Adjunct Proceedings of the 8th International Conference on Automotive User Interfaces and Interactive Vehicular Applications (AutomotiveUI '16), October 24–26, 2016, Ann Arbor, MI, US .

1st Workshop on Ethically Inspired User Interfaces for Automated Driving

Andreas Riener
University of Applied Sciences
Ingolstadt, Germany
andreas.riener@thi.de

Myounghoon "Philart" Jeon
Michigan Technological Univ.
Michigan, USA
mjeon@mtu.edu

Ignacio Alvarez
Intel Corporation
Hillsboro, OR
ignacio.j.alvarez@intel.com

Bastian Pfleging
Ludwig-Maximilians-University
Munich, Germany
bastian.pfleging@ifi.lmu.de

Alexander Mirnig and Manfred Tscheligi
Center of HCI at Salzburg
University, Austria
firstname.lastname@sbg.ac.at

Lewis Chuang
Max Planck Institute for
Biological Cybernetics,
Tübingen, Germany
lewis.chuang@tuebingen.mpg.de

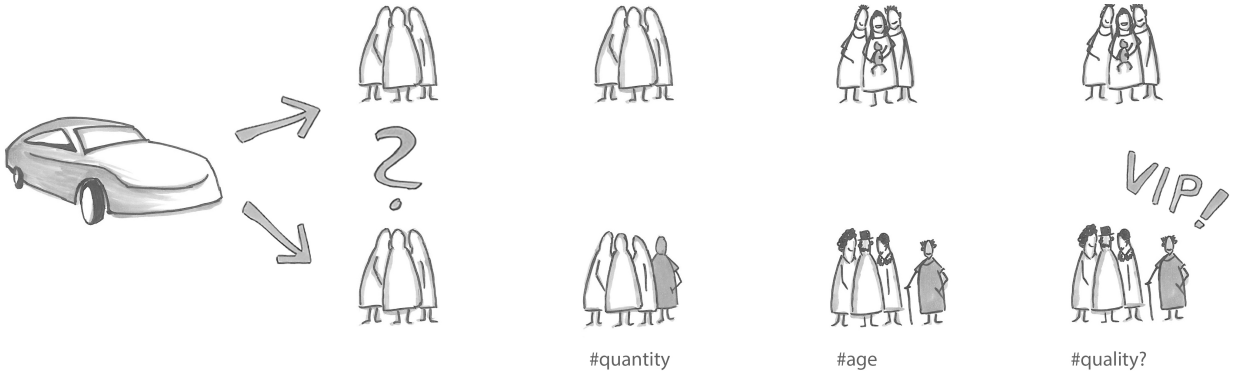

Figure 1: Which decision is, from an ethical point of view, the better one? This question will be discussed in detail at the workshop.

ABSTRACT

On July 1st 2016, the first automated vehicle fatality became headline news [9] and caused a nationwide wave of concern. Now we have at least one situation in which a controlled automated vehicle system failed to detect a life threatening situation. The question still remains: How can an autonomous system make ethical decisions that involve human lives? Control negotiation strategies require prior encoding of ethical conventions into decision making algorithms, which is not at all an easy task – especially considering that actually coming up with ethically sound decision strategies in the first place is often very difficult, even for human agents. This workshop seeks to provide a forum for experts across different backgrounds to voice and formalize the ethical aspects of automotive user interfaces in the context of automated driving. The goal is to derive working principles that will guide shared decision-making between human drivers and their automated vehicles.

Automotive'UI 16 Adjunct, October 24-26, 2016, Ann Arbor, MI, USA
ACM 978-1-4503-4654-2/16/10.
http://dx.doi.org/10.1145/3004323.3005687

Author Keywords

Automated driving; Asimov's laws; Trolley problem; Driver-vehicle interfaces; User acceptance and trust; Decision making; Negotiation algorithms.

ACM Classification Keywords

H.5.m [Information interfaces and presentation (e.g., HCI)]: Miscellaneous; K.4.1 [Computers and society]: Public Policy Issues – Ethics

MOTIVATION

Even with established ADAS technology, there is always a residual risk present. The past has many examples of system errors and accidents that were mainly caused by unknown system boundaries and limitations [6]. For example, many drivers are typically uninformed that an adaptive cruise control (ACC) system does not work properly in stop-and-go traffic or at sharp curvature [3]. Human misjudgments of system capabilities is the reason for numerous accidents. However, past events should not be an excuse to accept fatal accidents for the future. According to the World Health Organization (WHO), there is a stable

number of 1.25 million people who are annually killed in road traffic by driving manually. Thus, even if automated driving still carries the risk of fatal accidents, one could put forward the following pragmatic argument: "As long as automated vehicles eventually kill fewer people than are currently killed without automated vehicle features due to human error, the technology should be used".

