Philosophy & Technology (2021) 34:1475–1499
https://doi.org/10.1007/s13347-021-00468-1

RESEARCH ARTICLE

Autonomous Driving and Public Reason: a Rawlsian Approach

Claudia Brändle[1] · **Michael W. Schmidt**[1]

Received: 15 February 2021 / Accepted: 8 August 2021 / Published online: 21 August 2021

Abstract

In this paper, we argue that solutions to normative challenges associated with autonomous driving, such as real-world trolley cases or distributions of risk in mundane driving situations, face the problem of reasonable pluralism: Reasonable pluralism refers to the fact that there exists a plurality of reasonable yet incompatible comprehensive moral doctrines (religions, philosophies, worldviews) within liberal democracies. The corresponding problem is that a politically acceptable solution cannot refer to only one of these comprehensive doctrines. Yet a politically adequate solution to the normative challenges of autonomous driving need not come at the expense of an ethical solution, if it is based on moral beliefs that are (1) shared in an overlapping consensus and (2) systematized through public reason. Therefore, we argue that a Rawlsian justificatory framework is able to adequately address the normative challenges of autonomous driving and elaborate on how such a framework might be employed for this purpose.

Keywords Autonomous driving · Reasonable pluralism · Public reason · Reflective equilibrium · Trolley cases · Risk distribution

✉ Michael W. Schmidt
michael.schmidt@kit.edu

Claudia Brändle
claudia.braendle@kit.edu

[1] Karlsruhe Institute of Technology (KIT), Institute for Technology Assessment and Systems Analysis (ITAS), Karlstrasse 11, 76133 Karlsruhe, Germany

1 Introduction

Autonomous driving (AD) is an emerging technology that enables all of the dynamic driving tasks involved in operating a vehicle to be managed by its network- and sensor-supported computer system so that no human passenger is required to monitor the traffic.[1] This technology, if it (ever) comes to full fruition, promises immense advantages over our present transportation systems in which vehicles are still (mostly) operated by human drivers — advantages such as increased road safety, comfort, and access to personal mobility for previously excluded groups (e.g., the elderly, children, or people with impairments). However, AD also poses serious normative problems that must be solved if the technology progresses and is actually to be implemented as a mobility option in liberal democracies. Accordingly, there is an evolving philosophical and public debate on these issues.

While there are a number of serious fundamental normative challenges facing AD (AD challenges), we will take as an important example the specific challenge of determining the criteria that should govern an autonomous vehicle's decision-making process, especially when facing the threat of crashes or other accident scenarios. This problem can be exemplified by so-called trolley cases[2]: how should an autonomous vehicle (AV) react if it is faced with an accident scenario in which it is only possible to prevent serious harm to one group of persons by seriously harming another group of persons? We will also consider the related challenge of defining the acceptable risk of harm to fellow road users caused by an AV under normal traffic conditions.[3] In the current debate, these two challenges are given a wide variety of solutions, which are often grounded within specific comprehensive moral doctrines (Leben, 2017; Bonnefon et al., 2016; Wallach & Allen, 2008).[4]

In Sect. 2, our paper introduces an important aspect of dealing with AD challenges that — in our view — has not yet been adequately discussed in the debate: the problem of *reasonable pluralism*. This problem was identified by the political

[1] An alternative term for AD used in science and industry is "automated (and connected) driving." This term better captures the central notion of the emerging technology since the concept of autonomy in ethical and legal thought involves some features that are not and perhaps cannot be implemented in AD, such as (personal) agency or (moral) responsibility. Nevertheless, the terms "AD" and "autonomous vehicle" are often used, especially in the normative debates we will be focusing on. For this reason, we will use these terms to refer to vehicles with a high degree of automation (i.e. in terms of the automation levels defined in the *SAE International Standard J3016*: a vehicle with automation of level 4 [high driving automation] or level 5 [full driving automation] (SAE International 2018)). Since these levels of automation have not yet been achieved, our findings will primarily pertain to a yet still hypothetical future. When or if a level 5 type of driving automation will be fully developed in the future remains an open question for us. However, we are confident that level 4 automation — AD under specific conditions and in specific zones — can be established in the foreseeable future. While our findings might be partially applicable to normative challenges of existing levels of automation (levels 2 and 3), this will not be addressed in the present article.

[2] This refers to an ethical debate on the classic trolley problems: thought experiments introduced by Foot (1967) and Thomson (1976).

[3] For further detailed discussion of AD challenges, see Sect. 4.

