3D structure. Despite these advantages however, to our knowledge, there has not been any research into how blind and low vision users take or could be guided to collect videos.

Object recognition datasets are often crowdsourced from sighted contributors using their own devices and there is usually no consideration given to accessibility. For blind and low vision contributors, however, the data collection infrastructure needs to be accessible. This means that platform-specific applications with their respective accessibility functions may need to be developed for a data contribution infrastructure that is fit for purpose.

3.1.2 What data to collect? What objects: To be robust, a dataset should include a diversity of objects. To serve people who are blind and low vision, it should contain objects relevant to this community which are currently absent from most object recognition datasets. Directly involving the community as data collectors should ensure that relevant objects are captured. Yet, it is not without work for the data collectors to imagine the possibilities and potential scope, in terms of variety and types of objects that could be of interest. Data collectors may require support in ideating what objects might be of interest.
For example, will it be possible to distinguish between my keys and my partner's set of keys? Will it be able to identify my bus stop? Will it work for spotting whether bread is mouldy? Will it be able to locate a friend's car? The more variety that is encouraged, the more diverse the dataset, but the more complex the instructions for collecting good quality data.

How Many Objects: Realistically, a teachable object recogniser needs to be able to differentiate between multiple objects. To ensure the developed machine learning algorithms are robust, the dataset should comprise a minimum of a handful of objects per user, otherwise, the ML problem could become trivial (e.g., with 2 objects the recogniser has a 50% chance of being correct) and the model might not generalise well to more objects. This significantly increases the effort of the data contribution required from any single collector, as compared to generic crowdsourced datasets. As such, more thought is needed about how to support as well as incentivise participation.

What Examples: For a standard classification task, a dataset of labelled data is typically split into a training and a testing set to evaluate its ability to generalise. With teachable object recognisers, however, we need to explicitly capture training examples of the object on its own to "register" an object and then collect testing examples in realistic situations to test recognition ability in a scene. Understanding this distinction may be a challenge for data collectors not well-versed in how AI systems work. Moreover, machine learning models need to be trained with diverse data that captures

the object from different angles, with different backgrounds, and under varying light conditions. Otherwise, there is a risk that the model learns spurious correlations between the object and features frequently appearing with it, such as a common background. This requires extra attention and effort on the part of the data collector.

How Many Examples: Current low-shot systems require between 5–10 examples to work well when used in the real-world [18, 44, 49, 55]. This however also increases the need for good quality data which can be difficult for blind and low vision people.

3.2 Initializing the key constraints: a pilot study

To gain understanding of the above key constraints from the perspective of potential data collectors who are blind or low vision, we conducted a pilot study with eight blind and low vision participants (5 male, 3 female) in their homes. During a 90-minute visit, we asked participants to use their phone to record videos of five objects that they had been primed to think about before the research visit. These videos included two training and three testing videos per object using different filming techniques. The videos were sent to us and the visits were audio and video recorded for analysis.

A significant focus of the pilot was to investigate how to guide participants to confidently take good quality videos. Two filming techniques for training videos were explored: 1) zoom-out, and 2) rotate. For the zoom-out technique, participants were asked to start by recording the object from a very close-up view and slowly pulling the camera away from the object while videoing. In the rotate technique, they were asked to record a video so that different sides of the object were visible to the camera. The zoom-out technique was motivated by helping the user frame the object well. The rotate technique was intended to increase the chance of capturing discriminative parts of the object. For the testing videos, participants did not have to follow a specific filming technique.

3.2.1 Supporting Data Quality. We did not see any blurry or badly lit objects in the videos. This is perhaps an artefact of the quality of cameras available in iOS devices or the influence of the researcher. We found that both filming techniques were equally well suited to ensure objects were in frame; however, smaller size objects were generally better framed than medium to large ones.

The zoom-out technique was easy to apply and resulted in a good framing of the object. Pilot participants adapted the rotate technique depending on the size of the object. For most small objects, users recorded the video with one hand while rotating the object with the other, either by holding and manipulating the object at the same time, or by placing the object on a surface, and then rotating it only when needed. This latter technique provided less occlusion of hands in the data. Pilot participants often placed larger objects, such as a rucksack, on a surface and walked around the object while recording.

Some participants chose to use a new filming technique for testing videos, in which they panned the camera across a scene. When doing this, participants often panned the camera until the object was in frame and then kept the camera still for a few seconds or zoomed-in (with optical zoom) to get a closer view. This phenomenon was likely because participants were asked to position the object themselves and hence knew its location *a priori*. However, this resulted in videos always ending with the target object in frame which could introduce an unintended bias into a machine learning model; models may learn, for example, that objects only appear in the last few seconds of videos.

We integrated the following learnings from the pilot into our procedure:

- For the testing videos, we combined the zoom-out and rotate techniques into a 'zoom-rotate' technique asking participants to record the object by first zoom-out, then rotate, and then repeat three more times for different sides of the object to be visible in the video. We encouraged rotation on a surface, rather than in the hand.
- We developed different filming techniques for small to medium sized objects and for large or immovable objects.
- We changed the instructions for testing videos to encourage participants to record a scene with the object and other objects around, rather than their object per se.

3.2.2 What Data to Collect. Participants did not find it challenging to think of objects, so we chose to ask for 10 per participant in the procedure. Most objects were small/hand-held, for example keys, remotes, spectacles, wallets, and inhalers; however, some were medium-sized, for example bags, laptops, and canes. Figure 2 shows a word cloud of most frequently mentioned objects. Participants did not have difficulty distinguishing between training and testing filming techniques, so this was encouraging.

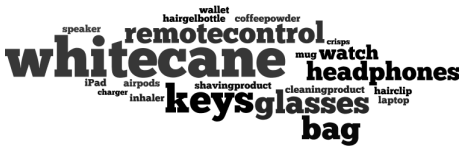

Figure 2: A word cloud showing the objects of interest that pilot participants mentioned. The size of the text shows the relative frequency. The white cane mentioned most frequently (17 occurrences), closely followed by keys (12 occurrences), bag (10), a remote control (9), and headphones (6). Less frequent objects were the following: speakers, wallet, iPad, airpods, mug, laptop, inhaler, coffee powder and others.

3.3 Initial Data Collection Procedure

Based on the learnings of the pilot, a data collection procedure was developed. It contained two main elements: a data collection protocol with a set of user instructions, and an accessible app (Figure 3) and supporting infrastructure to collect data. These worked in concert to support high quality data capture, reducing the burden on data collectors as much as possible, and to ensure the privacy of data. We tested the accessibility and usability of each of these components first through professional accessibility evaluation and then numerous iterative engagements with users.

