Academic Client Project, Fall 2017

USABILITY TESTING WITH VOICE USER INTERFACE

PROJECT OVERVIEW

- An esteemed insurance company has recently established a presence on the Alexa Skill platform, accessible through Amazon Echo devices. One of the features they have made available is the ability to initiate a homeowner's claim through their First Notice of Loss (FNOL) process.
- In order to test the overall efficacy of the FNOL process and to gauge users' level of trust and comfort using the process, usability tests were conducted
- The insurance company recognizes that users may not be "ready" for filing insurance claims through Alexa, but they hope to improve the process so that customers will consider their Alexa Skill presence reliable, easy to use, and a positive experience overall once adopted by users.

TESTING GOALS

To assess:

- **Overall efficacy**: Can users get through the FNOL process via Alexa Skill?
- **Concept validation and usability:** Are users confident and comfortable with using the Alexa Skill to file a claim?
- **Understand the concept of trust in voice user interfaces:** How can the Alexa Skill be enhanced in order to build trust?

THE CHALLENGE

- Once activated, the device registers and responds to every word the user says, so we could not have participants telling us their thoughts at every step during the process like in a Think Aloud method. We also couldn't have participants do a Cognitive Walkthrough during the process even if we deactivated the device because that would break the user's process flow which would hamper their ability to complete the task
- We thought of recording the interaction and replaying it to the participant, but from our own experience conducting the expert reviews we realized that listening to a recording of your own voice can be jarring and quite cringe worthy.
- We did not want to put our participants through any discomfort during our testing session, so we came up with **Voice Visualization technique** and **Retrospective Cognitive Walkthrough**

THE SOLUTION

- After the participant completes their task we asked them to draw a representation of their interaction with Alexa. This gave participants some time to decompress and channel their emotions into a drawing, a welcome break after having a potentially stressful conversation with Alexa
- The moderator then used the participant's drawing to take them through a "retrospective cognitive walkthrough" wherein the participant narrated their experience and their feelings about the interaction without increasing their cognitive load by having to accurately express their emotions solely from memory
- These techniques were used to discuss and get accurate feedback from the participants that was used to determine the overall usability of the process

****A deeper overview of Voice Visualization is added to the after the project slides*

METHODOLOGY

METHODOLOGY:

TESTING OVERVIEW

- We recruited 8 participants across a range of experience with both – filing an insurance and proficiency in voice user interfaces like Siri, Google Home, etc.
- Tests were conducted at Bentley's User Experience Centre with a moderator and note taker present in the session
- Each session consisted of:
  - Background questions to validate the participant's answers on the recruitment screener and to assess their level of expertise with insurance filing and voice user interfaces
  - Scenario based task to assess their ease of following through the process
  - Voice Visualization and Retrospective Cognitive walkthrough to assess their emotional journey during their interaction with Alexa
  - Discussion and feedback that concluded in a survey to triangulate qualitative data with quantitative data

METHODOLOGY:

TASK SCENARIO

In order to create a real life set up so that participants behave as closely as they would in a natural setting, we gave them a scenario based task:

The scenario: "You come back from your trip to find that your house is broken into. There is no sign of damage, but all your electronics - your TV, laptop, music system, etc. are gone. After spending several hours filing a report with the police, you now want to file a claim with Liberty Mutual using Alexa Skill."

We provided a proxy phone number to ensure the claim was not filed, but otherwise asked participants to use their own information or whatever responses felt natural.

METHODOLOGY:

DATA COLLECTED

Quantitative Measures:

- Time to complete task
- Comprehension errors
- Repetition errors
- Assists needed
- Survey ratings after task

Qualitative Measures:

- User background and experience
- Impressions of Alexa and the process
- Image of Alexa
- General feedback and suggestions

FINDINGS & RECOMMENDATIONS

VOICE VISUALIZATION FINDINGS

We found strong relationships between the voice visualization, participant's feedback, and the survey results:

- Participants' ratings of Alexa's friendliness were consistent with their drawing type: Users who considered Alexa simply a device rated her less friendly which correlated to the relatively higher errors encountered, while users who drew a person rated her as more friendly and experienced a fairly smooth interaction

"Alexa is happy to help" - P1

- Participant is a proficient voice interface user
- Participant completed the process relatively smoothly and encountered the least errors

"It still just seems like a computer…I put bows and eyes on it to make it more human" – P6

