A Conceptual Control System Description of Cooperative and Automated Driving in Mixed Urban Traffic With Meaningful Human Control for Design and Evaluation

Calvert, S.C.; Mecacci, G.

DOI
10.1109/OJITS.2020.3021461

Publication date
2020

Document Version
Final published version

Published in
IEEE Open Journal of Intelligent Transportation Systems

Citation (APA)
Calvert, S. C., & Mecacci, G. (2020). A Conceptual Control System Description of Cooperative and Automated Driving in Mixed Urban Traffic With Meaningful Human Control for Design and Evaluation. *IEEE Open Journal of Intelligent Transportation Systems*, *1*, 147-158.
https://doi.org/10.1109/OJITS.2020.3021461

IEEE Open Journal of
Intelligent Transportation
Systems

Received 24 April 2020; revised 9 July 2020 and 26 August 2020; accepted 30 August 2020. Date of publication 3 September 2020;
date of current version 1 October 2020.

Digital Object Identifier 10.1109/OJITS.2020.3021461

A Conceptual Control System Description of Cooperative and Automated Driving in Mixed Urban Traffic With Meaningful Human Control for Design and Evaluation

SIMEON C. CALVERT[1] AND GIULIO MECACCI[2]

[1]Department of Transport and Planning, Faculty of Civil Engineering and Geosciences, Delft University of Technology, 2600 GA Delft, The Netherlands

[2]Department of Artificial Intelligence, Donders Institute for Brain, Cognition and Behaviour, Radboud University, 6500 GL Nijmegen, The Netherlands

CORRESPONDING AUTHOR: S. C. CALVERT (e-mail: s.c.calvert@tudelft.nl)

This work was supported in part by the Research Programme Meaningful Human Control over Automated Driving Systems under Project MVI.16.044, which is financed by the Netherlands Organisation for Scientific Research (NWO), and in part by the Delft Transport Institute.

ABSTRACT The introduction of automated vehicles means that some or all operational control over these vehicles is diverted away from a human driver to a technological system. The concept of Meaningful Human Control (MHC) was derived to address control issues over automated systems, allowing a system to explicitly consider human intentions and reasons. Applying MHC to technological systems, such as automated driving is a real challenge, and the main focus of this article. An approach with mathematical elaboration has been developed that offers a first quantifiable operationalisation of MHC for the traffic domain and for use with automated vehicles. A major contribution lies in the taxonomification of control for MHC in the broader traffic environment, including consideration of the driver, the vehicle, the traffic environment, considering behaviour, moral standards and societal values, which are considered in a case study. The demonstration case shows the validity of the developed approach for an automated vehicle overtaking a cyclist on an urban street. This article is one of the first to operationalise MHC to such a level of detail and opens the door to further development of the concept for technological implementation.

INDEX TERMS Cooperative and automated driving, meaningful human control, automated vehicle design and evaluation.

I. INTRODUCTION

WITH the rise of automated vehicles in recent years and the expected introduction of cooperative, connected and automated vehicles (CAV) in the coming years, there has been much discussion in regard to safety and the required level of development for deployment on roads [1]. On high level roads, such as freeways and motorways, CAVs have a relatively safe traffic environment in which to be introduced with one directional uninterrupted traffic, relatively few conflict situations, and the absence of vulnerable road users (VRU). When allowing CAV's to drive in urban areas, many more challenges are present, such as interactions with vulnerable road users, greater and more heterogeneous interactions with other road users and the road infrastructure (e.g., traffic lights). There are also questions about gaps in control of CAVs [2], [3] and moral choices by CAVs [4]–[6] that are present for highway traffic, and these are all more potent and amplified in the urban traffic environment [7]. With certain aspects of CAV control being questionable from a purely operational and human acceptance perspective [8], [9], the concept of Meaningful Human Control (MHC) has been adopted for Cooperative and Automated Driving (CAD) to explicitly include human moral reasoning and ethical acceptability. MHC is a general concept that allows an automated driving system (ADS) to be designed and evaluated to a more morally responsible extent, while also considering driver and vehicle capabilities.

The review of this article was arranged by Associate Editor Claudia Campolo.

The notion of MHC, originated in the political debate on autonomous weapons systems [10], has been increasingly receiving attention from the engineering community in application to different kinds of automated systems, such as driving systems and surgical robots [28]. While MHC describes a control philosophy, in itself, it is not an operational control theory. Rather, it prescribes the conditions for a relationship between controlling agents and controlled system that preserves moral responsibility and clear human accountability even in the absence of any specific form of operational control from a human. In this article, the main objective is to take the concept of MHC and develop a description of CAV control in an urban traffic environment while considering the conditions laid out by MHC. Although we could consider this in many different traffic environments, the urban traffic environment offers unique challenges due to the extensive interactions and conflicts between road users. This has been widely acknowledged as an area in which automated vehicles will be tested to the full and require additional development for [3], [11]. The increased interaction with human beings (e.g., pedestrians, cyclists) also increases the necessity to consider this traffic environment from a broader point of view [12], [13] and to embrace the philosophical discussions on the issue of control and human intentions [14].

While control is often reduced to the technical aspects, the societal and moral aspects of control are just as important. In this article, we will consider the notion of Meaningful Human Control as conceptualized and successively developed by [14]. In their account [29], the authors distinguish two major conditions for *human* control to be *meaningful*, namely *tracking* and *tracing*. The tracking condition considers the responsiveness of a given system to the reasons to act of a certain (group of) agent(s). While classic theories of control focus on the causal relation between controller and controlled system, MHC is designed to deal with intelligent autonomous systems, and therefore considers how well its behaviour is (capable to be) attuned with a certain desired behaviour. The other condition, tracing, considers the degree of moral and practical involvement of the different agents with regard to the consequences of an automated system's behaviour. The aim of this condition is to allow the clearer and fairer possible identification of the roles automated of different agents in the chain of control, while also considering the ethical dilemma's involved.

In this article, the focus is on addressing issues of CAV control in the urban environment from the perspective of MHC. To do this, the concept of MHC needs to be operationalised in a formulation that allows it to be applied quantitatively to vehicle control. Operationalisation in this context refers to the translation of the concept into a tangible and process that can be applied in practical cases for a specific purpose. Therefore, the main contribution of this article lies in the description of CAV control in an urban environment from a human-centred, normative perspective of MHC. This, in turn, is meant to contribute to the design of automated systems that are more accountable and transparent, thereby allowing for a clearer understanding of the roles and responsibilities of the multiple agents involved in designing, deploying and supervising these systems. We must emphasise that the main focus is on the operationalisation of MHC and not on the development of control theory, which will be left to future work. In the following section, we start by describing the main system components that are required and how these relate to MHC and to the control approach. Section III describes the step towards operationalisation with the quantification of MHC in mathematical terms, which is demonstrated in a case experiment in Section IV. In Sections V and VI, we conclude with a discussion on the application of MHC in this context and the conclusions to the research.

II. URBAN CAD-V2X CONTROL DESCRIPTION

Here, we set the scope for the setup of the control system for automated vehicles in an urban environment with consideration of MHC. First, we touch upon the main components of the traffic system that should be considered. These give the scope for the design of the control system. The role of MHC in operational control is also discussed before we give the theoretical description of the control system thereafter.

A. SYSTEM COMPONENTS

The traffic system that we are investigating in this research refers to mixed traffic with cooperative automated vehicles (CAV) of various levels of technological development and human driven vehicles (HDV). We intentionally consider the control ability of this vehicle mix in an urbanised environment in which controlled intersections are present with intelligent traffic signals with the ability to communicate with CAV's. Vulnerable road users (VRU) are also present in the traffic system in the form of cyclists and on-road cyclists.

In Calvert *et al.* [15], a comprehensive overview of core components of the CAV traffic system is given, which describes the categories *driver, vehicle, infrastructure*, and *environment*, in which the categories driver and vehicles form *traffic*. We adopt the same division here in listing the relevant components for the control system. For environment, we consider the aforementioned urban environment with external environmental effects, such as weather conditions, to be 'regular', (entailing no effective precipitation, wind or any other type of disturbance) and consistent, and therefore uninfluential to traffic performance. Also, cyclists are also considered as additional actors in the traffic environment.

