An Honest Conversation: Transparently Combining Machine and Human Speech Assistance in Public Spaces

Thomas Reitmaier,[1] **Simon Robinson,**[1] **Jennifer Pearson,**[1] **Dani Kalarikalayil Raju,**[2] **Matt Jones**[1]

[1] Computational Foundry, Swansea University, UK
{ thomas.reitmaier, s.n.w.robinson, j.pearson, matt.jones } @swansea.ac.uk

[2] Studio Hasi, Mumbai, India
daniel@studiohasi.com

ABSTRACT

There is widespread concern over the ways speech assistant providers currently use humans to listen to users' queries without their knowledge. We report two iterations of the Talk-Back smart speaker, which transparently combines machine and human assistance. In the first, we created a prototype to investigate whether people would choose to forward their questions to a human answerer if the machine was unable to help. Longitudinal deployment revealed that most users would do so when given the explicit choice. In the second iteration we extended the prototype to draw upon spoken answers from previous deployments, combining machine efficiency with human richness. Deployment of this second iteration shows that this corpus can help provide relevant, human-created instant responses. We distil lessons learned for those developing conversational agents or other AI-infused systems about how to appropriately enlist human-in-the-loop information services to benefit users, task workers and system performance.

Author Keywords

Conversational agents, speech appliances, public space interaction, emergent users.

CCS Concepts

•**Information systems** → *Speech / audio search;* •**Human-centered computing** → Field studies; Interaction paradigms; Sound-based input / output; Interaction techniques;

INTRODUCTION

Human assistance with AI model training is increasingly common as data-driven services and appliances continue to expand in availability and scope. For some task domains, comprehensive and focused datasets are available that can be incorporated into any future models. Consider, for example, face recognition tools, which are routinely benchmarked against libraries of training and evaluation images. For other areas, such as spoken question-and-answer systems, training is constant and ongoing, with models continually being updated to incorporate the breadth and diversity of human speech interactions and responses. Here, then, there is a clear need for manual insight in *"complex edge cases that machines can't easily grasp, where you need humans to provide nuance and judgement"* [24].

While reliance on humans for data processing assistance is widespread, in many cases service providers only disclose the extent and specifics of their use of crowdsourcing as a reaction to journalistic enquiry [4, 11]. Consider, for example, the criticism recently attracted by smart speaker manufacturers such as Amazon, Apple and Google of their attempts to improve understanding and response quality via manual labelling [4, 8]. A key worry in these particular cases is that privacy might be violated: in effect, seemingly confidential utterances, spoken in homes and private spaces as if to another person in the room, are being eavesdropped upon by strangers. Unease about this situation is exacerbated by the fact that the insights given by human data labellers are not being used to directly respond to the questioner; instead, the response from the real neural network—the human—is being used to train a machine one for broader future use. Furthermore, the types of content that end up being sent to human classifiers are often by definition the *most* sensitive due to their context (e.g., mis-activations of the system, or difficult, uniquely identifiable questions [11]).

While concerns have been widely voiced, then, with regard to such practices in *private* spaces, in this article we build on our previous work that has shown that direct, overtly human-powered responses to questions posed to smart speakers in *public* settings can be of value [26]. Our earlier work explored the installation of voice assistant-type devices across Dharavi, a very large informal settlement in Mumbai, India. Residents of Dharavi typically have lower textual and technological literacy than mainstream users (in, say, Europe or USA). For these types of users spoken interaction has clear advantages. Many of the contextual assumptions made by smart speakers do not necessarily hold for slum residents—or emergent users [10]—making devices that are designed for mainstream users less appropriate. In addition, environmental factors and the conversational nuances of interactions that are made in public places can make giving correct and helpful responses even more challenging. In such settings, human interventions can be more relevant and contextually appropriate, with the trade-off of a necessarily slower response speed compared to the instant responses offered by conversational agents.

Previously we compared human-powered smart speakers against separate machine-powered ones. Here, we investigate whether *merging* machine and human responses in a single speech appliance can provide benefits in the same setting. We extended the open-source toolkit described in [26] to create

a combined machine+human system that we call TalkBack, and deployed ten instances of this in areas of Dharavi for a 25-day period. Unique to the TalkBack system is that forwarding questions to humans is an explicit, user-driven choice. Next, building upon deployment results, we explored whether libraries of previous spoken human responses could be used to assist with new queries. That is, we allowed users to query a corpus of human responses and assess whether these recorded answers have value as direct real-time responses to questions. We deployed a new system using this approach for 12 days in Dharavi, evaluating its ability to respond to new queries. Our results show how future speech appliances could nuance their approach and responses by being more open about (and making potentially better use of) their human help.

We do not explore here why users might or might not want to withhold their utterances from un-seen humans, but rather investigate how an open and transparent combination of AI and human help—as embodied in the two deployments—might be received. The paper answers this question, showing that people will often choose to forward queries to human helpers (and the results this affords); and, (in the second deployment), can get real value from previously recorded answers, in contrast to the conventional approach of using such prior queries primarily to tweak AI models and algorithms. Our work, then, builds on previous insights into smart speaker and human-in-the-loop systems by showing for the first time how a highly accessible advanced computational information retrieval system could be coupled with the richness of a human being in an explicitly transparent way for a more productive experience for users. In addition, in carrying out the work in a non-Western setting, beyond the "Californian" perspective, we aim to show the value of further "diversifying future-making" [27].

BACKGROUND

To help frame the work, we first turn our attention to the users and context; in our case, to so-called *emergent* users in Dharavi. Next, we draw lessons from related research in Interactive Voice Response (IVR) systems which have, to date, had a widespread positive impact within such resource-constrained communities, and summarise previous work on smart speakers in public resource-constrained settings. Given the dearth of research on such devices in *public* settings, we also consider relevant smart speaker and conversational agent research in more *private* settings. Finally, we contextualise our work within wider discourses that consider the role of humans in AI-enabled futures, particularly as parts of our deployment were enabled by a crowd of human question answerers.

Context and users

Dharavi, one of Asia's largest informal settlements (2.1 km^2), is a dynamic inner-city township in the centre of Mumbai. Located next to financial and business districts, the people living and working in Dharavi do so on some of the most expensive real estate on the planet. In many cases residents have never bought the land; *"they just got there first and made it their own"* [7]. It is no surprise then that there are complex *"conflicts and negotiations playing out in Dharavi between multinational corporate interests, state actors and residents [. . .] over land, development and rights to city space"* [25].

Within Indian HCI research and practice, residents of communities such as Dharavi are often referred to as 'emergent users' [10] to orient designers to the fact that technology users here *"may have less education"* (e.g., not reached college) and *"may be poor"* (e.g., marginal farmers, very small business owners), but at the same time are also beginning to get access to the sorts of advanced mobile devices and services that many users in developed regions are familiar with [10, 21]. In contexts such as Dharavi, and particularly for those unable to afford mobile devices, or who lack the textual or technological literacies to fully operate the device or utilise its services, technology usage patterns are 'intermediated' [31]. That is, the expanding reach of technologies is enabled by digitally skilled or financially better-off users. These users act as intermediaries by assisting persons for whom technology is inaccessible.