- Participant is not a proficient voice interface user
- Participant felt that the process was futuristic, but would prefer a human element

"… I wanted to create this idea of a labyrinth, in that you're finding your way through this… - P4

- Participant does not use voice interfaces, but has filed insurance for theft before
- Participant was lost in the process and required assistance to continue

TRIANGULATION OF FINDINGS

- Based on the qualitative and quantitative data we collected from the usability test sessions, we triangulated the results to represent the severity of the finding:

We assigned severity rating to the heuristics violated (determined during the expert review):

Category	Severity 1	Severity 2	Severity 3
Aesthetics	2	3	
Consistency		3	
Error prevention and recovery		9	4
Flexibility and efficiency	2	2	1
Match between system and real world	6	3	5
User control and freedom	5	7	6
Visibility of system status	4	1	
System capabilities	2		4
Prosody	3	4	3
Persona		3	

We then multiplied the rating with the frequency of errors to determine the "score":

Category	Score
Aesthetics	8
Consistency	6
Error prevention and recovery	30
Flexibility and efficiency	9
Match between system and real world	27
User control and freedom	37
Visibility of system status	6
System capabilities	14
Prosody	20
Persona	6

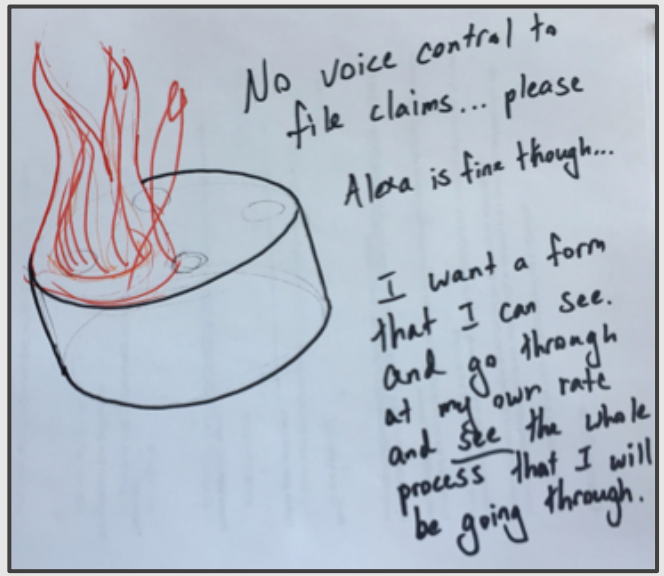
- Based on the score we organized the problems we identified into four categories to present to the client in order of priority:
  - **Error Prevention/Recovery:** The user's ability to successfully navigate the FNOL process.
  - **Flow:** The pacing and structure of the script and the FNOL process as a whole.
  - **Comprehension:** The system's ability to understand a user's input.
  - **Language/Prosody:** The system's use and articulation of language.

TRIANGULATION OF FINDINGS (CONTD.)

- Most of the severe problems we found had to do with a user's ability to successfully navigate the process on their terms. Lack of confirmation and an inability to adjust prior answers results in users having to exit and restart the process if the system does not understand them correctly.
- Most of the moderate and minor problems had to do with inconsistencies in the language used in the script or the system's ability to understand common conversational responses.

These findings were consistent with our expert review findings, however it was interesting to see the participants' reactions to these issues which was very well described by their voice visualizations –

"Alexa meant well but didn't quite get it" – P3



"I felt like I was just at her mercy" - Pilot

FINDINGS & RECOMMENDATIONS PRESENTED TO CLIENT

Finding 1: Damage vs. Loss Language

Severity: 3 - Major usability finding (results in task failure; causes significant delays/irritation for users)

Issue: The system uses "*damage*" and "*loss*" interchangeably (and occasionally in quick succession) throughout the process, despite a difference in meaning. Participants found the "*where did the damage occur*" and "*how many rooms were damaged*" questions particularly difficult when filing a claim related to stolen items.

Recommendations: Instead of using "*damage*," in the first question, Alexa could say, "*What kind of claim would you like to file?*" For later questions, the use of "*damage*" vs. "*loss*" should be dependent on which category was selected at the beginning.

"If I was talking with a person, I could say, *'Oh no no - there was no damage.'*"

-P2

"I felt I went down a little bit of a rabbit hole because somehow we were getting into damage. It wasn't damage it was theft."

