Is My Siri the Same as Your Siri? An Exploration of Users' Mental Model of Virtual Personal Assistants, Implications for Trust

Nathan L. Tenhundfeld, Hannah M. Barr, Emily H. O'Hear, and Kristin Weger

***Abstract*—Virtual personal assistants (VPAs) are becoming so widely available that considerations are being made as to whether to begin including them in self-driving vehicles. While research has been done exploring human interactions with single VPAs, there has been a little work exploring human interactions and human mental models with interconnected systems. As companies like Amazon consider whether to integrate Alexa in their self-driving car, research needs to be done to explore whether individual's mental model of these systems is of a single system or if every embodiment of the VPA (e.g., echo) represents a different VPA. Knowing this will allow researchers and practitioners to apply existing models of trust, and predict whether high trust in the Siri that exists in an iPhone will carry over into high (and potentially miscalibrated) trust in Apple's Siri-directed self-driving vehicle. Results indicate that there is not one consistent mental model that users have, and provides the framework for greater exploration into individual differences and the determinants that affect users' mental model.**

***Index Terms*—AI, mental models, trust, virtual personal assistants (VPA).**

I. Introduction

THE 2013 movie "her" depicts a future in which humans fall in love and date their virtual personal assistants (VPAs), assuming each one to be unique and distinct from every other VPA. With time it is revealed that these VPAs are in fact just different embodiments of the same system, leaving the humans to feel cheated and deceived. While an obviously fictional premise, it raises an interesting question: Do humans view the VPA in one device, as being the same as the VPA in another device? Said colloquially: does one view the Siri in his/her iPhone as being the same as the Siri in his/her iPad, or the Siri in a friend's iPhone?

Since Apple's release of Siri in April of 2011, the prevalence of VPAs has increased with competitors like Amazon's Alexa, Microsoft's Cortana, and Google Assistant joining the fold [1]. While the capabilities of such systems are everchanging, they are most commonly used for retrieving general information, such as the weather forecast, playing music/videos, and even telling jokes [2]. However, research suggests that people are reluctant in their adoption of such technologies [3]. Part of this reluctance may be low perceived usefulness or lower levels of enjoyment in the use of such a system, both of which can predict lower system usage [2], [4]. Higher satisfaction, on the other hand, predicts greater usage in a variety of settings [2]. These results are in accordance with the automation acceptance model which suggests that behavioral intention to use a system is a function of (among other things) perceived usefulness, perceived ease of use, trust, and attitude towards using the system [5], [6].

The design of these systems can directly affect the quality of experience one has. Research has shown the level of immersion [7], gender of the voice [8], [9], humanness [10], and even behavioral choices [11], can impact the interactions one has with a VPA system. Specific design choices can also be made to allow those with disabilities to interact with a system at all [12]. Design choices are well known to affect the trust of a user in a system, and the subsequent reliance on that system, through changing the mental model the user has of the system [13]–[17].

The question of whether users view the VPA in their phone as being the same as being the VPA in their computer, for instance, is ultimately a question of user mental models of interconnected systems. Understanding these mental models can have a substantial impact on the field's understanding of other cognitive constructs like trust. As such, this article briefly discuss mental models, interconnected systems, and trust.

A. *Mental Models*

This paper uses "mental models" to refer to a "cognitive representation of [a] system's internal mechanics" [18]–[20]. These internal mechanisms are at the heart of the present issue; do users hold that the VPA in his or her phone is simply a separate manifestation of the same entity that thus possesses the same internal mechanisms as the VPA in his or her tablet? In order for a user to believe that the VPA they interact with in both cases is the *same* entity, they would need to believe that they share a single internal mechanism. That belief of a shared internal mechanism

Manuscript received March 28, 2021; revised June 28, 2021 and August 4, 2021; accepted August 16, 2021. Date of publication September 15, 2021; date of current version May 17, 2022. This work was supported by a Grant from CRUSER N00244-20-2-0001 and the U.S. Army ARL W911NF1920209. This article was recommended by Associate Editor M. Monaro. (*Corresponding author: Nathan L. Tenhundfeld.*)

Nathan L. Tenhundfeld is with the Department of Psychology, University of Alabama in Huntsville, Huntsville, AL 35899 USA (e-mail: nlt0006@uah.edu).

Hannah M. Barr and Emily H. O'Hear are with N. L. Tenhundfeld's Laboratory, Department of Psychology, University of Alabama in Huntsville, Huntsville, AL 35899 USA (e-mail: hms0017@uah.edu; eho0002@uah.edu).

Kristin Weger is with the Department of Psychology, University of Alabama in Huntsville, Huntsville, AL 35899 USA (e-mail: kw0004@uah.edu).

Color versions of one or more figures in this article are available at https://doi.org/10.1109/THMS.2021.3107493.

Digital Object Identifier 10.1109/THMS.2021.3107493

would mean that violations of trust when interacting with VPA in one's phone should impact trust in the VPA across all devices. If, on the other hand, users believe that the VPA in their phone is distinct from the VPA in their tablet, they would need to believe there is not a shared internal mechanism. This is not to say that the mechanisms could not be identical between the two devices (i.e., one is a duplicate of the other), but rather that the two devices are not connected to the same singular mechanism.

Previous research has shown that individuals possess mental models about VPAs, even before they interact with them, which can dictate how users interact with them early on [21]. These early mental models can be a function of many things, such as the way in which media portrays systems [22], the physical embodiment of the system [23], extrapolations from knowing the system's language and origin [24], as well as recent experiences with similar technologies [25].

While users tend to develop a "sound" mental model of the way a system works over time, recent interactions with a system can influence their mental model as well [26]. For instance, those who win a game with AI assistance have better predictions of the AI's ability compared with those who lose a game [27]. When a system makes an error, the stochasticity and parsimony associated with the error boundary impacts users' mental model, and even team performance [28]. Similarly, when shortcomings of a system are not experienced, they are dropped from one's mental model, even when known to the user [29]. Despite these nuances, with sustained experience users seem to converge towards a more accurate mental model, regardless of their initial discrepancies [30].

As mentioned above, users' mental model of a system can have impacts on the interactions that s/he have with the systems. For example, (and of particular interest for this paper) one's mental model of a system can directly impact a user's trust in that system, subsequently affecting their use, disuse, and misuse of that system [31], [32]. While there has been substantial research examining users' mental model of individual systems, there has been surprisingly little examination into users' mental model of interconnected systems. This dearth of research on mental models of interconnected systems means there exists a shortcoming in the predictive capabilities of theories and models of constructs like trust. Understanding user mental models of interconnected systems can advance theory while helping practitioners better predict human-systems interactions, especially as these agents become more collaborative partners [33].

B. *Interconnected Systems*

The ability for iMac, MacBook, iPad, Apple TV, iPhone, and Apple Watches to all connect into a single ecosystem has made work and personal communications much more convenient for those that rely on the Apple ecosystem. However, Apple is far from the only technology ecosystem out there; similar interconnected systems are offered by Windows and Google, just to name a few. It is also important to note that while specific brands and products are mentioned throughout, the context of their use is to simply illustrate the real-world parallels. These conversational agents are represented by a host of physical embodiments these days ranging from headphones to cars, and from phones to televisions.