3.3.1 Data Collection Protocol. Supporting Data Quality: The data collection protocol was presented in a set of user instructions (available in the supplementary material). These instructions first motivated the use scenarios of teachable object recognisers within the blind and low vision community. This was followed by a brief overview of how machine learning works, with a special focus on the need for training and testing videos as well as variation in the data. The instructions gave a linear progression of what needed to be done with considerable detail. Examples were given throughout to provide context to data collectors, e.g. when explaining what a training video is, we gave the example of the white cane taken in different positions and environments - such as when it is folded up, when it is leaning against a wall, when it is outside or on the couch etc. Since there was significant detail in the instructions, they were sent to collectors before they started to collect the data and were also available in the help menu of the app.

Data collectors were given step-by-step instructions to implement the 'zoom-rotate' technique to capture training videos. They were asked to keep one hand on the surface next to the object as an 'anchor point' to aim the camera at the object, then holding the phone in the other hand and bringing it as close as possible to the object. After starting the recording, they were asked to slowly draw the phone away from the object until it reached their body at shoulder height, after which they should rotate the object so that a different side of it was facing them. Then they should return their phone's position to the anchor hand, repeat this process three times, and then stop the recording.

For testing videos, data collectors were asked to construct a realistic 'scene' in which the recogniser might be used, for example, a surface where they may look for their keys that were moved by somebody else. The scene needed to include the objects they wanted to film together with at least five other 'distractor objects' that had not been registered in the recogniser. Data collectors were then asked to take testing videos of a scene with two different techniques: 'zoom-out' and 'pan'. For 'zoom-out', they had to place an 'anchor hand' in the scene, and then draw their phone slowly towards their body, then stop recording. For a 'pan' video, we asked data collectors to face the scene and to make sure the selected object was not directly in front of them. To position the camera to face the scene, they could use an anchor hand but they were asked to remove it before panning over the scene at shoulder height from right to left, by turning their upper body.

In addition to the instructions, data collectors were supported in a number of other ways: 1) a Frequently Asked Questions (FAQ) document was available (available in supplementary material); 2) data collectors could request a phone call for assistance; 3) weekly emails were sent, providing participants with summary stats, encouraging them to submit more videos, as well as asking them to re-record any videos as necessary; and 4) two online training sessions via video calls were provided. Throughout the data collection period, we ensured that participants had a researcher as a personal contact point whom they could approach for help.

Data Requested: Data collectors were asked to select ten objects; two of these ten objects were requested to be large or immovable objects. They were then asked to take eight videos of each object, six "training" videos and two "testing" videos. For the six "training" videos, collectors were asked to record the object in isolation using the 'zoom-rotate' technique. Each of these training videos required a different background with variation in lighting also recommended. They were also asked to record two types of "testing" videos with the object situated in a scene, one each using the 'zoom-out' and 'pan' methods. Altogether, each collector should collect 80 videos.

3.3.2 Accessible Data Collection Infrastructure. We developed and distributed a fully accessible data collection iOS app for iPhone and iPad (Figure 3) so data collectors could use their own devices in their everyday environment to capture and label videos of objects. The app has three main screens. The Things screen (Figure 3.a) allowed the data collector to 1) add a new thing and 2) list the objects has chosen to video and the videos already taken for each object. By selecting a thing or adding a new object, the data collector was taken to the record and review screen. On the Record and Review screen (Figure 3.b), the collectors could choose whether they wanted to add a training, test-zoom or test-pan video and then add videos of the selected type to the collection. Once a video had been recorded, this screen also allowed the collector to review the collection of videos (Figure 3.c) in order, for example, to re-record a video that the collector was not happy with or that did not pass our validation criteria (as marked by a researcher) and could not be added to the dataset. The app supported data collection in a number of ways. First, it allowed the data collectors to organise the data collection process, providing structure for the various objects and types of videos that needed to be collected. Both the user interface and the interaction flow supported collectors in creating the range of videos required. Second, it provided audio feedback during recording. A tic sound was played every five seconds to give an indication of when the object should be rotated and a double tic and vibration to indicate when a video should be stopped. To prevent inadvertent long videos, recordings were automatically ended after two minutes.

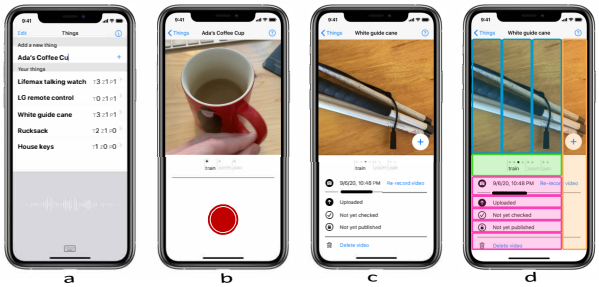

Figure 3: The data collection app: (a) the main 'Things' screen, (b) a 'Thing' screen adding a recording, (c) a 'Thing' screen after some recording activity, (d) the same 'thing' screen marked up with the adapted information hierarchy and touch-targets of the accessibility interface.

Eligible collectors were sent an accessible PDF document giving them details about the study, instructions for how to collect and label the videos using the collection app, and the App Store link to install the app. The document included information about data ownership, security and privacy and what would happen to the data beyond data collection, including how to withdraw participation and delete data. It was made clear that the resulting anonymised data would be released publicly in an open-source dataset.

Informed consent was obtained via the app and a record was created in our server infrastructure. Supporting privacy of data collectors was a key consideration in the design of the infrastructure. Audio data and meta-data from videos was never collected. We developed a validation process for videos submitted to the server via the app. Videos were individually checked by a researcher to ensure that: 1) they did not contain any Personal Identifiable information (PII) in the video or object name; 2) the video or the object name were not inappropriate or offensive and 3) the object was in-frame at least some of the time. The only bar in terms of data quality was that the object had to be visible in-frame at some point in the video; this was to encourage ML researchers to make their models robust to realistic data, an important long-term contribution to computer vision apps for the blind and low vision community. All the videos that did not meet the validation criteria above were considered not suitable for the dataset and were automatically deleted from the server. For example, videos that had a participant's first name in the label, pictures of people in the background, or names or QR codes on cards were sent back to the data collector for re-recording. The app helped users track the status of their videos and whether they had been accepted into the dataset.

The code for the data collection app and server infrastructure is available for download at https://github.com/orbit-a11y.

4 PROCEDURE VALIDATION AND EVOLUTION

In this section we report on the empirical validation of this initial procedure through two phases of data collection. For each phase, we detail the implementation of the procedure and assess the key learnings which led to further evolutions. We show the main differences between phases in Table 1. In addition to a description of the data and our reflections on the experience, we used small machine learning-driven experiments to motivate evolutions to the procedure.