-P7

FINDINGS & RECOMMENDATIONS PRESENTED TO CLIENT

Finding 2: Phrasing of Claims Categories

Severity: 2 - Moderate usability finding (can result in task failure for some users; causes moderate delays/irritation for users)

Issue: Early in the process, the system asks about the type of claim a user is interested in and says, "*You can choose from fire, plumbing, weather,* ***stolen or missing items or vandalism****, or other.*" Users were uncertain how to respond to this question, with predictable responses resulting in a lack of system recognition and in one case task failure.

Recommendations: Elicit the proper user response by incorporating more pronounced spacing between the articulation of claim options, and program a broader range of responses into the recognized grammar.

"I thought I could do shorthand. She wanted to hear this whole string of things that she clumped together as one thing, she wanted the whole thing . . . because I just said "[*stolen*]" and she didn't get it."

-Pilot

FINDINGS & RECOMMENDATIONS PRESENTED TO CLIENT

Finding 3: Process "Visibility"

Severity: 2 - Moderate usability finding (can result in task failure for some users; causes moderate delays/irritation for users)

Issue: Participants expressed discomfort in being unable to tell what questions were coming, or what information they would have to have on hand throughout the process. Participants also expressed frustration when they were unable to change their answers or correct Alexa. Several participants expected more questions, mentioning that it was not until the conclusion that they realized the full process would not be completed using the Skill, but that a representative would follow up with them.

Recommendations: The introduction to the script could provide a brief overview of the process and alert users early on of system capabilities and that it is okay if they don't have all of their information handy.

"I like to *see* the process from start to finish, and actually *see* where I'm going to go."
- Pilot

"So that the person is not walking in the dark, and feeling kinda like, out of control."
- P4

FINDINGS & RECOMMENDATIONS PRESENTED TO CLIENT

Finding 4: Lack of Assurance & Assistance

Severity: 2 - Moderate usability finding (can result in task failure for some users; causes moderate delays/ irritation for users)

Issue: Participants expressed discomfort with the conclusion of the FNOL process. Several participants mentioned the pacing was hard to understand as far as what the next steps in the process were, and a few missed that a representative would be reaching out to them. Many participants expressed a lack of confidence in their claim being filed correctly, or an inability to adjust the claim after completing the process.

Recommendations: A sound clip could be used to break up the question phase from the conclusion. The pacing of the conclusion could be improved by increasing the pauses between each chunk of information. The system could provide a claim number or send an email to a user after the process is completed, which they could reference online or by phone in order to check the status of/adjust their claim.

"Tomorrow if I call up, I have no idea if the claim was logged in or not. . . . Some confirmation would make it a much better experience."

- P1

"A claim number to me is absolutely key. You always worry you call and say, '*I'm* [*X*],' and they go, '*I don't see you.*'"

-P7

RECOMMENDATIONS

- Reduce the “corn maze” effect by providing a clear overview of the FNOL process at the beginning.
- Clearly communicate system capabilities and provide users with the capacity to revise answers and return to a main menu.
- Avoid the frustrations of the “Rabbit Hole” by creating specific dialogue flows that differentiate between property damage and loss.
- Reproduce everyday conversational styles at the level of prosody (intonation, spacing) and politeness (“*Please say, ‘Yes or No’*”). Increase the range of acceptable answers by anticipating common response alternatives.
- Implement a chime to break between the question phase and conclusion. Break up the conclusion into smaller chunks to increase the clarity of information.
- Provide a more definitive confirmation of the filing process through the use of confirmational cues (such as claim numbers or email affirmations).

THANK YOU

VOICE VISUALIZATION: A methodology to test voice user interfaces

ABOUT VOICE VISUALIZATION:

To get accurate and insightful data while testing a voice user interface we could not use testing best practices like the Think Aloud Method or Cognitive Walkthrough due to restrictions the voice interface posed. We came up with a method called **Voice Visualization**, wherein after the participant completed the task we set up for the study we pulled out crayons, markers and color pencils and asked the participant – **"Based on your interaction draw a picture of Alexa".**

By asking participants to draw, our intention was to provide them with an outlet to decompress after completing what was a potentially stressful task and use that time to transfer aspects of their experience into a physical artifact – a drawing - that the moderator could then use as a launchpad for discussion.

