Feess, Eberhard; Muehlheusser, Gerd

Working Paper

Autonomous Vehicles: Moral Dilemmas and Adoption Incentives

CESifo Working Paper, No. 9825

Provided in Cooperation with:
Ifo Institute – Leibniz Institute for Economic Research at the University of Munich

Suggested Citation: Feess, Eberhard; Muehlheusser, Gerd (2022) : Autonomous Vehicles: Moral Dilemmas and Adoption Incentives, CESifo Working Paper, No. 9825, Center for Economic Studies and ifo Institute (CESifo), Munich

This Version is available at:
https://hdl.handle.net/10419/263755

CESifo WORKING PAPERS

9825
2022

July 2022

Autonomous Vehicles: Moral Dilemmas and Adoption Incentives

Eberhard Feess, Gerd Muehlheusser

CESifo

Impressum:

CESifo Working Papers
ISSN 2364-1428 (electronic version)
Publisher and distributor: Munich Society for the Promotion of Economic Research - CESifo GmbH
The international platform of Ludwigs-Maximilians University's Center for Economic Studies and the ifo Institute
Poschingerstr. 5, 81679 Munich, Germany
Telephone +49 (0)89 2180-2740, Telefax +49 (0)89 2180-17845, email office@cesifo.de
Editor: Clemens Fuest
https://www.cesifo.org/en/wp
An electronic version of the paper may be downloaded
· from the SSRN website: www.SSRN.com
· from the RePEc website: www.RePEc.org
· from the CESifo website: https://www.cesifo.org/en/wp

CESifo *Working Paper No. 9825*

Autonomous Vehicles: Moral Dilemmas and Adoption Incentives

Abstract

In unavoidable traffic accidents, autonomous vehicles (AVs) face the dilemma of protecting either the passenger(s) or third parties. Recent studies show that most people prefer AVs following a utilitarian approach by minimizing total harm. At the same time, however, they would adopt an AV only if it prioritizes the passenger(s), i.e. themselves. As AVs exhibit a lower accident risk in the first place, a regulator therefore faces a trade-off: the harm-minimizing behavior of AVs (ex post efficiency) hampers the willingness to adopt them (ex ante efficiency). Using a game-theoretic model, we analyze how the second-best optimal level of AV passenger protection depends on (i) the AV safety advantage, (ii) the intensity of drivers' social preferences, and (iii) their reluctance to adopt AVs. A higher AV safety advantage may either increase or decrease the second best optimal level of passenger protection.

JEL-Codes: O310, K230, L510, L620, R410.

Keywords: autonomous vehicles, ethical dilemma, trolley problem, adoption of new technologies, game theory.

Eberhard Feess
Victoria University of Wellington
School of Economics and Finance
Pipitea Campus, Lambton Quay
Wellington / New Zealand
eberhard.feess@vuw.ac.nz

Gerd Muehlheusser
University of Hamburg
Department of Economics
Von-Melle-Park 5
Germany – 20146 Hamburg
gerd.muehlheusser@uni-hamburg.de

July 1, 2022
We gratefully acknowledge the hospitality and financial support of the Center of Interdisciplinary Research (ZiF) at Bielefeld University (Research Group "Economic and Legal Challenges in the Advent of Smart Products").

1 Introduction

Autonomous Vehicles (AVs) are widely believed to become available in the not too distant future, and this movement towards autonomous mobility will have a huge impact on the automobile industry, consumers, policymakers and regulators.[1] One currently topical issue concerns the ethical dilemma of whether AVs should protect the passengers or third parties when accidents are inevitable (e.g., Greene, 2016, Goodall, 2016, Awad et al., 2018, Bigman and Gray, 2020, Bonnefon et al. 2020).

Surveys show that most people prefer AVs following a utilitarian approach by minimizing total harm, but, at the same time, are more willing to buy an AV if they protect the passenger(s) under most circumstances (Bonnefon et al. 2016, Gill 2021, Liu and Liu 2021). A utilitarian regulator therefore needs to compromise between ex post and ex ante efficiency: Ex post efficiency requires that AVs protect third parties whenever their (expected) harm exceeds the passengers' harm. This, however, reduces ex ante efficiency by discouraging drivers from adopting the safer AV technology, thereby increasing the accident risk itself.[2]

We analyze this trade-off in a game-theoretic model where a benevolent regulator (who sets the level of AV driver protection) interacts with drivers (who decide whether or not to adopt the AV). Drivers care about the harm to third parties, but to a lower degree than about their own. This accounts for the robust insight of behavioral economics that most people are neither fully selfish nor fully altruistic (see e.g. Fehr and Schmidt, 1999,

[1] See e.g. European Commission (2018) for an outline for the road ahead. Tesla already offers a "Full Self Driving" package since several years; however, drivers must always be ready to immediately take over control. According to the classification system of the Society of Automotive Engineers (SAE), this corresponds to autonomy level 2 (out of 5), see SAE International (2021). In 2021 Mercedes introduced its "Drive Pilot" system, where the human driver is not obliged to monitor the driving at all times, but must only be ready to take over after being prompted by the system (level 3). The system is currently approved for motorways and with a speed of up to 60 km/h. In June 2022, the *United Nations Economic Commission for Europe* (UNCE) has extended the maximum speed to 130 km/h (effective as of 2023) for vehicles which satisfy the respective requirements, see `https://unece.org/sustainable-development/press/un-regulation-extends-automated-driving-130-kmh-certain-conditions`.

[2] Human driving is often impaired by poor sight or slow reaction times due to fatigue, distraction, alcohol or drug consumption. For example, according to the 2008 National Motor Vehicle Crash Causation Survey (NHTSA), more than 93 percent of the analyzed crashes were classified as being caused by driver errors. By contrast, it is widely held that, once matured, AVs will be much safer than human-operated vehicles. For example, a recent McKinsey report predicts a drop in the crash frequency per vehicle of up to 90 per cent once there are sufficiently many AVs on the streets (`https://perma.cc/V7ZC-VL2T`). See also the discussions in Fagnant and Kockelman (2015), Luetge (2017) and Geistfeld (2017).

2006, Andreoni and Miller, 2002, Charness and Rabin, 2002). Furthermore, we assume that drivers differ in their willingness to adopt AVs, which is their private information. Such heterogeneity is widely documented in surveys (see e.g., Kyriakidis at al., 2015), and a reluctance to adopt could, for example, be due to the fear of liability or a low level of trust in AVs (see e.g. Adnan et al., 2018, Cunningham at al., 2019) or a preference to be involved when deciding in critical situations (Basu et al., 2019).

