Getting things done in an autonomous vehicle

John Sherry, Richard Beckwith, Asli Arslan Esme, Cagri Tanriover, *Intel Corporation*

Abstract— **Despite the heavy industry emphasis on navigation and collision avoidance, autonomous vehicles, as many researchers have argued, should also be considered social robots. In this paper, we describe initial findings from research designed to gather data on human-vehicle interactions in a real, moving vehicle designed to simulate an actual autonomous vehicle. Qualitative analysis focusing on observation and user interview data suggests several promising areas for deeper investigation, including the embodiment of the in-vehicle agent and exploiting the structure of the trip or related tasks to inform vehicle-passenger dialogue. While we argue that AVs should include robust multimodal sensing of the interior, as well as the exterior of the vehicle to promote better vehicle-passenger interaction, we also explore potential risks.**

I. Introduction

This paper represents a preliminary report on research in the area of human interaction with autonomous vehicles (AVs). AVs have received a tremendous level of research attention and investment over the past few years. While most of this focuses on autonomous navigation and collision avoidance [14][17] it is clear that AVs also represent a form of social robot. This is not just in terms of social, legal or ethical consequences of AVs actions [3] [6] but, more directly, because “autonomy” will require vehicles to achieve a high level of cooperation with humans to operate effectively [10] [5]. Thus, a variety of projects have begun to address the issue of how human passengers, pedestrians, cyclists or others might interact with autonomous vehicles [15] [16].

Our research builds on this body of work, focusing on passengers in particular, and how they might need or prefer to interact with an autonomous vehicle from inside the cabin. Our interest is shaped by a long history of research that acknowledges the importance of human agency. People are not simply “cargo” in AVs, they will have things to do. This interest is also shaped in part by collaborations with colleagues in the domain of “smart spaces” or ambient computing. Given this orientation, our research represents an attempt to apply sense-making technologies that integrate data from multiple sensory modalities, to enable an AV to better understand passenger states, activities, gestures and utterances, and respond appropriately. We anticipate a variety of reasons why it might make sense to enable a vehicle with multimodal sensing technologies. Among them:

- Successful operation of the vehicle seems to depend on a rich communications channel with passengers. For instance, passengers may need to indicate where, at a given destination, they want a vehicle to pull over to let them out, or may need an easy method for changing a destination or requesting an unplanned stop during a trip.
- In the absence of a human driver, vehicles may need to track and respond appropriately to a variety of passenger states or activities for safety reasons
- Service providers may want to observe passenger responses to a variety of conditions over time to improve service.

II. Research Description

In support of this research agenda, we conducted a series of trials with end users “in the wild” – on the streets of Richmond, British Columbia (a suburb of Vancouver). For reasons of safety, technology readiness and regulatory compliance, we did not use an actual autonomous vehicle, but rather a variation on the Wizard of Oz style approach used by others [18]. In our case, we used a passenger van that we modified to partially hide the vehicle operator, as well as the human acting as in-cabin agent, from the passengers in back. The cabin featured an array of sensing equipment. Three different video cameras were placed in the cabin of the vehicle to enable observation of passengers from a variety of angles. Each camera featured an active microphone as well. Audio-visual signals were time synchronized and aggregated onto a computing device hidden in a custom-built console. Passengers were also outfitted with electrocardiograms to monitor physiological response to vehicle events.

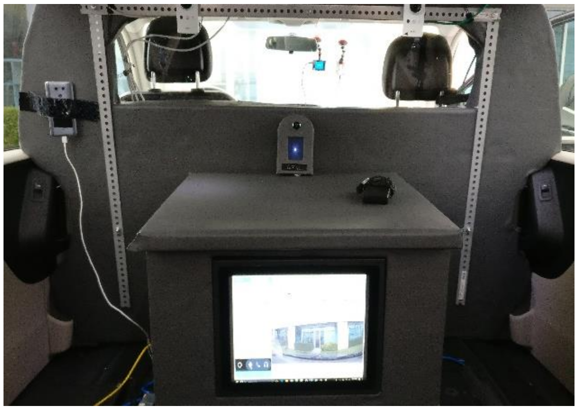

Figure 1. Passenger’s eye view of the AV cabin. At bottom is the display featuring a map and streaming video of external facing camera. In middle of photo, on top of the console, is the “face” of the AV agent. The human “actors” were seated in front of the barrier during sessions.