This method solved our main feedback concern – the level of accuracy of participants' recall of their experience. Given that there is no tangible representation of the interaction like a web or app page that the participant can refer to in order to provide feedback, we didn't want to resort to their memory of the interaction which we know is usually unreliable. Since It is difficult for people to give accurate accounts of their experience after time has passed, their feedback often results in participants making up feelings or projecting impressions. Thus, using the participant's sketch along with observations made by the moderator and note taker to transport the participant to a point in the interaction for further elaboration was very effective.

The voice visualization method proved to be an effective tool to dig deeper into the participant's experience, and look for non-verbal cues that were valuable in assessing the impact of the interaction on the participant. Our hope was that by using drawings, we might uncover emotional aspects that would not be covered by verbal communication alone.

We believed that using drawings in this way would also serve as an ice breaker and allow us to develop a rapport with our participants by letting them control and guide the discussion of their own experiences, rather than the other way around.

HOW IT PANNED OUT

Just as we hoped, participants put their feelings to paper and expressed emotions which we could turn into actionable insight that drove the recommendations we made to the client. The following examples illustrate the success of using the Voice Visualization method -

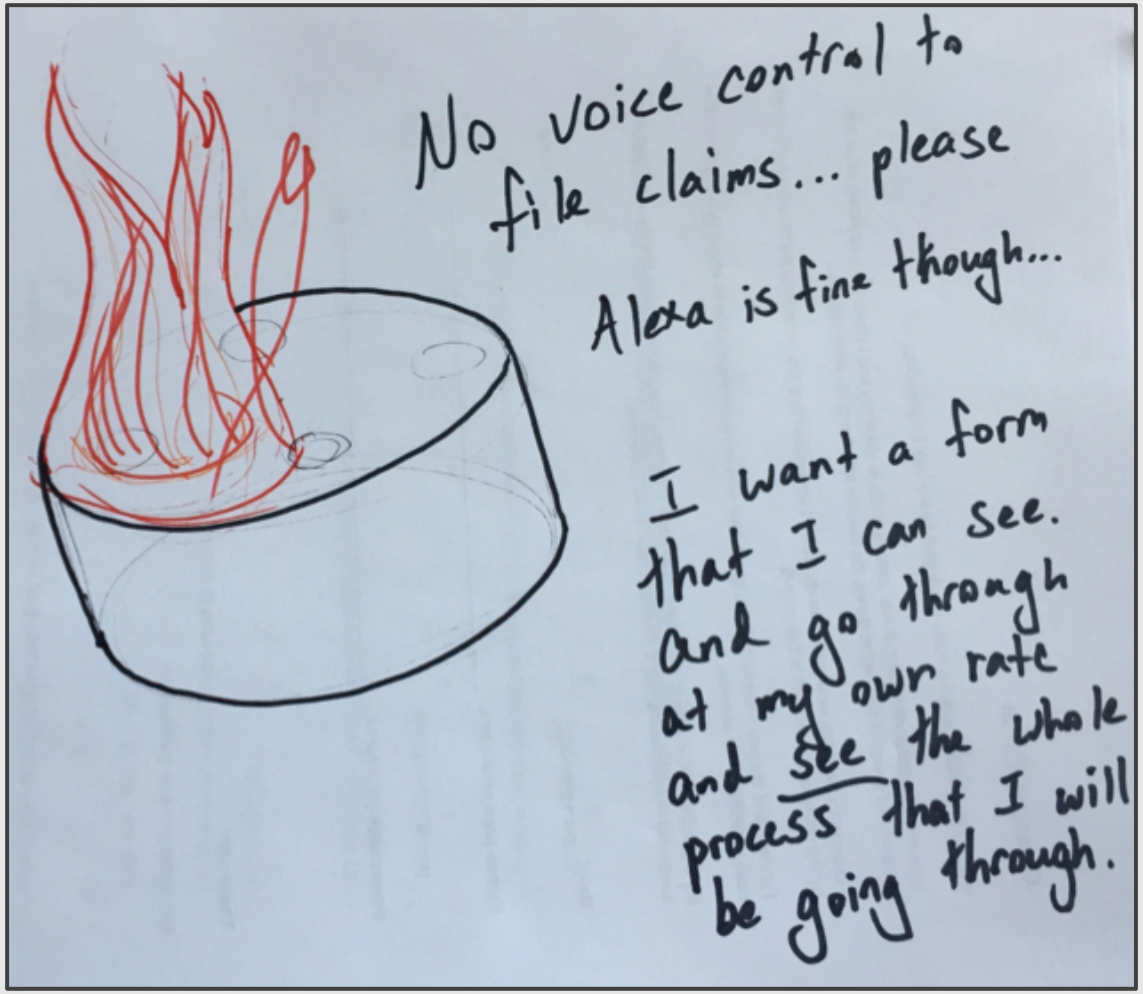
EXAMPLE 1

“Alexa is happy to help” - P1

Some participants drew Alexa through human representations. This participant drew a smiling face and described feeling trust, confidence, and friendliness from the experience.