Since people living in India enjoy the cheapest mobile data rates in the world [3] and through low-cost or second-hand mobile devices, it is now textual and media literacy, rather than financial constraints, that present the biggest barriers to information access. In fact, the emergent user technology adoption model developed by Devanuj and Joshi highlights the inability to input text as a key characteristic of basic users. Even for literate users, however, text input is more challenging for Dēvanāgarī scripts (used in Indic languages such as Hindi and Marathi) than Latin ones, because characters are not discrete but need to be typed as a combination of consonant and matra. Custom keyboards (such as Swarachakra [33]) have gone a long way to improve the usability, speed, and accuracy of Dēvanāgarī script input through good interaction design. Even so, this input requires users to understand and navigate information hierarchies, which, as Walton et al. argue, can themselves become media literacy barriers [36].

Voice user interfaces in resource-constrained settings

Addressing the above challenges, IVR systems have demonstrated widespread impact in emergent user communities for the past decade. Typically seen as an annoyance by mainstream users contacting customer service departments of businesses such as banks or mobile operators, IVR systems that allow users to create and share their own content have been tremendously popular elsewhere, particularly in India and Pakistan. For instance, Polly [29], which emphasises viral information spread, has been used extensively for entertainment and social contact in Pakistan. Sangeet Swara, the IVR equivalent of an internet forum, has been used by rural communities in India for songs, poems, jokes and cultural content [35].

More recently, advances in AI, natural language processing (NLP) and conversational agents (CA) have created new technological possibility to (dis)solve the problem of Dēvanāgarī input and search for emergent users. In previous work, we reported on design workshops and technology walks in Dharavi, where residents identified opportunities for providing smart speaker services in public areas [30]. Deploying a simple Wizard-of-Oz probe that allowed passers-by to ask and receive answers to Hindi and Marathi language questions, the research demonstrated the value of experimenting with publicly accessible smart speakers, which in turn promote awareness and community learning about speech interaction.

Leveraging the newly-released Hindi language version of Google Assistant (GA), our StreetWise system [26] extended this line of research to compare the *source* and *speed* of responses to spoken language questions. Depending on the location and device on which the questions were posed, responses were either provided in realtime through GA (machine-powered and instant) or given by a crowd of human question answerers (human-powered and delayed). Longitudinal, multi-sited deployments of the devices revealed that human answerers were better able to understand and relevantly answer questions. However, users also appreciated the instant response they received from GA and, at times, found that the delayed human responses took too long, particularly when question responders were working through an influx of questions. When questions were about basic facts—such as the time, the weather or, say, who the prime minister of India is—and in the instances where GA could accurately transcribe—and therefore 'understand'—the question, users appreciated the instant responses they received. On the other hand, for more philosophical or open-ended questions GA was rarely able to provide a relevant response. More generally it was unclear to users whether GA could not understand (for instance because of background noise) or did not have an answer, as it would offer a canned "I don't know" response variation in both cases. Human responders fared better, providing mechanisms to repair failed interactions, such as by providing suggestions that users retry or rephrase their questions when they could not understand them.

Voice interactions in private spheres

Turning to more private contexts, researchers have studied, and developed design implications for, conversational agents embedded into phones [23]—presumably individually owned as the research is located in Western contexts—and home appliances such as smart-TVs and smart speakers [22, 28].

It might be tempting to think of a home as quieter compared to a slum alleyway; however, there are parallels between our previous public smart speaker deployments and Porcheron et al.'s ethnomethodological study of UK domestic smart speaker use [28]. During dinner time, the home is a complex social context with multiple concurrent activities and multiple people interacting with the device (and with each other) simultaneously. To better cope with multi-party conversations, Porcheron et al. suggest that voice user interfaces should focus on request/response design, rather than conversational design. Looking at mobile assistants, Luger et al. also take issue with the conversational terminology, because *"user expectations of conversational agents systems remain far from the practical realities of use"* [23]. The interaction design of the prototypes presented here attempts to clearly manage user expectations, orienting the systems around a request/response format.

Academic literature has also explored issues uncovered by journalistic inquiry [4, 11]. In their study of smart speakers, Lau et al. [22] show that while users trade privacy for convenience, in making that trade they also show varying levels of privacy deliberation and even resignation. In their view, *"transparency about smart speakers' data practices"*—beyond End User Licence Agreements—is needed as well as good interaction design that helps *"users form accurate mental models of the smart speaker's functionality"* [22]. They further recommend that *"rather than hiding privacy information [. . .], smart speaker companies should leverage the main interaction capabilities of their products – voice – to integrate conversational privacy dialogues into the smart speaker user experience"* [22]; advice we follow in the prototypes deployed for the studies reported here.

Humans and artificial intelligence

Issues surrounding privacy are intimately linked to crowdsourcing practices and wider societal discourses surrounding the role of human labour in AI-infused products and services [1]. Particularity for NLP and computer vision, human intelligence is often integrated into what Kamar calls 'hybrid systems', which *"offload computational tasks to humans on demand to overcome the deficits of AI systems"* [20].

Through their study of crowdsourcing platforms, Gray and Suri lay bare the often harsh realities of 'ghost workers' performing on-demand, crowdsourced 'computation'. Despite this reality the authors are optimistic that there is possibility embedded in this hidden work; and, crucially, that we would all be better served if we knew how the *critical infrastructure* of the AI supply chain enabled through such human support actually functioned [15]. Beyond crowdsourcing and AI, research reveals that the visibility and material forms of critical infrastructure contribute to its socio-technical effectiveness [34]. The prototypes we built through in this work made clear the role of human helpers in the voice assistant service.

DESIGN PROCESS

We created two prototypes to embed, explore and deepen our understanding of the range of issues outlined above and that users are confronted with when they speak to a smart speaker in a public slum setting; and, to understand the extent to which human and machine interactions can be combined in ways that are meaningful and transparent for these users. To further this aim we adapted the existing open-source StreetWise toolkit[1] that was developed during our previous work (see [26]).

For the studies described here, as in our prior work, we chose to deploy and longitudinally evaluate prototypes in Dharavi, precisely because its residents have been engaged in envisaging how public speech-based systems could be useful in their contexts, and have already experimented with prototype devices of different capability and fidelity (e.g., [26, 30]). We are of course conscious of comments that *"HCI might have become too concerned with use in new places at the very time when a revolutionary technology is altering the basis of computing"* [17]. But the work here suggests that it is possible to both consider new contexts and contribute to the innovation of such technologies. We illustrate how Dharavi is exactly the sort of place to take on the work of designing, developing and deploying speech assistance interfaces, while also locating this research within a wider design ethos that disrupts existing technology mindsets—that humans are an excised or invisible aspect of AI-enabled futures [15]—by situating "future making" [27] outside of its traditional mainstream locations.

[1] https://github.com/reshaping-the-future/streetwise

Figure 1. A TalkBack appliance in-situ in Dharavi.

SYSTEM 1: MACHINE + HUMAN

We chose to keep the same physical design as StreetWise [26] in order to both build on potential user familiarity with the system's capabilities and to be able to compare our results more directly. We made several key changes to the software elements, however. Firstly, we extended the system to combine machine and human responses into a single speech appliance, named TalkBack. In the original human-powered StreetWise device, all questions were immediately sent to a human for answering, whereas in our revised design, questions are first sent to the Hindi version of Google's Assistant API [14]. The machine response (if one is given) is immediately played to the questioner. Following this they are asked, via an audio prompt, to rate its suitability (*"Did this answer your question?"*) by pressing 'Yes' or 'No' on the appliance's keypad.