The aforementioned trade-off between ex ante and ex post efficiency is reflected in the equilibrium behavior of our model: drivers will not adopt the AV if it protects them too little, and the regulator takes this into account by choosing a (second-best optimal) level of AV driver protection that is too high from the perspective of ex post efficiency. We then analyze how the equilibrium outcome is influenced by (i) the safety of the AV (measured by the reduction in the accident risk compared to human-driven vehicles), (ii) the intensity of drivers' social preferences, and (iii) their reluctance to adopt an AV. As a higher AV safety advantage ceteris paribus leads to a higher willingness to adopt them, the regulator can induce the same adoption rate with a lower, and hence more efficient, level of driver protection (thereby saving third parties more often). However, the optimal level of driver protection may even increase with the AV safety advantage, in order to reach an even larger rate of AV adoption. A similar reasoning holds for the impact of drivers' social preferences and the willingness to adopt AV: Both factors allow for the same adoption rate with less driver protection, but it may again be better to target a higher adoption rate.

The relevance of situations where AV algorithms decide upon life and death, thereby basically resembling a *trolley problem* (Foot, 1967, Thomson, 1985), is debatable.[3] Some researchers (e.g., Nyholm and Smids, 2016, Dewitt et al. 2019, De Freitas et al. 2020, 2021, Lundgren, 2021) as well as AV practitioners such as Iagnemma (2018) argue that trolley-like scenarios are far-fetched. Others show that people care a lot about these issues, which hence need to be decided upon even when they are unlikely (Greene, 2016, Bonnefon et al., 2019, Gill 2020, Keeling 2020). In any case, our approach is not confined

[3]See Greene (2013) for a survey of the different variants of the trolley problem. Krügel and Uhl (2022) experimentally study trolley problems in stochastic environments.

to trolley-like dilemmas, but relevant whenever getting closer to the ex post optimal AV behavior reduces drivers' ex ante willingness to adopt them. For instance, the AV algorithm needs to decide about the speed under adverse weather conditions, whether to stop when traffic lights turn yellow, or how defensively to act when meeting other vehicles at junctions. In such situations, there is typically a difference between a driver's preferred AV behavior and the socially optimal one.

The remainder of the paper is structured as follows: The theoretical framework is laid out in Section 2. Section 3 derives the main results. Section 4 provides further discussion and concludes. All proofs are in the Appendix.

2 The model

We study a dynamic game between two players: a regulator and a driver. There are two types of vehicles, human-driven (HV) and autonomous (AVs), both of which can be involved in accidents. Without loss of generality, we set the accident probability of the HV equal to 1, while for the AV it is $(1-\pi)$, so that $\pi \in [0,1)$ expresses the safety advantage of the AV. If the vehicle swerves in case of an accident, the third party remains unharmed, but the driver suffers harm 1. When the vehicle does not swerve, the driver remains unharmed, but the third party suffers harm h, a random variable that is continuously distributed on $[0, h^{max}]$ with distribution $F(h)$ and strictly positive density $F'(h)$ everywhere.[4] The realization of h is learned before the swerving decision. An HV swerves when the driver decides to do so, while an AV swerves if the third party harm exceeds a threshold $\tilde{h}$ set ex ante by the regulator and programmed into the AVs operating system. Throughout, we interpret $\tilde{h}$ as the level of driver (or passenger) protection provided by the AV.

The driver has social preferences and puts weight $\theta \in (0,1)$ on third-party harm. Furthermore, we capture the reluctance of (some) individuals to adopt the AV by assuming that driver type $\beta \in \{L, H\}$ faces a cost ω_β upon adoption, where $\omega_H > \omega_L = 0$. The driver's type is their private information, and the ex ante probabilities for the low and

[4] We could easily allow the driver to also suffer a harm upon swerving. All we need that is that this harm is smaller compared to the harm when not swerving.

the high cost type are α and $(1-\alpha)$, and common knowledge.

The (utilitarian) regulator aims at minimizing the expected harm from accidents. As swerving yields harm 1 to the driver and no harm to third parties, it is ex post optimal to swerve if and only if $h \geq \tilde{h}^f = 1$. Furthermore, as $\pi \geq 0$, ex ante efficiency requires that both driver types adopt the AV. These two requirements express the tension between ex post and ex ante efficiency discussed in the introduction, as setting $\tilde{h} = \tilde{h}^f$ may hamper the adoption of the AV.

The timing of the game is as follows: At stage 1, the regulator sets the level of driver protection $\tilde{h}$ that the AV must be programmed to follow. At stage 2, the driver chooses between the HV and the AV.[5] At stage 3, nature decides on the occurrence of an accident and on the corresponding harm h to the third party. At stage 4, the swerving decision is taken either by the driver (in the HV) or by the AV according to its pre-specified rule $\tilde{h}$.

3 Analysis

For determining the equilibrium behavior, we proceed backwards, starting with the swerving decision at stage 4.

3.1 Stage 4: Swerving

The AV swerves if the third party harm exceeds the pre-programmed threshold set by the regulator, i.e. if $h > \tilde{h}$. The HV driver (who puts weight θ on third party harm) compares the own harm of 1 upon swerving to the internalized third party harm θh when not swerving, so swerving occurs if $h > \hat{h} := \frac{1}{\theta}$. Note that the two driver types might choose different vehicles, but they do not differ in their swerving decision when choosing the HV. Importantly, due to $\theta < 1$, we have $\hat{h} > \tilde{h}^f = 1$, i.e. an HV driver swerves less often than would be socially optimal.

[5] We abstract from vehicle prices and other factors that might influence the driver's choice between the AV and the HV.

3.2 Stage 2: The driver's vehicle choice

Taking the swerving decision into account, the expected cost of driver type β is

$$c^H = \int_0^{\hat{h}} \theta h F'(h) dh + \left(1 - F(\hat{h})\right), \tag{1}$$

and

$$c^A(\beta) = (1-\pi) \cdot \left[\int_0^{\tilde{h}} \theta h F'(h) dh + 1 - F(\tilde{h})\right] + \omega_\beta \tag{2}$$

when choosing the HV and the AV, respectively. The first (second) term refers to those realizations of h where swerving does not occur (occurs). Define $\Delta_\beta(\tilde{h}) = c^H - c^A(\beta)$ as the difference in expected costs between the two vehicles, so driver type β adopts the AV iff $\Delta_\beta(\tilde{h}) \geq 0$. The following assumption keeps the subsequent analysis analytically tractable and yet sufficiently rich to derive our main points:

Assumption 1.

