Meaningful Human Control over Automated Driving Systems

Daniël D. Heikoop, Delft University of Technology, The Netherlands, *d.d.heikoop@tudelft.nl*, **Marjan Hagenzieker, Giulio Mecacci**, **Filippo Santoni de Sio, Simeon Calvert, & Bart van Arem,** Delft, The Netherlands

ABSTRACT

Human Factors issues with automated driving systems (ADS) are becoming more apparent with the increasing prevalence of automated vehicles on the public roads. As automated driving demands increased performance of supervisory skills of the driver, rather than vehicle handling skills, a mismatch occurs between the demand and supply of the drivers' skillset. Therefore, it has been suggested that drivers should at all times have *meaningful human control* (MHC) over ADS. The basic idea behind MHC is derived from the debate on autonomous weapon systems, and entails three essential components: human operators are (1) making informed, conscious decisions, (2) sufficiently informed about lawfulness of an action and its context, and (3) properly trained, to ensure effective control over the use of ADS. This paper presents definitions, components and potential human roles within ADS, from an interdisciplinary and a MHC perspective. The ideas presented in this paper are valuable to both designers, manufacturers, and road operators, as well as policy makers, driving licensing bodies, and lawyers and insurers, and our future research into these topics will deliver usable results for all stakeholders.

Keywords: Meaningful Human Control, Automated Driving Systems, psychology, philosophy, traffic safety.

1 INTRODUCTION

1.1 Background

The call to address Human Factors issues with automated driving systems (ADS) is becoming increasingly important, as market penetration of (partially) automated vehicles (SAE level 1 or 2; SAE, 2016) also increases. As a driver relinquishes tasks to an ADS, the driver is being ushered into a new role: that of a supervisor. During level 1 and 2 automation, longitudinal and/or lateral control is being taken over by advanced driver assistance systems (ADAS), which entails systems such as adaptive cruise control (ACC), lane keeping assist (LKA), or park assist, but also traction control and anti-lock braking system (ABS). Higher levels of automation (SAE level 3-5), will be able to take over the complete dynamic driving task, leaving the driver to merely watch the car drive itself. With this role change, the driver is subjected to out-of-the-loop difficulties (Gold et al., 2013; Louw, Merat, & Jamson, 2015). However, merely keeping the driver in the loop does not seem to be a sufficient solution, especially for prolonged periods of time (cf. Mackworth, 1948; Szalma et al., 2004). A more promising solution would be a case in which the human driver is asked to perform within their capabilities. This raises demands for *meaningful* human control (MHC; originated from the field of autonomous weaponry [Future of Life Institute, 2015]) over automated driving systems. The concept of MHC expresses the extent to which a human can maintain control over an automated system, even when not in (full) operational control (e.g., when an ADS rather than a driver performs operational actions; Santoni de Sio, 2016; Santoni de Sio & Van den Hoven, 2018). MHC is ultimately more demanding as well as more inclusive than the classic notion of "direct" operational

control, where a physical link is constantly required between human controller and controlled system. It is more demanding, as it prevents certain systems (such as vehicles) to be deemed under human control simply because somebody is 'in-the-loop'. It is more inclusive, as it can also include supervisory control, which entails monitoring an intelligent system that is in (full or partial) operational control, and gives the user the ability to undertake action if required. Moreover, MHC applies in principle also to automated systems without direct supervision.

1.2 Objectives

This paper will address the importance of incorporating MHC in the design, implementation and use of ADS. First, we will provide a definition of MHC, after which we will introduce the core components of an ADS that incorporates MHC. Consequently, based on these core components, the chain of control within an ADS will be discussed. Lastly, the implications an ADS with MHC has on the human driver will be investigated. The implications and impacts of MHC over ADS to various types of stakeholders will be discussed, and suggestions will be made in light thereof.