All authors are with Intel Labs, Intel Corporation, 2111 NE 25th Avenue, Hillsboro, Oregon, 97124 (john.sherry@intel.com) (Richard.beckwith@intel.com) (asli.arslan.esme@intel.com)

Our data gathering efforts involved what might be called a "scavenger hunt" protocol, as a means of engaging our subjects/ passengers. Our goal with the protocol was to create a situation where passengers were not preoccupied with either the novelty or the safety of the autonomous experience, but were nonetheless explicitly engaged in affecting the operations of the vehicle. In service of that, passengers were asked to perform such tasks as:

- Find a specific billboard and document the phone number on it
- Find a parking lot at a specific address and guide the vehicle to a parking stall near a particular landmark (a short tree)
- Navigate to a neighborhood "big box" retailer, as well as a nearby restaurant
- Get in and out of the vehicle
- Pick up a passenger at a designated location

In all, ten passengers traveling alone, as well as ten pairs of passengers traveling as dyads, executed these tasks, for a total of twenty sessions with thirty subjects. Each session lasted about one hour. During that time, passengers were in regular contact via their personal mobile phones with a remote facilitator who provided instructions for each task in successive fashion, after completion of the previous task. At two points in each session, the facilitator called the passenger to change a destination address. Each trip also featured two unexpected stops, initiated by the vehicle. This protocol was designed to elicit the kinds of behaviors mentioned above – passengers requesting, updating or changing destination requests; passengers giving specific directions (often with gesture) to the agent about where to pull over or park; passengers requesting the vehicle to speed up or slow down; and passengers requesting information / explanations from the vehicle. An additional goal was to elicit a number of observable passenger states and activities as training data for multimodal recognition across a variety of conditions with respect to ambient light and noise, vehicle motion, and other factors. Activities of interest included talking on the phone, eating, drinking, looking out the window, or interacting with other passengers. States of interest included primarily level of arousal and emotional valence, with a particular interest in whether the startle response might be detected, using both audio-visual data as well as ECG monitors worn by passengers during the session. While these latter data gathering efforts are central to our research, given the preliminary stages of research, this brief paper reports on results that were more readily attainable, as discussed below.

III. Preliminary Results

The remainder of this brief report will discuss four preliminary results based on initial observation of sessions, post-session interviews and qualitative analysis.

A. Efficacy of the "scavenger hunt" method and Wizard of Oz design

Observational evidence clearly demonstrated that the "scavenger hunt" approach is an effective method for creating situations that engage the attention and interest of passengers, and enable them to "naturally" engage in activities of interest for our machine learning training data efforts. Passengers similarly confirmed that the tasks represent a useful approximation – albeit in a very concentrated format – of the fluidity of daily vehicle use, which may include changes in destination, unplanned or improvised stops, negotiation of best route, and other common tasks associated with intra-urban transportation and travel. Similarly, observational evidence suggests that the WoZ approach is effective in helping passengers suspend disbelief and attribute vehicle operations to a non-human agent. Indeed, in post-session interviews, most of our passengers explicitly attributed vehicle actions to the presumed agent. As one passenger told us: "I was impressed that AMIE was able to know that my windows were fogged up and turn on the heater." We harbor no illusions, however, that passengers were able to completely forget or discount the presence of two humans in the front of the vehicle, only partially hidden by our partition. As will be discussed below (Finding 3), passengers tended to look toward the front of the vehicle during verbal interactions with the "automated" agent.