4.1 Phase 1

4.1.1 Recruitment. The initial procedure was used during the period between beginning of May and mid-July 2020. UK-based data collectors were recruited through advertising on social media, personal contacts, a charity for blind and low vision children and young adults, and a further education college for blind students. Later efforts were supported by a technical blog post by Microsoft research that highlighted the project, and an email to UK-based users of an assistive technology provided by the same company. Each collector who completed the data collection was paid £50 in Amazon vouchers. Finally, we introduced snowball recruitment in which existing data collectors recruited their personal contacts to participate. For every referred data collector who completed the study, referrers received an additional incentive of £10 in Amazon vouchers. Recruitment was more of a challenge than we expected given the money offered, resulting in 48 data collectors making contributions (goal of 100). We were surprised that our contributors seemed to be older, tech-savvy users of existing apps that incorporate AI for accessibility. We surmise that these collectors already understood the benefits of collecting data for drive ML innovation. This insight prompted us to stress the direct benefits of developing AI for accessibility for the *community* instead of individual contributors in Phase 2. We reinforced this focus on the wider benefit to the community by replacing the direct payment to participants with making donations to selected community organisations. We also saw a significant drop-out of potential data collectors: 64 people dropped out after initially showing interest in participating, 43 did not continue after receiving the user instructions, and 21 downloaded the app but did not take any recordings after downloading the app. The complexity of the instructions might have contributed to this high level of drop-out. The instructions explained the recording techniques in great detail, due to many data collectors being unfamiliar with taking videos. However, data collectors mentioned that they spent significant time studying the instructions but were still unable to remember all the information, e.g., the different techniques to record training videos depending on whether the object was small or large. As a result, we created a more generic set of instructions for Phase 2 (available in supplementary material) that stressed capturing different angles of objects, rather than explicitly defining steps to follow.

4.1.2 Protocol. While we collected 3448 videos (2453 training, 501 test-pan, 494 test-zoom) for 390 objects, many collectors struggled to complete the data collection. Despite the findings in the pilot, data collectors reported that they struggled to think of 10 unique objects. On average, data collectors contributed eight objects each (Figure 4a). Data collectors also struggled to collect all the examples required for each object. At a minimum, we required six training videos, one test-zoom and one test-pan for a 'complete' object. Only two data collectors submitted 10 complete objects, with 11 submitting eight or more, and 17 collectors submitting none (Figure 4b).

We used a machine learning-driven analysis to assess whether we could ease the burden on data collectors without compromising model generalisability. We first considered the number of objects to be collected. A state-of-the-art few-shot model [44] was trained on 1) the dataset as it was with up to 11 objects per user, 2) a capped version of the dataset by artificially lowering the number of objects per user to five. In line with typical accuracies of this model on other datasets, training and testing without caps gave an accuracy of 46.32%, whereas capping the training objects to five per user gave a test accuracy of 41.18% on the uncapped dataset. We believe this drop of 5% accuracy is relatively small and can possibly be reduced further in a number of ways such as training with varying number of objects or carefully merging users' data. We therefore reduced the required number of objects to five in Phase 2 of the data collection.

We then considered whether we could reduce the number of videos to be collected per object, and specifically training videos. We analysed the effect of increasing numbers of training videos per object on test video accuracy (Table 2), using the same machine learning setup. We saw that accuracy was highest for four training videos and the gain in accuracy surprisingly flattened off as the number of training videos increased. We decided that we were able to reduce the number of training videos to five without affecting the recognition abilities of the recogniser. Five also aligned with the number of objects data collectors were asked to collect, thus making it more memorable.

Table 1: Main differences between Phase 1 and Phase 2

	Phase 1	Phase 2
Number of objects	10 with at least 2 large/immovable	5
Number of testing videos	1 zoom-out, 1 pan	2 zoom-out
Number of training videos	6 rotate-zoom	5 rotate-zoom
Incentives	£50 Amazon voucher	£25 donation to charity
Collectors	UK-based	Global

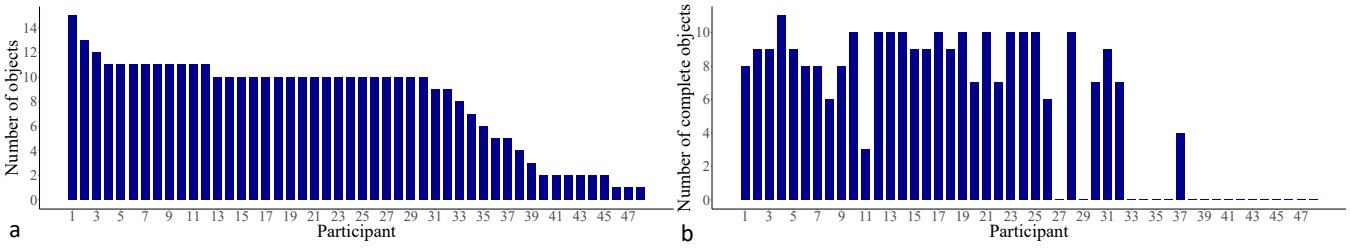

Figure 4: a) Number of objects collected per collector b) Number of complete objects (6 training, 1 test-zoom, and 1 test-pan) per collector.

Table 2: Effect of number of training videos per object on test accuracy - more training videos improves the ability to recognise the object in new environments.

# training videos per object	Test accuracy	Gain
1	36.21	-
2	43.57	+7.4
4	46.61 cap	+2.8
6	46.32	-0.3

We also re-visited the types of objects we asked data collectors to record. The most common objects submitted were keys (19), guide cane (11), wallet (9), AirPods (8), front door (7), headphones (7), mug (6), sunglasses (6), but we noted a long tail of objects that were unique to individuals. The data collected covered objects of different sizes, ranging from wallets and keys to house doors and a friend's car. Most data collectors, however, did not add the requested large objects, contributing on average one large object with just under half of users contributing none. Although having diverse objects naturally future-proofs the dataset, it comes at a cost of increased complexity of the user instructions. We therefore made the decision to remove the specific request for large or immovable objects.

Finally, we considered how we could reduce the number of video techniques to further reduce the complexity of the user instructions. To explore the relative value of the two types of testing videos, we trained a state-of-the-art low-shot model [44] using either training/test-zoom videos or using training/test-pan videos. Test-zoom videos yielded a recognition accuracy of 46.32%, while test-pan videos yielded an accuracy of 27.20%. This drop in accuracy can be explained by the absence of the objects in most portions of test-pan videos by virtue of collectors panning across a scene; thus, the model needs to predict not only what the object is but also guess whether the target object is in frame. Resolving this issue would require collecting frame annotations offline from sighted users, which would come at further time and cost. This analysis therefore highlighted that test-zoom videos were the most useful to collect also because object predictions would be made per frame and test- zoom videos will include more examples of the actual object.

Finally, we observed that 17% of data collectors did not collect any testing videos, making the evaluation of a trained model impossible. To encourage data collectors to record test-zoom videos, we asked them to record these videos before the training videos, in the place that they would normally keep the object. This removes the need to artificially construct a scene. They could then move to new surfaces with no other objects around to record the training videos. While we were still asking for two test-zoom videos to be recorded, we reduced effort even further by asking them to record them from a different angle rather than different location. These recording instructions were further supplemented by an online tutorial that walked data collectors through the process.