HOW IT PANNED OUT

EXAMPLE 2

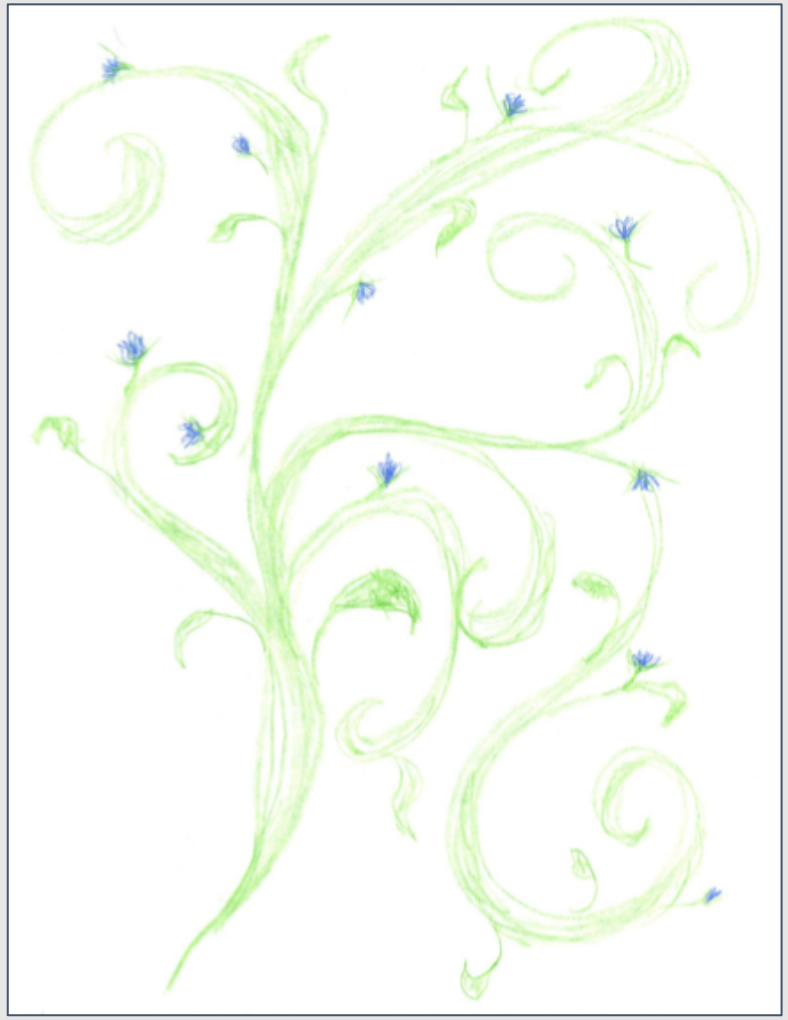


"I felt like I was just at her mercy" - Pilot

This participant was not so thrilled with the interaction. Still, this is exactly what we were looking for - drawings that could be used as a conduit for a participant's emotional journey through the experience