The question over how users view these interconnected systems is a function of the fact that these interconnected systems are more available than ever before. Understanding mental models of individual systems can provide meaningful insight into the interactions between a human and a machine. However, because the very nature of human interactions with VPAs is changing such that they are found not just in a single product from any given company, but rather across a host of platforms, means that the focus of mental models on a single device is insufficient, and instead there needs to be consideration of the totality of each ecosystem.

To date, there has been very little work examining the impacts of these technological ecosystems on user interactions with individual systems that comprise those ecosystems. For example, is the Alexa that one interfaces with through their own Echo dot, seen to be the same as the Alexa they may interact with through their neighbors' echo show? As companies begin to incorporate VPAs into a variety of platforms within a technological ecosystem (e.g., headphones, cellphones, computers, smart watches, smart speakers), it is imperative the field understand users' mental model of those VPAs. Mental models have a direct influence on user trust, and as such, understanding the mental models of interconnected systems is imperative for advancing theories of trust.

C. *Trust*

Trust is an attitudinal variable that encompasses a user's willingness to rely on a system in an environment characterized by vulnerability and uncertainty [16]. Research has shown that familiarity [34], [35], performance [36], [37], and design [38], [39] can all affect user trust in a system by altering user mental models of the system. Understanding the mental model that a user has about a system can directly inform predictions of that user's trust in the system [30].

It is important to understand user mental models of a system that can inform trust, as the level of trust a user has in a system can be highly predictive of his/her use of that system. Trust is considered an influential antecedent in predicting user reliance on a system [13]. One's expectations of a system is in part dependent on their initial learned trust (based on preexisting knowledge) of the system, along with the dynamic learned trust that can be influenced with repeated exposures and experience [13], [40]. Because these changes in trust can impact users' interactions, it is imperative to predict whether users of interconnected systems will apply a situation specific trust, in which they learn to trust the system in its various capacities [41], or whether they are more reliant upon a system-wide trust [42] in which failures of the system in one domain are generalized to other domains too.

The process of interacting with a system over time will result in calibration of trust [43]. This calibration of trust is imperative for highly effective human-machine teams. Trust that is too high for what is warranted by the capabilities of a system is referred to as "overtrust" and can result in detrimental behaviors. Some of these behaviors could be fatal, such as when supervising a

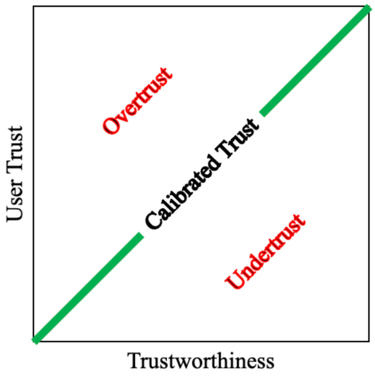

Fig. 1. Visual representation of what constitutes calibrated trust, versus regions of over- and undertrust. Figure is based on similar figure in [40].

self-driving car: complacency and failure to monitor the system [40], [44], [45]. Trust that is too low for what is warranted by the system's capabilities is referred to as "undertrust." Undertrust can result in disuse of a system or even a lack of adoption [34], [43] (see Fig. 1).

These sorts of miscalibrations in trust can be recalibrated, however. As people tend to employ social rules in their relationships with computers, actions, such as an apology or explanation from a virtual agent following a failure can repair undertrust [45]–[47], whereas overtrust can be dampened through conveying system limitations or lowering expectations [40]. With that said, interventions that aim to recalibrate miscalibrated trust can be costly and time consuming. It is therefore imperative to design systems which are accurately calibrated to the system's capabilities from the start, and thus do not require such interventions. This calibration requires the deliberate and effortful updating of one's mental model about the system [30]. However, questions remain: as one's mental model begins to calibrate to the capabilities of Siri in his/her phone, how does this change his/her trust in Siri in his/her Mac, HomePod, or Apple Watch? That set point of trust may be warranted for the tasks a user gives Siri when interfacing through a phone, but may be unwarranted for the capabilities of a HomePod or any other physical embodiment.

This concern obviously extends beyond commercially available products. For example, the National Aeronautics and Space Administration has indicated that Artificial Intelligence (AI) systems, such as Crew Interactive Mobile Companion (CIMON) and Astrobee, will play an ever-expanding role in the pursuit of greater spaceflight autonomy [45]–[48]. As such, will astronauts perceive these systems and their inclusion into different environments (i.e., spaceflight, operations of the International Space Station (ISS), etc.) as distinct systems, or as interconnected interfaces as part of a greater centralized intelligence (like say, Hal 9000 for example)? In other words, what are users' mental models of interconnected systems with which they regularly interact? Being able to answer this question is imperative for understanding trust in human-machine interactions, especially given the ever-increasing role of machines in collaborative and true teaming roles and the impact that mental models can have on team performance [33], [49], [50].

D. Our Study

Existing theory, which aims to predict user interactions with, and reliance on systems, does not provide a clear prediction of how interconnected AI systems will affect interactions with other devices within that same technological ecosystem, e.g., [13]. By understanding users' mental model of these interconnected systems, the field can start modifying existing theories to account for this new complexity imposed by technological advancements. Such theory refinement would have potentially wide-reaching effects on the design of systems, helping establish the necessary steps in order to prevent carry-over of mental models which would be miscalibrated to new devices within a certain interconnected ecosystem.

On the one hand, perhaps users view every embodiment of a device as being representative of a *different* system, and thus, for example, their trust in the Siri found in their phone would not carry over to the trust in the Siri found in a HomePod. Similarly, interactions with a VPA like CIMON that was in control during initial flight would have no bearing on expectations of CIMON that controlled some aspects of the international space station operation. On the other hand, perhaps users develop a mental model of interconnected systems as being all representative of the *same* system. This would mean that Siri's ability to dictate a text message on a phone, would affect the mental model of Siri, and subsequently the trust in any other Apple device. If this were true, this would have consequences for designers of systems; there could be carryover effects such that early experiences could result in either misuse or complete disuse of an otherwise unrelated device, leading to negative and possibly dangerous outcomes. It is the goal of this study to understand the nature of these mental models.

II. Methods

In order to better understand the nature of these mental models, a survey was conducted with users of various commercially available VPAs. This was done in order to ascertain their understanding and mental model of these systems. In addition, participants were probed to see whether changes in voice, gender, appearance, or physical embodiment would impact that mental model. Finally, participants were asked about their trust and liking of their VPAs.

A. Participants

Participants were pulled from the available participant pool at the University of Alabama in Huntsville. The survey was started 145 times, however 133 participants got to the open-ended questions. Each of these 133 participants finished the survey.