4.1.3 Accessible Data Collection Infrastructure. We revisited the design of the data collection app as well. One important aspect is to encourage minimum collection of examples for an object whilst discouraging 'over-collecting'. Over-collecting happened because data collectors were worried about data quality and responded by taking more or longer videos than necessary. For example, some data collectors recorded multiple videos of each object from different sides instead of rotating the object in one video. To address this challenge, the app was adjusted to highlight when all video slots for an object are filled. Another type of 'over-collecting' was recording very long videos; as a result, we reduced the video cut-off time to 30 and 20 seconds for training and testing videos respectively.

We also improved accessibility of the app to support better data collection. We provided stronger audio and haptic feedback during

recording and we added better VoiceOver markup that read out appropriate recording instructions in the right context. This improved the video quality as it helped collectors to collect the data using the right technique without the need to refer back to the instructions file.

A major consideration for Phase 2 was also to ensure the privacy and confidentially of the data. To communicate with anonymous collectors, we added a notification feature to the app that we used to inform collectors about their progress and the need to re-collect data. As mentioned above, a researcher watched all submitted videos to remove the small fraction that contained PII, no object, or inappropriate content. This validation process was time-consuming: on average, one hour was needed to validate 100 videos. In scaling up, this task should not be underestimated. For example, suppose we are aiming for 300 collectors, only a ten-fold increase in contributors, each recording five objects, and five training and two testing videos for each object. This would require more than 100 hours of researcher labour to watch and validate all videos, or nearly three weeks of full-time researcher work effort. While there is some research into detecting PII automatically [19], it is currently not robust enough to ensure complete anonymisation. We hence employed a contract researcher in Phase 2 for supporting this validation process.

4.2 Phase 2

4.2.1 Recruitment. Phase 2 data collection was carried out in two rounds from mid-October to November 2020 and again in January 2021 in multiple English-speaking countries around the world. In the first round, three podcasts aimed at blind and low vision technology users were created in addition to advertisement on social media. For each collector that completed their videos, a £25 donation was offered to one of five charities of their choice that supports blind and low vision communities. In the second round, we worked in close collaboration with one charity's prepare-for-work programme. The charity led recruitment and communication with the data collectors as well as provided technical support as needed. Each data collector who completed was given the equivalent of £25 by the charity.

New community engagement strategies were also tried in this phase. For round 1, we had added a pre-filmed filming instruction tutorial, detailed online instructions, and app notifications to simplify mass communication between the research team and data collectors. Round 2 was more interactive and involved more direct contact. To help collectors understand the purpose of the data collection, we offered an optional one-hour online interactive lecture by a field leader entitled: "Artificial Intelligence Basics: Understanding the tech in your pocket". The lecture was for background knowledge and participants were not required to collect data. Additionally, similar to the training sessions provided in Phase 1 to help participants complete their videos, we carried out two "Let's get it done" sessions where data collectors had the chance to complete their videos on a live video call with researchers present. However, participation in both sessions was relatively low. These touchpoints were an important part in engaging a community that may not understand why datasets are needed to drive the tools that many already have in their pockets.

Despite a significant focus on community engagement, recruitment remained a struggle with just 52 contributing data collectors (goal of 125, revised down from 300). In the first round, we got several complaints from data collectors that they wanted to get paid directly for their work. Yet, even when paid directly, the numbers were not as high as the charity had expected.

4.2.2 Protocol. In total, we received 1380 videos (965 training, 415 testing) for 271 objects. This included 423 videos of 96 objects by 21 people in round 1 and 957 videos of 175 objects from 31 people in round 2. In both rounds, we saw significant drop- out of potential data collectors: 18 people dropped out after downloading the app but without taking any recordings in round 1 and 14 people in round 2. The number of drop-outs was lower compared to Phase 1, which shows that the simplified version of the instructions might have helped in addressing this issue. Interestingly, while the number of collectors was smaller in round 2, the strategies for engagement better supported collectors in completing data collection.

Our efforts to engage users and support them to gather good quality data also paid off. The most common objects in Phase 2 were: keys (16), wallet (9), remote control (7), pen (5), sunglasses (4) and white cane (4). We also had some success in driving new types of objects to be collected. A few unexpected objects were collected this time, including a face mask, a guide dog, and dog waste, likely prompted by our additional efforts to help collectors think of different objects. Even though the new set of instructions was more generic rather than providing every detail and there was no specific request to record large or immovable objects, this did not stop users from collecting videos of a bus stop, a patio gate, a car as well as home appliance such as a washing machine, a fridge and a dryer. In total, the dataset included 21 large objects collected by nine users in Phase 2 while the remaining 43 users did not submit any videos of large objects. However, some of the data submitted were shot in very low lighting conditions, at night, and we needed to ask collectors to retake the video.

The fact that most of the data collectors completed their videos as well as the low participation in the additional training sessions shows that the new simplified version of the user instructions helped collectors to complete their videos. We also observed that collectors submitted both testing and training videos this time. This shows that the strategy of asking them to take the testing videos first, and in the place that they would normally keep the objects, was helpful.

4.2.3 Accessible Data Collection Infrastructure. We still faced significant challenges in getting the app to work well with some collectors, even with technical support provided by the charity in round 2. For example, we had unexplained crashes, audio problems and data which did not upload. These problems proved very difficult to troubleshoot and resolve because of the diversity of devices and settings, and lack of direct access to collectors.

For faster validation of the videos, we hired a contract researcher who was dedicated only to this job. This helped with over-collecting videos since there were no delays on notifying data collectors about the quality of their videos. In order to 'rescue' some of the videos that contained PII, we trimmed some of the videos to remove the PII. This was time consuming and was feasible only for a few videos,

for example, a video of a car where the licence plate was visible only at the start and end of the video.

5 DISCUSSION

There is significant scope for AI innovation to open up new possibilities for people with disabilities to access the world. However, such innovation requires large datasets that reflect the data of the people with disabilities. In this paper, we presented the development of a procedure, comprising a protocol and data collection infrastructure, that enables data collectors who are blind or low vision to contribute to a dataset useful for developing teachable object recognisers. We validated this procedure through two phases of data collection. Throughout the creation of a disability-first dataset, we reflected upon the tension of including people with disabilities as data collectors and the work that is consequently required of them.

In this discussion, we consider the broader implications of this work for others collecting disability-first datasets for AI innovation. We highlight eight orienting questions for researchers that address the engagement, data collection procedures, and accessible data collection infrastructure.

5.1 Engaging Data Collectors with Disabilities

- Do you have adequate resources for supporting data collection?