High-level rationalization of human life loss compared to a greater good, however, is hardly in line with the expectations that we generally have of future technologies. Aiming at developing machines that can kill people, just not as well as humans do, seems to be an unsatisfactory goal to aim for. Beyond this high-level issue, even more ethical problem areas need to be explored. One of the most elementary ones (and focus of this workshop) is to define clear rules and guidelines that allow for negotiation and conflict resolution between multiple parties (i.e., "driving algorithms" in automated vehicles). The primary objective would be to define an ethically fair set of rules, based on which all decisions are made even though it might not be possible at all. This leads into two other related questions of similarly elementary nature: "How can these rules and relevant decision-making factors in traffic situations be displayed in a vehicle?" and "Should a machine be allowed to act on (or even enforce) such rules when human lives are on the line?"

Asimov's Laws

A good starting point for related discussions is the field of human-robot interaction (HRI), and therein a set of established laws. Asimov's laws [1] are three rules devised by Isaac Asimov (science fiction author) in 1942. The laws, quoted as being from the "Handbook of Robotics, 56th Edition, 2058" are:

1. A robot may not injure a human being or, through inaction, allow a human being to come to harm.
2. A robot must obey the orders given it by human beings except where such orders would conflict with the first Law.
3. A robot must protect its own existence as long as such protection does not conflict with the first or second laws.

These rules have been discussed for a long time in HRI until now and been found to be a valuable starting point. In contrast to earlier discussions, when translated to the domain of automated driving, we might now run into trouble as the three rules can no longer be considered as being isolated. As soon as an automated vehicle is carrying a passenger, the vehicle can no longer choose to violate rule 3 in order to uphold rule 1. After all, if the vehicle crashes in order to save the life of someone outside the vehicle, it will inevitably harm its passenger, thus automatically violating rule 1. It seems that in automated driving, there is no "right" choice to make in many situations when we apply only these rules. As a consequence, we cannot simply adapt findings and regulations from HRI and be done with it. Instead, it seems that automated driving requires its own model or rule base for decision-making, which needs to be developed and validated (simulator, naturalistic driving studies). On the way to such a complete model, however, a number of problems with yet unclear solutions will show up.

For instance, consider the traditional trolley problem (Figure 1, [2]). There are people on the road who are unable to move. You can "pull the lever" (i.e., swerve) to avoid them, but then your passenger or you will be sacrificed. How would you make a decision? Does quantity or quality of people matter at that moment? We can even create a more dynamic situation where two automated vehicles with passengers each approach one another and sensors of both cars detect in advance that a severe crash is inevitable. How will they decide to go for a conflict resolution? As the decision-making algorithms are working in a probabilistic way based on clear rules, it is most likely that both cars end up with the same conflict resolution strategy – e.g., leave the road on the left and drive into the field. The result is at least two damaged cars and two people injured or dead. From a human resources (or economic) stand point, however, it would have been much better to only sacrifice one car-driver pair and save the other one. If so, which whom should be sacrificed? If we apply Asimov's first rule, then both cars should have given control back to the human, asking for help, as they cannot make decisions that would end in harming or killing humans. Given that they have enough time to deliberate on which option to take, people will decide based on experience, social context, culture, attitudes toward life, etc. and it is expected that they are ethically correct. This, however, is an unrealistic expectation of the capacities of the human mind, considering the mere seconds that a human agent would have available in such situations. It would also require a clear set of rules for human agents to make such decisions in an uncontroversial manner. As of yet, however, there is no such set of universally accepted rules (be they utilitarian or otherwise) for situations in which all decisions lead to violations of a basic norm (e.g., no killing).

Even if we presuppose that such rules do exist and are known and that we can display all ethically relevant decision-information to the driver in the small time frame they have before the crash, then what role does the human play? Is handing control back to the human, while providing information regarding ethical constraints, values, and factors (i.e., a recommendation on who to kill) not a bit too close to the machine itself making and executing the decision? If the rules are clear and the context is clearly analyzed (i.e., the ideal decision strategy is known), then why do we need to introduce an element of fallibility by handing control over back to the human? Who do we really want to decide in life or death situations? Humans lack the computational capacity of machines, but are adaptable, and possess a strong sense of reasoning. Machines, on the other hand, excel in computation, logic, and multitasking, yet

lack any interpretative capacity or empathy [7]. Aside from these factors, it simply seems wrong to allow humans to decide over the lives of others. However, giving this privilege to machines seems, at least, just as wrong if not more so.