B. Materials and Procedure

Data were collected online with Qualtrics, with participants' own computers. The survey consisted of 13 questions that asked about an individual's experience with a system, as well as their mental model of the systems. Questions and response options are can be given in Table I. Note that [VPA] was replaced with the name of the VPA they used the most often. For example,

TABLE I
QUESTIONS ASKED OF PARTICIPANTS

	Question	Answer
1)	"Which of the following do you have (select all that apply, if none are applicable, leave blank)?"	Forced choice: [A device with Alexa, A device with Cortana, A device with Google Assistant, A device with Siri, Other (specify below)]
2)	"Which do you use more often?"	Forced choice: [Alexa, Cortana, Google Assistant, Siri, Other (auto populated from question 1 if given)]
3)	"How many devices do you currently have that use [VPA]?"	Forced choice: [0, 1, 2, 3, 4, 5, 6+]
4)	"How often do you use [VPA]?"	Forced choice: [Multiple times a day, Once a day, A couple times a week, A couple times a month, almost never]
5)	"Does the [VPA] you talk to know everything that the [VPA] others talk to, know? (Elaborate as much as possible)"	Open Ended
6)	"If you were to get a new device and talked to [VPA] there, would [VPA] know the same amount about you that it does when you talk to it through other devices? (Elaborate as much as possible)"	Open Ended
7)	"Is [VPA] a single automated assistant that everybody has access to, or does every device have a different [VPA]?"	Open Ended
8)	"If [VPA] makes a mistake when you are interacting with it through one device, is that reason to not trust [VPA] when interacting with it in a similar way through a different device?"	Open Ended
9)	"If [VPA]'s voice changed, would it still be the same [VPA]?"	Open Ended
10)	"If [VPA]'s gender changed, would it still be the same [VPA]?"	Open Ended
11)	"If [VPA]'s appearance changed, would it still be the same [VPA]?"	Open Ended
12)	How much do you trust [VPA] based on your last interaction?	Sliding scale: 0 (Not at All) – 100 (Completely)
13)	How much do you like [VPA] based on your last interaction?	Sliding scale: 0 (Not at All) – 100 (Completely)

[VPA] was replaced by the answer selected for Question 2.

question five would have read (to a participant who indicated they used Alexa most often): "Does the Alexa you talk to know everything that the Alexa others talk to, know? (elaborate as much as possible)." The questions pertaining to the mental models were developed in order to try and understand how users are mentally representing a system's internal mechanisms. These questions aimed to directly assess their beliefs regarding whether the VPA is a single system or a collection of interconnected systems (as in questions 7, 9, 10, and 11) as well as to determine their expectations (questions 5, 6, and 8) based on changes in the physical embodiment of a system.

C. Procedure

After signing up for the study, participants were provided a link to the survey. They were required to provide consent before continuing on. After the participant gave consent, s/he was given the survey, which asked about his/her experience with a system, as well as his/her mental model of the system (Appendix A). If participants did not indicate that they had at least one device with a VPA, the experiment ended, otherwise, the experiment continued. Following completion, participants were debriefed and thanked. Chi-squared tests were analyzed using version 1.5.2 of the "scipy.stats" library for Python, with the exception of the chi-squared tests for independence, which were calculated with software developed by [51], the rest of the analyses were run with SPSS.

D. Coding

For the open-ended response questions, inductive content analysis was applied [52]–[54]. First, two experimenters separately reviewed and created codes to apply to the responses. The codes for each question were then consolidated, simplified, and agreed upon by the two experimenters. Interrater reliability was calculated by the number of responses for which the experimenters had the exact same code list. If one experimenter had one code, and the other had that same code, but another one in addition, that counted against the reliability. Reliability was 0.87. In a second step, each experimenter independently coded the responses for each question according to the agreed upon coding scheme. After both experimenters had finished, they went back and discussed any answer for which there was disagreement on their codes. In a second validation cycle of the coding scheme, experimenters then provided a new agreed upon code for that answer. This allowed for the codes to be further consolidated providing more clarity for analyses. This left every open-ended response answer with a single representative code of the response.

III. RESULTS

There were 133 participants submitted to the following analyses. Devices with Siri were the most commonly owned ($N = 105$), followed by Cortana ($N = 40$), Alexa ($N = 35$), Google Assistant ($N = 28$), and then one participant responded "other" to indicate owning a device with Samsung's Bixby. Participants

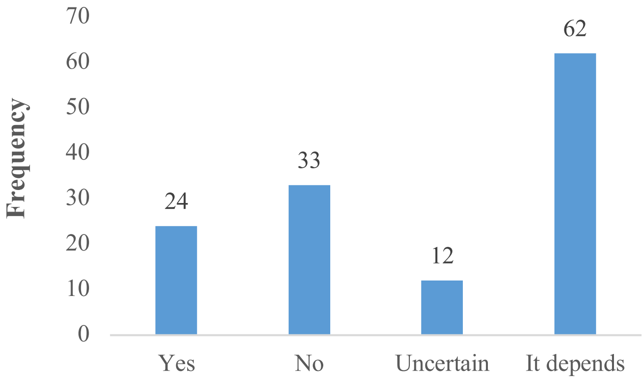

Fig. 2. Frequency of responses that were coded with each of the codes found on the *x*-axis.

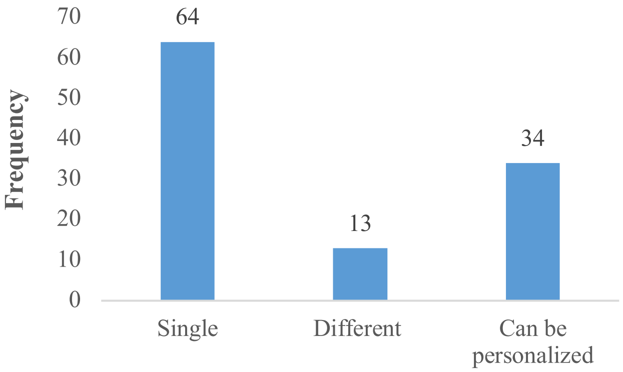

Fig. 3. Frequency of responses that were coded with each of the codes found on the *x*-axis.

reported using Siri most often ($N = 79$), followed by Alexa ($N = 20$), Google Assistant ($N = 18$), Cortana ($N = 3$), and Bixby ($N = 1$). An additional 12 participants did not select a device they used most frequently. Those who did not select a device were still included in subsequent analyses, provided they answered the question, even though it did not refer to a specific device. Note that "[VPA]" was replaced in the actual question by the system the participant reported using most frequently.[1]

For Section III-A–III-E nonparametric evaluations of frequency across the different codes were performed. These evaluated whether there were significant differences between how often a given response code was indicated for every question. Following, in Section III-F means comparisons are run in order to evaluate whether there are any observed differences in self-reported trust and liking of the VPAs by participants who identified using the respective VPAs most often.

A. Does the [VPA] You Talk to Know Everything That the [VPA] Others Talk to, Know?

For this question, participant responses were identified as either "yes,", "no," or "uncertain." There was no significant difference between the number of participants whose responses fell into each category (39, 47, and 27, respectively), $\chi^2(2, N = 113) = 5.38$, $p = .068$ (see Fig. 2). The frequency of responses did not vary as a function of the frequency of use of the system, $\chi^2(10, N = 112) = 10.59$, $p = .390$.