[4] On the general ethical debate about accidents involving AV, see Hevelke and Nida-Rümelin (2015) and N. J. Goodall (2014a); (2014b).

philosophy of John Rawls and is based on the fact that there exists a plurality of reasonable yet incompatible comprehensive moral doctrines (e.g., religions, philosophies, worldviews) within liberal democracies. The problem that arises due to reasonable pluralism is that acceptable answers to fundamental political questions cannot refer to only one of these comprehensive moral doctrines, since these answers must in principle be acceptable to every reasonable citizen. We argue that solutions to AD challenges are ultimately faced with the problem of reasonable pluralism. As stated above, there are several potential solutions to AD challenges, and especially those that pertain to the criteria for the automated decision-making process are usually grounded in specific comprehensive moral doctrines. At the same time, decisions about AD challenges are political in nature, which means that any politically legitimate solutions to these challenges cannot be based solely on any one reasonable comprehensive moral doctrine. Fortunately, the problem of reasonable pluralism has a well-known solution: solutions to AD challenges are acceptable when they are justified within the framework of *public reason* and respect the corresponding requirements. Interestingly, this applies not just to AD challenges but to all fundamental normative challenges posed by emerging technologies — thus our findings are significant for any kind of technology assessment or ethics of technology.

In Sect. 3, we outline a Rawlsian framework for public reason in which solutions to AD challenges must be justified by the method of *full reflective equilibrium*. This means (among other things) that all arguments that are crucial for the justification of a specific solution to AD challenges must have premises that are either shared by all reasonable citizens in a so-called overlapping consensus or supported by shared commitments.

A more detailed characterization of two important AD challenges is given in Sect. 4 in order to provide a more concrete background for our further discussion. We will focus on the relationship between trolley cases and real-world collision situations as well as on normative challenges of so-called mundane driving situations. The trolley problem and by extension the design of a multitude of different trolley cases — with the MIT moral machine experiment (Awad et al., 2018) providing abundant examples — has been a popular topic of research and discussion in the debate on AD for some time. We will address the rather controversial question whether the advent of AD will lead to "real-world trolley cases." Some scholars believe that AVs will be faced with dilemmatic decisions that are sufficiently similar to trolley cases (Lin, 2016; Wallach & Allen, 2008; Bonnefon et al., 2016), while critical voices are more skeptical about such situations arising and claim that other, more dire ethical challenges should take precedence (Himmelreich, 2018; Nyholm & Smids, 2016; N. Goodall, 2019; JafariNaimi, 2018). We will take a closer look at the contribution of the discussion of trolley cases and mundane driving situations to the establishment of an overlapping consensus regarding AD.

As the overlapping consensus in different societies might be developed to varying degrees, we distinguish in Sect. 5 two forms of overlapping consensus: (1) a substantial overlapping consensus and (2) a less substantial or insubstantial overlapping consensus. The solutions that can be given for AD challenges and their corresponding justified political decisions might depend on the form of overlapping consensus a given society has developed.

In Sect. 6, we discuss the need for a debate on the meta-level: there must be a discussion of whether it is in fact desirable to authorize the implementation of AD. This meta-challenge might depend partly on whether it is possible to find justified solutions for the AD challenges presented in Sect. 4.

We summarize and conclude in Sect. 7.

2 How to Justify Regulation for AD? A Middle Ground Between Political Realism and Political Moralism

The political sphere is obviously important for AD challenges: solutions to AD challenges (as well as to many fundamental normative challenges related to other emerging technologies) will have to take the form of feasible political regulation (which also informs further research and engineering). Why? The main reason is that the use of an actual AV imposes on persons certain serious risks. The use of an AV could lead to accidents, potentially harming the passengers and fellow road users and thereby impacting their right to bodily integrity or even their right to life. Since these rights are important political values that must be protected in a liberal democracy, such a democracy has to regulate technologies (and their use) that pose such risks. Thus, political regulation for AD will be necessary in any liberal democracy.

Now, not just any political regulation will suffice. Political regulation of AD can impose constraints on individuals' decisions and choices and on the development of the technology — on research, innovation, engineering and entrepreneurship, and many other areas of life that are also connected with certain rights and liberties. Especially when political regulation is concerned with fundamental rights, it must be legitimate. And in order to be legitimate, political regulation of AD — a solution to AD challenges — must be justified in a politically adequate way.

What does it mean for a political regulation such as a solution to AD challenges to be justified in a politically adequate way? Let us examine two exemplary positions on how proposed solutions to AD challenges are to be justified. These positions seem to contradict one another but each also has a valid point. We will then propose a middle ground that adopts the plausible aspects of both positions.

The first position is formulated by Himmelreich (2018), who points out that we should not try to justify such solutions by means of a comprehensive moral doctrine and by doing moral philosophy:

> […] we think that this locates the problem on the wrong level. Instead, solutions are called for on the level of politics. Whereas moral philosophy is a reflection on individual conduct, political philosophy is a reflection on social arrangements before the backdrop of substantive disagreement. (Himmelreich, 2018, 676).

So, AD challenges according to Himmelreich are to be solved primarily through collective reflection and political regulation and not through personal moral reflection and personal decisions. In this case, collective reflection must indeed deal with the fact of reasonable pluralism, which Rawls describes in *Political Liberalism* and associated works:

> A modern democratic society is characterized not simply by a pluralism of comprehensive religious, philosophical, and moral doctrines but by a pluralism of incompatible yet reasonable comprehensive doctrines. No one of these doctrines is affirmed by citizens generally. Nor should one expect that in the foreseeable future one of them, or some other reasonable doctrine, will ever be affirmed by all, or nearly all citizens. Political liberalism assumes that, for political purposes, a plurality of reasonable yet incompatible comprehensive doctrines is the normal result of the exercise of human reason within the framework of the free institutions of a constitutional democratic regime. (Rawls, 2005, xvi).