2 INCORPORATING MEANINGFUL HUMAN CONTROL IN AUTOMATED DRIVING SYSTEMS

2.1 Defining Meaningful Human Control

We approach the definition of Meaningful Human Control (MHC) over automated driving systems (ADS) by distinguishing two conditions, namely *tracking* and *tracing* (Santoni de Sio & Van den Hoven, 2018). In their article, Santoni de Sio and Van den Hoven (2018) identified tracking as the first necessary condition of MHC, and defined it as the ability for a decision-making system (such as ADS) to at all times be responsive to (i.e., 'to track') the human agent's (e.g., driver) relevant reasons to act. This can be at the level of a planned destination, or conventional social moral reasoning.

They identified tracing as the second necessary condition of MHC, and define it as the possibility to trace an automated system's behaviour back to some (responsible) human agent (e.g., operator, supervisor, designer, etc.). In order for that condition to be satisfied, there has to be at least one human agent in the system's design- and use history who is able to understand both the capabilities of the system to a sufficient extent, and their own role as targets of potential moral consequences for the system's behaviour.

With these two definitions in mind, we aim to identify core components involved in MHC over ADS, and the roles humans could or should maintain in order to achieve MHC over ADS.

2.2 Core components of automated driving systems with Meaningful Human Control

Driver, Vehicle, Infrastructure, and *Environment* have been identified as categories, wherein core components that are relevant to ADS have been identified (Calvert, Heikoop, & Van Arem, submitted). Based on an extensive literature search, the underlying components enable MHC to be analysed and integrated into the ADS in a very practical fashion. These core components play an important role in the chain of control within, and interaction between the driver and the ADS.

To give a few examples, some of the core *Driver* components identified are the sensory components necessary

for a driver to perceive stimuli (e.g., visual, auditory, and tactile components), and components involved in the cognitive and decision making process, such as interpretation and tactics. These (dynamic) components are commonly collated as driver behaviour, as an expression of a driver's state (as opposed to the usually static driver trait, like their personality). The (quality of) performance of these behaviours depend on whether they are automatized (i.e., skill-based), learned (i.e., rule-based), or 'on-the-spot' (i.e., knowledge-based) (cf. Rasmussen, 1983), and will be discussed further at section 2.3.

Some of the core *Vehicle* components are for instance sensors, such as the speed indicator and fuel gauge, and actuators such as the engine and clutch. Note that for an automated vehicle, for example its sensors take on a different role, from distributing information to the human driver during manual driving, to using this information itself when driving automatically, leading to completely different components for the same purpose.

For *Infrastructure* and *Environment* components, the authors identified physical and digital components, such as road surface materials and structure, and GPS and V2V communication, as well as weather components, such as rain and snow, and geographic components, such as urban or rural surroundings.

Furthermore, in their paper, Calvert, Heikoop and Van Arem (submitted) show how and why control is affected during the transition from manual to automated driving. They identified that in particular at higher levels of automation (i.e., SAE level 4-5), uncertainty increases over how this chain of control is (meaningfully) maintained, as operational control primarily lies with the ADS. Control and responsibility are shown to be even more unclear in the intermediate levels of automation (SAE level 2-3), where the chain of control is often shared between the driver and the ADS.

When aiming to adhere to the two necessary conditions of MHC for the design of ADS, following the chain of control, identified within the core components of ADS, allows explicit traceability of MHC to be performed.

2.3 Exploring human control over automated driving systems

The previous two sections explained two key conditions for MHC, and emphasized that following the chain of control within an ADS is important to apply MHC therein. This section explores the human role within ADS, and its relation towards MHC.

From a human perspective, controlling an (automated) vehicle requires skill-based, rule-based, as well as knowledge-based behaviour (Rasmussen, 1983). Loosely defined, skills are elements one is completely familiar with and can execute almost effortlessly (like steering with the wheel, negotiating a curve, or braking comfortably for a traffic light; basically anything one will be taught to do for getting your driver's license, and now can do automatically). Rule-based behaviour entails behaviour that is performed by remembering how to act given a specific situation (such as driving 50 kph at a 50 kph road, or stopping at a red light, but also not driving through a military convoy). Knowledge-based behaviour is addressed when a driver experiences a new, unfamiliar situation, and entails drawing conclusions from past experience and general knowledge, to be able to deal with this unfamiliar situation (like driving in snowy weather, or correcting while skidding). A lot of these situations can be trained, such as during advanced driver training courses.