(i) $\Delta_\beta(\tilde{h})$ *is strictly concave over* $[0, h^{max}]$.[6]

(ii) $\Delta_L(\tilde{h} = 0) < 0$ *and* $\Delta_H(\tilde{h} = \hat{h}) > 0$.

The assumption is illustrated in Figure 1 below. Part (i) ensures a unique interior solution for the optimal level of AV driver protection from the driver's viewpoint. Part (ii) ensures that both driver types strictly prefer the HV if the AV never protects the driver ($\tilde{h} = 0$), and that there exists $\tilde{h}$ such that both types strictly prefer the AV. Note that each of the two conditions in part (ii) is less restrictive for the respective other type. We can now state our main result concerning the driver's optimal vehicle choice:

Proposition 1.

(i) For any given level of driver protection, the incentive to adopt the AV is higher for the low adoption cost type than for the high cost type, i.e., $\Delta_L(\tilde{h}) > \Delta_H(\tilde{h})\ \forall \tilde{h} \geq 0$.

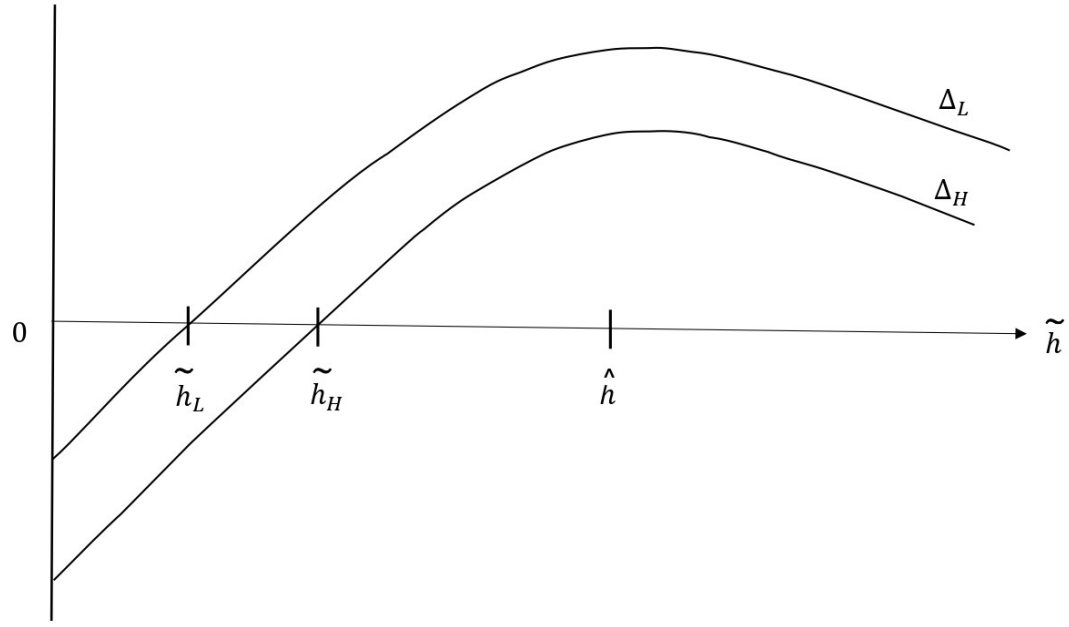
[6] This holds if $-(1-\pi)\left[F''(\tilde{h})\left(\theta\tilde{h} - 1\right) + F'(\tilde{h})\theta\right] < 0$ for all $\tilde{h}$. A sufficient condition would be that F is uniform.

(ii) For both driver types, the incentive to adopt the AV is highest when it exhibits the same level of driver protection as optimally chosen by the driver in the HV, i.e., both $\Delta_L(\tilde{h})$ and $\Delta_H(\tilde{h})$ are maximized at $\tilde{h} = \hat{h}$.

(iii) There exist two thresholds $\tilde{h}_L$ and $\tilde{h}_H$ satisfying $\Delta_L(\tilde{h}_L) = 0$ and $\Delta_H(\tilde{h}_H) = 0$, and $\tilde{h}_L < \tilde{h}_H < \hat{h}$.

(iv) The driver's incentive to adopt the AV is the higher, the larger the AV safety advantage and the stronger the driver's social preferences, i.e., $\tilde{h}_L$ and $\tilde{h}_H$ both decrease with π and θ. Moreover, type H's incentive to adopt the AV decreases in her adoption costs, i.e., $\tilde{h}_H$ increases with ω_H.

Figure 1: Optimal vehicle choice by driver type



The proposition is also illustrated in Figure 1. Part (i) follows immediately from the fact that type H has higher costs from adopting the AV, $\omega_H > \omega_L = 0$. Moreover, for both driver types the incentive to adopt the AV is largest when it mimics their own preferred swerving behavior (part (ii)). However, due to the AV's safety advantage, both driver types would adopt it even when it protects them less often, where $\tilde{h}_H < \tilde{h}_L$ follows again from $\omega_H > \omega_L = 0$ (part (iii)). The comparative statics properties of part (iv) are also intuitive: The larger the AV safety advantage, the more the driver cares about others,

and the lower the adoption costs ω_H, the lower is the level of driver protection required to induce adoption of the AV.[7]

3.3 Stage 1: The regulator's optimal choice of driver protection

When setting the level of driver protection $\tilde{h}$, the regulator faces a trade-off: The closer $\tilde{h}$ is to the ex post efficient level $\tilde{h}^f$, the lower is the expected harm from accidents involving the AV. However, such AV behavior reduces the demand for AVs because both driver types prefer a higher degree of driver protection. In our framework with two driver types who differ in their willingness to adopt the AV, the regulator hence chooses between only two options: either implementing an (inefficiently) high level of driver protection that leads to *full adoption* of the AV along the equilibrium path (i.e. both driver types adopt it), or a lower level of driver protection that leads to *partial adoption* by the low cost type only.[8]

From Proposition 1(iii), it follows that full adoption requires a level of driver protection $\tilde{h}^F \in [\tilde{h}_H, \hat{h}]$, leading to expected social costs of

$$SC^F(\tilde{h}^F) = (1-\pi) \cdot \left[\int_0^{\tilde{h}^F} hF'(h)dh + 1 - F(\tilde{h}^F) \right]. \tag{3}$$