Indeed, Himmelreich refers to Rawls's political liberalism as an alternative justificatory framework to a moral philosophy based only on specific comprehensive moral doctrines. However, there is also a notable difference between Rawls's position and Himmelreich's proposal. Rawls makes a distinction between a political justification that would adequately address a reasonable pluralism but is still based on the moral powers of the citizens (especially their sense of justice), and a metaphysical justification that would be inadequate (Rawls, 1985). Considered judgements concerning justice as elements of a political justification thus remain moral judgements to some degree. Rawls himself states that his political conception of justice "[…] is, of course, a moral conception […]" (Rawls, 1985, 389; see also Rawls, 2005, 11; Rawls, 2001, 26). In contrast, Himmelreich makes a distinction between politically and morally justified solutions to AD challenges. He stresses that the necessary political solutions to AD challenges would not be moral solutions:

> […] what we in fact face is a social choice. What seems needed is a kind of compromise to overcome disagreements over issues of value. Insofar as we value the moral diversity of our political community, it should be recognized that autonomous vehicles pose primarily a political problem, not a moral one. (Himmelreich, 2018, 676).

Now, Himmelreich's position is somewhat ambiguous: What does it mean that AD challenges and their solutions are primarily political and not moral? Is it simply a matter of finding a solution that is actually accepted through democratic decision-making, or are there still moral constraints on which solution might be acceptable? It is possible to read Himmelreich as if the only concern is the political *acceptance* of AD-solutions and not their *acceptability*.[5]

Adopting this interpretation of Himmelreich, Keeling (2020) criticizes this position and advocates for an alternative, which is the second position we want to present. Keeling acknowledges that social acceptance is a necessary condition for an adequately justified solution to AD challenges, but rejects Himmelreich's framing of AD challenges as merely a problem of social choice. If a solution is political simply by virtue of having an empirical social acceptance and providing stability, then this is not a desirable solution, since the

[5] Taebi (2017) raises the issue of a discrepancy between political or social acceptance on the one hand and moral or ethical acceptability on the other in the context of nuclear waste storage. Similar to our approach, the author applies the Rawlsian concept of the reflective equilibrium to this problem, albeit in the form of wide instead of a full reflective equilibrium (Taebi 2017). On the difference between wide and full reflective equilibrium, see Sect. 3 of the present paper.

aim is to find a solution that is morally correct or acceptable. Keeling argues that no solutions to AD challenges that are immoral or ignore morally relevant arguments or morally justified distinctions would be adequate:

> [T]here are reasons to accept or reject solutions to the moral design problem [of AD] which do not pertain to social choice. On one hand, if a collective judgement holds that AVs should act in accordance with immoral principles, then there is a moral reason to reject that solution to the moral design problem. On the other hand, if there is a moral difference between, for example, killing and letting die, then there is a pro tanto reason for this distinction to be reflected in AV decision-making algorithms. (Keeling, 2020, 304).

The arguments Keeling presents have some plausibility but raise different problems. What criteria should determine that a solution to AD challenges is immoral? What criteria should define a valid and relevant moral argument or distinction? It is plausible to interpret Keeling as if these moral aspects are determined without considering the political sphere at all by means of a comprehensive moral doctrine such as a form of utilitarianism or Kantianism, which would be problematic in the light of reasonable pluralism.

This is where the main thesis of our paper comes into play. We argue that both Keeling and Himmelreich make important points and that it is possible to combine the two positions while avoiding the more problematic aspects highlighted in the interpretations above. A middle ground between Himmelreich and Keeling would be to accept the need for moral justifications of solutions to AD challenges while acknowledging that these solutions must be political as well in a certain sense: the moral justification and the reasonable acceptance of this justification in the political sphere should not be seen as two distinct processes. It is not the case that one first justifies a solution to AD challenges morally and then, in a separate second step, shows that this justification is in fact accepted by all or the majority of reasonable citizens (or vice versa). This would not ensure that the solution is accepted for the right reasons (Rawls, 1997, 781). Rather, the political justification begins with shared moral (and non-moral) commitments — commitments that are already accepted in the political sphere by all reasonable citizens. Of course, this is only the starting point of the justificatory process, as these commitments still require systematization — only if a given commitment (e.g., a principle of justice or a solution to an AD challenge) is publicly shown to be part of the most plausible system of shared moral commitments is it shown to be justified. This presupposes that there are certain justified political propositions that form a subset of justified moral propositions, namely, shared and coherently systematized moral commitments concerning issues of justice. Solutions to AD challenges, then, must be justified by showing that they belong to this set of justified moral commitments that are shared as such by the members of a specific democratic society. In Rawlsian terms: solutions to AD challenges must be justified within the scope of public reason by showing that they are part of the set of considered judgements that form a full reflective equilibrium based on an overlapping consensus of a specific liberal society. We contend that this means the justification for the solution to AD challenges has been shown to be appealing and acceptable to every reasonable citizen.