With automated driving, many (if not all) of these behaviours will be taken over by the ADS, meaning that there will be little (if anything) left for the driver to perform within the context of operational control. But on the other hand, drivers of such vehicles will need to learn new skills too, such as driving (interacting) with ADS, and taking on a supervisory role. However, it is as of yet unclear to what extent the decrease and/or increase in behaviours occurs during automated driving.

This is currently being investigated by looking at what effects the levels of automation, as defined by the SAE, have on human behaviour, quantifying the number of tasks added or taken over by the ADS, depending on its level of automation (i.e., level 0, 1, 2, 3, 4 or 5).

3 IMPLICATIONS AND IMPACTS OF MEANINGFUL HUMAN CONTROL FOR AUTOMATED DRIVING SYSTEMS

3.1 Tracking and tracing

This paper briefly stresses the importance of having trackability and traceability incorporated in ADS in order to achieve a meaningful form of human control over such systems. Having an ADS tracking a human's intentions and moral standards avoids awkward situations in which the vehicle does something the human does not want, like missing an exit, or potentially in extreme cases, running over someone.

Being able to trace the status and actions of an ADS also helps the human understand the functioning of the ADS, and enables them to act appropriate, interacting as it were with the ADS, allowing for safe driving with a vehicle equipped with such a system.

Thus, when designing an ADS, adhering to the tracking and tracing principles will allow a human to have a vehicle at all times under meaningful control (i.e., not plain monitoring until something goes wrong, inevitably ending out-of-the-loop). For policy makers, this type of human control also allows for morally and legally sound ADS, as responsibility can always be tracked and traced back. Perhaps most important, the end user (i.e., driver/operator) of a vehicle equipped with ADS will face an acceptable and trustworthy form of control over their vehicle, and won't ever feel (nor actually be) completely out-of-the-loop.

3.2 Following the chain of control

An awareness of the importance of following the chain of control opens the gateway for safe and secure design and implementation of ADS, as it shows the core components involved within an ADS, and the components affected by a transition of control due to a transition in the level of automation. This allows ADS designers and legislators tracing back responsible components much easier, as the system is transparent. Moreover, ADS designers can understand the impact their systems have on the human driver as well as designing ADS such that a human driver can comfortably and safely (i.e., meaningfully) use that system. The extent to which the chain of control has to be, and is possible to be followed, will have to be further researched. For example, from a *Driver* perspective, would it suffice to follow it up to a behavioural level, or should we go as deep as a neurological level? From a *Vehicle* perspective, do we need to know all the nuts and bolts of a given vehicle? From an

Infrastructure level, will we come to a stage where we need to investigate individual pebbles to identify what went wrong? With higher levels of automation, we might need to be able to follow the chain of control further. These questions may be particularly interesting for lawyers and insurers to be answered, and for driver licensing bodies, it would be of interest to know what future drivers need to know and be able to do in various use cases.

3.3 Control challenges and impacts for a human driver

In order to maintain meaningful human control over an ADS, understanding the effect the transition of control has on the human driver in terms of trackability and traceability could be achieved by identifying what skills, rules and knowledge is being taken over by the ADS. Unexpected or unprecedented shifts in the remaining tasks for a human driver during higher levels of automation could prove disastrous for the human driver's ability to act as a fall-back in case of emergency. Given the core components involved with manual driving, from a human perspective, mirrored against how much influence they still have during (fully) automated driving, calls for an overhaul of the current transition of control over the various levels of automation. Moreover, the fact that a driver gets ushered into the unfamiliar, unknown role of a supervisor, entails that the demand for skilled behaviour is increasingly replaced by the demand for knowledge-based behaviour. Driving licensing bodies could step into this caveat, addressing the now unfamiliar situations, incorporating those into driver training, to allow future drivers to have the appropriate amount of skills to meaningfully control an ADS.