If the machine-provided response is rated negatively, the user is informed that they can press '0' on the keypad if they would like to send their question to a person for answering. If they choose to do so, their question is submitted to a human answerer and a retrieval token (a four-digit number) is provided. Finally, after a questioner enters this number (into any deployed appliance) to listen to the human-provided answer, they are asked to respond 'Yes' or 'No' as to whether this new response answers their question. All interaction prompts have a timeout period of 15 seconds, after which the system returns to an idle state and registers that there was no user response. Figure 1 shows an appliance in-situ in Dharavi, and Fig. 2 outlines the interaction flow through the system.

Longitudinal deployment

We deployed 10 TalkBack instances in Dharavi for a period of 25 days in April 2019. We recruited 10 local business owners to host the devices and act as caretakers, keeping the appliances powered, secure and in working order throughout the deployment (one device per caretaker business). The host businesses selected ranged from a stationery store to a milk shop to a picture framer, aiming to cover a wide range of potential users and public community areas. Caretakers were asked to display the devices in a public-facing area of their shop at all times during normal opening hours (approx. 10am–10pm

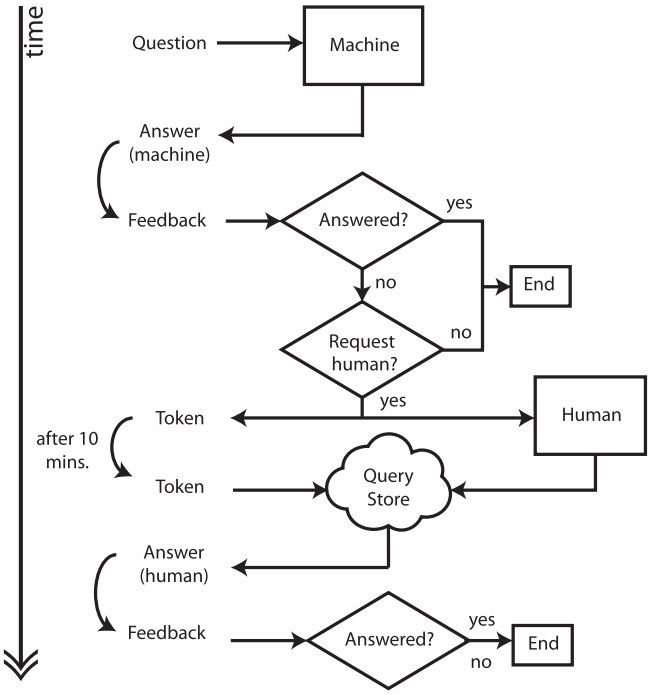

Figure 2. The interaction flow of the first TalkBack prototype (System 1). Machine-provided answers that the user states do not answer their question can optionally be forwarded to a human for a second answer.

Tuesday to Sunday). Each caretaker was compensated with ₹7,500 (~$100) for their assistance during the deployment.

We also recruited a pool of 10 people (6M; 4F) from various suburbs across Mumbai to answer questions via an accompanying Android app. There were seven college students (average age 21) and three working professionals (average age 28). Seven answerers were from the Lower Parel suburb of Mumbai (7 km from Dharavi) and three from North Mumbai (15 km).

All incoming questions were sent to a common pool, where people could answer them in whatever context they happened to find themselves. We gave each answerer an instruction document that outlined the process and protocols they should follow around politeness and interaction style. This instruction manual also gave suggestions about how to research answers they did not know, and how to respond to inaudible, inappropriate or unanswerable questions. We created an instruction video that showed the design of the TalkBack appliance, and demonstrated how questions were asked and retrieved. Answerers appreciated the added context the video provided and, as later interviews revealed, were motivated by the fact that the questions they answer might help someone in Dharavi. This contrasts with how transcribing tasks that enable conversational agents to 'become smarter' are typically discretised, decontextualised, and ultimately articulated into 'Human Intelligence Tasks' on platforms such as Amazon's Mechnical Turk [15]. We also followed best crowdsourcing practices (as detailed in [15]), by encouraging collaboration through a WhatsApp group, and providing certificates to each answerer (which could for example be used on a resume). Each answerer was paid ₹10,000 (~$140) for their help.

Aims and research questions

Our prior work has demonstrated that, when compared to conversational agents, humans are not only able to give more relevant answers, but are also able to give answers in cases where machines are unable to help [26, 30]. However, with TalkBack's rating functionality we were able to gather user- rather than researcher-generated insights about the sorts of questions that a machine can or can not answer well. Further, with the explicit prompt that asks users if they want to send their question to a human for answering, we can (a) explore the sorts of questions that users are willing to submit to a human; and, (b) assess the quality of the human responses received.

We were particularly interested in learning more about human referral patterns and exploring whether this information, along with the resulting corpus created in partnership with users and question answerers, could be used to improve future human+machine systems. However, the main research question that motivated this first deployment was: *in situations where machine-powered conversational agents are unable to give a satisfactory answer, will questioners choose to forward their questions to humans in order to get a better answer?*.

Local adaptation

We engaged with graphic designers familiar with the Dharavi context to create and refine text-light illustrations. On the main front panel of the device the illustration showed that user questions would initially be directed to and answered by Google's voice assistant (see Fig. 1). Mounted above the device, we further illustrated that users would be given the subsequent choice to send their question to a human assistant for an answer, in case it was not responded to satisfactorily, as well as how and when to subsequently retrieve the human answer. Voice prompts were further refined from previous deployments to ensure that they matched the graphical depictions.

Data capture and consent

We captured and logged the following data from each deployed TalkBack appliance:

- Question audio and Google Assistant API transcripts;
- Machine response audio and transcripts;
- Human response audio recordings;
- Ratings (one per question/answer pair for machine answers; one per retrieval for human answers);
- Interaction events (e.g., question, answer and retrieval times; button presses; errors).

We analysed the question and answer audio only for questions where users were informed that their questions would be heard and responded to by humans (i.e., those that were explicitly chosen to send to a person for answering). This analysis was done using the same method as in our previous study [26] (and by the same native English-speaking researcher) by transcribing and translating questions and answers into English as is common practice for emergent-user smart speaker research [5]. All analysed audio (i.e., questions as well as machine and human responses) was fed through Google's Speech-to-Text [12] and translation APIs [13]. Human answer translations were quality-checked by a native speaker, and all resulting data was

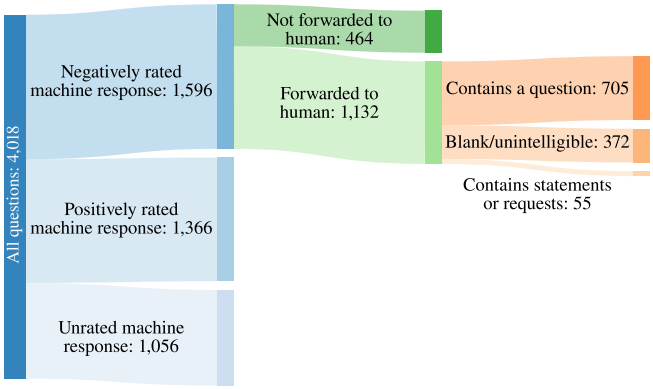

Figure 3. A breakdown of all questions, machine response ratings and human forwarding choices from the first TalkBack deployment.

placed into a tagging system, which used the same process, conventions and categories as in [26]. That is, we classified each interaction into broad types (as shown in Table 2) and identified the general demographic of the questioner (i.e., male, female or child).