Keeping all these considerations in mind, there are three central problem fields surrounding ethical issues in HMI for automated driving:

1. How do we arrive at ethically sound decisions in everyday situations as well normative conflicting ones? How can such decision rules be translated into decision making algorithms for automated vehicles?
2. How can ethically relevant decision making parameters be visualized in a vehicle to support decision making for humans (or in other words: Could there be a "speedometer" for ethics in automated vehicles?)
3. Who should be the one to make these decisions, especially in situations with potentially fatal consequences? Humans (strong intuition, unreliable, consistent with Asimov's first rule) or Machines (good at calculations, reliable, inconsistent with Asimov's first rule)?

OBJECTIVES FOR THE WORKSHOP

In a brainstorming session during a recent seminar on future automotive user interfaces (Dagstuhl 16262, http://www.dagstuhl.de/16262), a large number of mentions belonged to the broader field of trust, acceptance and ethics related to automated driving. In that seminar, we did not have time to discuss all of these topics, but we realized that there is no common understanding on terms and definitions and how to develop them for the future. As it turned out to be highly controversial but, also, of high importance for the success of automated driving, we agreed to propose this workshop to consolidate thoughts on the topic.

Potential **topics to be discussed at the workshop** include, but are not limited to, the following ethical issues:

- Based on which moral rule set do (or should) humans make decisions in driving situations?
- How can such rules be formalized?
- How can information relevant to ethical decision making be displayed inside the vehicle?
- Should automated vehicles be allowed to violate Asimov's rules 1 and 2 in some cases (e.g., to choose the lesser of two evils)?
- Different cultural regions have different moral norms. How can we make sure that cars with different normative systems can interoperate in a "friendly way"(i.e., driving and negotiation functions automatically adapted to local conditions based on GPS location)?
- What are accepted conflict resolution strategies, i.e. when reaching a deadlock (two options with 50% probability each)?
- Should the vehicle even be allowed to make decisions on its own accord? Why (not)? And what are the consequences?
- Should we allow for rule-bending according to political, cultural or religious difference-s? E.g. as of today, females are less rewarded in Saudi Arabia. Should this gender inequality be translated to algorithms in extreme decision making?
- What is the right time to present ethically relevant information and how long should the system wait for human intervention/scored weighting in a no-win scenario?

WORKSHOP SUMMARY

After two invited/introduction talks by Ignacio Alvarez (Intel Corporation) and Andreas Riener (THI Ingolstadt), workshop submitters will get a chance to present their approach and issues in the field. After Q&A (which hopefully provokes lively discussions), a brainstorming session will be followed to identify common terms, issues, problems, and challenges related to ethics and ethically inspired UIs for automated driving. The workshop organizers will group the collected PostIts on a brainstorming wall and compile, from the most mentioned keywords, questions for the interactive part.

After a short break, workshop participants will be divided into smaller subgroups and, based on interest, each group will be assigned a topic. Groups will discuss related questions for more than an hour, create a poster/PPT, and finally present their point of view on the topic to the auditorium. After discussions in the large group, organizers will summarize the interactive part and collect the material created in the group works (posters, PPTs, prototypes). All documents will be provided to the workshop participants/contributors on a secured area of the website (http://www.andreasriener.com/AutoUI16WSEthical/).

In the concluding session, both participants and organizers will discuss future plans, e.g., to continue with this workshop series at AutoUI, CHI and related conferences or b) to try to develop a research agenda (mid-term goal), etc.

Summary of contributions

The paper "Ethical Automated Vehicles: Considerations and Plausible Directions" by R. Khan, E. Vasey, S. Landry, and M. Jeon gives a basic introduction to the field of automated driving algorithms and discusses the question, why these vehicles need to behave in a moral or ethical manner. In particular, the paper provides a good summary of related work and ongoing projects in the field. It further provides a table with possible ethical approaches and considerations to follow.

P. Wintersberger and A. Riener present in the paper "Determining the importance of fate to create publicly accepted moral agents" a solution for ethical decision making in fully automated driving scenarios. The authors

present the concept of "fateful decision making" when available variables cannot be easily rationalized. The fateful decision logic is composed of a randomized decision that they propose to test in a study to learn user preferences and acceptance as well as overall impact to ADAS.