B. Subjects' responses to vehicle passenger sensing

Given the centrality of multimodal data gathering to our efforts, a key element of our qualitative research was to gauge user reactions to such data gathering. As indicated above, there are a variety of reasons why AV service providers might want to enable sensing and sense-making in the cabin, including the use of audio-visual signals. However, there are equally valid reasons to limit vehicle observation of passengers, particularly out of respect for their privacy [11] [13]. We were surprised, in our observation of sessions, the extent to which passengers accepted without concern the presence of multiple, obvious cameras in the vehicle cabin. This might well be attributable to the experimental nature of the event. In post-session interviews, passengers did express some resistance to the idea of constant surveillance – several passengers expressed the desire to have sensing limited under certain circumstances. For instance, one subject mentioned that he would be uncomfortable with audio-visual sensing during sensitive business calls. Others were uncomfortable with the observation of more intimate personal conversations, whether these were on the phone or with other co-present passengers. Subjects discussed a number of means by which they may want to control in-cabin observation, but the most common (and perhaps simplest) was the idea of providing passengers with an explicit control whereby they could allow or block in-cabin audio-visual sensing. When pressed, some subjects recognized that this method might be problematic – for instance, in shared ride situations where passenger safety might be a concern. Another possibility may be the use of entirely local analytics and sense-making on audio-visual data, so that some of the benefits of observation (safety, richer interaction channel) might be available, while keeping the data unavailable to observers outside the vehicle. This is clearly an area that requires deeper thought and exploration.

C. *Embodiment of the in-cabin agent*

In early design discussions, we explored a variety of alternatives for embodying our in-cabin agent, providing a "face" of the autonomous vehicle. As mentioned above, we separated the embodiment of our agent from other passenger displays (map and car's eye view) by placing it nearer the front of the vehicle, reasoning that passengers would naturally expect an agent to be near where the traditional vehicle controls are located. Observational data suggests that passengers did not actually orient their attention to our design with any appreciable frequency. We attribute this to two contributing factors. First, the face of our autonomous vehicle was too small, and not placed in a way that made its significance obvious. That is, it was *toward* the front of the vehicle, but still behind our partition, and thus not *in* the front of the vehicle. Second, as indicated above, it was clear that despite a general suspension of disbelief, passengers still tended to look toward the front of the vehicle (and the humans seated there) in some circumstances. For example, during each session passengers would receive a phone call from our facilitator, informing them of a change of address for their next destination. One passenger, while on the phone receiving instructions, paused the phone conversation to address the in-cabin agent, looking vaguely toward the front of the vehicle. Other passengers did so as well. This ostensibly negative result is suggestive, however. The passenger on the phone actually subtly adjusted her posture and gaze, making it obvious to human observers that the intended recipient was the (human) agent in the front seat. A similar situation may arise when multiple passengers are riding together in the vehicle: it may be easier for passengers to indicate that the vehicle / agent is the intended recipient of an utterance if they have some salient reference point to physically orient. This may seem like an investigation that would be easier in a simulator, we believe that testing in the wild made this issue more obvious. Travel in an actual moving vehicle, while attending to a set of tasks and instructions, provided passengers with a more realistic environment in terms of demands on attention. The busy-ness of the environment enabled us to see passengers "naturally"- even unconsciously – adopting strategies of communication that might have been less obvious in other settings.

Our exploration of embodiment raised a second interesting question: what exactly the embodiment represents. Will passengers identify the embodied agent with the vehicle itself, like the character KITT, from the popular 1980s television show *Knight Rider*? Or will they alternatively think of it as somehow separable from the vehicle, more like a taxi or Uber driver? Impressionistically, the latter seemed to be the case in our study, as evidenced by the greater than chance frequency with which passengers looked toward the front of the vehicle when addressing the agent. This may have been an artifact of how our design was implemented, with the driver and agent operator not completely invisible. It may also represent passenger habit. Regardless, this issue is far from trivial: current technology architectures manage what is considered to be core AV functionality separately from any on-board agent that might interact with passengers. It is unclear what effects this separation of functionality may have on the behavior of these systems, and on passengers' expectations of their own ability to either understand or influence vehicle behavior. These issues could in turn deeply affect passengers' feelings of safety and trust in the vehicle. For these and other reasons, therefore, we intend to continue with additional research on embodiment.