Infrastructure

- Road geometry
- Intersection design (# lanes, cyclists stop area, etc)
- Road sensors (e.g., loop detectors, motion sensors, camera sensors, etc)
- Traffic signals (traffic light controller, I2V communication, V2I prioritisation, etc)

Vehicle

- Primary vehicle sensors (e.g., speedometer, etc.)
- Perception sensors (radar, camera, ultrasonic, etc.)
- Connected and cooperative information (for other vehicle and infrastructure)
- Primary vehicle control (e.g., pedals, steering wheel, gears)
- Automated Driving System (ABS, (C)ACC, LKA, etc)
- NB: Power- and drivetrain actuation is presumed to be present and work effectively without explicit consideration in the control system.

Driver

- Driver attributes and driver state (e.g., stressed, inattentive, distracted, etc.)
- Driver perception
- Cognition and decision processing (incl., e.g., goals, plans, reasons, moral understanding, etc.)

B. ROLE OF MHC IN CONTROL SYSTEM

The concept of MHC in such a control system and within the traffic environment is a logical, but nevertheless intriguing, one that has not been considered or developed previously in literature. The concept explicitly addresses aspects of human behaviour and ethical awareness that has not previously been encompassed in technical systems.The application of the concept of MHC in vehicle automation is logical as humans must maintain generic control over such an ADCS that is there to aid mobility, but also has the potential to cause undesirable, unsafe or even dangerous situations, especially in an environment with VRU's. MHC relies on two formal conditions called tracking and tracing. The tracking condition considers the responsiveness of a system's behaviour to human (moral) reasons and intentions to act. Automated systems should be designed to recognize -and eventually respond to- different reasons in favour or against certain behaviour, given a certain situation. The nature of these reasons might be arbitrary and subjective, and pertain to single individuals, as well as highly general and intersubjective, and reflect societal values. E.g., different legal systems or social contexts might or might not want to allow more decisional freedom to drivers rushing to the emergency room, and perhaps less to those that have to rush to work. The tracking condition strives to achieve a form of control where direct operational control might not be available by design, such as in highly automated systems. The tracing condition demands the possibility to identify one or more human agents (e.g., ADCS designers, drivers, etc.) in the system's design and operation, who are able to: (i) appreciate the capabilities of the system and (ii) understand their own role as targets of potential moral consequences for the system's behaviour. This could be the driver, but does not have to be. MHC defines conditions for control that do not depend on whether a particular agent is performing specific tasks, e.g., exercising direct, operational control. Rather, those conditions regard certain capacities of the system as a whole. In such a way, it is clear that operational control by a qualified driver can lead to the system being under MHC, not just because a driver engages in driving tasks, but because the system satisfies the two fundamental conditions. In this way, the chain of control in the system can be traced to human users, while the intentions of the human users can be tracked in the automated system, as shown in Fig. 1.

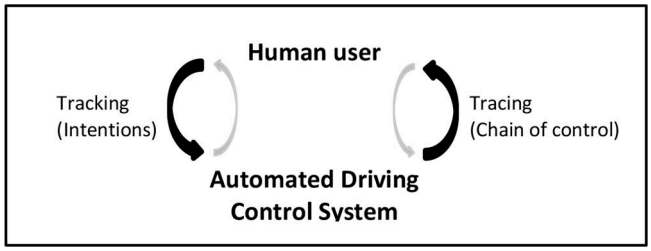

FIGURE 1. Influence of tracking and tracing on automated vehicle control.

Inclusion of MHC in traffic control systems can be viewed on different levels of abstraction. We mention two here, and will place the greater focus in this article on the second. The first level is that of the *global traffic system* in which the entirety of interactions between actors and components can be considered in regard to the extent to which they can be classified as *under MHC*, i.e., complying with the normative requirements and constraints expressed by the tracking and tracing conditions. The global traffic system that is considered is that of the highest level, which obviously includes aspects such as traffic (drivers and vehicles), and the infrastructure, but also regulatory and societal aspects such as traffic laws, driver training and experiences, trip planning, and general acceptable driving behaviour. Such a system is vastly expansive and enormously complicated. For this reason, we will by no means try to capture it in its entirety, but will be selective in describing some of the main factors of influence and we will make simplifications to the system to achieve this.

The second level considers MHC *over individual CAV's*. The controller in the control system for a fully automated vehicle is the automated driving control system (ADCS) of a CAV, while in a partially automated vehicle, the controller of the system is an interaction between the ADCS and the driver (see Calvert *et al.* [15] for a detailed description). In both cases, the ADCS performs tasks based on input from sensors, and results in behaviour that fulfil the following two conditions to the greatest extent:

- It corresponds to an identifiable set of relevant reasons to act of one or more designated human controllers. This is the tracking condition.
- It can be traced back through the chain of control to the appropriate human agents, i.e., those who display the relevant capacities and moral awareness. This is the tracing condition.

Furthermore, MHC is determined as a ratio rather than in binary terms: for any given system under consideration, the extent to which the tracking and tracing conditions are fulfilled indicates the extent to which the system is, or can be, under MHC. The two conditions are expressed through

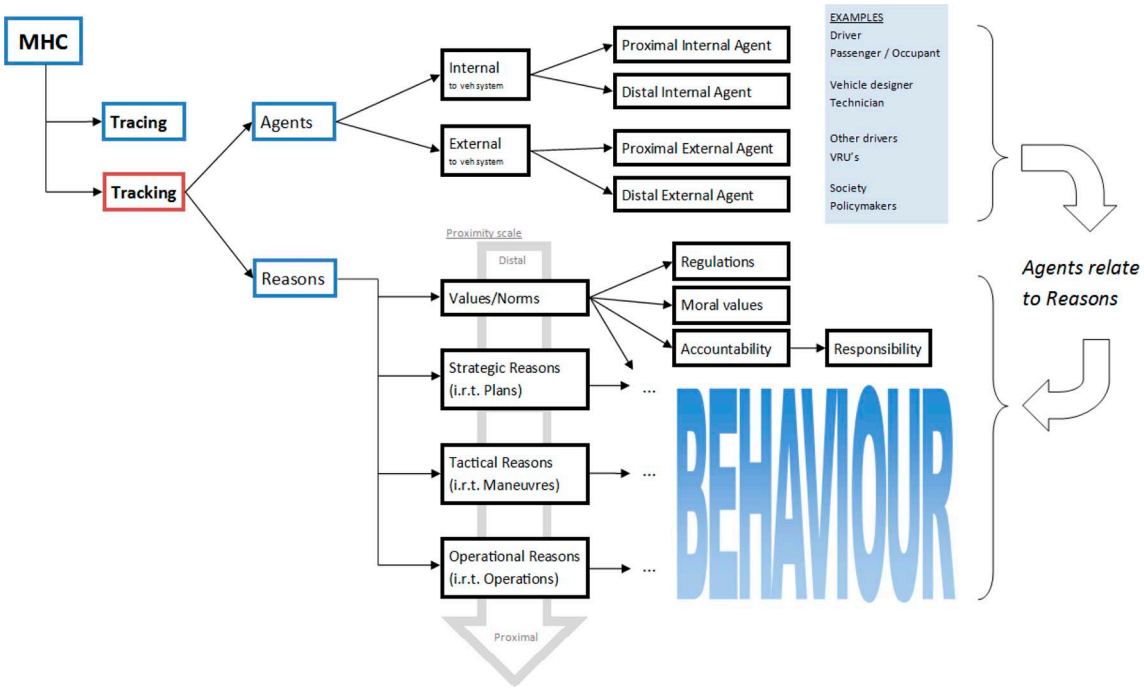

FIGURE 2. Taxonomy of the tracking condition for automated driving.

numerous sub-conditions that need to be disaggregated in a clear fashion before an operationalisation is possible. To aid understanding of the global traffic system, a decomposition of some of the main elements of MHC is given (see Fig. 2 and Fig. 3). In the flow charts, the aim is to decompose down to the level of core components of the traffic system, such that it starts to become feasible to consider how the concept of MHC can be transferred to a control strategy.