HOW IT PANNED OUT

EXAMPLE 3



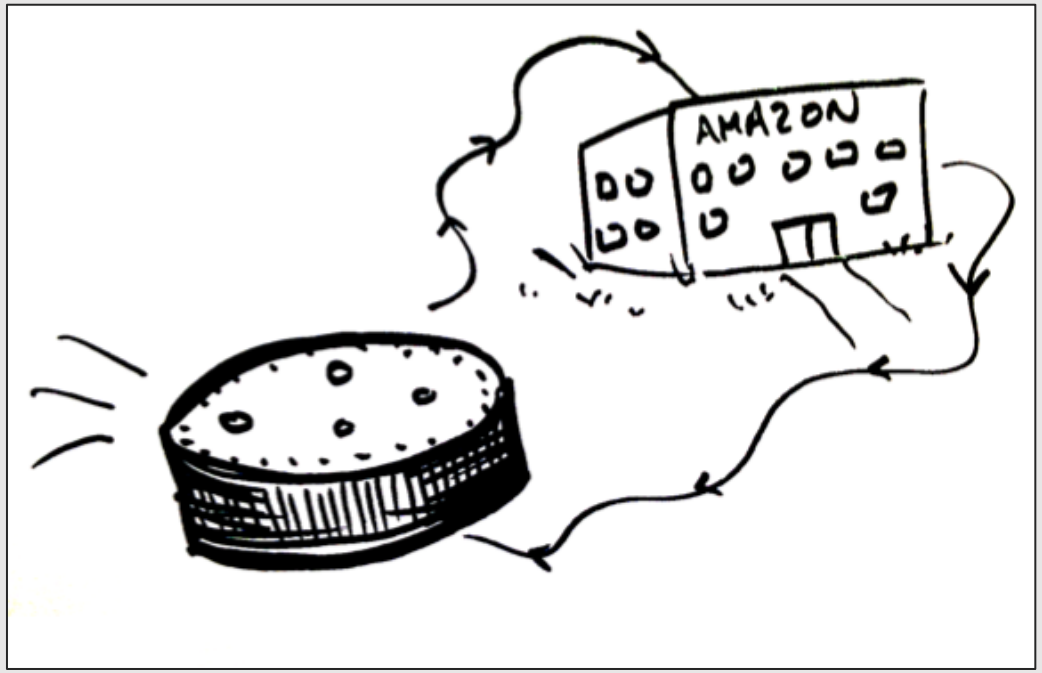
"… I wanted to create this idea of a labyrinth, in that you're finding your way through this…" - P4

- Instead of representing Alexa as something human or something technological, this participant drew a labyrinth. She described feelings of genuine anxiety about not knowing what the system would ask and what kind of information she might need to have on hand. She felt like she was being chased through a maze, quickly having to decide which way to proceed.
- We were able to translate these feelings into several recommendations for our client, including a brief overview of the system at the beginning of the interaction, as well as adding encouraging language throughout the dialogue to inform users they are on the right track.

This drawing and the discussion that followed was not something we were anticipating at all, and we may have missed this data had we gone with other methods.

HOW IT PANNED OUT

EXAMPLE 4



- The participant described this drawing as a feedback loop between the Echo Dot we were testing with and Amazon's headquarters. But as this participant walked us through their drawing, we noticed something very important was missing from the feedback loop: our client!
- This participant was visualizing their interaction as a conversation between themselves and Amazon, rather than a conversation between themselves and our client. If you were to take out your iPhone and the Target app you wouldn't think of that interaction as between yourself and Apple. So why was this participant thinking differently?
- Aside from the novelty of voice user interfaces, we think there's something else at play here. On a web page or an app screen, companies can prominently display their names and logos to remind customers of where they are – a luxury you don't have with a voice interface. We used this insight to recommend that our client incorporate branding chimes and language customers would recognize from their other company channels in order to remind them where they were

Again, this is the kind of non verbal cues that we would have completely missed if we had not had our participants draw out their experiences

TAKEAWAYS

- Voice Visualization can be used to test any voice interaction
- The true value of this tool is not in the quality of the drawing, but in the discussion that follows, where participants explain their drawings
- Voice Visualization provided us a path for exploring feelings and emotion that could be hard to express verbally
- It uncovered data we otherwise wouldn't have known to ask about, especially for an experience as abstract as voice
- When paired with other tools, it created a triangulation of study data that was extremely helpful in developing actionable insights and providing recommendations to our client
- Voice Visualization is a good tool to deliver data that determines the proof of concept – one of our clients' main goals for the study

NEXT STEPS

My team – Cameron Cross, Amanda Holmes and I are presenting a 10 minute student talk at the UXPA Boston 2018 on the 10th of May on voice visualization. We hope to have inspired you, and we encourage you to adapt, expand, and tweak this method as you approach testing with voice.

Keep us in the loop if you discover something exciting or develop a more effective method to test voice interfaces. We love being challenged and are always eager to learn.

Amanda Holmes
@aholmes8

Cameron Brown-Cross
@cameronbcross

Stuti Jhaveri
@veg_stew