There are many perspectives on how data collectors could be or should be incentivised or compensated for their time. One could compare data collection to crowd work and consider the payment rates in that context [23]. However, even when accounting for the implicit efforts of data collection (e.g. reading training materials), in our case some people communicated with us that they expected higher amounts. The amount varied by location, depending on what others had offered previously, a problem well-known in anthropology [19, 50]. An alternative approach would be to consider implicit incentives to motivate data collectors beyond monetary rewards [32]. In our work, we worked hard to appeal to intrinsic motivations around learning new skills and helping the community, making it very clear that there would be a tangible product outcome. Gamification is another approach, as used in [11] to capture sign language data. The data needed for few-shot learning, however, does not easily lend itself to gamified elements that are based on quantity. Nor could we provide a prototype service as state-of-few-shot learning algorithms are not accurate or computationally efficient enough to achieve a teachable object recognisers at the time of data collection. It can be seen as a "chicken or egg" problem for an initial dataset that allows for more contextualised data collection in the context of machine teaching applications. Other approaches are to use non-disabled crowd-workers as in [26] to build an initial base for applications relevant to the disability community. This is in opposition to initial views that motivated this work to take a disability-first perspective. In our experiences, the nature, amount and recipients of incentives are one of the key points to consider. We suggest that researchers consider implications on resources available within the project early on to achieve successful data collection.

- What strategies will you use to cultivate informed contributors?

We found that collecting a dataset for ML works best when collectors have a basic understanding of how ML works. This helps them focus on certain aspects of the data that they collect which are crucial to being able to drive ML innovation. In our case, for example, participants often omitted test videos for objects which meant that we could not evaluate the models we developed, and thus rendered all the training videos that the data collector had provided useless. To counteract this, we needed to spend more effort on educating our collectors, through giving information about what ML is and how it works in written instructions and interactive lectures. It may also be important to cultivate an understanding of any potential ethical issues that might be present in collecting the data as well [5]; we endeavoured to highlight the removal of any PII prior to the dataset release. There are currently some efforts to make AI transparent to a wider audience [34, 40], as well as educate end-users about AI [17, 36, 52]. However, there is little research to investigate how to make AI for accessibility transparent to people with disabilities [3] and to our knowledge there are no curricula for teaching about AI for accessibility to the disabled community. Your project will need to consider how to overcome these challenges and educate your target community.

- Are you inspiring your participant community to think about their role in ML innovation?

Creating a dataset to support specific applications, while an important part of engaging a community, is very challenging. As Yang et al. [59] discuss, it is difficult to specify an AI experience in advance of having a working prototype trained on data. In our project, we tried to motivate the data collection by describing how an app that embeds a teachable object recogniser might be used. However, these efforts could be improved to engage the target community, either by supplying non-working 'demo' versions, or providing initial prototypes that demonstrate its use, even if they are not robust for everyday use [43]. Another option is to release experimental 'lab' areas in existing apps that can employ 'donate to science' approaches. A different approach would be to give space for building experiences alongside users. This could be achieved through co-designing [14, 47, 57] with the communities and integrating multiple iterations of dataset collection into prototypes. These choices will play a crucial role in inspiring your participant community.

With so many decisions and demands to get data suitable for ML research, it is easy to lose sight of the important role that disability-first dataset collection efforts play in engaging a target community. This may be the first time a person with disability experiences AI and what it may have to offer them. What is offered in return may be critical to the relationship that participants form with AI. Moreover, the way recruitment, task, and support are set up will speak strongly about the power relationship between research and disability communities [7]. Any work done by collectors might be seen to go into a 'dark hole', without seeing a direct benefit in the future and not connecting how individual work shapes a ML innovation for the community. 'Drive-by' data collection is likely to have a very negative impact on the relationship with a disability community. Including researchers from your disabled community

and ensuring that the community receives useful technology in return for its effort are potential approaches to this question.

5.2 Data Collection Procedures

- Are you striking the right balance between ML needs and demands on collectors?

A key contribution of the work presented here was a switch from images to videos. While much of the ML work for teachable object recognition is still done on images, videos give people who are blind or low vision more opportunities to capture good images of an object. They also enrich the data by providing temporal information that could be used to make object recognition from real-world data more robust. We encourage researchers to radically rethink the data collection needs to suit collectors.

To do so requires close collaboration between HCI, accessibility and ML researchers. It might feel that the considerations that enable robust ML development are relatively fixed, and that data collectors simply need to 'get on with it'. However, as our work showed, probing these touch points in a collaborative setting can help prioritise the most critical aspects of the data collection. In our work, ML driven data evaluation at intermediate points in the process helped us better balance data needs with demands on the collectors. For example, we were able to reduce the number of objects, while keeping the number of training videos more or less the same. With the help of HCI and accessibility experts, we were also able to situate the ML needs in the protocol and supporting infrastructure. We believe that inter-disciplinary teams are necessary to drive ML innovation in disability communities.

- How can ML support your community in data collection?

While collecting datasets is often about the development of new ML techniques, existing techniques may also be useful in enabling participation. In the object recognition domain, there is some work that is using ML models to help users take good images, for example by indicating whether any object is detected at all or where it is located in the current frame, or using the presence of hands as an indication of where the object is located [4, 9, 37]. This work could be extended to videos, also perhaps indicating low lighting conditions or the possible presence of PII. What is important is to consider in this question is how the burden on collectors can be further eased.

5.3 Accessible Data Collection Infrastructure

- How closely are the collection procedure and the infrastructure entwined?

As part of our project, we are sharing the code of our data collection infrastructure (https://github.com/orbit-a11y) so other researchers can extend the data collection we started. However, the app is tightly bound to the procedure. While we attempted in Phase 1 to build a relatively generic app to record videos of objects, we came away from that in Phase 2 as we found that collectors were better supported by specifying the number of videos needed, and giving recording instructions in context. Embedding further support, for example, object in-frame detection would entwine the app even more with our use case. This of course prevents reuse of the app, and significant effort would have to be expended to create the right infrastructure for collecting other data. This means a careful balance has to be struck between development effort of the supporting infrastructure and intended re-use between data collection projects.

- How do you support collectors across the world?

When collecting data in a disability-first way, accessibility of the data collection infrastructure becomes very important. In order to support disabled collectors best, it is often necessary to make use of platform-specific accessibility functions. In our project, we designed for iOS 13.2 devices such as iPhones and iPads to provide access through VoiceOver, and made use of Apple Accessibility Guidelines in designing the app experience. Because we collected videos, we also required significant storage space for collected data, and a stable Internet connection. All these aspects meant that collectors had to own or have access to high-spec, expensive devices. However, this cuts out a large swathe of collectors who use other platforms (e.g. Android), lower-spec devices or do not have broadband. This then can lead to further marginalisation of potential collectors, due to their socio-economic status, age, technology-affinity, or their location. Reflecting on this question will need to address how to extend the reach of the infrastructure to people beyond WEIRD (Western, Educated, Industrialised, Rich, Democratic) societies [25]. In turn, this will also affect decisions to support cultural adaptations of the data collection, for example, translation of components or markup for collectors in non-English speaking countries.