The tracking condition considers the responsiveness of a system to act according to human reasons and is decomposed in Fig. 2. The decomposition starts with the identification of agents (in the broadest sense) whose (moral) reasons can influence the system and MHC, and the corresponding reasons themselves. We can define external and internal agents to the system, such as the driver, on one extreme, to society on the other extreme. These agents can again also be defined in regard to how closely and directly ('proximally') their intentions influence the vehicle while in operation [14].

The human reasons are defined on a proximity scale: reasons of different agents are correlated to the behaviour of the controlled system. For example, a tactical reason may be that a driver wishes to take an off-ramp and will plan a manoeuvre to change lanes. This manoeuvre is therefore subject to the driver's behaviour to do so. At the top end of distal reasons, regulations play an important part, but also softer moral values and norms from which those regulations normally stem.

The tracing condition of MHC is decomposed and shown in Fig. 3. Tracing is defined as the ability to identify, within a system's design history or use context, one or more human agents who appreciate the system's capability and understand their own role as a target of potential moral consequences of the systems behaviour [29]. Therefore, 'Knowledge and Capacity' and 'Moral awareness' are taken as the highest level parts. *Knowledge* is the product of a driver's training and own past experience, which can be formal, e.g., driving lessons, or informal, e.g., reading a vehicles manual [10], [16]. *Experience* can also be active, passive, regular or irregular depending on the type of traffic events that are experienced [16], [17]. A driver's *capacity* relates to the type of task, which can be defined on a scale that goes from highly general strategic plans through to subconscious operations. This particular taxonomy, classically proposed by [30], regards the decision making process and the drivers' ability to make those decisions in relation to behavioural driving tasks. Reference [14] discussed how this distinction could be integrated with insights from classic theory of action to expand it to encompass a larger set of agents and their related (moral and ethical) reasons, intentions, values. There, it is proposed to discriminate the manifold of (moral) reasons and the agents of the MHC chain according to their relative 'proximity' to the behaviour of the controlled system, i.e., to how closely in time and space they influence the system's behaviour. For more details on this, we point to [14]. *Moral awareness* requires an agent who controls the vehicle (generally the driver) to be aware of generally accepted values, while their own role must be clear to themselves, in the sense that they have to understand and accept to bear moral and legal responsibility for the consequences of the behaviour of the system that is (partially) under their control [14], [29].

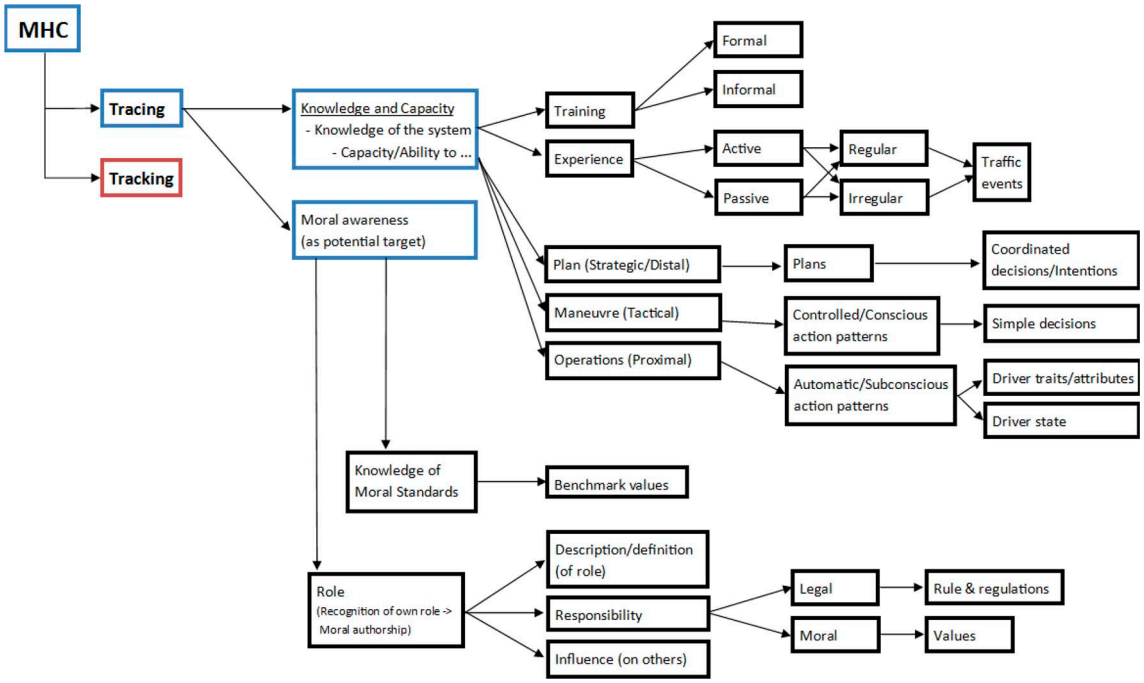

FIGURE 3. Taxonomy of the tracing condition for automated driving.

III. QUANTIFICATION AND APPLICATION OF MHC CONTROL APPROACH

The conceptual control approach of CAVs in an urban environment that is optimised for MHC was given in the previous section. The application and quantification of a confined case of this operational of control is given in this section. It is necessary to restrict the expanse of the quantitative description, as giving a full generic description is not feasible. To illustrate, this would need to include all aspects of Fig. 2-3 worked out to the fullest. If we take society's values as an example, this is just one element that would require extensive research and derivation and even then would only be able to convey a small subset of reality. The theoretical description is therefore given as a function of a precisely defined case with certain constraints put in place to allow quantification and demonstration of the control approach in a model. This addresses the aspects of MHC as indicated in Fig. 4. To this extent, we first give a short description of the case we consider. This is followed by the mathematical description of the quantified control approach with MHC. In Section IV, we describe the setup of the demonstrative case and give the results of modelled scenarios of the case using the applicable theory.

It should also be noted that the presented approach to operationalise MHC is unique due to the underlying human moral reasons and behavioural considerations that are based on ethical reasoning. The approach aligns with the domain of Responsible Innovation and Value-Sensitive Design research, which focusses on the embedding and expression of societal values into technical and socio-technical systems [18]–[20]. The fundamental question posed by these approaches is

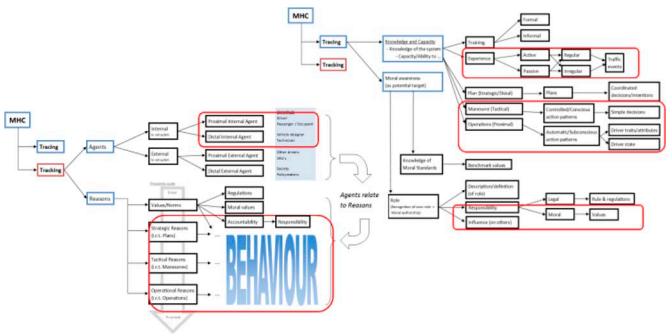

FIGURE 4. Considered areas of MHC in approach (in red), taken from Figures 2 and 3 combined.

how to design technical and socio-technical systems that satisfy normative societal requirements and values such as transparency, accountability, explainability and so on. For example, the upcoming and increased use of Artificial Intelligence (AI) in socio-technical systems is often seen as an opportunity to include most explicit and implicit aspects of a system control and optimisation [21], [22], including human-centred qualitative aspects [23]. Friedman *et al.* [18] recently performed an extensive review of current AI and ethics literature and concluded that "risk assessments, while valuable, do not capture important ethical risks which may be unquantifiable, qualitative, or unobservable" [18], which includes the objectives set out by MHC. For this reason, while being potentially valuable if achievable, we have not performed a comparison against any benchmark approach, as it would not be valid to do so. Moreover, the goal of the