D. *Travel and the structuring of talk*

A final preliminary finding reinforces insights from other research on human-computer interaction and computer-mediated communications, particularly work that is informed by ethnomethodology and sociolinguistics: context, including the structure of extra-linguistic activity, is vital to structuring turns at talk and comprehension of utterances [4][8][9]. For instance, passengers frequently used gesture, along with deictic reference, to attempt to control the vehicle – for instance, saying "Can you pull over by that mailbox?" while pointing out the window of the van. We note that in doing so, virtually no passengers pointed to the "car's-eye-view" display inside the cabin. All passengers who gestured did so in reference to the actual, physical instantiation of the object outside the vehicle. This type of gesture suggests the need for a close connection between sensing inside the vehicle with sensing of the external environment. With sufficiently accurate pose detection and speech recognition in the vehicle, along with object recognition in the external environment, it may be possible to identify the target of such a pointing gesture. This insight was clearly only possible in the rich environment of the "real world."

Finally, utterances were not only structured by the external physical environment; travel through space also introduced a temporal structuring to talk that current approaches to automatic speech recognition (ASR), with a reliance on "wake words" seem poorly designed to manage. As a simple example: one user task involved finding a parking spot near an intentionally ambiguously described landmark ("Park near the shortest tree in the parking lot.") Most passengers working on this task resorted to a fairly exhaustive search of the parking lot to ensure they found the shortest tree – indeed, they were thorough to an extent we never intended. During this search, in addition to asking for the vehicle's assistance in identifying the shortest tree, or asking the vehicle to advise whether one shrub was a tree or "just a bush," passengers typically uttered a series of instructions to the vehicle ("please turn right and go along that row"; "can you go to the back of the building?") Each instruction received a spoken confirmation of receipt by the agent, and subsequently resulted in some form of maneuver by the driver. Sometimes these maneuvers were what passengers desired, sometimes not. When expectations were violated, corrective instructions followed. In cases where instructions were correctly followed, passengers would then wait to see if they'd found their tree. In both cases, silences of ten seconds or more often transpired between utterances, and yet passengers clearly conceived of the interaction as a contiguous and coherent dialogue. Current ASR systems would require the use of a wake word after such intervals.

Other strategies – such as active solicitation of input if the system expected instructions, would be equally annoying. Indeed, the only way to understand these dialogues is to understand the structure introduced by the task and its temporality. If natural language is to be a part of human-AV interactions, it will clearly need to recognize and take advantage of such structures as well.

IV. Conclusion

As these preliminary results hopefully show, despite the fact that our social robot was entirely simulated, interacting with it in the wild was essential in providing us with a number of key insights about the embodiment of the in-cabin agent, and the structuring of human-machine interactions, as well as the potential value of multimodal sensing. Adding a bit of artificial structure in the form of tasks for users to complete proved equally useful in helping us understand how people might need to interact with their AVs to get things done, rather than sit passively as AV cargo. As mentioned, these are early returns on a much broader research agenda, in particular, the gathering and characterization of multimodal data across a variety of conditions, to determine the utility of such data for training machine learning algorithms. Ultimately, our goal is to test whether vehicles that can take advantage of multimodal data to recognize passengers, along with some of their states and activities, will be able to deliver a better ride experience. Restricting ourselves only to explicit passenger-vehicle communications, the answer seems to be “yes.”

Acknowledgment

The authors are deeply indebted to the members of GlobalMe, Inc., particularly Rick Lin and Sophie Salonga, for their extensive efforts in organizing and executing this research in collaboration with our team at Intel Corporation.