Conversely, for the subset of questions that were only posed to Google Assistant and not subsequently forward to a human, we only looked at interaction data, and not question/answer audio recordings or transcripts. Finally, at the end of the study we also interviewed the caretaker shopkeepers and human answerers in order to be able to assess and respond to their insights about the deployment.

Results and analysis

Over the 25-day deployment period we received a total of 4,018 questions across the 10 appliances installed. The minimum number of questions per appliance was 171, and the maximum was 912 (average: 402).

Of the 4,018 questions received, 2,963 were given ratings for their associated machine answers, with positive ratings given 1,366 times (34 %) and negative ratings 1,596 (40 %) times. The remaining 1,056 answers (26 %) were not rated by the user (see Fig. 3). This is confirmatory to our previous work [26], where 41 % of machine questions received relevant answers. As a caveat we note that the previous work did not include blank questions or those recorded during accidental activation in the 41 % aggregate. In addition, in that work *researchers* rather than *users* assessed whether an answer was relevant, as there was no rating step in the StreetWise system design. As noted above, because one of the prime drivers of this work was to be transparent about when we would listen in to queries, we did not analyse questions that were not sent for a human response. However, as the system flow in Fig. 2 illustrates, the first interaction users have with the system is through the Hindi version of Google Assistant just as in the machine-powered appliances of the previous work, and we have no reason or evidence to believe that users' patterns of interaction would be any different in the TalkBack system. Therefore, we refer the reader to the earlier work to for a more thorough analysis of the types of questions that users asked and the range and relevance of responses they received.

Questions forwarded to a human answerer	**1,132**
Unique requests for answer *	324
Repeated requests for answer *	498
Answer not requested	808
Total user requests (sum of *)	**822**
Unique successful retrievals	187
Repeated successful retrievals	204
Unsuccessful retrievals	431

Table 1. A breakdown of user requests for, and retrievals of, human answers from the first TalkBack deployment. Retrieval requests made before a human answer had been provided are classed as unsuccessful. Some users made multiple requests until an answer was available; others tried only once – success in this case depended on the speediness of the human answer. Many answers were successfully retrieved repeatedly.

Query category	n	%	MPI / HPD	
Question:	**705**	**93**	**86**	**96**
Basic facts	428	56	50	58
Contextual questions	112	15	19	24
Directed at machine	47	6	12	4
Philosophical questions	118	16	5	10
Not a question:	**55**	**7**	**14**	**4**
Statements	32	4	8	3
Requests	23	3	6	1

Table 2. Types of queries received in the first TalkBack deployment, showing the percentage of valid interactions in each category (using the same classifications as detailed in [26]). Interactions are categorised as either a *question* or *non-question*, then into further sub-categories. The percentage distributions of machine- (MPI) and human-powered (HPD) interactions from the previous work are also shown for comparison.

Of the 1,596 interactions where the user rated the machine response negatively, the questioner chose to send their query to a human 1,132 times (71 %). Hereafter we focus solely on these 1,132 interactions, and the machine and human responses provided. TalkBack users were explicitly aware that these utterances would be heard and responded to by humans. The average time taken for a human response was 55 minutes (min: 30 s; max: 10h 34min (overnight)). A majority of these questions were submitted by children (≈61 %), with the remainder ≈33 % male, ≈4 % female and ≈2 % unknown.

There were 822 requests for answers, with the keenest being 19 seconds after asking, and the largest interval being almost 19 days after asking (see Table 1). People often requested answers before they were available (e.g., the 19 s request, above), which led to a large number of requests being unsuccessful. In total the 822 requests were for 324 questions (29 % of those submitted to a human) and therefore around 61 % of the answer requests were duplicates. That is, users either attempted to request an answer a second time after their previous request was unsuccessful; or, wanted to listen to an existing answer again. There were 391 successful retrievals of human answers (including repeated retrievals) for 187 unique questions (≈17 % of those sent). This number also includes 35 answer retrievals that were initiated on a different TalkBack appliance from that on which the question was asked. We received 260 human answer ratings for 143 unique questions (with the remainder not rated before the system's 15 s timeout). 172 ratings (66 %) were positive and 88 (33 %) were negative.

After analysis of the 1,132 human-forwarded questions, we were able to identify a question in 705 interactions. In a further 55 interactions we identified statements or requests that were valid sentences, but were not a question, or that did not warrant a response. The remaining 372 questions were blank, background noise or unintelligible (see Fig. 3). Table 2 shows the type classifications of all valid interactions received.

Human-machine question insights and reflections

For the majority of categories, the percentage of questions and non-questions fell somewhere between those reported in previous work for each type of device [26]. There were two exceptions to this pattern. Philosophical questions accounted for a higher percentage of total valid questions than both the machine and human-powered devices of our earlier work. The other exception was for contextual questions, which were recorded less often in this deployment than the previous one.

In line with our previous work's findings, human answerers were able to give relevant responses to the 760 total valid interactions in 642 cases (84 %). In comparison, only 143 questions (19 %) received a response from the machine that was classified as relevant, though it must be noted again that all of these questions are solely where the user rated the machine response as inadequate in order to progress to sending their question to a human (see Fig. 2). The remaining answers were either irrelevant in 16 cases (2 %) or were "I don't know" responses in 102 cases (13 %). Figure 4 illustrates answer relevance for both human and machine across question analysis categories.

Unlike our previous work, we *are* now able to make direct comparisons between machine and human answers to the same query, which leads to further interesting insights about how machines are able to handle complex interactions. For example, the machine often struggles with contextual queries which typically expect a certain level of background knowledge. Questions such as *"Who is playing today?"* fail unanimously by machine, but are usually answered correctly by humans who fill in the contextual gaps to understand that this question most likely refers to Indian Premier League cricket, and therefore can answer appropriately: *"Today is IPL of Rajasthan Royals and Chennai Super Kings"*. Another topical and often variably-constructed query theme was local elections (25 questions), where again humans prevailed over the machine response by inferring vital context (e.g., *"What is the exit poll?"*, *"When is the election?"* or *"Who is going to win in 2019?"*). Furthermore, locally specific questions, such as: *"Document required for making PAN card?"* or *"Who is the corporator of ward number 184?"*, typically unknown by the machine, were usually answered appropriately by humans.