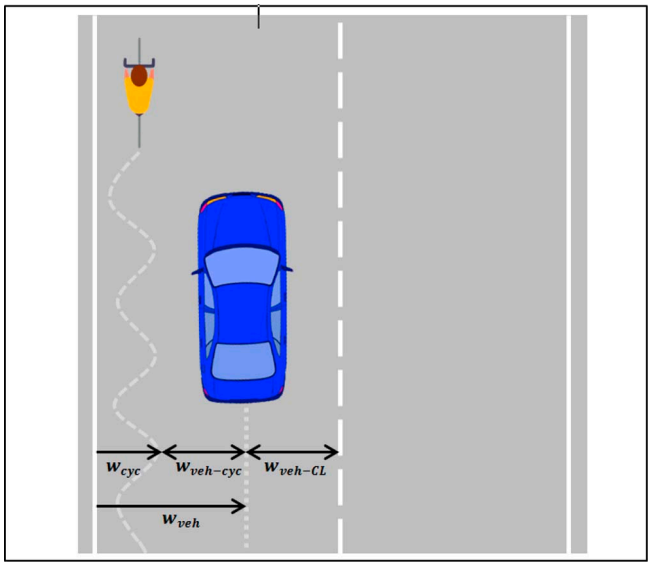

FIGURE 5. Graphical representation of the experimental case, showing the trajectory and positioning of the cyclist and AV.

use case is not exclusively present an irrefutable methodology for vehicle automation, but rather to offer an arbitrary demonstration of the application of the presented approach for automated driving. Therefore, the choices of parameters values, etc, are also arbitrary, but nevertheless logical ones.

A. USE CASE DESCRIPTION

The use case considers the interaction between a CAV and a VRU to demonstrate the technical and meaningful levels of the control approach for MHC in an urban environment. We reiterate that the objective of the case is to demonstrate the developed approach and that the case should not be considered as a standalone methodology, but rather an arbitrary exemplification. The case considers an ego-CAV on an urban road with bi-directional traffic. The ego-CAV encounters a cyclist on the road moving in the same direction as itself. There is limited, but sufficient, space to pass the cyclist and the CAV must decide when to overtake the cyclist, while taking the lateral distance to the cyclist and to the traffic on the opposite side of the road into account.

The cyclist maintains a constant longitudinal speed, but also has a lateral deflection described by a sinusoidal function with random error that enhances the lateral movements in a semi-predictable fashion. The ego-CAV can be considered to be a SAE level 4 vehicle, such that the vehicle is in control of operations and manoeuvres and must make the decision how to overtake and when to overtake. The model is setup, such that the CAV performs a large number of these overtaking manoeuvres so that the control system can learn from each experience and can adapt its strategy to optimise MHC. The described case is depicted in Fig. 5. The mathematical description of the approach based on MHC is given in the remainder of this section.

B. MHC CONTROL APPROACH

1) EVALUATION OF MHC

The translation of MHC and the control approach is made such that MHC is optimised in a workable model based on the control approach described in Section II and in accordance with a correct understanding on MHC and its components. The application of the control approach to a model is given in Fig. 6. The objective is to optimise MHC, thus ensuring that the vehicle performs actions that are as in line with MHC as possible. MHC is defined as a unitless variable:

$$MHC = Tracking + Tracing \tag{1}$$

As **tracking** considers the extent to which human reasons are met by the system, we define two levels of reasons for this case. The first is a tactical reason: a general human moral intention to not cause an accident or create a feeling of unsafety and therefore maintain a sufficiently high enough level of safety: $R^T_{safe}(t)$. The second is a strategic reason: the intention of the human driver to traverse their route with a minimal duration: $R^S_{dur}(t)$. The operational reasons related to the actual performance of the overtaking manoeuvre are presumed to be met and are not considered. As the duration should be minimised to create positive MHC and safety maximised, tracking is defined in the case as:

$$Tracking = R^T_{safe}(t) + R^S_{dur}(t) \tag{2}$$

Safety is defined here as a bi-logarithmic function of the lateral distance from the ego-vehicle to the cyclist $w_{veh_{ped}}$ or the centre line of the road w_{CL}, and of the inverted speed of the ego-vehicle $v(t)$, such that safety is written as:

$$safety = \frac{\log\left(a_1.w_{veh_{cyc}}(t)/v(t)\right) + \log(a_2.w_{CL}(t)/v(t))}{b} \tag{3}$$

$$R^T_{safe}(t) = \max(\min(safety, 1), 0) \tag{4}$$

The log function is chosen to capture the potential severity of close proximity driving to another object, which relents quickly as the distance increases. Here, a_1, a_2, and b are sensitivity parameters that allow each part of the function to be calibrated if required. Furthermore, the eventual value of the R^T_{safe} is constrained to keep it in the range [0, 1] as seen in Eq. (4).

The function for the duration to pass the cyclist is setup such that the longer the ego-vehicle takes to pass the cyclist, the lower the 'reason' to pass promptly is satisfied. Also, if the current speed $v(t)$ is further away from the desired speed v_0, then the situation is further from the human driver's reasons. This is given by the equation:

$$R^S_{dur}(t) = \frac{1}{\sqrt[c]{T_{cyc}}}.c_1 + \frac{v_0 - |v(t) - v_0|}{v_0} \cdot (1 - c_1) \tag{5}$$

Here, T_{cyc} is the time to pass the cyclist, c is a sensitivity parameter for this time, and c_1 is a sensitivity parameter that distributes the importance of the duration versus the speed sensitivity.

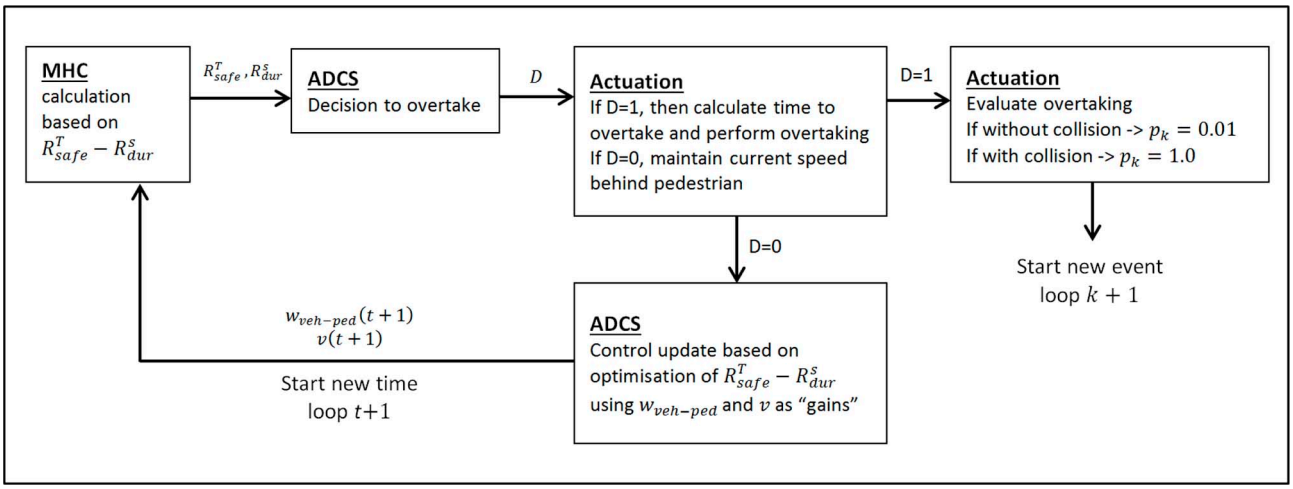

FIGURE 6. Implementation of control approach in the model.