To clarify further our proposal, we will relate these findings to the debate on *political realism* and *political moralism*. The starting point of this debate was a distinction made by Bernard Williams in his (posthumously published) book *In the Beginning Was the Deed* (Williams, 2005). Political moralism, according to the understanding of Williams and most participants of the debate, means that morality is prior to politics in the sense that it either sets the limits of legitimate political authority or sees politics as purely instrumental what makes the quest for political legitimacy in both cases a chapter of applied morality or ethics (Williams, 2005, 2). Political realists, like Williams, claim that this is not adequate in light of the normative autonomy and independence of the political sphere as well as in light of the fact of reasonable pluralism. Political philosophy (or theory) is thus not seen as dependent on morality abstracting from political reality but essentially involving critical reflection of political history, culture, and practice. Participants who see their positions attacked by these claims (but do normally not use the term "political moralist" as self-designation) object that either political realism cannot distinguish properly legitimate from illegitimate authority or has to rely on basically moral criteria for doing so (e.g., Erman & Möller, 2015, 220–26).

We present the Rawlsian justificatory framework as finding an adequate middle ground between political realism and political moralism.[6] For readers familiar with Williams, this might seem a little odd, since he presents Rawls's theory as one of the two paradigmatic examples for political moralism (the other example is utilitarianism) (Williams, 2005, 1–2). However, according to our interpretation, at least in his later works, Rawls should not be seen as a political moralist — Williams does not fully acknowledge Rawls's "political turn" leading to *Political Liberalism* (Rawls, 2005; on this topic see also Gledhill, 2012; Jubb, 2015; Thomas, 2017; Rossi, 2019, 639). Here Rawls does grant the political sphere at least a "basic autonomy" (Erman & Möller, 2015, 228–31) from comprehensive moral doctrines: His justificatory framework for matters of justice and legitimacy is described as a "freestanding" political module consisting of shared commitments within an overlapping consensus of *actual* citizens (Rawls, 2005, 10, 12, 28, 140–41, 154–55). These shared commitments have to be seen as political not metaphysical (Rawls, 1985, 2005, 10, 95–99; Rawls, 2001, 14, 181–83), so a simple reduction to moral doctrines that are thought to be metaphysically true or correct by some citizens is blocked. Rawls provides a historical analysis, which political developments did lead to the prevalence of corresponding political commitments (Rawls, 2005, xxi–xxvii). Additionally the reflection of political reality and culture is a necessary element in trying to achieve a full reflective equilibrium, since this requires a holistic view on matters relevant to the inquiry (Rawls, 2005, 28; Rawls, 2001, 5–6, 29–32). With regard to the two last points, Alan Thomas shares our conclusion that it is inadequate to present Rawls as a political moralist:

> This is to ignore both Rawls's explicit commitment to reflective equilibrium and his detailed discussion of the nature of political legitimacy in a modern society

[6] We agree with Charles Larmore when he writes: "Political philosophy must be a more complex enterprise than either of the customary positions assumes […]" (Larmore 2013, 5).

> developed in Political Liberalism. […] Specifically, it ignores its explanation of why Rawls believed that no political conception of justice could be based on any single comprehensive conception of the good. Any such conception would fail to acknowledge the particular, practical task of 'reconciliation' that political philosophy faces in our historically determined situation of reasonable pluralism (Rawls, 2001: 29, 41). […] Addressing the reflexive precondition of its own possibility is the task of political liberalism which, whatever its ultimate plausibility, can hardly be accused of being a theory which ignores its historical situatedness. Realists may find Rawls's historical narrative thin and unconvincing, but that is a different criticism from the claim that he does not provide one. (Thomas, 2017, 5).

However, we also concede that Rawls should not be seen as political realist: The elements of the freestanding political module still have to be seen as a special sort of moral commitments and the module has to be integrated by reasonable comprehensive doctrines, each of them having their own rationale of doing so (Rawls, 2005, 11–12, 140–41). It is not morality on its own that sets the limits for legitimate political authority; it is moral normativity in combination with political normativity.[7] This is why we claim that a Rawlsian justificatory framework provides an adequate middle ground between political moralism and political realism, which are exemplified by the positions of Himmelreich (2018) and Keeling (2020) in the debate on AD challenges.

Now, even if it is granted that a middle ground between political realism and political moralism is needed and that Rawls's idea of justification by public reason provides such a middle ground, it does not automatically follow that it has to be adopted. There are other approaches, which specify the relation between ethics and politics in an alternative way. Rodríguez-Alcázar (2017) is such an alternative account which is recently also applied to AD challenges (Rodríguez-Alcázar et al., 2020).[8] It will therefore be necessary to adduce additional arguments for preferring the public reason approach. Of course, it is not possible to give a comprehensive justification of the public reason approach in this paper, and we do not claim to introduce new arguments for it. Rather, we will briefly outline — according to our understanding — the two most important considerations that speak in favor of the public reason approach.