As we previously discussed in [26], it is clear that relatively minor inaccuracies within questions still prove difficult for machines to interpret. For example, several people asked about the Chief Minister of India, a question which is flawed in itself: India has a Prime Minister and a President, but only the states within the country have Chief Ministers. This exact question was received six times during our deployment; five times the machine responded with *"I don't know"*. In the final interaction, the machine returned the Wikipedia definition of Chief

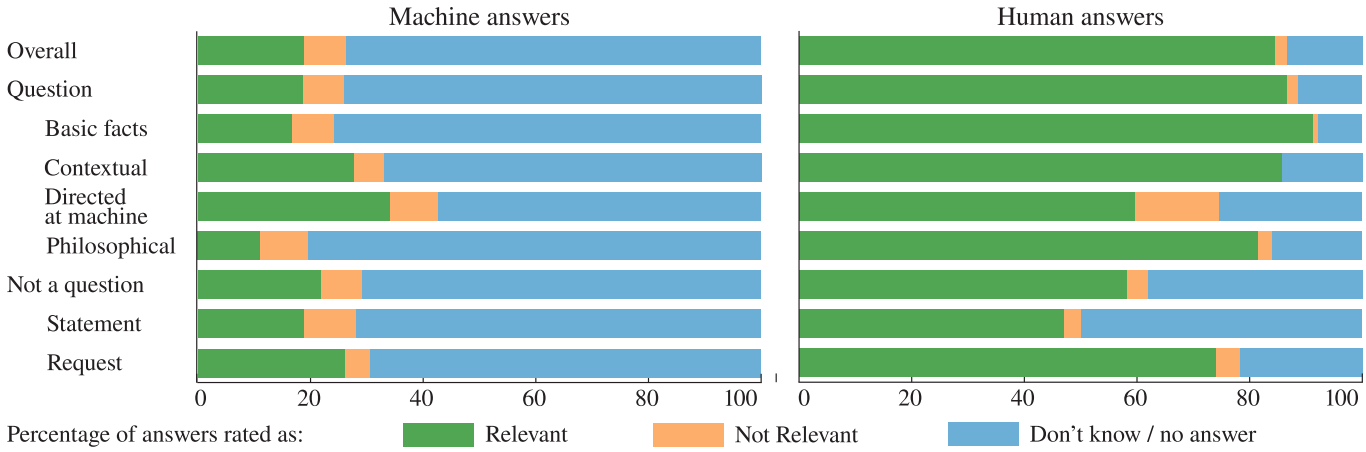

Figure 4. Answer relevance distribution across analysis categories (see Table 2) for all 760 valid interactions that were responded to by both machine and human in the first TalkBack deployment.

Minister, perhaps suggesting the invisible hand of a crowdsourced human in the network. We note, however, that the answer is still far below the standard of that provided by the human answerers, who almost always try to repair questions such as this by providing appropriate answers – for example: *"The country doesn't have Chief Ministers, they have a Prime Minister and the Prime Minister is Modi"* or *"Narendra Modi is the Prime Minister of India and Maharashtra's Chief Minister is Devendra Fadnavis"*.

Background noise and a mismatch between when the speed at which users wanted human answers to be ready and when question answerers were actually able to deliver them were major factors that affected system performance. While the background noise in many of the question recordings presents challenges to the natural language processing algorithms used by Google Assistant to accurately transcribe and act upon the questions users asked, this challenge is by no means unique to Dharavi or public settings more generally, as Porcheron et al.'s study of smart speaker use in British homes reveals [28].

Users did not attempt to retrieve answers on the TalkBack devices as often as with our previous human-powered systems: 17 % vs. 45 % [26]. While this result might lead one to conclude that adding a human-powered step does not significantly improve the system, it is important to understand the differences between the two studies. In our previous work, users of the human-powered device would receive no answer at all if they did not retrieve it at a later time. In the current study, users of the TalkBack appliances received an immediate machine response, even if it was often not relevant. We posit that users may have been less patient because their experiences were shaped by a machine that at least *tried* to answer immediately. It is clear that from the perspective of Dharavi users, human answers need to come more quickly, and that with such improvement the value of the service would be higher.

Shopkeeper and answerer insights

We interviewed shopkeepers and answerers after the deployment to learn from their insights into the experience. Shopkeepers told us that most users were happy that TalkBack was a free service, but noted, as suspected, that waiting for 10 minutes for human answers is too long. They also highlighted the pull of the device for children. For them, when a parent leaves the house, their access to the internet—in the form of a shared mobile phone—leaves with them. Children identify the caretaker shops as the ones with the 'speech box' and make sure to visit the shop to buy things, even if there are other shops to buy from. Some shopkeepers reported helping first-time users, either by demonstrating the device, or by offering them suggestions for how they might rephrase their questions so that the TalkBack device is more likely to respond to it. Some shopkeepers also assisted by writing down the four-digit retrieval token for users.

All of the answerers used their own knowledge in conjunction with web search engines and services such as Wikipedia to effectively answer questions. One answerer mentioned also enlisting the help of family members. Although the systems were deployed in public settings, questions requiring a personalised answer were common, such as *"When is my exam?"* or *"When is my result?"* which they felt were unusual questions, and difficult to answer usefully. Answerers were familiar with general contextual information about Dharavi, such as the frequency of trains, buses and so on, but they reported that correctly answering questions about specific locations in Dharavi was difficult as many local businesses do not have a digital footprint. They also noted that it could be difficult for a computer to understand the dialect of Dharavi and the way residents phrased the questions. One criticism was that they found it demotivating to listen to empty audio while expecting questions (i.e., accidental appliance button presses), and responding to repetitive questions was tiring. However, the answerers all found value in helping residents of Dharavi who do not have internet access, and suggested improvements for future versions of the app, such as, for example, the ability to sign in or out of the task based on their availability.

SYSTEM 2: MACHINE + HUMAN + SEARCH

Following the deployment of the initial TalkBack prototype, we developed a further iteration of the system. This second revision added an intermediate step between the instant machine and delayed human response stages which searches a corpus of

previous human question/answer pairs in an attempt to provide a relevant and instant human-powered response. The aim of this prototype was to determine whether we could achieve a balance of human and machine capabilities while reducing the delay encountered from the human side and alleviating the strain on human answerers from repetitive questions.

The question we had was: can we effectively provide instant human responses from a repository of previous question/answer pairs? Further, how might this affect how users enlist human answerers? As with most information retrieval systems, success of a system of this type is highly dependent upon the depth and breadth of its corpus. We might reasonably expect, then, that many of the questions asked of a small corpus will return answers that are not entirely related to the original query. However, in order to gauge broad responses and assess the value that such a system might have to a community—and to those providing information services to that community—we decided to present the system to shopkeepers and TalkBack users in Dharavi in a 'raw' form. After all, if software and the data that software acts on is the material of design [6], we wanted to find out—in collaboration with users—what kind of material are we working with?

User interaction and system design

The physical hardware of the second TalkBack iteration was identical to that of the first deployment, but the appliances differed in a key part of their interaction, as illustrated in Fig. 5. Users begin with their question and receive an instant machine-powered response. The system then prompts the user for feedback on this answer, asking, as before, whether it acceptably answers their question. Selecting 'Yes' will end the interaction, while choosing 'No' causes the system to search its human answer corpus and provide the stored previous response to the closest matching question in the repository. As we were unsure of the quality and relevance of the answers the corpus might surface, we prefaced the playback of each question with a prompt explaining the provenance of the answer that was about to be played back: *"Here is a previous, related answer that you might find interesting . . . "*. Next, users are asked if this response answered their question, and their selection is used to either end the interaction (via a positive response) or ask if they would like to listen to another answer from the human corpus (a negative response, repeated up to a maximum of three times). Finally, if none of the stored answers prove useful, users are given the option to send their question to the human answer team for a delayed but tailored response (providing, as before, a 4-digit numerical token to be entered after 10 minutes). At any point in the process, if the user does not provide feedback after 20 seconds, the system times-out and the interaction ends.