- What strategies will be used to ensure privacy and confidentiality?

Considerable thought has to be given to ethical issues in a disability-first data collection, especially if the data is to be made publicly available to drive ML innovation on a large scale. Our experiences and that of the VizWiz project [21] show that data might be frequently intrinsically private, such as bank cards, or that personally identifying information (PII) might also slip in inadvertently, such as 'selfies' from reflections on shiny surfaces. We have tried to address this through educating collectors about PII and a researcher manually checking all videos. However, as we discussed, a manual process can considerably escalate the effort and resource required during a collection process and might also necessitate a feedback loop for collectors to re-collect data. It would be much better if the detection and elimination of PII could be automated, while preserving the data already collected. For example, this could extend to blurring of addresses or car license plates while still using the remainder of the image as data. This is an active research concern that is necessary to drive disability-first ML innovation.

6 CONCLUSION

AI for accessibility is a rapidly developing space, and dataset collection will be essential to creating many novel and useful applications for disability communities. While it is critical to ensure that disabled people are at the centre of the ML innovation process, this is also effortful for disabled data collectors. In this paper, we have presented key constraints that needed to be balanced when collecting a disability-first dataset to innovate in AI for accessibility. Of particular importance were:

- engaging with and supporting target communities,
- ensuring the quality of the data collected,

- balancing the usefulness versus the effort of data to be collected,
- providing a suitable data collection infrastructure.

We presented and evaluated a procedure for collecting data to develop teachable object recognisers from blind and low vision collectors. We found that:

- How to engage and support blind and low vison collectors needs to be carefully considered and adequately resourced. We found that recruitment strategies need to be adapted to increase community involvement, and that extra effort has to be spent on supporting collectors and addressing privacy considerations.
- We proposed a switch from images to videos to support collectors who are blind or low vision. Good quality examples can be ensured by developing clear and simple filming instructions that use hands as anchor points. We developed a rotate-zoom technique for training videos and a test-pan and test-zoom technique for testing videos.
- Thought needs to be given to the number and kinds of objects and examples to be collected to balance the effort required by collectors. We ran a number of ML-driven data analyses to help us to refine our procedure.
- We developed a data collection infrastructure that supported the collection efforts of a large number of people who are blind and low vision across the English-speaking world.

We presented eight orienting questions to other researchers creating datasets for AI for accessibility to help them reflect on the needs of their participant community in these dataset collection efforts. Our work, which resulted in a disability-first dataset [38] is a significant step in contributing to driving ML innovation for the disabled community [39].

ACKNOWLEDGMENTS

This work was supported by a gift through the Microsoft AI for Accessibility program. Luisa Zintgraf is supported by the 2017 Microsoft Research PhD Scholarship Program, and the 2020 Microsoft Research EMEA PhD Award. We thank all the participants and data collectors for their contribution to ORBIT. We also thank VICTA, the Royal National College for the Blind (RNC), the Royal National Institute of Blind People (RNIB), the Canadian National Institute of Blind People (CNIB), Humanware, Tekvision School for the Blind, Blind South Africa (BlindSA), the National Federation of the Blind (NFB), Galloways and AbilityNet for their help in the project. We also thank Emily Madsen for helping with video validation.

REFERENCES

[1] Dustin Adams, Sri Kurniawan, Cynthia Herrera, Veronica Kang, and Natalie Friedman. 2016. Blind photographers and viz snap: A long-Term study. *ASSETS 2016 - Proceedings of the 18th International ACM SIGACCESS Conference on Computers and Accessibility*, 201–208. https://doi.org/10.1145/2982142.2982169

[2] Dustin Adams, Lourdes Morales, and Sri Kurniawan. 2013. A qualitative study to support a blind photography mobile application. *ACM International Conference Proceeding Series*, 1–8. https://doi.org/10.1145/2504335.2504360

[3] Subeida Ahmed, Cecily Morrison, Harshadha Balasubramanian, Abigail Sellen, Simone Stumpf, and Martin Grayson. 2020. Investigating the Intelligibility of a Computer Vision System for Blind Users. (2020), 11. https://doi.org/10.1145/3377325.3377508

[4] Dragan Ahmetovic, Daisuke Sato, Uran Oh, Tatsuya Ishihara, Kris Kitani, and Chieko Asakawa. 2020. ReCog: Supporting Blind People in Recognizing Personal Objects. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–12. https://doi.org/10.1145/3313831.3376143

[5] Cynthia L Bennett. 2020. *Authentic Intelligence: A Blind Researcher Bringing Wisdom to the Future of Technology Innovations*. https://www.bennettc.com/wp-content/uploads/2020/07/NFB_2020_Speech_Final.pdf Transcript of a talk at the NFB, Retrieved September, 2020.

[6] Cynthia L. Bennett, E. Jane, Martez E. Mott, Edward Cutrell, and Meredith Ringel Morris. 2018. How teens with visual impairments take, edit, and share photos on social media. *Conference on Human Factors in Computing Systems - Proceedings* 2018-April, 1–12. https://doi.org/10.1145/3173574.3173650

[7] Cynthia L Bennett and Daniela K Rosner. 2019. The Promise of Empathy: Design, Disability, and Knowing the" Other". In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. 1–13. https://doi.org/10.1145/3290605.3300528

[8] Jeffrey P. Bigham, Chandrika Jayant, Hanjie Ji, Greg Little, Andrew Miller, Robert C. Miller, Robin Miller, Aubrey Tatarowicz, Brandyn White, Samuel White, and Tom Yeh. 2010. VizWiz: Nearly real-time answers to visual questions. *UIST 2010 - 23rd ACM Symposium on User Interface Software and Technology*, 333–342. https://doi.org/10.1145/1866029.1866080

[9] Jeffrey P. Bigham, Chandrika Jayant, Andrew Miller, Brandyn White, and Tom Yeh. 2010. VizWiz::LocateIt - Enabling blind people to locate objects in their environment. *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Workshops, CVPRW 2010*, 65–72. https://doi.org/10.1109/CVPRW.2010.5543821

[10] Erin L Brady, Yu Zhong, Meredith Ringel Morris, and Jeffrey P Bigham. 2013. Investigating the appropriateness of social network question asking as a resource for blind users. In *Proceedings of the 2013 conference on Computer supported cooperative work*. 1225–1236. https://dl.acm.org/doi/abs/10.1145/2441776.2441915

[11] Danielle Bragg, Naomi Caselli, John W Gallagher, Miriam Goldberg, Courtney J Oka, and William Thies. 2021. ASL Sea Battle: Gamifying Sign Language Data Collection. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–13.