4 ASSESSING AND IMPLEMENTING MEANINGFUL HUMAN CONTROL IN AUTOMATED DRIVING SYSTEMS

4.1 The challenges in implementing Meaningful Human Control

Having MHC over an ADS implies full awareness (or traceability) of the systems' status and actions, and also the presence of an ADS that knows what you want (i.e., trackability). Several, if not all, core components within such an ADS are involved when aiming to maintain MHC, and trying to address all these components in order to achieve the ability for a human driver to maintain MHC over an ADS is no easy task. From a human perspective, the role transfer comes with an unavoidable and often undesired transfer in required skill- and rule-based behaviour. A transfer of control over various levels of automation that incorporates both defined conditions implies a smooth transition for the human driver to the extent that they will not be asked to perform a task they are not qualified or capable of doing.

4.2 Future research

This paper introduces the notion of MHC over ADS by means of theoretical reasoning. Empirical assessment of the key elements discussed in this paper regarding MHC are therefore yet to be assessed. In order to be able to incorporate the notion of MHC in ADS, one could think of a human-machine interface that provides the human driver with up-to-date status and action reports, and allows the driver to actively interact with the ADS. Future research could investigate whether an ADS that overtly 'decides', allowing for a trackable system (in contrast to a 'black box'), increases the likelihood of MHC over such an ADS. Furthermore, for researchers exploring this

domain, it is suggested to view the transition of control from a human perspective, rather than from a technical perspective (like the SAE levels of automation). Having an ADS incrementally take over tasks from a driver to such an extent that the driver is capable of performing their new role, could set another step closer to meaningful human control over an automated driving system. As a final, hereto related, suggestion, future research could investigate which tasks lend themselves best for ADS take-over without the risk of losing driver skills.

REFERENCES

Calvert, S. C., Heikoop, D. D., & Van Arem, B. (submitted). Core components framework of automated driving systems with meaningful human control. *IEEE Transactions on Human-Machine Systems.*

Future of Life Institute. (2015). ”Autonomous Weapons: An open letter from AI & Robotics Researchers”. from https://futureoflife.org/open-letter-autonomous-weapons/.

Gold, C., Damböck, D., Lorenz, L., & Bengler, K. (2013). ”Take over!” How long does it take to get the driver back into the loop? *Proceedings of the Human Factors and Ergonomics Science 57th Annual meeting,* 1938-1942. San Diego, CA, USA.

Louw, T., Merat, N., & Jamson, H. (2015). Engaging with Highly Automated Driving: To be or not to be in the loop. *Proceedings of the 8th International Driving Symposium on Human Factors in Driver Assessment, Training and Vehicle Design,* 190-196. Snowbird, UT, USA.

Mackworth, N. H. (1948). The breakdown of vigilance during prolonged visual search. *Quarterly Journal of Experimental Psychology, 1*, 6-21. doi:10.1080/17470214808416738 Rasmussen, J. (1983). Skills, rules, and knowledge; signals, signs, and symbols, and other distinctions in human performance models. *IEEE Transactions on Systems, Man, and Cybernetics, 13*, 257-266.

Rasmussen, J. (1983). Skills, rules, and knowledge; Signals, signs, and symbols, and other distinctions in human performance models. *IEEE Transactions on systems, man and cybernetics, 13*, 257-266.

SAE International. (2016). Taxonomy and definitions for terms related to driving automation systems for on-road motor vehicles. Washington, DC: SAE International.

Santoni de Sio, F. (2016). Ethics and Self-Driving Cars: A White Paper on Responsible Innovation in Automated Driving Systems. Dutch Ministry of Infrastructure and Environment Rijkswaterstaat.

Santoni de Sio, F., & Van de Hoven, J. (2018). Meaningful Human Control over autonomous systems: A philosophical account. *Frontiers in Robotics and AI*, *5*, 1-14. doi:10.3389/frobt.2018.00015

Szalma, J. L., Warm, J. S., Matthews, G. S., Dember, W. N., Weiler, E. M., & Meier, A. (2004). Effects of sensory modality and task duration on performance, workload, and stress in sustained attention. *Human Factors: The Journal of the Human Factors and Ergonomics Society, 46*, 219-233. doi: 10.1518/hfes.46.2.219.37334