References

[1] Arminen, Ilkka, Petra Auvinen, and Hannele Palukka. 2007. "Repairs as the last orderly provided defense of safety in aviation." Journal of Pragmatics 42: 443-465. (http://www.sciencedirect.com/science/article/pii/S037821660900174X)

[2] Banavar, G. 2016. "What will it take for us to trust AI." Harvard Business Review https://hbr.org/2016/11/what-it-will-take-for-us-to-trust-ai.

[3] Bennefon, Jean-Francois, Azim Shariff, and Iyad Rahwan. 2016. "The social dilemma of autonomous vehicles." Science 24 (352) 1573-1576.

[4] Button, G, and W Sharrock. 1995. "Simulacrums of a Conversation." In The Social and Interactional Dimensions of Human-Computer Interfaces, by P Thomas. Cambridge: Cambridge University Press.

[5] Casner, Steven, Edwin Hutchins, and Donald Norman. 2016. "The Challenges of Partially Automated Driving." Communications of the ACM 70-77.

[6] Elish, Madeleine, and Tim Hwang. 2015. "When your self-driving car crashes, you could still be the one who gets sued." Quartz https://qz.com/461905/when-your-self-driving-car-crashes-you-could-still-be-the-one-who-gets-sued/.

[7] Garfinkel, H. 1984. Studies in ethnomethodology. Cambridge: Cambridge University Press.

[8] Goodwin, C. 1981. Conversational organization: Interaction between speakers and hearers. New York: Academic Press.

[9] Heath, Christian, and Paul Luff. 1992. "Collaboration and control: crisis management and multimedia technology in London Underground line control rooms." Proceedings of Computer Supported Cooperative Work (CSCW). Toronto: ACM. 69-94.

[10] Johnson, Matthew, Jeffrey Bradshaw, Robert Hoffman, Paul Feltovich, and David Woods. 2014. "Human Machine Teamwork: Examples from the DARPA Challenge." IEEE Intelligent Systems 74-80.

[11] McCreary, Faith, Heather Patterson, and Alex Zafiroglu. 2016. "The Contextual Complexity of Privacy in Smart Homes and Smart Buildings." Proceedings of HRI International .

[12] McFedries, Paul. 2014. "The inescapability of ambient computing." IEEE Spectrum, Nov 20: https://spectrum.ieee.org/computing/it/the-inescapability-of-ambient-computing.

[13] Patterson, H, and H Nissenbaum. 2013. "Context-Dependent Expectations of Privacy in Self-Generated Mobile Health Data." Privacy Law Scholars Conference http://www.law.berkeley.edu/plsc.htm.

[14] S. Shalev-Shwartz, S. Shammah and A. Shashua. On a Formal Model of Safe and Scalable Self-driving Cars.ArXiv:1708.06374 Aug., 2017

[15] Sirkin, David, Kerstin Fisher, Lars Jensen, and Wendy Ju. 2016. "Eliciting Conversation in Robot Vehicle Interactions." AAAI Spring Symposium on Enabling Computing Research in Socially Intelligent Human-Robot Interaction. Palo Alto: AAAI.

[16] Stayton, Erik, Melissa Cefkin, and Jinyi Zhang. 2017. "Autonomous Individuals in Autonomous Vehicles: The Multiple Autonomous of Self-Driving Cars." EPIC2017 Proceedings. Montreal: American Anthropology Association. 92-110.

[17] The Policy Lab, University of Washington (2017) Driverless Seattle: How Cities Can Plan for Automated Vehicles. University of Washington. http://mic.comotion.uw.edu/wp-content/uploads/2017/02/TPL_Driverless-Seattle_2017.pdf

[18] Wang, Peter, Srinath Sibi, Brian Mok, and Wendy Ju. 2017. "Marionette: Enabling On-Road Wizard-of-Oz Autonomous Driving Studies." HRI'17. Vienna, Austria: ACM. 234-243.