The first is that a political system and political regulations that are justified within the scope of public reason receive a reasonable acceptance that secures political stability. If citizens accept a political regulation because it coheres well with their convictions about justice, there is no need to try to change this regulation and they have intrinsic reasons to comply with it. This would be different if the political

[7] Thomas describes this as a form of co-originality (Thomas 2017, 4). And although he ascribes this idea to Williams's position as an instance of a concessive form of political realism, this concessive political realism seems very close to Rawls and to our notion of a middle ground between both idealized camps (Thomas 2017, 15 n5, 18 n22).

[8] We think the approach proposed by Rodríguez-Alcázar might share with non-concessive forms of political realism the inability to discriminate properly between legitimate and illegitimate political authority, but a due discussion of this position is beyond the scope of this paper and must be provided in the future.

regulations were not based on an overlapping consensus but on some kind of compromise between different political powers. According to Rawls, such a compromise would be merely a type of *modus vivendi* that would be abandoned should the balance of political powers change (Rawls, 2005, 147; Rawls, 2001, 192, 194f.). Thus, political regulation for which no justification can be given that satisfies the conditions of a justification within public reason tends to be instable. But stability is needed in cases where the basic structure of society or elements of it are at stake: this is the case when an interpretation of fundamental rights is needed and priority rules must be determined in cases of conflicts between fundamental rights, such as the right to bodily integrity and the right to the freedom of movement and mobility. We hold that since regulation for AD touches upon the interpretation of fundamental rights and must refer to some kind of priority rules, the justification for that regulation should be stable. This is important not only from a political perspective but also from the point of view of engineers and entrepreneurs, who have to work in compliance with those regulations.

The second consideration is that only a public reason approach takes seriously the political autonomy of all reasonable citizens. Anyone might agree, irrespective of the comprehensive moral doctrine they support, that it is of great value — inherently or instrumentally — for citizens as moral persons to be able to use their own reasoning capability to determine the fundamental principles of justice and corresponding interpretations of them (including political regulations that touch upon these fundamental principles). Only when persons can see the political principles and regulations of their society in a certain sense as the principles and regulations that they would implement themselves, were they to reflect upon their beliefs, are they in a political sense autonomous. The public reason approach ensures that this is the case by requiring that one must appeal to beliefs every reasonable citizen shares and propose the most plausible systematization of those shared beliefs relevant to a given topic in order to justify a principle of justice or important political regulation. If there is in fact an agreement in liberal societies that in an important political sense being free means being politically autonomous and that human dignity demands this kind of freedom (without presupposing a specific metaphysical explanation of why this is the case), this would make a strong rationale for the public reason approach.

Now, these two arguments certainly might not convince someone who has given serious thought to the matter and supports a different approach to political legitimacy. But the above discussion at least should make clear that the public reason approach — which is by no means a niche position — deserves due attention in the debate on AD challenges. If one can agree with this modest claim, our further findings are relevant.

3 Details of the Rawlsian Justificatory Framework for Public Reason

In this section, we offer a more concrete account of the justificatory framework associated with the public reason approach. Although we will use explicitly Rawlsian terminology, we hold that this framework is at least in the most relevant respects

compatible with and useful in the context of other variations of the public reason approach, such as a Habermasian approach to deliberative democracy.

In the following, we will not focus on Rawls's original position and its potential to inform solutions for AD challenges, for two reasons.[9] First, we do not think that one is obliged to adopt Rawls's original position in order to justify claims within the scope of public reason. Second, it is an open empirical question whether the original position is shared in the overlapping consensus of a specific liberal society. Only if this is the case can one reasonably adopt it in order to support solutions to AD challenges (or any other political claim).

As mentioned above, we instead focus on the notion of *full reflective equilibrium* as the general method for justification within public reason (Rawls, 1995, 141; Rawls, 2001, 31f.; Daniels, 1996, 144–75; Daniels, 2020). A full reflective equilibrium is a special case of a *wide reflective equilibrium* (Rawls, 1999, 43; Rawls, 1974; Daniels, 1979, 2020) with the additional condition that the starting and resulting commitments are shared within an overlapping consensus. With this method, a solution to AD challenges is justified *iff* it is shown to be part of a full reflective equilibrium, i.e., the most plausible systematization of shared political commitments concerning the topic. In order to have the potential to achieve a full reflective equilibrium, certain rules must be followed[10]:

1) According to the public reason approach, one begins with the commitments that are shared by all reasonable doctrines in an overlapping consensus, and so one must first determine which of the relevant commitments are actually shared and thus part of the overlapping consensus. This is at least in part an empirical task and requires a combination of sociological and philosophical research. Since all of the elements of a full reflective equilibrium must be shared commitments or at least supported by shared commitments, these commitments are the foundation of the justification.[11]
2) One must identify all relevant commitments for the topic at hand and attempt to systematize them. Which commitments might be relevant for AD challenges? There are likely to be straightforward relevant commitments, e.g., that there is a right to mobility, but there might also be commitments that arise only if we employ (thought) experiments or through the systematization itself, e.g., that killing is different than letting someone die. If the commitments are systematized and thus seen as connected, conflicts between commitments can arise, e.g., between

[9] If the original position were employed for this topic, it would have to be modified according to the four-stage sequence at stage 2 or 3 (Rawls 1999, 171–76; 2001, 48, 173f.). For the claim that one should indeed use the original position, see Leben (2017; 2019).