It should be noted that the previous human corpus will in most cases return a list of results. Their relevance to the query will vary, however. The system provides the closest answer based on standard search index data-structures, which operate similar to the index found in the back of a physical book; and, query algorithms that act on that index. We used the Apache Solr/Lucene software platform for this purpose, in large part because of its popularity and extensibility. We tailored the

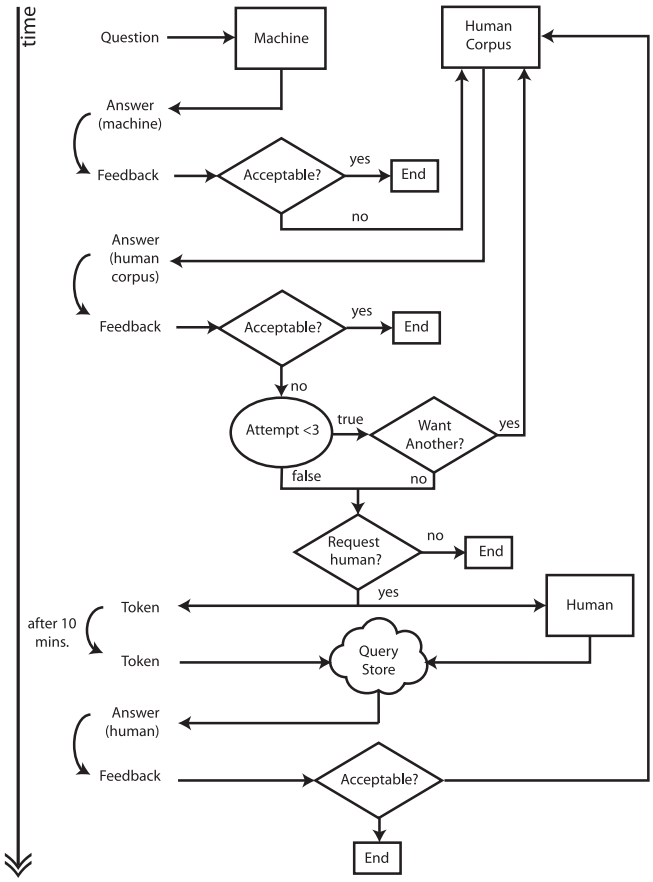

Figure 5. The interaction flow of the second TalkBack prototype (System 2). Interaction is identical to System 1 except that between negatively rating a machine answer and requesting a new human answer, the user is offered up to three previous human-provided answers that a search query system ranks as related to the current question.

query system with a Hindi language stemmer and by removing stop-words from the search index and from each query [2].

The corpus of questions we indexed were those obtained from previous deployments: 4,543 questions in total. As users first interact with Google Assistant, which provides a live transcript, we use this question transcript as the search query and are therefore able to immediately return search results. To better preserve user privacy, even though the retrieved answer is matched based on the similarity of the currently and previously asked questions, only the answers are played back (rather than, say, also the associated question).

Based on answerer feedback (as reported above), we also refined the answering architecture to a) speed up responses to common or inaudible questions (by providing boilerplate answers); b) assign questions to individuals by default rather than using a common question pool; and, c) allow answerers to sign in and out, where signing in indicates that the answerer is ready and able to respond to questions.

Deployment

Five months after the first TalkBack deployment, we deployed ten devices of the second design iteration in Dharavi at the

same shops and in the care of the same shopkeepers as in previous deployments. As before, we recruited a team of local human answers from around the city of Mumbai. These were a subset of five (3F) from the previous deployment, so required no further training in how to respond to queries, and minimal instructions regarding the modified accompanying app.

While the overall deployment lasted longer, we encountered various environmental, technical, and socio-technical challenges, which mean that we are only able to report on twelve days of data, with a fully-operational set of ten devices for nine days in total. An unusually intense monsoon resulted in many of the now discontinued VoiceKit [18] microphones used in the original appliances degrading due to the high humidity in shopkeepers' storage areas, which are not climate controlled. We substituted general purpose components—a USB-sound card and 3.5 mm microphone—which required rebuilding the appliances and re-architecting the embedded application. We were therefore only able to install two devices on the first day of the deployment period. Flooding on the second day meant that we were unable to reach any of the shops. On the third and fourth day we were able to install six and two devices, respectively. Finally, we experienced technical problems with the remotely-hosted search server, leading to a period of downtime during which the system automatically reverted to operation as in the first deployment reported earlier in this paper. In the following reporting we have excluded all question/answer pairs received during this timespan.

Results and discussion

1,042 questions were received during the 12 days of deployment. As illustrated in Fig. 5, corpus answers are only played if the user first negatively rates the machine answer and their query subsequently generates a search result. 363 questions met these criteria and matched previous results in the human answer library. Of these, users indicated that they were satisfied with the search result that they received 123 (34 %) times, and that the result did not address their question 167 (46 %) times. The remaining 73 (20 %) times did not receive a response (i.e., the system timeout was reached).

To analyse the results, we chose to compare user-generated human corpus ratings with researcher-assessed answer relevance[2]. This analysis took place using question and answer transcripts using the same relevance criteria as in our earlier study. Of the 363 human corpus answers, we found 99 answers (27 %) to be relevant and 264 answers (73 %) not relevant to the question asked (see Table 3). It is important to note that we analysed transcripts of *all* search results provided, so these results also include the 73 human corpus answers that the actual user did not rate. Drilling into the 123 positively user-rated corpus answers, our assessment of relevance agreed with the user's positive rating on 44 occasions (36 %). For the 167 negatively user-rated corpus answers, our assessment of relevance agreed with the user's negative rating on 130 occasions (78 %). Finally, for the 73 human-corpus answers the users did not rate, we found 18 responses (25 %) to be relevant.

[2]To maintain consistency, each answer was rated as 'relevant' or 'not relevant' to the new query by the same single researcher as in [26].

Answer corpus responses	363	100 %
a) User ratings:		
Positively rated	123	34 %
Negatively rated	167	46 %
Unrated	73	20 %
b) Researcher assessments:		
Relevant response	99	27 %
Irrelevant response	264	73 %

Table 3. a) User ratings of the human answer corpus responses they received. b) Researcher's assessment of the relevance of these responses to the associated (i.e., new) query.

We might learn a bit more by unpacking a sample positive and a sample negative interaction with the search system. When a user asked the question *"When will I get married"*, Google Assistant offered the response *"I think you are no less than anyone, if you want we can find dating tips"*, which the user rated negatively. Searching through the corpus,[3] this query matched with an identical previous one, to which the question answerer had offered the response: *"Soon"*. After this answer was played to the user, they indicated that they were satisfied with the answer they received, which ended the interaction.

On a different device, another user asked the question *"How many died in Azamgarh,"* to which the machine responded: *"Sorry I don't understand"*. In the human corpus, this query matched on the word 'die' with the question *"When did Gandhiji die,"* to which the question answerer of the previous deployment had offered the correct response: *"Gandhiji died on January 30, 1948"*. However, when this answer was played back to the current deployment's user, they naturally found that it did not satisfy their request. The user could have opted to send the question to a human for answering, and likely would have received a relevant response, given that 84 % of valid questions such as this one did so in our earlier deployment. However, they did not choose to do so. And while we certainly do not know the user's line of reasoning, we can draw evidence from the fact that only 27 % of questions were relevantly responded to through the human corpus (as assessed by a researcher), and that like this particular user, only 18 % subsequently opted to forward their question to a human.