The *tracing* condition is constrained in this case by the knowledge *Kn* and capacity *Cap* of the vehicle/driver to perform their tasks, and is given by:

$$Tracing = Kn \cdot Cap \tag{6}$$

We make a simplification to presume that the driver and/or vehicle designers are aware of their role allowing us to omit this part from the function. The ego-vehicles are presumed to have a base level of knowledge (or experience) d_1 based on previously acquired experiences, while the level of knowledge increases with each time the ego-vehicle performs an iteration of this case and attempts to pass a cyclist. The knowledge component is given by:

$$Kn = d_1 + \left(1 - \frac{1}{k^h}\right) \cdot (1 - d_1) \tag{7}$$

where k is the current number of iterations/experiences, and h is the sensitivity parameter for experiences.

The capacity, or capability, is a-posteriori variable, determined from previous performances. This entails that the number of poor overtaking manoeuvres is weighed against the number of correct overtaking manoeuvres. This is given by:

$$Cap = d_2 + \left(min\left(\left[1 - \frac{q. \sum_k^n (p_k > 0)}{k}\right]; 1\right)\right) \cdot (1 - d_2) \tag{8}$$

Here, d_2 is the distribution parameter for the base level of capability, e.g., 0.5, p_k is a binary penalty value for bad experiences, and q is the influence parameter for those bad experiences.

2) DECISION AND ACTUATION

The decision of the ego-vehicle's ADCS to perform an overtaking manoeuvre is based on the capacity to manoeuvre, and willingness to, based on experience. This entails a trade-off between reasons for safety coupled with ability on one side against the reasons for progression (short travel times). The safety reasons to perform the overtaking manoeuvre must be sufficiently high, while the capacity to perform the manoeuvre must also be sufficiently high. However, a decreasing level of the reasons to have a short travel time, may act as a trigger to perform the overtaking manoeuvre, hence why a lower R^s_{dur} will lead to an overtaking manoeuvre sooner. We make the fair assumption here that capacity to perform a manoeuvre is connected to the level of desire. This is all captured in the decision equation *D*:

$$D = Tracing \cdot R^T_{safe} > R^s_{dur} \tag{9}$$

$$D = 2 \cdot Kn \cdot Cap \cdot R^T_{safe} - R^s_{dur} > 0 \tag{10}$$

Once safety above the desired duration reaches a critical value (computed by the ego-vehicle capacity), then the driver overtakes: $D = 1$. The 2 at the start of Eq. (10) is there to balance the unitless variables and has no physical meaning. Once the decision to overtake is made, then the manoeuvre is blindly carried out without intermediate re-evaluation of the decision. This process is carried out in three steps:

The *lateral position* $w_{veh-cyc}$ and the current *vehicle speed* $v(t)$ are applied, as calculated in that control iteration, to perform the overtaking manoeuvre.

The time required to overtake is then calculated. This is presumed to be the required time to cover a static distance of 12 m (5 m following distance plus 5 m vehicle length plus 2 m buffer) plus the additional dynamic distance covered while overtaking is in progress ($v - v_{cyc}$).

The speed and lateral position of the cyclist are reviewed after the overtaking manoeuvre to update the ego-vehicle's experience and capacity values. If the cyclist moves into the vehicle's path, then this is recorded as a bad experience and is given a penalty: $p_k = 1.0$. If the vehicle passes without interaction, then this is positive experience and is recorded without penalty: $p_k = 0$. The corresponding R^T_{safe}, R^s_{dur} values are also recorded.

The control iteration update runs in time within a single experience k and updates the desired distance between the vehicle and cyclist $w_{veh-cyc}$ and the desired speed v of the ego-vehicle. This update is performed by maximisation of the reasons to act (R^T_{safe}, R^s_{dur}) to aim to maximise MHC. A further corrective factor c_{pk} is applied, which penalises bad experiences, as described above, and leads to a small increase in future $w_{veh-cyc}$ values to avoid the negative impact on MHC due to reduced future capacity Cap. The applied optimisation is given as:

$$Tracking\left(\left(w_{veh-cyc}(t) + c_{pk} \cdot \sum_{k}^{n} (p_k > 0)\right), v(t)\right) \rightarrow max \tag{11}$$

To clarify, tracing is not applied in the optimisation equation, as this relates to variables that are updated after each experience iteration and are not optimised within a single iteration k.

IV. USE CASE SCENARIOS AND RESULTS

A. USE CASE SETUP

A concise description of the case study is given in Section III-A along with a graphical example of the considered overtaking case of a cyclist by an ego-vehicle with high automation (see Fig. 5). The model, as described in the previous section, is implemented in MATLAB with certain constraints and values for the various parameters. The setup of the model for this case and the applied parameter values are now given.

A total of 200 experience iterations are performed, each with a maximum time duration of 600 time-steps, such that the ego-vehicle always has enough time to attempt an overtaking manoeuvre. A single time step is set at 0.1 s, meaning that the ego-vehicle has 60 s to pass the cyclist. To clarify, each experience iteration is a single encounter of the cyclist by the ego-vehicle in which it will attempt to pass the cyclist. Once an overtaking manoeuvre is initiated, the ego-vehicle will always proceed with the manoeuvre until completion regardless of the proximity to the cyclist. The decision to overtake the cyclist is made using the current lateral spatial gap available for the ego-vehicle. However, the cyclist while in motion, also has a dynamic lateral deviation that the ego-vehicle cannot account for. For this reason, in some cases the cyclist may move into the path of the ego-vehicle while an overtaking manoeuvre is in operation, which leads to a 'bad experience' for the ego-vehicle and the ADCS. The movement of the cyclist is governed by a sinus function expanded by a random factor to resemble the, sometimes unpredictable, cycling behaviour of a cyclist. This is applied by:

$$w_{cyc}(t) = m_1.sin\left(\frac{t}{-5}\right) + m_2 \cdot M(t) \tag{12}$$

Here, $m_1 = 0.6$, while $m_1 = 0.2$ and M is a set of random numbers from a uniform distribution in the range [0, 1]. This results in the cyclist maintaining a default position $w_{cyc.0}$ on the road of 0.5 m from the curb, with oscillations in both directions, shown in Fig. 7.

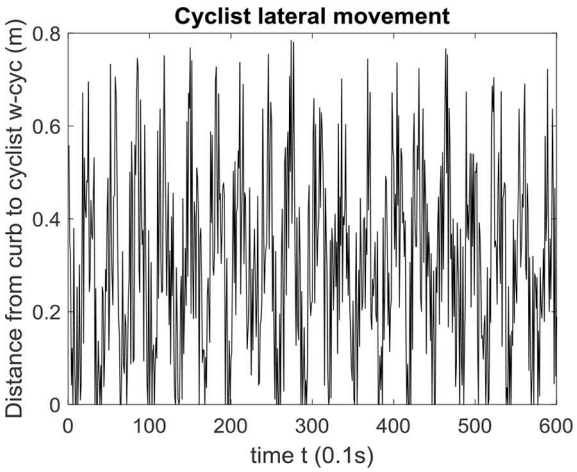

FIGURE 7. Lateral movement of the cyclist (example).

The nominal width of the lane is set at 1.2 m, which is taken as the distance of a regular lane minus the width of an average car and minus the width of a cyclist. This allows us to assume the cyclist and the ego-vehicle as single points rather than have to consider their lateral dimensions (see Fig. 5). The cyclist has a stable longitudinal speed of 4 m/s, while the default desired speed of the ego-vehicle is set at 12 m/s, which is also the default speed for the first time iteration.

The parameter settings applied in the case are calibrated to ensure the model is stable, starting from arbitrary assumptions of appropriate values. This resulted in values of $b = 5$ and $a_1 = a_2 = 1$ for the safety reason sensitivity parameters, $c = 3$ and $c_1 = 0.8$ for the duration reason sensitivity parameters. $d_1 = 0.4$ for the base level of driver/vehicle knowledge, $d_2 = 0.5$ for the base level of driver/vehicle capacity, and $q = 2$ for the influence parameter for bad experiences. Furthermore, the optimisation correction factor for the $w_{veh-cyc}$ distance is set at $c_{pk} = 0.01$ metres.