[12] Danielle Bragg, Oscar Koller, Mary Bellard, Larwan Berke, Patrick Boudreault, Annelies Braffort, Naomi Caselli, Matt Huenerfauth, Hernisa Kacorri, Tessa Verhoef, Christian Vogler, and Meredith Ringel Morris. 2019. Sign Language Recognition, Generation, and Translation: An Interdisciplinary Perspective. In *The 21st International ACM SIGACCESS Conference on Computers and Accessibility* (Pittsburgh, PA, USA) *(ASSETS '19)*. Association for Computing Machinery, New York, NY, USA, 16–31. https://doi.org/10.1145/3308561.3353774

[13] John Bronskill, Jonathan Gordon, James Requeima, Sebastian Nowozin, and Richard Turner. 2020. Tasknorm: Rethinking batch normalization for meta-learning. In *International Conference on Machine Learning*. PMLR, 1153–1164.

[14] Morrison Cecily, Edward Cutrell, Martin Grayson, Thieme Anja, Alex S. Taylor, Geert Roumen, Camilla Longden, Rita Marques, Abigail Sellen, and Sebastian Tschiatschek. 2021. Social Sensemaking with AI: Designing an Open-ended AI experience with a Blind Child. In *The 2021 CHI Proceedings on Human Factors in Computing Systems*. 1–12.

[15] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. 2017. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE transactions on pattern analysis and machine intelligence* 40, 4 (2017), 834–848.

[16] Microsoft Corporation. 2016. *SeeingAI*. https://www.microsoft.com/en-us/ai/seeing-ai Retrieved September, 2020.

[17] deeplearning.ai. 2019. AI For Everyone | Coursera. https://www.coursera.org/learn/ai-for-everyone?utm_source=gg&utm_medium=sem&utm_content=08-AIforEveryone-ROW&campaignid=9727679885&adgroupid=99187762066&device=c&keyword=artificial%20intelligence%20and%20machine%20learning&matchtype=b&network=g&devicemodel=&adpostion=&creativeid=428167449287&hide_mobile_promo&gclid=CjwKCAjw_qb3BRAVEiwAvwq6Vm_0ODrr8pKM8y3rrBvUoBlD7920KDsNb-JC1ajRb9uzeEEwbAFHtxoCzIQQAvD_BwE#syllabus

[18] Chelsea Finn, Pieter Abbeel, and Sergey Levine. 2017. Model-agnostic meta-learning for fast adaptation of deep networks.

[19] C Grady. [n.d.]. Ethical and practical considerations of paying research participants. *Department of Clinical Bioethics Clinical Center/NIH* ([n. d.]).

[20] Anhong Guo, Ece Kamar, Jennifer Wortman Vaughan, Hanna Wallach, and Meredith Ringel Morris. 2019. Toward Fairness in AI for People with Disabilities: A Research Roadmap. (7 2019). http://arxiv.org/abs/1907.02227

[21] Danna Gurari, Qing Li, Chi Lin, Yinan Zhao, Anhong Guo, Abigale Stangl, and Jeffrey P. Bigham. 2019. VizWiz-Priv: A Dataset for Recognizing the Presence and Purpose of Private Visual Information in Images Taken by Blind People. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

[22] Danna Gurari, Qing Li, Abigale J Stangl, Anhong Guo, Chi Lin, Kristen Grauman, Jiebo Luo, and Jeffrey P Bigham. 2018. Vizwiz grand challenge: Answering visual questions from blind people. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 3608–3617.

[23] Kotaro Hara, Abigail Adams, Kristy Milland, Saiph Savage, Chris Callison-Burch, and Jeffrey P Bigham. 2018. A data-driven analysis of workers' earnings on

Amazon Mechanical Turk. In *Proceedings of the 2018 CHI conference on human factors in computing systems*. 1–14.

[24] Susumu Harada, Daisuke Sato, Dustin W. Adams, Sri Kurniawan, Hironobu Takagi, and Chieko Asakawa. 2013. Accessible photo album: Enhancing the photo sharing experience for people with visual impairment. *Conference on Human Factors in Computing Systems - Proceedings*, 2127–2136. https://doi.org/10.1145/2470654.2481292

[25] Julia Himmelsbach, Stephanie Schwarz, Cornelia Gerdenitsch, Beatrix Wais-Zechmann, Jan Bobeth, and Manfred Tscheligi. 2019. Do We Care About Diversity in Human Computer Interaction: A Comprehensive Content Analysis on Diversity Dimensions in Research. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. 1–16.

[26] Jonggi Hong, Kyungjun Lee, June Xu, and Hernisa Kacorri. 2020. Crowdsourcing the Perception of Machine Teaching. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–14.

[27] Chandrika Jayant, Hanjie Ji, Samuel White, and Jeffrey P. Bigham. 2011. Supporting blind photography. *ASSETS'11: Proceedings of the 13th International ACM SIGACCESS Conference on Computers and Accessibility*, 203–210. https://doi.org/10.1145/2049536.2049573

[28] Hernisa Kacorri. 2017. Teachable machines for accessibility. *ACM SIGACCESS Accessibility and Computing* 119 (2017), 10–18.

[29] Hernisa Kacorri, Utkarsh Dwivedi, Sravya Amancherla, Mayanka K. Jha, and Riya Chanduka. 2020. IncluSet: A Data Surfacing Repository for Accessibility Datasets. *ASSETS 2020 - 22nd International ACM SIGACCESS Conference on Computers and Accessibility*, 6.

[30] Hernisa Kacorri, Kris M. Kitani, Jeffrey P. Bigham, and Chieko Asakawa. 2017. People with visual impairment training personal object recognizers: Feasibility and challenges. *Conference on Human Factors in Computing Systems - Proceedings* 2017-May, 5839–5849. https://doi.org/10.1145/3025453.3025899

[31] Amlan Kar, Nishant Rai, Karan Sikka, and Gaurav Sharma. 2017. Adascan: Adaptive scan pooling in deep convolutional neural networks for human action recognition in videos. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 3376–3385.

[32] Nicolas Kaufmann, Thimo Schulze, and Daniel Veit. 2011. More than fun and money: Worker motivation in crowdsourcing-a study on Mechanical Turk. 11 (2011), 1–11.

[33] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. 2012. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*. 1097–1105.

[34] Todd Kulesza, Margaret Burnett, Weng Keen Wong, and Simone Stumpf. 2015. Principles of Explanatory Debugging to personalize interactive machine learning. *International Conference on Intelligent User Interfaces, Proceedings IUI* 2015-January, 126–137. https://doi.org/10.1145/2678025.2701399

[35] Brenden Lake, Ruslan Salakhutdinov, Jason Gross, and Joshua Tenenbaum. 2011. One shot learning of simple visual concepts. In *Proceedings of the annual meeting of the cognitive science society*, Vol. 33.