We can sympathise with the user in this interaction that after receiving two unsatisfactory responses they simply gave up. Perhaps, in some cases, a perceptive or regular appliance user might simply indicate that a response answered their question just to end the interaction, even if it did not do so. This would help explain the surprising result that we only agreed with users' positive ratings 36 % of the time.

LIMITATIONS AND FUTURE WORK

Turning first to the people who used the TalkBack system. Our findings reproduce those of emergent user studies in India and Pakistan of IVR systems [29, 35] as well as smart speaker studies [5, 26] in that women—as far as we could tell in our analysis—were in the minority of users (4 %). Future work should focus on their non-use [32] and, if appropriate, could

[3]Although we translate the query into English, it is important to note that the actual search query and corpus index are in Hindi.

pivot back into the home where female users in Bhalla's study indicated they prefer to use mobile voice assistants [5].

In our second study, by presenting the corpus in its 'raw' form we have learned that deliberate choices (such as removing time-sensitive answers from the corpus), better information retrieval techniques, and perhaps also larger corpora are needed to improve performance. Without such steps, the intermittent stage we introduced in the second TalkBack iteration might eventually serve only to frustrate users. In future work, an alternative use of the corpus could integrate the human answer library as a knowledge base for the conversational agent to draw upon; or, it could be presented as a resource for question answerers to consult – for instance, to avoid having to re-record an answer to popular questions.

A further limitation of this work is brought about by asymmetric power relations between researchers and emergent users, which make it notoriously difficult to utilise typical instruments (e.g., questionnaires and surveys) to assess the usability and utility of new technologies [9]. In the longitudinal, exploratory studies presented here, we focused instead on unprompted usage at the expense of obtaining and analysing more granular demographic data on, for instance, the 'technology savviness' of particular users. Situating future deployments into diverse home environments in Dharavi might afford an opportunity to drill down into more granular levels to determine if there are mediating or moderating relationships between particular demographics and usage patterns.

While it may be interesting to consider why a proportion of the queries received were not forwarded to a human, we intentionally opted not do this, as it would undermine the key basis of the paper – that is, that questions should only be listened to when the user is aware that this will happen. In the cases where users did not forward their queries, they assumed—rightly—that no further analysis of their queries would take place. Asking users if we could analyse such non-forwarded queries to a researcher as part of a scientific study—immediately after they had indicated they did not want to send it to a human—would have led to a far more complicated interaction flow and potentially confusion over how trustworthy the box was in terms of conforming to the user's wishes.

In 29 % of cases, users opted not to have their questions listened to by a human. Given that our deployment took place in a *public* setting where questions were already likely to be overhead by shop keepers or passers-by this is a substantial percentage. As we already know from the growing corpus of academic literature [22] and journalism [24, 11] there are many occasions inside more private contexts such as the home, or the car, where users would not want their questions listened to by unknown others. In future work that pivots from public into more *private* contexts such as the home or car we would expect this number to grow from the 29 % public space baseline that we report in this work.

CONCLUSIONS

We have built upon our previous work [26] that compared two types of publicly situated voice-based assistants: one that used AI-based software in an attempt to provide an immediate answer; the other, human-powered, that drew on remote question answerers who supplied responses. Those studies showed that humans could provide better query responses than the AI-powered version in a range of diverse cases, but that their response time (many minutes, typically) was inadequate.

In this new work, we were interested in combining these approaches through two iterations of the TalkBack system. Deploying both versions longitudinally in Dharavi, a large informal settlement in Mumbai, India, we provided answers to two questions: i) would users request human help if the machine answer was insufficient?; and ii) could previously provided human answers satisfy new queries and at the same time reduce the response time?

The first study provided an answer to the initial question: where given a choice, if not satisfied with the automatic machine answer, users requested a human response in 71 % of cases. In the second study, we demonstrated the potential of immediately providing a human answer by reusing responses to previous questions. While users only rated such results positively for 34 % of queries, we note that this was while using a relatively small corpus of previous answers and a simple retrieval system. We would expect this rate to rise with larger corpora and with further sophistication in terms of integrating advanced data science techniques and interaction design.

These results also point to the value of building AI systems that can benefit from the creativity [16], general knowledge [15] and ability to understand and tell [19] of in-the-loop task workers. Recently, there have been widespread concerns about the use of humans to review input to improve voice assistant performance. Through the prototypes developed and user studies conducted here, we have shown that it is possible, practical, and beneficial to transparently combine machine and human intelligence. In our systems, it was clear to the user when and how a human's help would be enlisted; and, the in-the-loop query responders were not 'ghost workers' [15] but given a clear sense that they were directly engaging with community members who were requesting their help. The clear recommendation from this work, then, is to consider how such a design ethic might improve other AI-infused products [1] that rely on such human-in-the-loop work.

ACKNOWLEDGEMENTS

We would like to thank Manik, Shashank, Aparna, Manjiri and Nayeem for their help with the deployments described here. This work was supported by EPSRC grants EP/M00421X/1, EP/M022722/1 and EP/R511614/1.

REFERENCES

[1] Saleema Amershi, Dan Weld, Mihaela Vorvoreanu, Adam Fourney, Besmira Nushi, Penny Collisson, Jina Suh, Shamsi Iqbal, Paul N. Bennett, Kori Inkpen, Jaime Teevan, Ruth Kikin-Gil, and Eric Horvitz. 2019. Guidelines for Human-AI Interaction. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (CHI '19)*. ACM, New York, NY, USA, Article 3, 13 pages. DOI: https://doi.org/10.1145/3290605.3300233

[2] Apache. 2019. Apache Solr Reference Guide 8.1 | Language Analysis. Retrieved 2019-09-20 from https://lucene.apache.org/solr/guide/8_1/language-analysis.html#hindi

[3] BBC News. 2019a. Mobile data: Why India has the world's cheapest. Retrieved 2019-12-26 from https://www.bbc.com/news/world-asia-india-47537201

[4] BBC News. 2019b. Smart speaker recordings reviewed by humans. Retrieved 2019-09-20 from https://www.bbc.com/news/technology-47893082

[5] Apoorva Bhalla. 2018. An Exploratory Study Understanding the Appropriated Use of Voice-based Search and Assistants. In *Proceedings of the 9th Indian Conference on Human Computer Interaction (IndiaHCI' 18)*. ACM, New York, NY, USA, 90–94. DOI: https://doi.org/10.1145/3297121.3297136

[6] Eli Blevis and Eric Stolterman. 2006. Regarding software as a material of design. In *Wonderground Design Research Society Conference*. Design Research Society, London, UK, Article 68, 18 pages. https://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.364.9981

[7] Joseph Campana (Ed.). 2013. *Dharavi: the city within*. Harper Collins India, New Delhi, India.