B. USE CASE RESULTS

The case is applied to demonstrate how the developed control approach can work in simulating MHC and can be applied to aid AV design considering MHC. The case results are evaluated based on the optimised $w_{veh-cyc}$ value with the occurrence of bad experiences (see Fig. 8a), the applied speed of the ego-vehicle together with the number of time-steps before the ego-vehicle decides to initiate an overtaking manoeuvre (see Fig. 8c) and the resulting values of the tactical reasons for safety, the strategic reasons for duration, and the calculated value of MHC per experience iteration (see Fig. 8b). In Fig. 9, we also give an example of how the values for MHC, R^T_{safe}, R^s_{dur}, $w_{veh-cyc}$ change within a single experience iteration as the ego-vehicle waits for a suitable opportunity to overtake the cyclist.

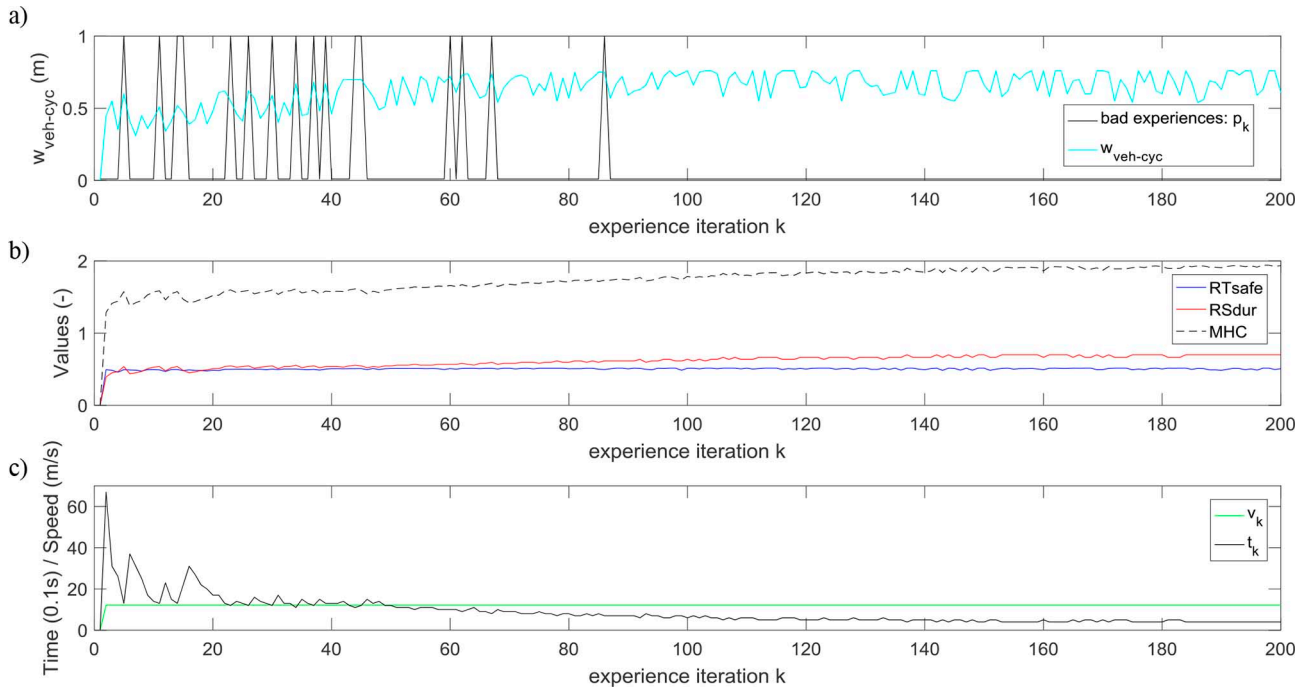

FIGURE 8. Results of model for a) lateral passing distance $w_{veh-cyc}$ at decision, bad experiences, b) level of MHC and corresponding values for strategic and tactical reasons, c) required time to make overtaking decision and ego-vehicle speed during overtaking.

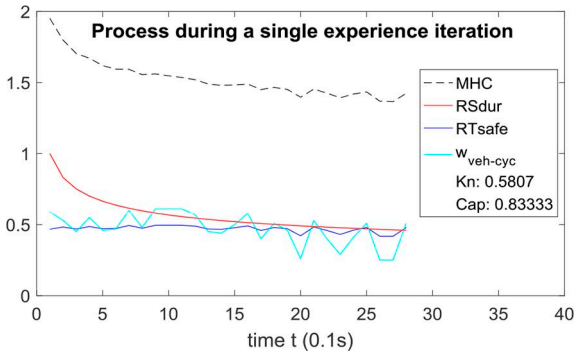

FIGURE 9. Results of model for a single experience iteration.

The stochastic behaviour of the cyclist's lateral movement naturally results in deviations in the lateral distance at which the ego-vehicle attempts to pass the cyclist. This is clearly visible from Fig. 8a as well as in Fig. 9 in which the resulting $w_{veh-cyc}$ per time step within a single experience iteration is shown prior to a decision to overtake. As the cyclist's lateral deviation sometimes continues to move in the direction of the ego-vehicle while the overtaking manoeuvre is taking place, a conflict can occur in which the distance between the vehicle and cyclist becomes zero (and would result in a collision without compensatory measures from the ego-vehicle). This is recorded as a bad experience by the ego-vehicle as shown in 8a. This also shows that the more experience the ego-vehicle gains, the fewer bad experiences occur. As safety, short travel time duration and also vehicle/driver capacity and knowledge contribute positively to MHC, we also see a trend that with increasing experience, the value of MHC also increases (8b). This increasing trend is not due to a single factor, but due to an increase in all three of these aspects that comprise the calculation of the MHC calculation. Also, with a greater degree of experience, the ego-vehicle is able to pass the cyclist with a shorter time duration. This positively impacts on the strategic duration reason and is made possible by a greater degree of safety and capacity of the ego-vehicle in the initial time intervals. Throughout the experiment, it emerged that the choice of ego-vehicle speed v_k remained inferior to the distance to the cyclist in the optimisation process and did not deviate from the desired speed for acceptable parameter values.

Within a single experience iteration, the process towards a positive decision to overtake can be graphically seen in Fig. 9. Initially when the ego-vehicle approaches the cyclist, the immediate desire to overtake is not high (inverse to R^s_{dur}). However, as the strategic reason starts to quickly decrease, the chance of an overtaking manoeuvre increases to prevent the overall MHC value dropping too far (as this is what is optimised in the decision to overtake and at which distance). As the cyclist moves laterally in the lane, an opportunity to overtake will occur at a certain point when the tactical reason for safety is sufficiently high and the desire to overtake is also sufficiently high. The knowledge and capacity (ability) of the vehicle does not change within a single experience iteration, as no additional knowledge or abilities are created. However, for a following iteration k, these values will change and will influence the decision to overtake (as seen in Eq. (10)). From Fig. 9, it is also clear why a prompt overtaking decision

can lead to a higher MHC value, as the strategic reason for a short duration is met much quicker, but only if the combined tactical reason for safety and vehicle experience and capacity are sufficiently high enough.

V. DISCUSSION

A. CASE AND MODEL DISCUSSION

The concept of MHC is an important one in the scope of automated systems, however operational generalisation of the concept in quantitative terms is challenging, if not impossible, as it basically encompasses the entire world of human values and reasons, something that in itself is difficult to quantify [4], [7], [24]. The main components, derived from literature found in psychology, automotive domain and philosophy as given in Fig. 2 and Fig. 3, give a good reflection of the system, while offering a framework to build a model from. The framework is intentionally left at a higher level of abstraction, as when one branches the framework out further, many cross connections, unproven theories and new abstract concepts of various psychological and philosophical areas open up, which might end up being too vast or insufficiently understood to be operationalized in any meaningful way. Some of those notions, e.g., moral responsibility and accountability, have been keeping generations of philosophers busy (for a useful taxonomy in relation to automated systems and control, see, e.g., de Sio and Mecacci [25]). Comprising a generic taxonomy is not realistic and also not the purpose of this research. Also for this reason, when we considered a case to demonstrate the developed mathematical control model, we have also specified and constrained the case description to adhere to acceptable and well defined components that allow an operationalisation and demonstration of the concept. The focus is therefore on explicit components that are internal to the case's system and that can be quantified, such as a driver's experience, tactical and operational manoeuvres and driver behaviour. External components and those that are funded more in societal and philosophical constructs, such as norms, values and moral standards, as well as legal aspects, are excluded due to their immense complexity when considering a system in operational sense. We do plan, and also challenge other researchers, to explicitly investigate these more distal components in the future and how they work and influence the system.