For all $\tilde{h}^P \in [\tilde{h}_L, \tilde{h}_H)$, we get an equilibrium with partial adoption, where driver type L chooses the AV, while type H prefers the HV and swerves if $h \geq \hat{h}$. The resulting expected social costs are

$$SC^P(\tilde{h}^P) = \alpha(1-\pi) \cdot \left[\int_0^{\tilde{h}^P} hF'(h)dh + 1 - F(\tilde{h}^P) \right] + (1-\alpha) \cdot \left[\int_0^{\hat{h}} hF'(h)dh + 1 - F(\hat{h}) \right]. \tag{4}$$

Whether the regulator prefers an equilibrium with full or partial adoption depends on the model's parameters. There are three cases that differ by which driver types adopt an

[7] For notational convenience, the dependence of $\tilde{h}_H$ and $\tilde{h}_L$ on π, θ and ω is often suppressed.

[8] When determining the second-best optimal level of driver protection, we can safely confine attention to the range $\tilde{h} \leq \hat{h}$. The reason is that, in our framework, providing a higher level of driver protection than the driver wants can never be optimal.

AV with the ex post optimal level of driver protection, $\tilde{h} = \tilde{h}^f$ (see also Table 1 below). In case I, the safety advantage π of an AV is so large that both driver types adopt it for $\tilde{h} = \tilde{h}^f$; notwithstanding that the level of driver protection is smaller than the driver's preferred level $\hat{h}$. For intermediate values of π (case II), only the low adoption cost type L chooses the AV for $\tilde{h} = h^f$. Finally, for π small (case III), neither driver type adopts the AV with $\tilde{h} = \tilde{h}^f$.

Formally, the three cases are delineated as follows: Define π_L and π_H such that $\tilde{h}_L = \tilde{h}^f$ at $\pi = \pi_L$ and $\tilde{h}_H = \tilde{h}^f$ at $\pi = \pi_H$, where $\pi_L < \pi_H$ holds (this follows from $\tilde{h}_L < \tilde{h}_H$ and the fact that these thresholds are decreasing in π, see parts (iii) and (iv) of Proposition 1). Cases I, II and III are then given by the intervals $\pi \geq \pi_H$, $\pi \in (\pi_L, \pi_H]$ and $\pi < \pi_L$, respectively. To ensure a non-empty parameter range for each of the three cases, we assume $\tilde{h}_H(\pi = 1) < \tilde{h}_f < \tilde{h}_L(\pi = 0)$. This yields the following results concerning the regulator's optimal policy (see Table 1 for an illustration):

Proposition 2.

(i) If the AV safety advantage is large ($\pi \geq \pi_H$, case I), the regulator sets $\tilde{h} = \tilde{h}^f$ and both driver types adopt the AV. For $\pi < \pi_H$, the regulator either induces full adoption by setting $\tilde{h} = \tilde{h}_L$ or partial adoption (by type L only) by setting $\tilde{h} = \tilde{h}^f$ in case II and $\tilde{h} = \tilde{h}_H$ in case III.

(ii) Inducing full adoption is the more attractive for the regulator, the lower the frequency of the low adoption cost type (α). The effect of the AV safety advantage (π) and the driver's social preferences (θ) on the optimal choice between full and partial adoption are ambiguous.

Part (i) summarizes the regulator's candidate optimal regulations for the three cases. Case I is straightforward, as the AV safety advantage is sufficiently large to induce full AV adoption even for $\tilde{h}^f$, which yields both ex ante and ex post efficiency.[9] In cases II and III, ex ante and ex post efficiency cannot be reached at the same time. Then, the

[9]Note that the optimal level of driver protection inducing partial adoption would be $\tilde{h}_H - \epsilon$ where $\epsilon > 0$, which is (by definition) dominated by the efficient choice $\tilde{h}^f$.

Table 1: Optimal AV regulation depending on AV safety advantage and type of adoption

Case	AV safety advantage	Adopt AV with $\tilde{h} = \tilde{h}^f$?	Optimal AV regulation:	
			Full adopt.	Partial adopt.
I: $\tilde{h}_H \leq \tilde{h}^f$	large $\pi \geq \pi_H$	both types	$\tilde{h} = \tilde{h}^f$	dominated
II: $\tilde{h}_L < \tilde{h}^f \leq \tilde{h}_H$	intermediate $\pi \in (\pi_L, \pi_H]$	only type L	$\tilde{h} = \tilde{h}_H$	$\tilde{h} = \tilde{h}^f$
III: $\tilde{h}^f < \tilde{h}_H$	small $\pi < \pi_L$	neither type	$\tilde{h} = \tilde{h}_H$	$\tilde{h} = \tilde{h}_L$

regulator either induces full adoption by setting the lowest level of driver protection that ensures adoption also by the high cost type ($\tilde{h} = \tilde{h}_H$), or partial adoption by cost type L only. By definition of case II, type L adopts the AV for $\tilde{h}^f$, which is hence implemented by the regulator. However, for lower levels of AV safety advantage, the regulator needs to adjusts the level of driver protection upwards to $\tilde{h}_H$ in order to induce adoption by type L (case III).

Part (ii) considers the impact of the key model parameters on the regulator's optimal choice between partial and full adoption.[10] The only straightforward effect concerns the frequency of driver types α: The more prevalent type H (i.e. the lower α), who does not adopt the AV under partial adoption, the more desirable is full adoption for the regulator.

By contrast, the impact of the AV safety advantage (π) is clear-cut only in case II, where full adoption becomes more desirable as π increases. The reason is that, in case II, the optimal level of AV driver protection $\tilde{h}^f$ (which is independent of π) is chosen anyway. Under partial adoption, the only positive impact of higher π is hence the lower accident probability, which matters only for the low cost type (who adopts the AV). By contrast, under full adoption, the benefit of an increase in π is higher for two reasons: First, it affects both driver types. Second, it allows for a lower (and hence closer to the ex post efficient) level of driver protection $\tilde{h}_H$ (recall from Proposition 1(iv) that $\tilde{h}_H$ is decreasing in π). In case III, however, the optimal level of driver protection under partial adoption, $\tilde{h}_L$, also decreases with π, thereby moving closer $\tilde{h}^f$ and reducing the expected

[10] Thereby, we can restrict attention to cases II and cases III, as case I yields full adoption with the ex post optimal driver protection $\tilde{h}^f$.

accident costs under partial adoption. This countervailing force renders the overall effect of π ambiguous.