Examples of representative participant responses are below with their code category (in bold).

1) "Yes. I believe it's the same system, so they each know the same things." - Yes
2) "No, because the Siri I talk to has data on what I have asked it and what I'm interested in (by looking at data on my phone), while the Siri others talk to have the data of their respective user." – No
3) "I think so, I really only ask it the local weather and what time it is so I'm not sure beyond that." – Uncertain

B. If You Were to Get a New Device and Talked to [VPA] There, Would [VPA] Know the Same Amount About You That It Does When You Talk to It Through Other Devices?

For this question, four themes were identified that yielded four codes: "yes;" "no;" "uncertain;" or "it depends." Responses coded "it depends" involved participants saying things revolving around whether or not the new device would be linked with their existing account. There was a significant difference between the number of participants who gave each response, $\chi^2(3, N = 131) = 41.61$, $p < .001$ (see Fig. 2). The frequency of responses did not change as a function of the frequency of use of the system, $\chi^2(15, N = 130) = 15.27$, $p = .432$.

Examples of representative participant responses are below with their code category (in bold).

1) "Yes. She knows the same things through different devices." – Yes
2) "No I do not believe so. I think Alexa is a system that begins learning when you turn Alexa on and only knows what you tell it." – No
3) "I'm not sure! I am bad with technology and do not know if it carries over through phone number or device or account or what." – Uncertain
4) "If the new Alexa is registered under the same Amazon account then Alexa would know the same things about me, like the songs I like to play and my grocery list etc." – It depends

C. Is [VPA] a Single Automated Assistant That Everybody Has Access to, or Does Every Device Have a Different [VPA]?

For this question, responses were coded as either "same," "different," or "the device can be personalized." There was a significant difference between the number of participants whose responses fell into each category, $\chi^2(2, N = 111) = 35.51$, $p < .001$ (see Fig. 3). The frequency of responses did not change as

[1] Accordingly, this means that for those 12 participants who did not identify a VPA that they used most frequently, the questions did not identify a specific VPA. While this may have led to confusion on the part of the participants we decided (*a priori*) to keep these data for several reasons. First, we felt as though this could potentially provide more insight that we would otherwise be missing, given the nature of the free response questions. Second, because responses were coded, any participant who seemed unclear about how to answer was coded as such and their data was removed from that given analysis. Finally, we verified, after the fact, that excluding these participants would not change the results; even when excluded, all results and interpretations remained the same.

a function of the frequency of use, $\chi^2(10, N = 110) = 8.30$, $p = .600$.

Examples of representative participant responses are below with their code category (in bold).

1) "Siri is the same on all Apple devices." – Single
2) "Different but identical." – Different
3) "Everybody who has an Apple product has access to Siri, but each is altered in some way to cater towards the user's information." – Can be personalized

D. If [VPA] Makes a Mistake When You are Interacting With It Through One Device, Is That Reason to Not Trust [VPA] When Interacting With It in a Similar Way Through a Different Device?

For this question, responses were coded as either "yes," "no," "uncertain," or "depends on the mistake." There was a significant difference between the number of participants whose responses fell into each category (38, 76, 3, and 14 respectively), $\chi^2(3, N = 131) = 95.72$, $p < .001$. The frequency of responses did not change as a function of the frequency of use, $\chi^2(15, N = 130) = 16.78$, $p = .332$.

Examples of representative participant responses are below with their code category (in bold).

1) "Yes, Google assistant is a single automated assistant that everybody has access to so why would my trust differ from device to device." – Yes
2) "No because all computers and computer programs can glitch or mess up from time to time." – No
3) "I don't know. I don't typically ask Siri to do things that have any real grounds that would make me lose trust in it." – Uncertain
4) "Possibly. If it makes a mistake because of the access it has to common info, like news articles, I'd be more dubious about its answers." – It depends

E. If [VPA]'s Voice/Gender/Appearance Changed, Would It Still Be the Same [VPA]?

For these three questions, responses were coded as either "same," "different," or "uncertain." There was a significant difference between the number of participants whose responses fell into each category for voice change, $\chi^2(2, N = 129) = 197.16$, $p < .001$, gender, $\chi^2(2, N = 128) = 205.42$, $p < .001$, and appearance, $\chi^2(2, N = 131) = 200.26$, $p < .001$, (see Fig. 4).

Examples of representative participant responses are below with their code category (in bold):

1) "[Y]es, it would just have a different voice, nothing about the programming would be different, just a different voice output." – Same
2) "Personally I associate the voice that both siri and Alexa have with the devices that I use them on so yes, if the voices changed it would be different." – Different
3) "Intuitively, I would say yes; psychologically, I'm not entirely sure." - Uncertain

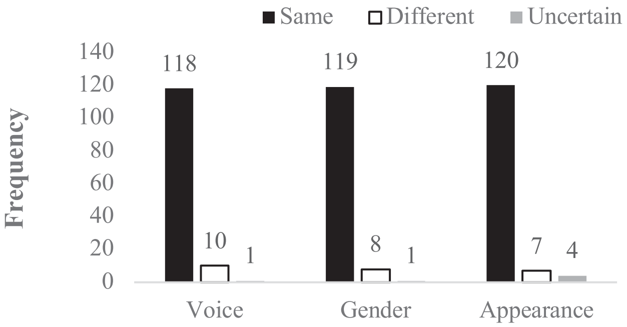

Fig. 4. Frequency of responses for each question (x-axis). The response code is differentiated between bars by their color.

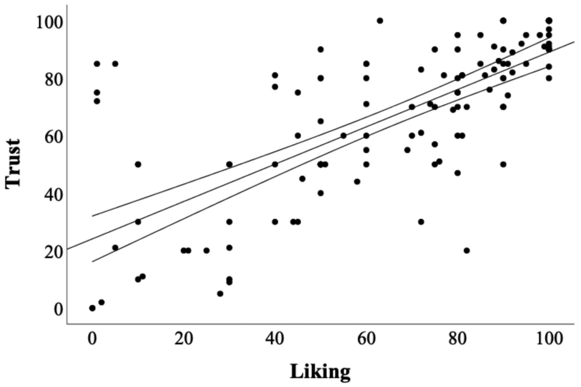

Fig. 5. Correlation between participants trust and liking scores given with a best fit line regressed upon the data. The bordering lines represent the 95% confidence interval.

F. Trust and Liking

Self-reported levels of trust ($M = 65.90$, SD $= 27.76$) was highly correlated with self-reported levels of liking ($M = 65.71$, SD $= 30.10$), $r(127) = .723$, and $p < .001$, see Fig. 5.

The trust and liking data for participants who indicated that they used Alexa, Google Assistant, or Siri most was then submitted to two separate one way ANOVAs. Data for Cortana and Bixby were excluded given the small number of participants who indicated using those most frequently (3 and 1, respectively). There was no significant difference for the amount that users trusted the system between the three VPAs, $F\,(2, 114) = 0.40$, $p = 0.674$. There was a significant difference, however, between the amount that users liked the system between the three VPAs, $F\,(2, 114) = 3.73$, $p = 0.027$. Trust and liking can be found for each of the three most commonly used VPAs in Fig. 6.