The control system is optimised for MHC, which makes sense as the premise is to demonstrate that automated vehicles control can be applied considering MHC. MHC was defined as a function of the two main conditions, tracking and tracing. The choice to assume a summation of these rather than a multiplication, and to assume equal weights is arbitrary. There is no evidence to indicate for or against this assumption. A similar case applies to the definition of tracking as a function of strategic and tactical reasons, and for tracing as a function driver knowledge and capacity. In regard to that last point, we do take the product of knowledge and capacity, rather than the summation, as without any knowledge, we deem driver's capacity to not be relevant and vice versa. For example, if a driver has zero capacity to drive (e.g., they cannot move there body or give instructions), then having knowledge of the system of how to drive is irrelevant. In a similar way to these logical and thought through, but nonetheless in many cases arbitrary, decisions, many of the other formulation have also been made in the model. The formulations are argued in the descriptions, and we do encourage other scholars to examine, adjust and propose altered or new formulations based on stronger theoretical foundations from the related fields of philosophy, psychology and alike.

The case is chosen as a simple and clear example of how MHC can work in practice. Again, this has been well constrained to allow the mechanism to be clearly presented and avoid unfounded assumptions and complexity beyond what can be demonstrated. The case successfully showed how MHC can be applied to lead to a system that generates a greater fulfilment of human reasons; in the case, a reason for safety and a reason for short travel time. With increasing iterations, fewer bad experiences occurred and we see the travel time increase as well, which both directly lead to a higher level of MHC due to the reasons being met to a greater degree. It goes without saying that the absolute values of the case should not be over analysed as the case is about demonstrating the mechanism and MHC principle in practice rather than being able to model exactly how many bad experiences an AV might encounter.

Finally, in recent years the development of automated driving systems has accelerated, with increasing on road testing of higher level systems. Currently, only low-level AVs exist on roads (e.g., with ACC and LKA technology), which are governed by simpler control mechanisms and encompass a limited decision making process. Higher level systems, that may be closer to autonomous driving, will require more complex decision making technology, of which much is expected to come from Artificial Intelligence (AI) [21], [26], [27]. This in itself is a broad and yet uncertain discussion that is ongoing in scientific and technical communities and is also of relevance. We have chosen not to engage directly in this discussion in the paper, as the focus is not on the decision making process here. Nevertheless, the existence of this discussion needs to be highlighted and should also relate to the concepts presented in this article.

B. BROADER MHC DISCUSSION

MHC is a notion of control that crucially includes normative elements and constraints. It configures as a radical reconceptualization of the notion of control in engineering, to embrace psychological and political intuitions. Systems where AI plays an increasing role, can less and less be controlled as simple machines, but must be dealt with in a more sophisticated way. Intelligent systems are by design "out of control". The challenge of MHC is to find new ways to keep human agents involved in, and responsible for, the behaviour of AI driven systems. Although its origin is the

political discussion about robotization, and specifically the military context, in such debates, societal values and interests are core elements and points of departure. There remains a challenge for MHC is to find its way into a technical and institutional discussions and implementation. This article represents an audacious attempt at moving some steps towards a technical operationalization of fundamentally normative notions, e.g., those of moral responsibility and accountability in automated intelligent systems. It also provides insights into conceiving and designing systems that are more transparent in the sense that the roles and stakes of the different agents who contribute to determine their behaviour are more clearly identifiable. Challenges to this extent that have been confronted in this research focus on three main areas.

First, operationalisation requires simplification. The concept of MHC is not only inherently normative, but it is meant to apply to a socio-technical system (a broad notion of system characterized by a complex interdependence of normative, social and technical elements) in its entirety. This means including societal infrastructures, e.g., legal systems as much as technical infrastructures, e.g., the engineered environment. Simplifications are required to gain focus and encompass those areas that are most influential, relevant and changeable.

Second, technical operationalisation requires some form of quantification of the considered values. This is particularly hard because normative notions are intersubjective at best. In order to be able to attribute a numerical value to some of the elements that MHC considers, we might need to commit to some arbitrary choices or rely on psychometric methods. Moreover, it is unlikely and unrealistic that we can obtain absolute and time-stable numerical quantities for certain values, due to their inherent subjectivity and variability. A more modest aim is to be able to establish the relative value of two or more instances of a certain element, e.g., whether a certain agent is *more or less* morally aware of their role when compared to a second agent, *ceteris paribus*.

Third, and connected to the second point, MHC is never conceived as an all or nothing feature of a system. Even in a case where there is a perfect intersubjective agreement on the meaning and relative value of the components of control, it is impossible in principle to establish a theoretical maximum of MHC. Hence, MHC is not a matter of binary presence or absence, but rather an intersubjectively established relative quantity, that always comes in degrees, and makes the most sense in a comparative scenario (e.g., one can establish, in a control model, whether a certain element at time T has increased or decreased relatively to a previous point in time T-1)

An important aspect of MHC is the realisation that it is not a concept that can be considered universally absolute. No scale can exist that can allow MHC to be 100% present, as it exists of many components that are related to very abstract and collective normative aspects, such as human values and reasons. These will nearly always have some element of conflict between agents and can never be achieved to perfection. Therefore, we speak of the degree to which MHC can be achieved on a relative scale, rather than talking about MHC being present or absent as a binary notion. Translation to a quantitative operational variable requires some refining of this, but should also consider the aforementioned characteristic of the concept, such that various factors will results in a greater or weaker degree of MHC (in our case this is the trade-off between the tactical reason for safety and the strategic reason for duration).

VI. CONCLUSION

With challenges of increased vehicle automation and questions regarding sufficient control of these vehicles, this article has presented an elaboration of the operationalisation of Meaningful Human Control (MHC), as a control concept that can address many control issues in Cooperative and Automated Driving (CAD). Operationalisation of a philosophical control concept is challenging for a number of reasons that have been addressed in this article. Building on past and present developments, the paper has identified key components of the control system and has developed a first mathematical underpinning of how these can be operationalised for analysis and design in real life cases. This is captured in a control approach that focusses on operationalisation in the field of transportation, rather than systems control, and has taken a corresponding approach. A major contribution lies in the taxonomification of control for MHC in the broader traffic environment, including consideration of the driver, the vehicle, the traffic environment, considering behaviour, moral standards and societal values, such as safety. The presented case is applied as demonstration of the approach and considers the control process of a highly automated vehicle attempting to pass a cyclist on an urban road, while considering trade-offs between human desires, moral standards, traffic efficiency and safety. These interactions are highly critical and will pose challenges for CAVs as they continue to develop and be implemented. The developed approach demonstrates that MHC can be applied as the core control concept to allow the system to interact with other road users (a cyclist in this case) and to learn to improve performance through multiple iterations. The importance of MHC at the core of these control developments further means that human reasons and values are explicitly considered internally, rather than nominally and externally to the system, which has demonstrated up to this point, to lead to breaches in acceptability and causality when considering safety and control with automated vehicles. Design and evaluation of automated vehicles system performance with consideration of MHC is taken a step closer through the research in this article and the broader automotive and transportation communities are urged to take heed of the developments.

ACKNOWLEDGMENT

The authors would like to thank Konstantinos (Kostas) Ampountolas, previously of the University of Glasgow and

now at the University of Thessaly, for the fruitful discussions on the topic during a visit to Glasgow.

REFERENCES

[1] G. Meyer and S. Beiker, *Road Vehicle Automation*. Cham, Switzerland: Springer, 2014.