[10] This characterization of the method of full reflective equilibrium is compatible with the characterization given by Daniels (1996, 144–75; for his characterization of the method in general terms, see 1979; 2020). See also Schmidt (forthcoming).

[11] One could call this the weak foundationalist requirement, since these commitments are not necessarily basic beliefs that one could claim are reliably connected with the truth of the propositions (this would already constitute a moderate foundationalism; for this epistemological terminology, see BonJour 1985; Elgin 2005).

the right to mobility and the commitment that no one has the right to impose a serious health risk on another. There are various possibilities for adjusting these conflicting commitments on all levels of generality in order to bring them into agreement with each other: the different possible adjustments lead to different possible systems of commitments — all of these systems are candidates for a full reflective equilibrium. If there are mutually exclusive solutions for AD challenges, they will be part of different candidate systems. Because all relevant commitments have to be systematized (so far as this is possible given restrictions of time and resources) and because candidate systems of commitments as a whole (and not isolated commitments) are evaluated, this could be called the (moderate) holistic requirement. In Rawlsian terms: one must always find a *wide reflective equilibrium*, not a narrow one (Rawls, 1999, 43; Rawls, 1974; Daniels, 1979, 2020).

3) In the process of adjustment, no commitment is immune to revision, which results in the emergence of many quite different candidate systems for a full reflective equilibrium. This requirement is connected to the notion of fallibilism and is a crucial element of the method of reflective equilibrium in general.
4) From the various candidate systems, we should choose the one that is — in light of all relevant commitments and their inferential connections — the *most plausible*. This could be called the minimalistic rationality requirement, since it would be irrational to adopt a system of commitments that seems to be less plausible than other systems. A solution to AD challenges would be justified if it is shown that it is part of this most plausible candidate system of commitments and thus of a full reflective equilibrium. Once achieved, a full reflective equilibrium is not static and settled for all time, but always open to new considerations — further reflection or experiences might lead to new commitments and a new evaluation of the emerging possible candidate systems.

The use of the method of full reflective equilibrium to justify solutions to AD challenges (or any other concrete political problem) might differ somewhat from the justification of principles of justice in Rawls's *A Theory of Justice* or *Political Liberalism*. Instead of relating judgements about justice to abstract principles of justice, a justification of a concrete political problem in an actual liberal democracy relates concrete political judgements to interpretations of constitutional essentials, human rights, fundamental rights, or civil rights.[12]

One difficulty in achieving not just a wide reflective equilibrium (a system of commitments that are justified for a specific person) but a full reflective equilibrium (a system of commitments that are justified for all reasonable citizens of a liberal democracy) might be the possibility of being mistaken about the content of the overlapping consensus (especially since this consensus itself is also dynamic, not static). One could additionally be mistaken about differences in the weight that reasonable

[12] Of course, these are included in Rawls's principles of justice and a more abstract reflection that includes principles of justice might be necessary to solve persisting disagreements on lower levels of abstraction.

citizens assign to each commitment even if they share it, and consequently about which adjustments seem most plausible. At this point, again, empirically informed research is required.[13]

4 Taking a Closer Look at Some AD Challenges: Principles for the Decision-making Process

In this section, we take a closer look at some AD challenges. AD challenges involve a wide range of issues, from data privacy and security[14] to moral and legal responsibility for harm caused by AVs,[15] regulation of import and export of AD technology,[16] questions of sustainability,[17] and fair access to AD,[18] as well as other long-term effects of AD on various parts of society.[19] Technically, all solutions to these normative questions should be justified through the method of full reflective equilibrium insofar as they have consequences for human or fundamental rights. For the present paper, we set these issues aside and instead relate our findings specifically to one question that is currently a central focus of the debate: how should liberal democracies choose and subsequently implement principles to govern the automated decision-making process of an AV?

Following a distinction made by Himmelreich (2018), this question can be broken down into two interrelated challenges that require solutions justified within a full reflective equilibrium if AD is to be implemented in the transport system of a liberal democracy:

1) Which principles should guide decisions in real-world trolley cases?
2) Which principles should guide decisions about risk distribution in so-called mundane driving situations?

In the following subsections, we first explore the controversial question whether real-world trolley cases are possible, as well as the associated moral challenges and

[13] We stressed already that we restrict our discussion to liberal democracies: One reason is, that according to our proposed justificatory framework it would matter which political regime is in power because this has strong influence on the political culture and therefore on the political commitments of the citizens and their ability to share their political convictions freely. The latter is a prerequisite for the use of the method of full reflective equilibrium (citizens in illiberal societies, of course, still can make personal or semi-public use of the critical potential of the method of wide reflective equilibrium, but the public discursive forum that could give space to a full reflective equilibrium is restricted).

