APPENDIX: BEST PRACTICES FOR VR USER STUDIES IN CONTEXT OF AUTONOMOUS DRIVING

In this section, we detail best practices and details related to VR user studies which are not specific to VISTA, but may be helpful as a resource to autonomous driving researchers.

Questionnaires

Evaluation of Sickness with SSQ. Evaluation of VR systems is necessary since user immersion and sickness directly correlate to the realism of the virtual simulation.

The Simulator Sickness Questionnaire (SSQ) is a well-validated standard index for evaluation of simulator sickness that is both distinguished from the evaluation of motion sickness and simple to administer [3, 30].

The SSQ measures simulator sickness via self-evaluation of physical symptoms on a Likert scale. Such symptoms include general discomfort, fatigue, sweating, increased salivation, etc.

During evaluation of results from the SSQ, it is important to interpret SSQ results in a manner which is not misleading. Bimberg, Weissker, and Kulik had recently established guidelines on calculating scores for the SSQ as well as best practices for interpretation of such scores [5].

These guidelines include (1) including of missing brackets from final score computation, (2) it is necessary to administer the SSQ before and after exposure to VR, (3) not automatically attributing low scores to the simulator being "bad", (4) practitioners should report statistical means, medians, and standard deviations of total score as well as subscales.

Point (2) further supports the design choice of having questionnaires administered both before and after VR components, to ensure any increase in sickness is related to simulator exposure and not already preexisting within participants.

Evaluation of Immersion with SUS. The System Usability Scale (SUS) is a broad evaluation of technology usability in general [7]. Originally created by John Brooke in 1996, the SUS is a short 10-item questionnaire for subjectively evaluating the usability of a system. Obviously, this short questionnaire should be given only after exposure to VR.

Driver Characterization Metrics. Behavioral driving research has also produced well-validated indices for driver characterization. Relating established driver characterization questionnaire results to quantitative sensor data from simulation is not only an interesting area for future work, but also provides objective evaluation of otherwise subjective self-evaluations of driving behavior.

Those conducting user studies in VR for driving may also want to consider established metrics for driver characterization, most commonly in the form of questionnaires. We name a few classical methods for characterizing driver behavior, which have been heavily used and modified for current behavioral studies in driving and transportation.

Reason et al. established the Driver Behavior Questionnaire (DBQ) in order to categorize behaviors into violations, errors, and lapses [45]. The DBQ has often been used to predict self-reported motor vehicle accident involvement. More recently, Martinussen et al. have also established an abridged version of the DBQ which has results consistent to the full-length DBQ [35].

Other widely-used and popular self-reported driver characterization questionnaires include Lajunen and Summala's "The Driver Skill Inventory" (DSI) for assessing self-reported driving skill [34], as well as Taubman et al.'s "Multidimensional Driver Personality Inventory" (MDPI) for the categorization of human driving personality into eight distinct factors [54].

While simulator sickness and the quality of the simulation has the ability to disrupt or bias data collected in an experiment. Other factors may also contribute to low fidelity data.

Other Biases and Bias Mitigation

Biases can affect the results of user studies, and there are various sources of it. For instance, while sharing their thoughts with researchers, participants' responses might reflect social desirability (i.e., the tendency to provide researchers with perceived preferred answers) [19].

Moreover, humans have been found to have an inherent bias in spatial attention, with the left-side of the space tending to be favored over the right [4]. This sort of inherent tendency or pseudoneglect cannot be avoided but should be acknowledged [6] as it may affect reactions that were previously not predicted.

Furthermore, the order in which simulated scenarios are presented to study participants as well as the order of the responses can influence unwanted associations in the minds of participants [27]. So it is important to randomize the orders of scenarios when performing the study.

The timing and form of the questionnaires can also create unwanted biases. Most questionnaires are offered in a conventional paper or computer based form, and thus requires participants to remove their VR headsets to complete them. Studies have shown that requesting responses to questionnaires too soon before or after participants are required to wear VR headsets leads to a greater loss of self-control and is more likely to lead to temporal disorientation as they switch from one domain to another [2, 28]. However it is simple to remedy this, questionnaires can be administered and answered within the virtual environment or after sufficient time has passed for the participant to re-orient themselves in their real-world surroundings [2].

Ethical Considerations

The VISTA system of performing user studies requires the ethical procurement of data. Informed consent must be sought by the researchers to make sure that participation of the user study is voluntary.

Current datasets may exist containing a variety of everyday traffic behaviors from safe to risky,but human behaviors extend far beyond that of the expected. A virtual environment grants researchers a unique chance at collecting otherwise difficult to procure data. Virtual simulations can not only allow us to collect data from accident prone incidences, but also allows us to collect data from populations who are not equally represented [58]. Those with less than 20/20 visual acuity, or suffering from diseases such as Alzheimer's, now may have the ability to test response and motor skills through VR simulators, and help to create a more diverse set of human responses for autonomous vehicles to benefit from. This can also extend to modeling the behaviors of criminal or negligent driving, such as that of someone who is inebriated or fleeing arrest. While we outline many scenarios of participants driving in cars, a larger variety of vehicles will be sharing the roads with autonomous ones. VR setups can be extended and modified to emulate vehicles such as bicycles, tractors, trailers, etc.

Although there are many benefits from conducting VR user studies for autonomous driving research, conducting users studies that involve participants using VR headsets and interacting with other agents in a virtual environment raises various ethical concerns. For example, from a privacy perspective, users might feel uncomfortable while wearing the VR headset as it will prevent them from sensing if someone unwanted is watching them while they are in the virtual environment. Additionally, some VR applications require the use of various types of sensors, such as haptic gloves. These devices have data gathering capabilities and have the potential to be intrusive. Participants may not fully understand the capabilities of the interactive technologies they are using, and thus be less willing to use them to avoid more data being collected from them.

Researchers should be careful to evaluate the different privacy risks associated with each information state. This means, it is not

enough to take precautions to maintain the data confidentiality and integrity while it is stored, but also when it is being transferred or processed. Contextual integrity, a framework for understanding privacy [40], could be useful for researchers as it can allow them to identify the different information flows that are generated during all stages of their user study and that involve participants' data. With this information, researchers can reflect on how to minimize the risk of participant information exposure.

Also, due to the nature of VR, user studies simulating driving has the potential to be harmful or (re)traumatizing [12]. These experiences are not only bad for the participants but can also create negative connotations in this domain. This can be exacerbated by the involvement of deception (i.e., misguiding users to obtain results that could not be possible to obtain otherwise) [53].

In addition, researchers should seek to follow the ethical principles established in ethical frameworks such as the Belmont Report [20] and the Menlo report [29].

Storing Human Data in the Long Term.

It is key to provide participants with control over their data. Thus, researchers must take into account that participants might change their preferences on their data being part of a public dataset over time, so there needs to be a mechanism that facilitates the request of data removal from a participant perspective. Based on the specific type of data gathered from the studies, locations of the studies, various privacy regulations might apply (e.g., HIPAA, GDPR, etc.), and researchers need to be aware of them to violations. Additionally, when consolidating the dataset, there is a need to apply anonymization techniques to avoid the risk of a participant being re-identified. Researchers need to be careful while doing this task, as it has been evidenced how previously-believed anonymized datasets allow re-identification of people, due to the linking of different types of publicly-available datasets [51].