[8] Matt Day, Giles Turner, and Natalia Drozdiak. 2019. Amazon Workers Are Listening to What You Tell Alexa. Retrieved 2019-09-01 from https://www.bloomberg.com/news/articles/2019-04-10/is-anyone-listening-to-you-on-alexa-a-global-team-reviews-audio

[9] Nicola Dell, Vidya Vaidyanathan, Indrani Medhi, Edward Cutrell, and William Thies. 2012. "Yours is Better!": Participant Response Bias in HCI. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '12)*. ACM, New York, NY, USA, 1321–1330. DOI: https://doi.org/10.1145/2207676.2208589

[10] Devanuj and Anirudha Joshi. 2013. Technology Adoption by 'Emergent' Users: The User-usage Model. In *Proceedings of the 11th Asia Pacific Conference on Computer Human Interaction (APCHI '13)*. ACM, New York, NY, USA, 28–38. DOI: https://doi.org/10.1145/2525194.2525209

[11] Flanders News. 2019. Google employees are eavesdropping, even in your living room. Retrieved 2019-12-23 from https://vrtnws.be/p.DxW6YZ49y

[12] Google. 2019a. Cloud Speech-to-Text. Retrieved 2019-09-20 from https://cloud.google.com/speech-to-text/

[13] Google. 2019b. Cloud Translation. Retrieved 2019-09-20 from https://cloud.google.com/translate/

[14] Google. 2019c. Google Assistant SDK | Google Developers. Retrieved 2019-09-20 from https://developers.google.com/assistant/sdk

[15] Mary L. Gray and Siddharth Suri. 2019. *Ghost work: how to stop Silicon Valley from building a new global underclass*. Houghton Mifflin Harcourt, Boston, MA, USA.

[16] Elizabeth Hallam and Tim Ingold (Eds.). 2007. *Creativity and cultural improvisation*. Berg, New York, NY, USA.

[17] Richard H. R. Harper. 2019. The Role of HCI in the Age of AI. *International Journal of Human–Computer Interaction* 35, 15 (Sept. 2019), 1331–1344. DOI: https://doi.org/10.1080/10447318.2019.1631527

[18] Lucy Hattersley. 2017. AIY Voice Essentials. Retrieved 2018-03-05 from https://www.raspberrypi.org/magpi/issues/essentials-aiy-v1/

[19] Tim Ingold. 2013. *Making: anthropology, archaeology, art and architecture*. Routledge, London, UK.

[20] Ece Kamar. 2016. Directions in Hybrid Intelligence: Complementing AI Systems with Human Intelligence. In *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence (IJCAI-16)*. AAAI Press, New York, NY, USA, 4070–4073.

[21] Vivek Kant and Anirudha Joshi. 2018. Challenges In Supporting The Emergent User. In *Proceedings of the 9th Indian Conference on Human Computer Interaction (IndiaHCI '18)*. ACM Press, Bangalore, India, 67–70. DOI: https://doi.org/10.1145/3297121.3297131

[22] Josephine Lau, Benjamin Zimmerman, and Florian Schaub. 2018. Alexa, Are You Listening?: Privacy Perceptions, Concerns and Privacy-seeking Behaviors with Smart Speakers. *Proc. ACM Hum.-Comput. Interact.* 2, CSCW, Article 102 (Nov. 2018), 31 pages. DOI: https://doi.org/10.1145/3274371

[23] Ewa Luger and Abigail Sellen. 2016. "Like Having a Really Bad PA": The Gulf Between User Expectation and Experience of Conversational Agents. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems (CHI '16)*. ACM, New York, NY, USA, 5286–5297. DOI: https://doi.org/10.1145/2858036.2858288

[24] Madhumita Murgia. 2019. AI's rise means work for army of data labellers. Financial Times. https://www.ft.com/content/56dde36c-aa40-11e9-984c-fac8325aaa04

[25] Sheela Patel, Jockin Arputham, Sundar Burra, and Katia Savchuk. 2009. Getting the information base for Dharavi's redevelopment. *Environment and Urbanization* 21, 1 (April 2009), 241–251. DOI: https://doi.org/10.1177/0956247809103023

[26] Jennifer Pearson, Simon Robinson, Thomas Reitmaier, Matt Jones, Shashank Ahire, Anirudha Joshi, Deepak Sahoo, Nimish Maravi, and Bhakti Bhikne. 2019a. StreetWise: Smart Speakers vs Human Help in Public Slum Settings. In *CHI Conference on Human Factors in Computing Systems Proceedings (CHI '19)*. ACM, New York, NY, USA, Article 96, 13 pages. DOI: https://doi.org/10.1145/3290605.3300326

[27] Jennifer Pearson, Simon Robinson, Thomas Reitmaier, Matt Jones, and Anirudha Joshi. 2019b. Diversifying Future-Making Through Itinerative Design. *ACM Trans. Comput.-Hum. Interact.* 26, 5, Article 33 (July 2019), 21 pages. DOI: https://doi.org/10.1145/3341727

[28] Martin Porcheron, Joel E. Fischer, Stuart Reeves, and Sarah Sharples. 2018. Voice Interfaces in Everyday Life. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (CHI '18)*. ACM, New York, NY, USA, Article 640, 12 pages. DOI: https://doi.org/10.1145/3173574.3174214

[29] Agha Ali Raza, Rajat Kulshreshtha, Spandana Gella, Sean Blagsvedt, Maya Chandrasekaran, Bhiksha Raj, and Roni Rosenfeld. 2016. Viral Spread via Entertainment and Voice-Messaging Among Telephone Users in India. In *Proceedings of the Eighth International Conference on Information and Communication Technologies and Development (ICTD '16)*. ACM, New York, NY, USA, Article 1, 10 pages. DOI: https://doi.org/10.1145/2909609.2909669

[30] Simon Robinson, Jennifer Pearson, Shashank Ahire, Rini Ahirwar, Bhakti Bhikne, Nimish Maravi, and Matt Jones. 2018. Revisiting "Hole in the Wall" Computing: Private Smart Speakers and Public Slum Settings. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (CHI '18)*. ACM, New York, NY, USA, Article 498, 11 pages. DOI: https://doi.org/10.1145/3173574.3174072

[31] Nithya Sambasivan, Ed Cutrell, Kentaro Toyama, and Bonnie Nardi. 2010. Intermediated Technology Use in Developing Communities. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '10)*. ACM, New York, NY, USA, 2583–2592. DOI: https://doi.org/10.1145/1753326.1753718

[32] Christine Satchell and Paul Dourish. 2009. Beyond the User: Use and Non-use in HCI. In *Proceedings of the 21st Annual Conference of the Australian Computer-Human Interaction Special Interest Group: Design: Open 24/7 (OZCHI '09)*. ACM, New York, NY, USA, 9–16. DOI: https://doi.org/10.1145/1738826.1738829

[33] Swarachakra Team. 2019. Swarachakra Hindi Keyboard. Retrieved 2019-12-26 from https://play.google.com/store/apps/details?id=iit.android.swarachakra

[34] Karen Taylor and Andrew C.K. Wiedlea. 2007. In Defense of Ugliness: The Role of Technical Presence in Critical Infrastructure System Endurance. In *2007 IEEE International Symposium on Technology and Society*. IEEE, New York, NY, USA, 1–6. DOI: https://doi.org/10.1109/ISTAS.2007.4362236

[35] Aditya Vashistha, Edward Cutrell, Gaetano Borriello, and William Thies. 2015. Sangeet Swara: A Community-Moderated Voice Forum in Rural India. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems (CHI '15)*. ACM, New York, NY, USA, 417–426. DOI: https://doi.org/10.1145/2702123.2702191

[36] Marion Walton, Vera Vukovic, and Gary Marsden. 2002. 'Visual literacy' as challenge to the internationalisation of interfaces: a study of South African student web users. In *CHI '02 Extended Abstracts on Human Factors in Computing Systems (CHI '02)*. ACM Press, Minneapolis, Minnesota, USA, 530–531. DOI: https://doi.org/10.1145/506443.506465