All the data were combined together again for the following analyses. There was no significant difference in self-reported trust between those in the different coded categories for the question "Does the [VPA] you talk to know everything that the [VPA] others talk to, know,?" $F(2, 108) = 2.51$, $p = .086$. Similarly there was no significant difference in self-reported trust between those in the different coded categories for the question "if [VPA] makes a mistake when you are interacting with it through one device, is that reason to not trust [VPA] when interacting with it in a similar way through a different device?," $F(3, 123) = 1.89$, $p = .135$.

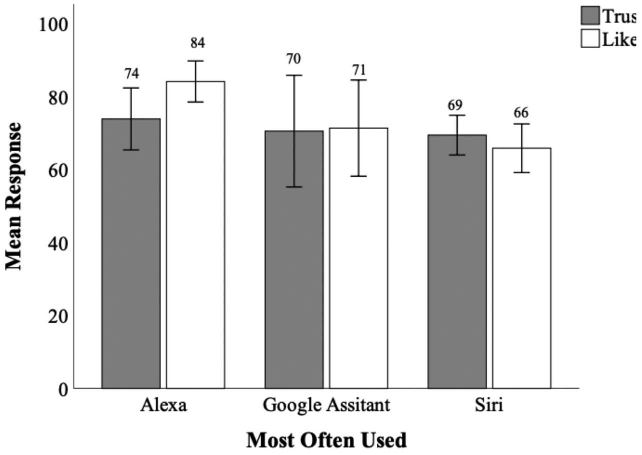

Fig. 6. Mean responses of trust and liking across participants with Alexa, Google Assistant, and Siri. Error bars represent the 95% confidence interval.

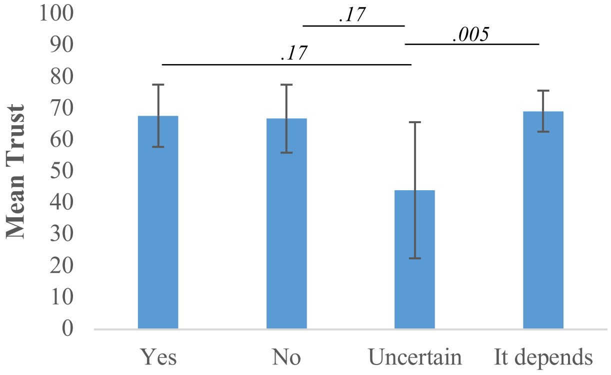

Fig. 7. Mean responses of trust across VPAs, broken down by coded response for the question "If you were to get a new device and talked to [VPA] there, would [VPA] know the same amount about you that it does when you talk to it through other devices?" Connecting lines and numbers above represent the significant differences and their corresponding p-values, as indicated by a post-hoc LSD test. Error bars represent the 95% confidence interval.

There were, however significant differences in self-reported trust for those in the different coded categories for the questions about getting a new device, $F(3, 123) = 2.80$, $p = .043$ (see Fig. 7), and about whether the VPAs are a single system or distinct, $F(2, 105) = 3.47$, $p = .035$ (see Fig. 8).

IV. Discussion

Humans are provided the opportunity to interact with highly advanced VPAs on a regular basis. As these VPAs begin integrating into more of society's daily routine, it is imperative the field understand the relationship between the human and the machine. Previous research has examined aspects of human interaction with VPAs [2], [4], [7], as well as user mental models of the systems [21]. However, to date there has been little exploration into users' mental model of the interconnected aspects of these systems.

While trust often relies on users' mental models, existing models of constructs like trust do not directly account for users' mental model of interconnected systems. Therefore, it is unclear whether interactions with VPAs affect trust in only the device through which the interaction takes place, or if the trust is generalized across all VPAs within that ecosystem. Such an understanding is needed in order to update models that serve to predict, among other things, user reliance on a system, i.e., [13]. Having this understanding extends beyond theory, however, and into considerations that need to be made by designers of such systems.

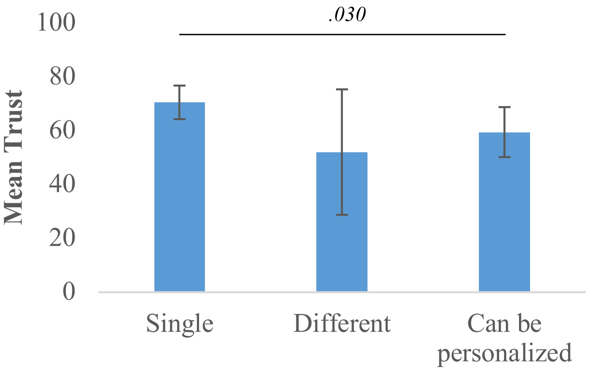

Fig. 8. Mean responses of trust across VPAs, broken down by coded response for the question "Is [VPA] a single automated assistant that everybody has access to, or does every device have a different [VPA]?" Connecting lines and numbers above represent the significant differences and their corresponding p-values, as indicated by a post-hoc LSD test. Error bars represent the 95% confidence interval.

If individuals view the Siri in their phone as being the same as the Siri in their iPad, or in someone else's phone, violations of trust in one domain could carry over into others. This means that Alexa's inability to play the right song when asked at home, could affect a user's trust in an Amazon self-driving vehicle [55]. If, on the other hand, individuals believe that they are different systems, then the user will learn to calibrate trust to each specific system. However, despite what their mental model of these systems is, previous research has suggested there is substantial user support for the VPAs which users have at home to be included into self-driving vehicles [56]. It stands to reason that this preference would extend beyond just self-driving vehicles.

Given the importance for researchers and practitioners alike, this article aimed to evaluate user's mental model of interconnected VPA systems. Data were collected on participants' self-reported views about the nature of interconnected VPA systems through a series of forced-choice and open-ended questions. Results paint somewhat of a mixed picture. There was no clear consensus among participants regarding whether the VPA they interact with is part of a single system that everyone interacts with, or whether their system is unique. Interestingly, some answers seemed to give contradictory results. For example, when asked about whether the VPA they would interact with in a new device would be the same (see Fig. 2), the pattern of results was different from the next question in which participants were asked whether the VPA was representative of a single system that everyone has access to, or if every device had a different version of the VPA (see Fig. 3). One possible explanation for this discrepancy is the question itself, however it is also possible that users have not thought about these issues in the way they are being presented. As such, discrepant views may be held. While there was no consensus on whether the VPAs are a single system, there seemed to be substantial agreement that no change in the voice, gender, or appearance of the system would indicate a change in how they thought of the system itself.

These results present an interesting glimpse into the complexity of users' mental model of interconnected systems. Whereas users seem to differentiate between the embodiment features of a VPA and the system's operations uniformly, there does not appear to be a cohesive understanding about how these systems work with regards to whether they are manifestations of the same central system, or if they are each distinct. What is more, this relationship does not appear to be affected by the frequency of use of the system. Given the lack of uniformity, there is a unique opportunity for greater exploration.