The impact of the driver's social preferences (θ) on the optimal choice between partial and full adoption is also ambiguous. Intuitively, an increase in θ reduces the expected social costs under both partial and full adoption: On the one hand, a higher θ makes the AV more attractive, in particular for driver type H (recall from Proposition 1(iv) that $\tilde{h}_H$ decreases in θ). This brings the level of driver protection under full adoption, $\tilde{h}_H$, closer to the ex post efficient level, $\tilde{h}^f$. On the other hand, a higher θ also increases the degree of ex post efficiency of the driver's swerving decision in the HV ($\hat{h} = 1/\theta$), which reduces the social costs under partial adoption.

A straightforward next question is whether improvements in AV safety lead to a higher or lower level of driver protection under the optimal regulation. Our previous results suggest that this is ambiguous as well: Consider first a (small) increase in π which does *not* affect the regulator's optimal choice between partial or full adoption. In this case, the increase in π will lead the regulator to set a (weakly) lower level of driver protection (recall from Proposition 1(iv) that both $\tilde{h}_H$ and $\tilde{h}_L$ are decreasing in π). Then, third parties benefit from a higher π in two ways: there are fewer accidents and they are more often protected in case of an accident. However, we know from Proposition 2 (i) that an increase of π may trigger a regime change from partial to full adoption, which then entails a higher level of driver protection. For example, in case II, the regime change would lead to a switch from $\tilde{h}^f$ under partial to $\tilde{h}_H > \tilde{h}^f$ under full adoption. We summarize this discussion as follows:

Corollary 1. *An increase in the AV safety advantage (π) can increase or decrease the second best optimal level of driver protection.*

4 Conclusion

In moral philosophy, the question whether Autonomous Vehicles (AVs) should protect its passengers or third parties is often characterized as an ethical dilemma in the tradition

of the *trolley problem* (Fort, 1967). From a utilitarian perspective, the problem can be interpreted as the trade-off between ex ante and ex post efficiency: Ex post efficiency requires that, if an accident is unavoidable, an AV does not protect its passenger(s) when their (expected) harm is lower than that of third parties. By contrast, ex ante efficiency increases in the adoption rate of the (safer) AV, which tends to be higher when the AV prioritizes the owner/passenger(s) even in situations where the third party harm would be larger. We have developed a simple game-theoretic model to analyze how the second best optimal level of AV driver protection, taking both ex ante and ex post efficiency into account, depends on the AV safety advantage, individuals' social preferences (i.e., their concern about the harm suffered by third parties), and their willingness to adopt a new technology such as an AV, which comes along with a loss of control in critical situations.

Our first and most straightforward finding is that the second best optimal level of driver protection increases with the frequency of drivers who are reluctant to adopt the AV. The higher this frequency, the more important is it that also these drivers adopt the AV, and the detrimental impact of the required higher level of driver protection on the harm actually realized in case of an accident is then likely to be overcompensated by the overall AV safety advantage (which makes such accidents less likely in the first place). While we derive this result in a model with two driver types only, it should be robust also with a continuous framework. The impact of the other model parameters, however, is ambiguous: On the one hand, if drivers care more about harm to other people, then they are ceteris paribus more willing to adopt the AV, which allows the regulator to implement a lower (and hence more efficient) level of AV driver protection without hampering adoption incentives. On the other hand, it may now be welfare-enhancing to induce adoption also by individuals with high adoption costs, and achieving this may require to increase the level of driver protection. A similar result emerges for the impact of a higher AV safety advantage. Again, this allows ceteris paribus for a efficiency-improving reduction in driver protection, but even increasing the level of driver protection may now be superior if this leads to a substantially higher adoption rate.

Our model focusses on (dilemma) situations where accidents are unavoidable, thereby

taking as given the level of AV safety (i.e. the probability of an accident). This is a limitation of our model, as the regulatory environment for AVs will also affect producers' incentives to make (costly) investments in AV safety which, in turn, will also affect consumers' adoption decisions. In this respect, using a dynamic model framework, Dawid and Muehlheusser (2022) show that the incentives to invest in AV safety is positively related to the number of AVs sold. In our model, taking this aspect into account would presumably reinforce the regulator's incentive to choose a level of AV driver protection, thereby enhancing the industry's innovation incentives via a higher rate of adoption.

The criteria for setting the level of protection for the driver/passenger in an AV in case of an accident are currently not only discussed by moral philosophers, but also at the regulatory level. In a recent expert report for the European Commission (see Horizon 2020), it is argued that fundamental principles of fairness require that no subgroup should be more vulnerable with AVs than with HVs. Subgroups refer both to demographics such as age and the role in traffic (car drivers, pedestrians and cyclists). However, the report keeps silent about an issue that seems important in the light of our analysis: Does the notion of fairness apply solely to the ex post perspective (i.e., when accidents are unavoidable) or to the overall perspective (i.e. including the ex ante safety advantage of AVs which makes accident s less likely to occur)? To see why this distinction might matter, suppose that AVs are only predominantly used if their safety advantage is relatively large. In this case, even more vulnerable road users will might benefit overall from the presence of AVs. Hence, from this overall perspective, requiring that AVs do not impose a higher expected harm on any subgroup seems unproblematic. However, from the perspective of ex post efficiency (i.e. when accidents are unavoidable and the only the decision to be taken is who will be harmed), the overall harm may well be lower when it accrues to someone from a vulnerable subgroup (e.g. a cyclist or pedestrian) rather than the AV passenger(s). This ex post perspective may hence give rise to a trade-off between the fairness perspective emphasized in the expert EU report and the utilitarian objective of minimizing overall expected harm. This trade-off can be analyzed similarly to the one between ex ante and ex post efficiency discussed in our paper.

References

Adnan, N., S. Nordin, M. A. bin Bahruddin, and M. Ali (2018): "How trust can drive forward the user acceptance to the technology? In-vehicle technology for autonomous vehicle," *Transportation Research Part A: Policy and Practice*, 118, 819–836.

Andreoni, J. and J. Miller (2002): "Giving according to GARP: An experimental test of the consistency of preferences for altruism," *Econometrica*, 70, 737–753.

Awad, E., S. Dsouza, R. Kim, J. Schulz, J. Henrich, A. Shariff, J.-F. Bonnefon, and I. Rahwan (2018): "The moral machine experiment," *Nature*, 563, 59–64.