[14] See Hillerbrand, Milchram, and Schippl (2019); Pype et al. (2017); Chaturvedi (2020); Rannenberg (2016); and Dhar (2016).

[15] See Hevelke and Nida-Rümelin (2015) and Baumann et al. (2019).

[16] This is a less discussed topic up until now but might be especially complicated with regard to different interpretations of human or basic rights in the international sphere, which might lead to incompatible regulatory requirements for AD.

[17] See Schreurs and Steuwer (2016) and Brimont, Saujot, and Sartor (2017).

[18] See Faber and Lierop (2020).

[19] For a general overview, see Grunwald (2016).

their relevancy to a full reflective equilibrium with respect to AD. In a second step, we take a closer look at the decisions arising from mundane driving situations: what, if any, moral challenges occur in these cases and (how) do they differ from those associated with real-world trolley cases?

4.1 Trolley Cases and Real-world Trolley Cases

Trolley cases are part of a popular philosophical thought experiment commonly known as the trolley problem,[20] which has become a familiar staple within the debate on AD. In one of its more classical forms, it consists of two very similar cases of moral dilemmas that evoke contrary moral intuitions concerning the right course of action: In one case, a runaway trolley is heading towards five people on the tracks who will be killed by a collision. By pulling a lever, you could redirect the trolley onto a different track where one person is present, thereby saving the five but killing the one. This trolley case is compared to another, similar trolley case: A runaway trolley is on a collision course with five people on the tracks. You could push a fat man from a bridge onto the tracks to stop the trolley, thereby saving the five but killing the fat man. Although the two cases seem analogous in their outcome in terms of the number of lives saved (saving five at the expense of one), they evoke contrary moral intuitions about what the right course of action is: in the former case, pulling the lever to save the five seems to many people to be at least morally permissible, if not a moral obligation, while in the latter case, pushing the man onto the tracks seems to most people to be morally prohibited (Foot, 1967; Thomson, 1976). The function of trolley cases like those described above is to pinpoint exactly the morally relevant differences of these seemingly analogous situations, e.g., by shedding light on the difference between "positive" and "negative" duties, between killing and letting die, or between consequentialist and non-consequentialist approaches to morality (Keeling, 2020; Nyholm & Smids, 2016).

While the "fat man" trolley case is firmly rooted in the fictional realm of thought experiments, the same cannot so readily be said about the original trolley case: prima facie, it seems entirely possible that situations analogous to the original trolley case could arise for an AV in the real world (Lin, 2016; Wallach & Allen, 2008; Bonnefon et al., 2016). Such real-world trolley cases could take the form of the following scenarios: An AV could be faced with the decision to stay its course and hit a pedestrian suddenly stepping on the street, thereby killing her, or swerving and hitting a wall, thereby killing its passenger. In another variant, an AV would have to choose between staying the course, hitting and thereby risking killing a group of pedestrians, or swerving and thereby sacrificing its single passenger. Indeed, there are almost endless variations of such cases. If these cases are possible and if they are sufficiently similar to trolley cases, the differences illuminated by trolley cases as philosophical thought experiments would be relevant to this AD challenge as well. However, an increasing number of scholars

[20] On the distinction between trolley cases and trolley problems, see Himmelreich (2018).

question whether it is possible for trolley-case-like scenarios to occur in the real world and are speaking out against centering the normative debate about AD on such cases. In general, arguments against the premise of real-world trolley cases can be divided into (1) technology-based arguments for the impossibility of trolley cases or sufficiently similar collision scenarios occurring in real life, and (2) arguments focused on showing fundamental disanalogies between real-world collision scenarios and trolley cases, which would render the discussion of the latter irrelevant to the discussion of AD.