One avenue of exploration would be the role of individual differences, which could explain the results. Existing work exploring the role of individual differences in trust of automated systems [57], [58] has shown promise. One potential avenue for exploration of individual differences includes controlling for differences in users' perfect automation schema (PAS) that refers to beliefs about the performance of automation. Users' PAS has demonstrated the ability to predict differences and changes in trust through an assessment of users' "all-or-none" thinking [59]. There are also differences in users' propensity to trust which may explain results in the article, or may interact with user mental models in affecting user behavior [60]–[62]. Given the existing validated measures, and predictive abilities of PAS and propensity to trust, this may prove to be a fruitful place to start. Additionally demographic factors like age and gender should be explored as potential individual differences in these mental models [63], [64]. This can provide a foundation upon which to explore this relationship. Research on individual differences can also help to explain the mechanism that undergirds users' mental model of VPAs. Such an understanding could allow for designers to directly influence the users' mental model of the system, in order to promote safe and efficient human-machine teaming.

Another avenue pertains to one of the principal limitations of this study; it is unclear how participants interpreted some questions. For the questions of "Does the [VPA] you talk to know everything that the [VPA] others talk to, know?" and "Is [VPA] a single automated assistant that everybody has access to, or does every device have a different [VPA]?" 18 and 22 responses were thrown out (respectively) by the experimenters coding responses for being indicative that the participants did not understand the question (no other question had more than 5). The first five responses that were thrown out for "Is the [VPA] a single automated assistant that everybody has access to, or does every device have a different [VPA]?" were as follows.

1) "Different devices have different assistants to talk to, so yes, it is all somewhat of a Siri. just with different names."
2) "Our family doesn't have any listening device, we don't even use Siri on any of our cellphones. I have Cortana on my laptop, but we don't have anything like an Alexa device in our living room or kitchen listening for commands."
3) "It's likely there is an assistant framework built into Alexa devices that gains additional capability with internet access."
4) "Only Apple products such as the iPhone or iPad have access to Siri. Other devices use assistants such as Cortana or Google Assistant."
5) "Siri is only available on Apple products."

While these responses are elucidative of participants' mental models, they deviate from the intention of the question in a way that made it difficult to make meaningful comparisons across responses for any given question. The answers seem to indicate that some of the questions given to participants were insufficiently clear.

It is possible that the lack of clarity from the results is indicative not of lack of clarity of participants' mental models, but instead of questions that elicited different thoughts across participants. This should be re-examined moving forward. Similarly, whereas the scale used in the present experiment appears to have high face validity, data were not collected that can assess this scale's psychometric properties, nor that can evaluate whether this scale truly assesses mental models of VPAs in line with recommendations of [65]. Given these limitations it is suggested that future research improve upon the wording used in the present scale in order to help mitigate issues of confusion (as detailed above) or imprecision, while also rigorously testing for validity and other important psychometric properties. Before being used again, research should assess whether the scale used here sufficiently adheres to principles needed for scale development.

Future research should also consider the overlap between the users' stated mental model, and their behaviors. For example, it is possible that a user says violations of trust from one VPA is reason to not trust the same VPA in another device, but their interactions with the VPA may indicate an actual loss of trust in the VPA found in that second device.

The contexts of VPA use could significantly drive the interactions as well. These different contexts may affect the expectations of the users. Interacting with a VPA in a car may result in profoundly different expectations than a user may have if he/she were to interact with the same VPA in the kitchen. These expectations may interact with the embodiment in ways that further complicate users' mental models in ways already demonstrated in cases such as the form function attribution bias (FFAB) [66]. The FFAB suggests that the design of a system directly impacts user expectations of its capabilities. For example, a large industrial robot may be believed to be strong enough to lift tens if not hundreds of kilograms, however in the case of once such robot (Baxter) the capacity is around 2.3 kg. It therefore may be important to consider the actual physical embodiment of a system whose capacities is limited. The environmental differences inherent to the different contexts may also result in discrepancies in expectations. Ambient environmental noise in a vehicle may result to poorer performance by a VPA which could be forgiven where similarly poor performance is not forgiven when a VPA falters in a quieter, private environment. In order to better predict this aspect of environment, it is important to understand user expectations of performance in varied situations. The degree to which these expectations map onto designers aspirations could provide actionable insight for designers and developers. Similarly, understanding user mental models in the context of system continuity (for example, knowing what ingredients need to be purchased as a function of the recipe that will be cooked) represents another fruitful avenue for research that could enhance the user's experience. Current approaches to providing this continuity could conflict with user mental models, which may result in disuse of a system simply because of design choices. In order to help mitigate the effects of these discrepancies, designers should consider the mental models inherent to not simply the system itself, but the environment and context in which these systems will be used. However, more research is needed in order to understand the role that context and environment may play in users' mental models.

Finally, research should explore the aspects of design and interaction that may influence users' mental model of these systems. For instance, there is already compelling evidence that in some cases users develop situation specific trust, such that they do not even generalize to the rest of the system [41]. By drawing from that literature, the field may be better able to understand the circumstances under which a user may generalize trust from one Siri to another, and those in which they are not likely to do so.

V. Conclusion

VPAs are becoming more widely available, and users indicate that they would like the same VPA they are familiar with to be used in new technologies, such as self-driving cars. However, no research to date has examined user's mental model of these interconnected systems. A survey here provided mixed results, showing that there is no universal mental model held by users. These findings need to be explored further in order to provide clarity for existing theories and designers of these systems, alike.

References

[1] A. Mutchler, "Voice assistant timeline: A short history of the voice revolution," 2017, Accessed: Dec. 10, 2020. [Online]. Available: https://www.voicebot.ai/2017/07/14/timeline-voice-assistants-short-history-voice-revolution

[2] M. Dubiel, M. Halvey, and L. Azzopardi, "A survey investigating usage of virtual personal assistants," 2018, *arXiv:1807.04606*.

[3] N. Gronau *et al.*, "Trust in smart personal assistants: A systematic literature review and development of a research agenda," in *Proc. 15th Int. Conf. WirtschaftsinformatikAt, Potsdam*, 2020, pp. 99–114.

[4] H. Yang and H. Lee, "Understanding user behavior of virtual personal assistant devices," *Inf. Syst. E-bus. Manage.*, vol. 17, no. 1, pp. 65–87, 2019.

[5] M. Ghazizadeh, J. D. Lee, and L. N. Boyle, "Extending the technology acceptance model to assess automation," *Cogn. Technol. Work*, vol. 14, no. 1, pp. 39–49, Mar. 2012.

[6] L. Matsuyama *et al.*, "Determinants that influence the acceptance and adoption of mission critical autonomous systems," in *Proc. AIAA SciTech Forum*, 2021.

[7] U. Saad, U. Afzal, A. El-Issawi, and M. Eid, "A model to measure QoE for virtual personal assistant," *Multimed. Tools Appl.*, vol. 76, no. 10, pp. 12517–12537, 2017.