Basu, C., M. Madan, and B. Sankaranarayanan (2019): "Are driverless cars truly a reality? A technical, social and ethical analysis," *Issues in Information Systems*, 20, 56–64.

Bigman, Y. and K. Gray (2020): "Life and death decisions of autonomous vehicles," *Nature*, 579, E1–E2.

Bonnefon, J.-F., A. Shariff, and I. Rahwan (2016): "The social dilemma of autonomous vehicles," *Science*, 352, 1573–1576.

——— (2019): "The trolley, the bull bar, and why engineers should care about the ethics of autonomous cars," *Proceedings of the IEEE*, 107, 502–504.

——— (2020): "The moral psychology of AI and the ethical opt-out problem," in *Ethics of Artificial Intelligence*, ed. by S. M. Liao, Oxford, UK: Oxford University Press, 109–126.

Charness, G. and M. Rabin (2002): "Understanding social preferences with simple tests," *Quarterly Journal of Economics*, 117, 817–869.

Cunningham, M., M. Regan, T. Horberry, K. Weeratunga, and V. Dixit (2019): "Public opinion about automated vehicles in Australia: Results from a large-scale national survey," *Transportation Research Part A: Policy and Practice*, 129, 1–18.

Dawid, H. and G. Muehlheusser (2022): "Smart products: Liability, investments in product safety, and the timing of market introduction," *Journal of Economic Dynamics and Control*, 134, 104288.

De Freitas, J., S. E. Anthony, A. Censi, and G. A. Alvarez (2020): "Doubting driverless dilemmas," *Perspectives on Psychological Science*, 15, 1284–1288.

De Freitas, J., A. Censi, B. W. Smith, L. Di Lillo, S. E. Anthony, and E. Frazzoli (2021): "From driverless dilemmas to more practical commonsense tests for automated vehicles," *Proceedings of the National Academy of Sciences*, 118, e2010202118.

Dewitt, B., B. Fischhoff, and N.-E. Sahlin (2019): "'Moral machine' experiment is no basis for policymaking," *Nature*, 567, 31–32.

European Commission (2018): "On the road to automated mobility: An EU strategy for mobility of the future," *Report COM(2018) 283*.

Fagnant, D. and K. Kockelman (2015): "Preparing a nation for autonomous vehicles: opportunities, barriers and policy recommendations," *Transportation Research Part A: Policy and Practice*, 77, 167–181.

Fehr, E. and K. Schmidt (2006): "The economics of fairness, reciprocity and altruism–experimental evidence and new theories," *Handbook of the Economics of Giving, Altruism and Reciprocity*, 1, 615–691.

Fehr, E. and K. M. Schmidt (1999): "A theory of fairness, competition, and cooperation," *Quarterly Journal of Economics*, 114, 817–868.

Foot, P. (1967): "The problem of abortion and the doctrine of the double effect," *Oxford Review*, 5, 1–7.

Geistfeld, M. A. (2017): "A roadmap for autonomous vehicles: State tort liability, automobile insurance, and federal safety regulation," *California Law Review*, 105, 1611–1694.

Gill, T. (2021): "Ethical dilemmas are really important to potential adopters of autonomous vehicles," *Ethics and Information Technology*, 23, 657–673.

Goodall, N. J. (2016): "Can you program ethics into a self-driving car?" *IEEE Spectrum*, 53, 28–58.

Greene, J. (2013): *Moral tribes: Emotion, reason, and the gap between us and them*, Atlantic Books, London.

Greene, J. D. (2016): "Our driverless dilemma," *Science*, 352, 1514–1515.

Horizon 2020 (2020): *Commission Expert Group to advise on specific ethical issues raised by driverless mobility (E03659). Ethics of Connected and Automated Vehicles: Recommendations on road safety, privacy, fairness, explainability and responsibility*, Publication Office of the European Union, Luxembourg.

Iagnemma, K. (2018): "Why we have the ethics of self-driving cars all wrong," *World Economic Forum, https://www.weforum.org/agenda/2018/01/why-we-have-the-ethics-of-self-driving-cars-all-wrong*.

Keeling, G. (2020): "Why Trolley Problems Matter for the Ethics of Automated Vehicles," *Science and Engineering Ethics*, 26, 293–307.

Krügel, S. and M. Uhl (2022): "Autonomous vehicles and moral judgments under risk," *Transportation Research Part A: Policy and Practice*, 155, 1–10.

Kyriakidis, M., R. Happee, and J. de Winter (2015): "Public opinion on automated driving: Results of an international questionnaire among 5000 respondents," *Transportation Research Part F: Traffic Psychology and Behaviour*, 32, 127–140.

Liu, P. and J. Liu (2021): "Selfish or Utilitarian Automated Vehicles? Deontological Evaluation and Public Acceptance," *International Journal of Human–Computer Interaction*, 37, 1231–1242.

Luetge, C. (2017): "The German ethics code for automated and connected driving," *Philosophy & Technology*, 30, 547–558.

Lundgren, B. (2021): "Safety requirements vs. crashing ethically: what matters most for policies on autonomous vehicles," *AI & SOCIETY*, 36, 405–415.

Nyholm, S. and J. Smids (2016): "The ethics of accident-algorithms for self-driving cars: An applied trolley problem?" *Ethical Theory and Moral Practice*, 19, 1275–1289.

SAE International (2021): "Taxonomy and Definitions for Terms Related to Driving Automation Systems for On-Road Motor Vehicles," *Report J3016-202104*.

Thomson, J. J. (1985): "The trolley problem," *The Yale Law Journal*, 94, 1395–1415.

Appendix

A Proof of Proposition 1

Part (i). $\Delta_\beta(\tilde{h}) = c^H - c^A(\beta)$, with c^H and $c^A(\beta)$ as given in (1) and (2), is the (driver type dependent) difference in expected costs with the HV and AV. As only type H faces a cost $\omega_H > 0$ from adopting the AV, we get

$$\Delta_H(\tilde{h}) - \Delta_L(\tilde{h}) = c^H - c^A(H) - (c^H - c^A(L)) = \omega_H > 0 \quad \forall \tilde{h} \geq 0.$$

Part (ii). Consider $\Delta_L(\tilde{h})$, and observe from (1) and (2) that only c^A depends on $\tilde{h}$ while c^H does not. Moreover, $\frac{\partial \Delta_\beta(\tilde{h})}{\partial \tilde{h}} = -(1-\pi)F'(\tilde{h})(\theta\tilde{h} - 1)$ is zero at $\tilde{h} = \frac{1}{\theta} = \hat{h}$ for all $\beta = L, H$. As $\Delta_L(\tilde{h})$ is assumed to be strictly concave on $[0, h^{max}]$ (see Assumption 1(i)), it attains a unique maximum at $\tilde{h} = \hat{h}$. The proof for $\Delta_H(\tilde{h})$ is analogous, as $\Delta_L(\tilde{h})$ and $\Delta_H(\tilde{h})$ differ only by a constant, ω_H.