ORGANIZERS

The organizers of the workshop have either a background in automotive user interface design, human factors, psychology/psychophysiology or a combination of these areas. Organizers are as follows:

Andreas Riener
is a professor for HMI and VR at Ingolstadt University of Applied Sciences, Germany with co-appointment at CARISSMA (Center of Automotive Research on Integrated Safety Systems and Measurement Area). His research interests include human factors in driver-vehicle interfaces, driving ergonomics, driver state estimation from physiological measures, and (over)trust, acceptance, and ethical issues in automated driving.

Myounghoon "Philart" Jeon
is Associate Professor of Cognitive Science and Computer Science at Michigan Tech. He directs the Center for Human-Centered Computing at Tech. His research focuses on driver emotion modeling. He received his PhD from Georgia Tech.

Ignacio Alvarez
is Research Scientist at Intel Labs, USA. He obtained his PhD in Computer Science at University of the Basque Country, Spain. His background is in Human Computer Interaction. His research interest is on future intelligent transportation systems and the practical application of cognitive sciences to affective computing and ADAS.

Bastian Pfleging
is a researcher at the chair for Human-Machine Interaction at LMU Munich. His general research interests are multimodal and natural user interfaces. In particular, he explores novel human-computer interaction techniques in the automotive context. He received his Diploma in Computer Science from TU Dortmund.

Alexander Mirnig
is a Research Fellow at the Center for Human-Computer Interaction, University of Salzburg, Austria. He holds a Master's degree in Analytic Philosophy and his research interests include driver space design patterns, handovers in semi-automated vehicles, and ethical issues in automated driving.

Lewis Chuang
is Research Scientist at Max Planck Institute for Biological Cybernetics. He holds a PhD in Neural and Behavioral Sciences by University of Tübingen. His background is in experimental psychology. His research focuses on cognition and control for Human-Machine Systems; understanding how humans seek out and process information to operate in control environments.

Manfred Tscheligi
is professor for HCI & Usability at the University of Salzburg (directing the Center for Human-Computer Interaction) and is heading the Business Unit Technology Experience at AIT. He brings in expertise for experience innovation projects a for a variety of application domains. He is very much involved in driving experience activities (e.g. as an national initiative on Car Interaction Safety) and has been shaping the discussion on autonomous driving and human robot-interaction. He has been involved in several conferences (e.g., co-chairing CHI'04, ACE'07 and AUI'11) and co-organizing workshops and SIGs (e.g., CHI'15, AUI'14, AUI'15, and Interact'15). He will be also Conference Co-Chair for HRI 2017 as well as MobileHCI 2017.

ACKNOWLEDGMENTS

This proposal is based, in part, on discussions with participants of the Dagstuhl Seminar 16262 "Automotive User Interfaces in the Age of Automation", http://www.dagstuhl.de/16262.

REFERENCES

1. Isaac Asimov (1950), I, Robot, novel.
2. Bonnefon, J-F., Shariff A., Rahwan I. (2016). The social dilemma of autonomous vehicles, Science, Vol. 35, June 24, 2016, 10.1126/science.aaf2654.
3. Boyle, D. A. (2009). Drivers' Understanding of Adaptive Cruise Control Limitations. Proceeding of the HFES Society Annual Meeting.
4. Fitts, P. M. (1951). Human engineering for an effective air navigation and traffic control system. National Research Council, Washington, DC.
5. T. Helldin, G. F. (2013). Presenting system uncertainty in automotive UIs for supporting trust calibration in autonomous driving. Proceedings of AutomotiveUI 2013, ACM, S. 210-217.
6. Norman, D. A. (1990). The Problem with automation: inappropriate feedback and interaction, not over-automation. Philosophical Transactions of the Royal Society of London, S. 585-593.
7. Parasuraman Raja, V. R. (1997). Humans and automation: Use, misuse, disuse, abuse. Human Factors: The Journal of the Human Factors and Ergonomics Society 39.2, S. 230-253.
8. Payre, W. A. (2015). Fully Automated Driving Impact of Trust and Practice on Manual Control Recovery. Human Factors: The Journal of the Human Factors and Ergonomics Society.
9. Vlasic, B. and Boudette, N. (2016). A Tesla Driver Using Autopilot Dies in a Crash, The New York Times, July 1, 2016, p. A1.
10. Wintersberger, P., Frison, A.-K., Riener, A., Boyle, L. (2016). Towards a Personalized Trust Model for Highly Automated Driving. Proc. of M&C, pp. 8.