[8] P. Bazilinskyy, S. M. Petermeijer, V. Petrovych, D. Dodou, and J. C. F. De Winter, "Take-over requests in highly automated driving: A crowdsourcing survey on auditory, vibrotactile, and visual displays," *Transp. Res. Part F, Traffic Psychol. Behav.*, vol. 56, pp. 82–98, 2018.

[9] G. R. Arrabito, "Effects of talker sex and voice style of verbal cockpit warnings on performance," *Hum. Factors J. Hum. Factors Ergonom. Soc.*, vol. 51, no. 1, pp. 3–20, 2009.

[10] P. R. Doyle, J. Edwards, O. Dumbleton, L. Clark, and B. R. Cowan, "Mapping perceptions of humanness in speech-based intelligent personal assistant interaction," in *Proc. 21st Int. Conf. Hum.-Comput. Interact. Mobile Devices Serv.*, 2019.

[11] M. Bonfert, M. Spliethöver, R. Arzaroli, M. Lange, M. Hanci, and R. Porzel, "If you ask nicely: A digitial assistant rebuking impolite voice commands," in *Proc. Proc. Int. Conf. Multimodal Interact.*, 2018, pp. 95–102.

[12] A. Abdolrahmani, R. Kuber, and S. M. Branham, "Siri talks at you: An empirical investigation of voice-activated personal assistant (VAPA) usage by individuals who are blind," in *Proc. 20th Int. ACM SIGACCESS Conf. Comput. Accessibility*, 2018, pp. 249–258.

[13] K. A. Hoff and M. Bashir, "Trust in automation : Integrating empirical evidence on factors that influence trust," *Hum. Factors*, vol. 57, no. 3, pp. 407–434, 2015.

[14] E. J. de Visser *et al.*, "Almost human: Anthropomorphism increases trust resilience in cognitive agents," *J. Exp. Psychol. Appl.*, vol. 22, pp. 331–49, 2016.

[15] E. J. de Visser, M. Cohen, A. Freedy, and R. Parasuraman, "A design methodology for trust cue calibration," in *Virtual Augmented and Mixed Reality. Designing and Developing Virtual and Augmented Environments*, Berlin, Germany: Springer, 2014, pp. 251–262.

[16] J. D. Lee and K. A. See, "Trust in automation: Designing for appropriate reliance," *Hum. Factors J. Hum. Factors Ergonom. Soc.*, vol. 46, no. 1, pp. 50–80, 2004.

[17] E. Phillips, S. Ososky, J. Grove, and F. Jentsch, "From tools to teammates: Toward the development of appropriate mental models for intelligent robots," in *Proc. Hum. Factors Ergonom. Soc.*, 2011, pp. 1491–1495.

[18] F. G. Halasz and T. P. Moran, "Mental models and problem solving in using a calculator," in *Proc. SIGCHI Conf. Hum. Factors Comput. Syst.*, 1983, pp. 212–216.

[19] D. E. Kieras and S. Bovair, "The role of a mental model in learning to operate a device," *Cogn. Sci.*, vol. 8, no. 3, pp. 255–273, 1984.

[20] D. A. Norman, "Some observations on mental models," in *Mental Models*, Hove, U.K.: Psychology Press, 2014, pp. 15–22.

[21] J. Cho, "Mental models and home virtual assistants (HVAs)," in *Proc. Conf. Hum. Factors Comput. Syst.*, 2018, pp. 1–6.

[22] J. Banks, "Optimus primed: Media cultivation of robot mental models and social judgments," *Front. Robot. AI*, vol. 7, May 2020.

[23] S. Kiesler and J. Goetz, "Mental models robotic assistants," in *Proc. Extended Abstracts Conf. Hum. Factors Comput. Syst.*, 2002, pp. 576–577.

[24] S. L. Lee, I. Y. M. Lau, S. Kiesler, and C. Y. Chiu, "Human mental models of humanoid robots," in *Proc. IEEE Int. Conf. Robot. Automat.*, 2005, pp. 2767–2772.

[25] J. Walden, E. H. Jung, S. S. Sundar, and A. C. Johnson, "Mental models of robots among senior citizens: An interview study of interaction expectations and design implications," *Interact. Stud.*, vol. 16, no. 1, pp. 68–88, 2015.

[26] T. Kulesza, S. Stumpf, M. Burnett, and I. Kwan, "Tell me more? The effects of mental model soundness on personalizing an intelligent agent," in *Proc. SIGCHI Conf. Hum. Factors Comput. Syst.*, 2012, pp. 1–10.

[27] K. I. Gero *et al.*, "Mental models of AI agents in a cooperative game setting," in *Proc. CHI Conf. Hum. Factors Comput. Syst.*, 2020, pp. 1–12.

[28] G. Bansal, B. Nushi, E. Kamar, W. S. Lasecki, D. S. Weld, and E. Horvitz, "Beyond accuracy: The role of mental models in human-AI team performance," in *Proc. AAAI Conf. Hum. Comput. Crowdsourcing*, 2019.

[29] M. Beggiato, M. Pereira, T. Petzoldt, and J. Krems, "Learning and development of trust, acceptance and the mental model of ACC. A longitudinal on-road study," *Transp. Res. Part F, Traffic Psychol. Behav.*, vol. 35, 2015.

[30] M. Beggiato and J. F. Krems, "The evolution of mental model, trust and acceptance of adaptive cruise control in relation to initial information," *Transp. Res. Part F, Traffic Psychol. Behav.*, vol. 18, pp. 47–57, 2013.

[31] G. Matthews, J. Lin, A. R. Panganiban, and M. D. Long, "Individual differences in trust in autonomous robots: Implications for transparency," *IEEE Trans. Hum.-Mach. Syst.*, vol. 50, no. 3, pp. 234–244, Jun. 2020.

[32] R. Parasuraman and V. Riley, "Humans and automation: Use, misuse, disuse, abuse," *Hum. Factors J. Hum. Factors Ergonom. Soc.*, vol. 39, no. 2, pp. 230–253, 1997.

[33] B. W. Israelsen and N. R. Ahmed, " 'Dave …I can assure you …that it's going to be all right …' a definition, case for, and survey of algorithmic assurances in human-autonomy trust relationships," *ACM Comput. Surv.*, vol. 51, no. 6, pp. 1–37, 2019.

[34] N. L. Tenhundfeld, E. J. de Visser, K. S. Haring, A. J. Ries, V. S. Finomore, and C. C. Tossell, "Calibrating trust in automation through familiarity with the autoparking feature of a tesla model X," *J. Cogn. Eng. Decis. Making*, vol. 13, no. 4, pp. 279–294, 2019.

[35] R. Gulati, "Does familiarity breed trust ? The implications of repeated ties for contractual choice in alliances," *Acad. Manage.*, vol. 38, no. 1, pp. 85–112, 1995.

[36] R. Parasuraman and C. D. Wickens, "Humans: Still vital after all these years of automation," *Hum. Factors*, vol. 50, no. 3, pp. 511–520, 2008.

[37] J. Sauer, A. Chavaillaz, and D. Wastell, "Experience of automation failures in training: Effects on trust, automation bias, complacency and performance," *Ergonomics*, vol. 59, no. 6, pp. 767–780, 2016.