Part (iii). By Assumption 1(ii), $\Delta_L(\tilde{h} = 0) < 0$ and $\Delta_H(\tilde{h} = 0) < 0$, so both driver types prefer the HV for sufficiently small $\tilde{h}$. Recall from part (i) that $\Delta_H(\tilde{h})$ and $\Delta_L(\tilde{h})$ attain their unique maximum at $\tilde{h} = \hat{h}$, and are hence strictly increasing in $\tilde{h}$ to the left of $\hat{h}$. Moreover, also by Assumption 1(ii), $\Delta_H(\tilde{h} = \hat{h}) > 0$. It then follows from the intermediate value theorem that there exists a unique $\tilde{h}_H < \hat{h}$ where $\Delta_H(\tilde{h} = \tilde{h}_H) = 0$. Hence, in the interval $[0, \hat{h}]$ the HV is chosen for all $\tilde{h} < \tilde{h}_H$ and the AV for all $\tilde{h} \in [\tilde{h}_H, \hat{h}]$. An analogous argument applies to $\Delta_L(\tilde{h})$. In particular, as shown in part (i), $\Delta_L(\tilde{h}) > \Delta_H(\tilde{h})$ $\forall \tilde{h}$. Therefore, the value $\tilde{h}_L$ where the low type is indifferent between the two vehicles is strictly to the left of $\tilde{h}_H$.

Part (iv). We first show that $\tilde{h}_L$ and $\tilde{h}_H$ decrease with π. Consider first $\tilde{h}_L$. From the identity $\Delta_L(\tilde{h}_L) \equiv 0$ and the implicit function theorem, we get

$$\frac{d\tilde{h}_L}{d\pi} = -\frac{\frac{\partial \Delta_L}{\partial \pi}}{\frac{\partial \Delta_L}{\partial \tilde{h}_L}} = -\frac{\int_0^{\tilde{h}_L} \theta h F'(h) dh + 1 - F(\tilde{h}_L)}{(1-\pi) \cdot \left[F'(\tilde{h}_L)(1 - \theta\tilde{h}_L)\right]} < 0.$$

The numerator is positive. The denominator is also positive since $\tilde{h}_L < \hat{h} = \frac{1}{\theta}$ (see part (iii)) and, hence, $1 - \theta\tilde{h}_L > 0$. Thus, if the safety advantage of the AV increases, the AV is already preferred for smaller levels of AV driver protection $\tilde{h}$. The proof for $\tilde{h}_H$ proceeds analogously.

Consider next the impact of θ on $\tilde{h}_L$ and $\tilde{h}_H$, starting again with $\tilde{h}_L$. From $\Delta_L(\tilde{h}_L) \equiv 0$ and using $\hat{h} = 1/\theta$, we get

$$\frac{d\tilde{h}_L}{d\theta} = -\frac{\frac{\partial \Delta_L}{\partial \theta}}{\frac{\partial \Delta_L}{\partial \tilde{h}_L}} = -\frac{\int_0^{1/\theta} hF'(h)dh - (1-\pi)\int_0^{\tilde{h}_L} hF'(h)dh}{(1-\pi)\left[F'(\tilde{h}_L)(1-\theta\tilde{h}_L)\right]} < 0.$$

From $\frac{1}{\theta} = \hat{h} > \tilde{h}_L$ (see part (iii)), we know that the first integral is larger than the second, which, together with $\pi \leq 1$, ensures that the numerator is positive. The denominator is also positive since $\frac{1}{\theta} = \hat{h} > \tilde{h}_L$ implies that $1 - \theta\tilde{h}_L > 0$. The proof for $\tilde{h}_H$ proceeds analogously.

Finally, consider the impact of the AV adoption cost $\omega_H > 0$ for type H. Using again the implicit function theorem, we get

$$\frac{d\tilde{h}_H}{d\omega_H} = -\frac{\frac{\partial \Delta_H}{\partial \omega_H}}{\frac{\partial \Delta_H}{\partial \tilde{h}_H}} = \frac{1}{(1-\pi)\cdot\left[F'(\tilde{h}_H)(1-\theta\tilde{h}_H)\right]} > 0$$

which is due to $\frac{1}{\theta} = \hat{h} > \tilde{h}_H$. ■

B Proof of Proposition 2

Recall first from Proposition 1(iv) that $\tilde{h}_H$ and $\tilde{h}_L$ are strictly decreasing in π. This yields a threshold π_β such that driver type $\beta \in \{L, H\}$ adopts an AV with $\tilde{h} = \tilde{h}^f = 1$ iff $\pi \leq \pi_\beta$. Moreover, it follows from Proposition 1 that $\pi_L < \pi_H$. This gives rise to the three cases summarized in Table 1.

Part (i). In case I, setting $\tilde{h} = \tilde{h}^f$ yields full adoption and hence both ex post *and* ex ante efficiency, which minimizes the expected accident costs by definition of optimality.

Part (ii). In cases II and III, to induce full adoption, the regulator optimally chooses

the smallest feasible (i.e. closest to the ex post optimal) level of driver protection $\tilde{h} = \tilde{h}_H$. The expected social costs are

$$SC^F(\tilde{h}_H) = (1-\pi) \cdot \left[\int_0^{\tilde{h}_H} hF'(h)dh + 1 - F(\tilde{h}_H) \right].$$

To induce partial adoption, the regulator's optimal choice is $\tilde{h} = \tilde{h}^f$ for case II and $\tilde{h} = \tilde{h}_L$ for case III. The expected social costs are

$$SC^P(H) = \alpha \cdot (1-\pi) \cdot \left[\int_0^{H} hF'(h)dh + 1 - F(H) \right] + (1-\alpha) \cdot \left[\int_0^{\hat{h}} hF'(h)dh + 1 - F(\hat{h}) \right]$$

where $H \in \{\tilde{h}^f, \tilde{h}_L\}$.