[36] MIT Open Learning. 2019. *aik12-MIT*. https://aieducation.mit.edu/ Retrieved April, 2021.

[37] Kyungjun Lee, Jonggi Hong, Simone Pimento, Ebrima Jarjue, and Hernisa Kacorri. 2019. Revisiting blind photography in the context of teachable object recognizers. *ASSETS 2019 - 21st International ACM SIGACCESS Conference on Computers and Accessibility*, 83–95. https://doi.org/10.1145/3308561.3353799

[38] Daniela Massiceti, Lida Theodorou, Luisa Zintgraf, Matthew Tobias Harris, Simone Stumpf, Cecily Morrison, Edward Cutrell, and Katja Hofmann. 2021. ORBIT: A real-world few-shot dataset for teachable object recognition collected from people who are blind or low vision. https://doi.org/10.25383/city.14294597.v1 City, University of London, Dataset.

[39] Daniela Massiceti, Luisa Zintgraf, John Bronskill, Matthew Tobias Harris, Edward Cutrell, Cecily Morrison, Katja Hofmann, and Simone Stumpf. 2021. ORBIT: A Real-World Few-Shot Dataset for Teachable Object Recognition. *arXiv preprint arXiv:2104.03841* (2021).

[40] Tim Miller. 2019. Explanation in artificial intelligence: Insights from the social sciences. , 38 pages. https://doi.org/10.1016/j.artint.2018.07.007

[41] Meredith Ringel Morris. 2020. AI and accessibility. *Commun. ACM* 63 (5 2020), 35–37. Issue 6. https://doi.org/10.1145/3356727

[42] Joon Sung Park, Danielle Bragg, Ece Kamar, and Meredith Ringel Morris. 2021. Designing an Online Infrastructure for Collecting AI Data From People With Disabilities. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. 52–63.

[43] Jennifer Pearson, Simon Robinson, Thomas Reitmaier, Matt Jones, Shashank Ahire, Anirudha Joshi, Deepak Sahoo, Nimish Maravi, and Bhakti Bhikne. 2019. StreetWise: Smart speakers vs human help in public slum settings. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. 1–13.

[44] James Requeima, Jonathan Gordon, John Bronskill, Sebastian Nowozin, and Richard E Turner. 2019. Fast and Flexible Multi-Task Classification using Conditional Neural Adaptive Processes. In *Advances in Neural Information Processing Systems*.

[45] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. 2015. Imagenet large scale visual recognition challenge. *International journal of computer vision* 115, 3 (2015), 211–252.

[46] Manaswi Saha, Alexander J. Fiannaca, Melanie Kneisel, Edward Cutrell, and Meredith Ringel Morris. 2019. Closing the Gap: Designing for the Last-Few-Meters Wayfinding Problem for People with Visual Impairments. In *The 21st International ACM SIGACCESS Conference on Computers and Accessibility* (Pittsburgh, PA, USA) *(ASSETS '19)*. Association for Computing Machinery, New York, NY, USA, 222–235. https://doi.org/10.1145/3308561.3353776

[47] Elizabeth B-N Sanders and Pieter Jan Stappers. 2008. Co-creation and the new landscapes of design. *Co-design* 4, 1 (2008), 5–18.

[48] Ashley Shew. 2020. Ableism, Technoableism, and Future AI. *IEEE Technology and Society Magazine* 39 (3 2020), 40–50+85. Issue 1. https://doi.org/10.1109/MTS.2020.2967492

[49] Jake Snell, Kevin Swersky, and Richard Zemel. 2017. Prototypical networks for few-shot learning. In *Advances in Neural Information Processing Systems*.

[50] Adrianna Surmiak. 2020. Ethical Concerns of Paying Cash to Vulnerable Participants: The Qualitative Researchers' Views. *The Qualitative Report* 25, 12 (2020), 4461–4480.

[51] Mingxing Tan and Quoc Le. 2019. EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks. In *International Conference on Machine Learning*. 6105–6114.

[52] David Touretzky, Christina Gardner-McCune, Cynthia Breazeal, Fred Martin, and Deborah Seehorn. 2019. A year in K-12 AI education. *AI Magazine* 40 (12 2019), 88–90. Issue 4. https://doi.org/10.1609/aimag.v40i4.5289

[53] Shari Trewin, Sara Basson, Michael Muller, Stacy Branham, Jutta Treviranus, Daniel Gruen, Daniel Hebert, Natalia Lyckowski, and Erich Manser. 2019. Considerations for AI fairness for people with disabilities. *AI Matters* 5, 3 (2019), 40–63.

[54] Eleni Triantafillou, Tyler Zhu, Vincent Dumoulin, Pascal Lamblin, Utku Evci, Kelvin Xu, Ross Goroshin, Carles Gelada, Kevin Swersky, Pierre-Antoine Manzagol, et al. 2019. Meta-dataset: A dataset of datasets for learning to learn from few examples. *arXiv preprint arXiv:1903.03096* (2019).

[55] Oriol Vinyals, Charles Blundell, Timothy Lillicrap, Daan Wierstra, et al. 2016. Matching networks for one shot learning.

[56] Meredith Whittaker, Meryl Alper, Cynthia L Bennett, Sara Hendren, Liz Kaziunas, Mara Mills, Meredith Ringel Morris, Joy Rankin, Emily Rogers, Marcel Salas, et al. 2019. Disability, Bias, and AI. *AI Now Institute, November* (2019).

[57] Stephanie Wilson, Abi Roper, Jane Marshall, Julia Galliers, Niamh Devane, Tracey Booth, and Celia Woolf. 2015. Codesign for people with aphasia through tangible design languages. *CoDesign* 11, 1 (2015), 21–34. https://doi.org/10.1080/15710882.2014.997744 arXiv:https://doi.org/10.1080/15710882.2014.997744

[58] Shaomei Wu and Lada Adamic. 2014. Visually impaired users on an online social network. *Conference on Human Factors in Computing Systems - Proceedings*, 3133–3142. https://doi.org/10.1145/2556288.2557415

[59] Qian Yang, Aaron Steinfeld, Carolyn Rosé, and John Zimmerman. 2020. Re-examining Whether, Why, and How Human-AI Interaction Is Uniquely Difficult to Design. (2020), 1–13. https://doi.org/10.1145/3313831.3376301

[60] Chen Zhu, Xiao Tan, Feng Zhou, Xiao Liu, Kaiyu Yue, Errui Ding, and Yi Ma. 2018. Fine-grained video categorization with redundancy reduction attention. In *Proceedings of the European Conference on Computer Vision (ECCV)*. 136–152.

[61] Luisa Zintgraf, Kyriacos Shiarli, Vitaly Kurin, Katja Hofmann, and Shimon Whiteson. 2019. Fast context adaptation via meta-learning. In *International Conference on Machine Learning*. PMLR, 7693–7702.