[38] K. Weitz, D. Schiller, R. Schlagowski, T. Huber, and E. André, " 'Do you trust me?': Increasing user-trust by integrating virtual agents in explainable AI interaction design," in *Proc. 19th ACM Int. Conf. Intell. Virtual Agents*, 2019, pp. 7–9.

[39] P. Kulms and S. Kopp, "More human-likeness, more trust? The effect of anthropomorphism on self-reported and behavioral trust in continued and interdependent human–agent cooperation," in *Proc. ACM Int. Conf. Proc. Ser.*, 2019, pp. 31–42.

[40] E. J. de Visser *et al.*, "Towards a theory of longitudinal trust calibration in human–robot teams," *Int. J. Soc. Robot.*, vol. 12, pp. 459–478, 2020.

[41] M. Cohen, R. Parasuraman, and J. Freeman, "Trust in decision aids: A model and its training implications," in *Proc. Command Control Res. Technol. Symp.*, 1998, pp. 1–37.

[42] S. Rice, S. R. Winter, J. E. Deaton, and I. Cremer, "What are the predictors of system-wide trust loss in transportation automation?," *J. Aviation Technol. Eng.*, vol. 6, no. 1, 2016.

[43] N. L. Tenhundfeld, E. J. de Visser, A. J. Ries, V. S. Finomore, and C. C. Tossell, "Trust and distrust of automated parking in a tesla model X," *Hum. Factors*, vol. 62, no. 2, pp. 194–210, 2020.

[44] A. R. Wagner, J. Borenstein, and A. M. Howard, "Overtrust in the robotic age," *Commun. ACM*, vol. 61, no. 9, pp. 22–24, 2018.

[45] L. Fl, K. Browne, B. Coltin, J. Fusco, T. Morse, and A. Symington, "Astrobee robot software: Modern software system for space," 2018, Online. [Available]: https://ntrs.nasa.gov/archive/nasa/casi.ntrs.nasa.gov/20180003515.pdf.

[46] M. Williams and M. Braddock, "AI case studies: Potential for human health, space exploration and colonisation and a proposed superimposition of the Kubler-Ross change curve on the hype cycle," *Stud. Humana*, vol. 8, no. 1, pp. 3–18, 2019.

[47] J. D. Frank, K. McGuire, H. R. Moses, and J. Stephenson, "Developing decision aids to enable human spaceflight autonomy," *AI Mag.*, vol. 37, no. 4, pp. 46–54, 2016.

[48] J. D. Frank, "Artificial intelligence: Powering human exploration of the moon and mars," 2019, [Online]. Available: https://arc.aiaa.org/doi/abs/10.2514/6.2020-4164.

[49] E. K. Chiou and J. D. Lee, "Trusting automation: Designing for responsivity and resilience," *Hum. Factors*, pp. 1–29, 2021, Online. [Available]: https://doi.org/10.1177/00187208211009995.

[50] W. B. Rouse, J. A. Cannon-Bowers, and E. Salas, "The role of mental models in team performance in complex systems," *IEEE Trans. Syst. Man Cybern.*, vol. 22, no. 6, pp. 1296–1308, 1992, Online. [Available]: https://doi.org/10.1109/21.199457.

[51] K. J. Preacher, "Calculation for the chi-square test: An interactive calculation tool for chi-square tests of goodness of fit and independence [Computer software]," 2001.

[52] L. Schamber, "Time-line interviews and inductive content analysis: Their effectiveness for exploring cognitive behaviors," *J. Amer. Soc. Inf. Sci. Technol.*, vol. 51, no. 8, pp. 734–744, 2000.

[53] S. Elo and H. Kyngäs, "The qualitative content analysis process," *J. Adv. Nursing*, vol. 62, no. 1, pp. 107–115, 2008.

[54] H. Kyngäs, K. Mikkonen, and M. Kääriäinen, *The Trustworthiness of Content Analysis*, New York, NY, USA: Springer 2020.

[55] A. Walker, "Vehicle and driver : How voice supports the transition from manual to autonomous experiences in the car," Amazon, Seattle, WA, USA, 2019. [Online]. Available: https://developer.amazon.com/en-US/blogs/alexa/alexa-auto/2019/09/vehicle-and-driver-how-voice-supports-the-transition-from-manual-to-autonomous-experiences-in-the-car

[56] A. Walker, "J. D. power: Voice is a deciding factor vehicle purchase decision," Amazon, Seattle, WA, USA, 2019, Accessed: May 2, 2021. [Online]. Available: https://developer.amazon.com/blogs/alexa/post/4688f6ea-40ab-4cc0-a5e7-313f5735366a/j-d-power-voice-is-a-deciding-factor-in-the-vehicle-purchase-decision

[57] T. B. Sheridan, "Individual differences in attributes of trust in automation: Measurement and application to system design," *Front. Psychol.*, vol. 10, no. 1117, pp. 1–7, 2019.

[58] M. A. Rupp, J. R. Michaelis, D. S. McConnell, and J. A. Smither, "The role of individual differences on perceptions of wearable fitness device trust, usability, and motivational impact," *Appl. Ergonom.*, vol. 70, pp. 77–87, 2018.

[59] S. M. Merritt, J. L. Unnerstall, D. Lee, and K. Huber, "Measuring individual differences in the perfect automation schema," *Hum. Factors*, vol. 57, no. 5, pp. 740–753, 2015.

[60] S. M. Merritt *et al.*, "Automation-induced complacency potential: Development and validation of a new scale," *Front. Psychol.*, vol. 10, no. 225, pp. 1–13, 2019.

[61] S. M. Merritt and D. R. Ilgen, "Not all trust is created equal: Dispositional and history-based trust in human-automation interactions," *Hum. Factors*, vol. 50, no. 2, pp. 194–210, Apr. 2008.

[62] R. Parasuraman, R. Molloy, and I. L. Singh, "Performance consequences of automation induced complacency," *Int. J. Aviation Psychol.*, vol. 3, no. 1, pp. 1–23, 1993.

[63] M. G. Morris, V. Venkatesh, and P. L. Ackerman, "Gender and age differences in employee decisions about new technology: An extension to the theory of planned behavior," *IEEE Trans. Eng. Manag.*, vol. 52, no. 1, pp. 69–84, Feb. 2005.

[64] E. Rovira, A. C. Mclaughlin, R. Pak, and L. High, "Looking for age differences in self-driving vehicles: Examining the effects of automation reliability, driving risk, and physical impairment on trust," *Front. Psychol.*, vol. 10, no. 800, pp. 1–13, 2019.

[65] T. R. Hinkin, "A brief tutorial on the development of measures for use in survey questionnaires," *Org. Res. Methods*, vol. 1, no. 1, pp. 104–121, 1998.

[66] K. S. Haring, K. Watanabe, M. Velonaki, C. C. Tossell, and V. Finomore, "FFAB-the form function attribution bias in human-robot interaction," *IEEE Trans. Cogn. Devlop. Syst.*, vol. 10, no. 4, pp. 843–851, Dec. 2018.