Whether the regulator prefers an equilibrium with full oder partial adoption depends on the parameter setting and the distribution $F(h)$. Define the difference in expected social accident costs as $Z(H, \pi, \alpha, \theta) := SC^F(\tilde{h}_H) - SC^P(H)$. The higher $Z(\cdot)$, the more attractive is partial compared to full adoption. We next consider the impact of the AV safety advantage (π), the share of H-types in the population (α), and the intensity of drivers' social preferences (θ).

Effect of probability for H-type (α) In case II, we have $H = \tilde{h}^f$ and get

$$\frac{\partial Z(\tilde{h}^f, \cdot)}{\partial \alpha} = -(1-\pi) \cdot \left[\int_0^{\tilde{h}^f} hF'(h)dh + 1 - F(\tilde{h}^f) \right] + \int_0^{\hat{h}} hF'(h)dh + 1 - F(\hat{h}) > 0,$$

where the sign follows from $\pi \leq 1$ and $\hat{h} > \tilde{h}^f$ (where expected social harm is minimized) by definition of the parameter range for case II.

In case III, we have $H = \tilde{h}_H$ and hence get

$$\frac{\partial Z(\tilde{h}_L, \cdot)}{\partial \alpha} = -(1-\pi) \cdot \left[\int_0^{\tilde{h}_L} hF'(h)dh + 1 - F(\tilde{h}_H) \right] + \int_0^{\hat{h}} hF'(h)dh + 1 - F(\hat{h}) > 0,$$

since $\pi \leq 1$ and $\tilde{h}^f < \tilde{h}_H < \hat{h}$ by definition of case III, i.e. $\tilde{h}_H$ is closer to the ex post efficient level $\tilde{h}^f$ so that expected social costs there are lower than at $\tilde{h} = \hat{h}$.

Effect of AV safety advantage (π) In case II, we have $H = \tilde{h}^f$ and $\frac{\partial Z(\tilde{h}^f,\cdot)}{\partial \pi}$ is given by

$$-\left[\int_0^{\tilde{h}_H} hF'(h)dh + 1 - F(\tilde{h}_H)\right] + (1-\pi)\left[F'(\tilde{h}_H)\tilde{h}'_H(\pi)(\tilde{h}_H - 1)\right] + \alpha\left[\int_0^{\tilde{h}^f} hF'(h)dh + 1 - F(\tilde{h}^f)\right] < 0$$

Note first that, in case II, $\tilde{h}_H > \tilde{h}^f = 1$ holds. Together with $\alpha < 1$, this implies that the (negative) first term is larger in absolute terms than the (positive) third term. Moreover, the second term is also negative since $\tilde{h}'_H(\pi) < 0$ (as shown in Proposition 1(iv)) and, again, $\tilde{h}_H > \tilde{h}^f = 1$.

In case III, we have $H = \tilde{h}_L$ and hence get

$$\begin{aligned}&\tfrac{\partial Z(\tilde{h}_L,\cdot)}{\partial \pi} = -\textstyle\int_0^{\tilde{h}_H} hF'(h)dh + 1 - F(\tilde{h}_H) + (1-\pi)\cdot\left[F'(\tilde{h}_H)\tilde{h}'_H(\pi)(\tilde{h}_H - 1)\right]\\ &+\alpha\left[\textstyle\int_0^{\tilde{h}_L} hF'(h)dh + 1 - F(\tilde{h}_L)\right] - \alpha(1-\pi)\cdot\left[F'(\tilde{h}_L)\tilde{h}'_L(\pi)(\tilde{h}_L - 1)\right] > 0.\end{aligned}$$

Analogous to case II, the sum of the first three terms is strictly negative. However, the additional fourth term is positive since $\tilde{h}'_H(\pi) < 0$ (see Proposition 1(iv)) and $\tilde{h}_H - 1 > 0$ (since $\tilde{h}_H > \tilde{h}^f = 1$ by definition of case III). This renders the total expression ambiguous in case III.

Effect of the intensity of drivers' social preferences (θ) Taking into account that $\tilde{h}_L = \tilde{h}_H(\theta)$ and $\hat{h} = \frac{1}{\theta}$, we get for case II:

$$\frac{\partial Z(\tilde{h}^f,\cdot)}{\partial \theta} = (1-\pi)\cdot\left[F'(\tilde{h}_H(\theta))\tilde{h}'_H(\theta)(\tilde{h}_H(\theta) - 1)\right] - (1-\alpha)\left[F'(\hat{h})(-\frac{1}{\theta^3} + \frac{1}{\theta^2})\right] \lesseqgtr 0,$$

where the first term is negative due to $\tilde{h}'_L(\theta) < 0$ (see Proposition 1(iv)) and $\tilde{h}_H(\theta) - 1 > 0$ (by definition of case II). As the second term is positive due to $-\frac{1}{\theta^3} + \frac{1}{\theta^2} > 0$, the sign of $\frac{\partial Z(\cdot)}{\partial \theta}$ is ambiguous.

For case III, note that $\tilde{h}_L = \tilde{h}_L(\theta)$ and $\tilde{h}_H = \tilde{h}_H(\theta)$, and that $\tilde{h}_L(\theta) > \tilde{h}^f = 1$ by

definition of this case. Hence, we get

$$\begin{aligned}\frac{\partial Z(\tilde{h}_L,\cdot)}{\partial\theta} &= (1-\pi)\cdot\left[F'(\tilde{h}_H(\theta))\tilde{h}'_H(\theta)(\tilde{h}_H(\theta)-1)\right]-\alpha(1-\pi)\cdot\left[F'(\tilde{h}_L(\theta))\tilde{h}'_L(\theta)(\tilde{h}_L(\theta)-1)\right]\\ &\quad -(1-\alpha)\left[F'(\hat{h})(-\frac{1}{\theta^3}+\frac{1}{\theta^2})\right]\lesseqgtr 0.\end{aligned}$$

As in case II, the first term is negative, but the (additional) second and the third term are positive. Moreover, the sign of the difference between the first and second term depends, among the other parameters, on the value of the density $F'(\tilde{h})$ at the points $\tilde{h}=\tilde{h}_L(\theta)$ and $\tilde{h}=\tilde{h}_H(\theta)$, and on the derivatives $\tilde{h}'_H(\theta)$ and $\tilde{h}'_L(\theta)$. Again, the sign of $\frac{\partial Z(\cdot)}{\partial\theta}$ is ambiguous.■