[2] S. C. Calvert, G. Mecacci, B. van Arem, F. S. de Sio, D. D. Heikoop, and M. Hagenzieker, "Gaps in the control of automated vehicles on roads," *IEEE Intell. Transp. Syst. Mag.*, early access, Jan. 31, 2020, doi: 10.1109/MITS.2019.2926278.

[3] J. Guanetti, Y. Kim, and F. Borrelli, "Control of connected and automated vehicles: State of the art and future challenges," *Annu. Rev. Control*, vol. 45, pp. 18–40, 2018.

[4] G. Contissa, F. Lagioia, and G. Sartor, "The ethical knob: Ethically-customisable automated vehicles and the law," *Artif. Intell. Law*, vol. 25, no. 3, pp. 365–378, 2017.

[5] P. Lin, "Why ethics matters for autonomous cars," in *Autonomous Driving*. Berlin, Germany: Springer, 2016, pp. 69–85.

[6] S. Nyholm and J. Smids, "The ethics of accident-algorithms for self-driving cars: An applied trolley problem?" *Ethical Theory Moral Pract.*, vol. 19, no. 5, pp. 1275–1289, 2016.

[7] S. Nyholm and J. Smids, "Automated cars meet human drivers: Responsible human–robot coordination and the ethics of mixed traffic," *Ethics Inf. Technol.*, pp. 1–10, Jan. 2018.

[8] J. Axelsson, "Safety in vehicle platooning: A systematic literature review," *IEEE Trans. Intell. Transp. Syst.*, vol. 18, no. 5, pp. 1033–1045, May 2017.

[9] S. M. Casner, E. L. Hutchins, and D. Norman, "The challenges of partially automated driving," *Commun. ACM*, vol. 59, no. 5, pp. 70–77, 2016.

[10] M. C. Horowitz and P. Scharre, "Meaningful human control in weapon systems," Center for a New American Security, Washington, DC, USA, Working Paper, 2015.

[11] A. M. Pereira, H. Anany, O. Přibyl, and J. Přikryl, "Automated vehicles in smart urban environment: A review," in *Proc. Smart City Symp. Prague (SCSP)*, 2017, pp. 1–8.

[12] M. P. Hagenzieker *et al.*, "Interactions between cyclists and automated vehicles: Results of a photo experiment," *J. Transp. Safety Security*, vol. 12, no. 1, pp. 94–115, 2020.

[13] J. N. Velasco, H. Farah, B. van Arem, M. P. Hagenzieker, "Interactions between vulnerable road users and automated vehicles: A synthesis of literature and framework for future research," in *Proc. Road Safety Simulat. Int. Conf.*, vol. 2017, 2017, pp. 16–19.

[14] G. Mecacci and F. S. de Sio, "Meaningful human control as reason-responsiveness: The case of dual-mode vehicles," *Ethics Inf. Technol.*, vol. 22, no. 2, pp. 103–115, 2020.

[15] S. C. Calvert, D. D. Heikoop, G. Mecacci, and B. van Arem, "A human centric framework for the analysis of automated driving systems based on meaningful human control," *Theor. Issues Ergonom. Sci.*, vol. 21, no. 4, pp. 478–506, 2020.

[16] D. D. Heikoop, M. Hagenzieker, G. Mecacci, S. Calvert, F. S. de Sio, and B. van Arem, "Human behaviour with automated driving systems: A quantitative framework for meaningful human control," *Theor. Issues Ergonom. Sci.*, vol. 20, no. 6, pp. 711–730, 2019.

[17] F. Flemisch, D. Abbink, M. Itoh, M.-P. Pacaux-Lemoine, and G. We, "Shared control is the sharp end of cooperation: Towards a common framework of joint action, shared control and human machine cooperation," *IFAC PapersOnLine*, vol. 49, no. 19, pp. 72–77, 2016.

[18] B. Friedman, P. H. Kahn, Jr., A. Borning, and A. Huldtgren, "Value sensitive design and information systems," in *Early Engagement and New Technologies: Opening Up the Laboratory*. Dordrecht, The Netherlands: Springer, 2013, pp. 55–95.

[19] J. V. den Hoven, "ICT and value sensitive design," in *The Information Society: Innovation, Legitimacy, Ethics and Democracy in Honor of Professor Jacques Berleur SJ*. Boston, MA, USA: Springer, 2007, pp. 67–72.

[20] J. V. den Hoven, "Value sensitive design and responsible innovation," in *Responsible Innovation: Managing the Responsible Emergence of Science and Innovation in Society*, vol. 47. Hoboken, NJ, USA: Wiley, 2013, pp. 75–83.

[21] J. J. Bryson and A. Theodorou, "How society can maintain human-centric artificial intelligence," in *Human-Centered Digitalization and Services*. Singapore: Springer, 2019, pp. 305–323.

[22] R. H. Wortham and A. Theodorou, "Robot transparency, trust and utility," *Connection Sci.*, vol. 29, no. 3, pp. 242–248, 2017.

[23] A. Theodorou and V. Dignum, "Towards ethical and socio-legal governance in AI," *Nat. Mach. Intell.*, vol. 2, pp. 10–12, Jan. 2020.

[24] A. V. Wynsberghe, "Artificial intelligence: From ethics to policy," in *Panel for the Future of Science and Technology*. Brussels, Belgium: Eur. Parliamentary Res. Service, 2020.

[25] F. S. de Sio and G. Mecacci, "Four responsibility gaps with automated systems: Why they matter and how to address them," White Paper, 2020.

[26] F. Biondi, I. Alvarez, and K.-A. Jeong, "Human–vehicle cooperation in automated driving: A multidisciplinary review and appraisal," *Int. J. Human–Comput. Interact.*, vol. 35, no. 11, pp. 932–946, 2019.

[27] S. Noh and K. An, "Decision-making framework for automated driving in highway environments," *IEEE Trans. Intell. Transp. Syst.*, vol. 19, no. 1, pp. 58–71, Jan. 2018.

[28] F. Ficuciello, G. Tamburrini, A. Arezzo, L. Villani, and B. Siciliano, "Autonomy in surgical robots and its meaningful human control," *Paladyn, J. Behav. Robot.*, vol. 10, no. 1, pp. 30–43, 2019.

[29] F. S. de Sio and J. V. den Hoven, "Meaningful human control over autonomous systems: A philosophical account," *Front. Robot. AI*, vol. 5, p. 15, Feb. 2018.

[30] J. A. Michon, "A critical view of driver behavior models: What do we know, what should we do?" in *Human Behavior and Traffic Safety*. Boston, MA, USA: Springer, 1985, pp. 485–524.

[31] F. S. de Sio, "Ethics and self-driving cars: A white paper on responsible innovation in automated driving systems," Dept. Ethics Philosophy Technol., TU Delft, Delft, The Netherlands, White Paper, 2016.

SIMEON C. CALVERT received the M.Sc. and Ph.D. degrees in civil engineering, specialized in transport and planning, from the Delft University of Technology, The Netherlands, in 2010 and 2016, respectively, and where he is currently an Assistant Professor of traffic and network management and co-leads the Delft AI Lab on Urban Mobility Behaviour: CiTy-AI. From 2010 to 2016, he was a Research Scientist with TNO, Netherlands Organization for Applied Scientific Research. His research has focused on ITS, impacts of vehicle automation, traffic management, traffic flow theory and network analysis. Much of his recent research has involved various leadership roles in national and European research projects focusing on the application and impacts of vehicle automation and cooperation.

GIULIO MECACCI received the M.A. degree in philosophy of mind from the University of Siena, Italy, and the Ph.D. degree from Radboud University Nijmegen, Donders Institute for Brain, Cognition and Behavior, in the field of ethics of neurotechnology. After his Ph.D. studies, he spent three years with TU Delft working with psychologists and engineers on the multidisciplinary project "Meaningful Human Control over Automated Driving Systems." He is currently an Assistant Professor of ethics and philosophy of intelligent technology with the Department of Artificial Intelligence, Donders Institute for Brain, Cognition and Behaviour, Radboud University.