Himmelreich offers an argument of the first type. He states that the assumptions necessary to formulate trolley cases seem to be inconsistent with the technological possibilities of AD. These two features or assumptions of trolley cases are (1) a collision is "imminent and unavoidable" and (2) "a meaningful choice is nevertheless possible" (Himmelreich, 2018, 673). He claims that, from an engineering perspective, these two features are difficult to reconcile: If a meaningful choice between two actions is possible, the collision itself might also either still be avoidable or present a different type of choice from that in a trolley case, e.g., between killing a person and merely injuring another. On the other hand, if a collision is in fact unavoidable, it is unclear whether an AV would still be able to make a meaningful choice between two actions. Based on a variation of the idea that "ought implies can" applied to an autonomous machine, this line of thinking contends that the two conditions necessary to constitute a trolley case or a sufficiently similar collision scenario cannot both be true at the same time for an AV. Himmelreich himself is very cautious about the soundness of this argument and about drawing any definite conclusions about the (im)possibility of real-world trolley cases (Himmelreich, 2018, 673f.). However, even if one grants this impossibility for the sake of argument, as does Keeling (2018), and accepts that an AV would never encounter a real-world trolley case, it does not necessarily follow that philosophical reflection on trolley cases is irrelevant to what Keeling calls the "moral design problem" of AD (Keeling, 2020, 293). While it might be true that trolley-case-like collision scenarios will not happen in real life, trolley cases as philosophical thought experiments try to isolate and illuminate morally relevant properties that are still present in problems of risk distribution in mundane driving situations, such as the difference between positive and negative duties. Additionally, the idealized thought experiments, understood as moral dilemmas, highlight and thereby exemplify the problem of reasonable pluralism, which must be dealt with if liberal democracies are to find acceptable solutions to AD challenges. Relying on comprehensive doctrines like utilitarianism or Kantianism to inform the automated decision-making process of an AV can — prima facie and in specific cases — lead to incompatible solutions to such dilemmas. The challenge that then arises is identifying political beliefs that these doctrines might (perhaps for different reasons) nevertheless share in an overlapping consensus and thus establishing the basis for solutions to real AD challenges that are acceptable to every reasonable citizen. Therefore, an empirically informed philosophical reflection on idealized trolley cases is still useful in the attempt to achieve a full reflective equilibrium, even though idealized trolley cases do not themselves present an AD challenge. Even so, we would like to stress that the question whether the "impossibility argument" is

sound and sufficiently similar real-world trolley cases do not exist is empirical — it can only be answered based on relevant results from engineering research.

Presenting the second type of argument, Nyholm and Smids (2016) assert that there is a categorical difference between the moral concerns of trolley cases and the moral concerns of real-world collision scenarios: the latter take place under conditions of risk or uncertainty that are not present in trolley cases. In contrast to trolley cases, real-world collision scenarios involving AVs will always include a substantial element of risk and uncertainty, since they are impacted by complex factors such as environmental conditions (e.g., weather, the state of the road) and the behavior of other road users or bystanders. The decisions AVs have to make are thus not a matter of certainly killing one person or another, or one person or a group of people. Instead, there will be cases where steering towards one affected party will carry, for example, a 99% risk of killing (or severely injuring) this party, while staying the course will carry 99% risk of killing (or severely injuring) another affected party. Because of this different type of reasoning, it is — according to Nyholm and Smids — unclear whether any conclusions that can be drawn from the philosophical discussion about trolley cases are reasonably applicable to the case of AD.[21] Nevertheless, we would agree with Keeling that, despite this difference, the moral considerations that trolley cases showcase, such as a relevant distinction between killing and letting die, might still be an important moral aspect of real-world collision scenarios even with the added dimension of risk and uncertainty. Which moral principles should determine a machine's decision about which risky action to take? Should an AV risk killing a person or a group of people by swerving into the other lane, or risk letting a person or a group of persons die by staying the course?

To summarize: Including discussion of trolley cases in the debate on AD might still be a useful approach, even though real-world collision scenarios occur under risk or uncertainty, because the moral considerations they illuminate can tell us something about what type of outcome automated decision-making ought to favor and why. Furthermore, as long as AD-collision scenarios with very high risks of harm as the outcome for each possible decision are technically possible, they seem sufficiently similar to trolley cases insofar as these real-world scenarios pose a (dramatic) moral challenge or dilemma, where — at least prima facie — some of the same moral distinctions are highlighted. We would grant that when the additional dimension of risk or uncertainty is taken into account, it is possible that we also have to accept new moral distinctions or principles in an ethics of risk or political philosophy of risk (Hansson, 2013).[22]

[21] Davnall (2020) takes this third argument as a basis for her claim that the best course of action for any AV will be to engage in emergency braking over any kind of swerving action and that AVs will not face trolley like cases as such.

[22] The difference between the ethics of risk and the political philosophy of risk is, according to our position, that we seek a personal wide reflective equilibrium for the former and an interpersonal full reflective equilibrium for the latter.

4.2 Risk Distributions in Mundane Driving Situations

Apart from extreme situations like real-world trolley cases, AVs will have to make a multitude of mundane decisions during even a short drive. These "mundane driving situations" include among other things "approaching a crosswalk with limited visibility, making a left turn with oncoming traffic, and navigating through busy intersections" (Himmelreich, 2018, 678). In any type of driving situation, an AV will have to continuously update and alter its behavior according to numerous different and constantly changing parameters (surrounding objects, other road users, environmental conditions) and this turns out to be a very challenging task for an AV: mundane driving situations iterated over time can lead to injuries and deaths. In this way, such situations are not all that different from trolley cases or real-world collision scenarios. The only difference lies in the degree of risk and uncertainty: death or harm to at least one party is unavoidable in trolley cases, and almost unavoidable in real-world collision scenarios, while most mundane driving situations have a significantly lower risk of harm. Despite this difference, mundane driving situations are rendered ethically challenging in part precisely because of their structural similarity to trolley cases and real-world collision scenarios: just like these, mundane driving situations will involve decisions about risk distribution between AV users and other road users. Mundane decisions about risk distribution likewise contain implicit moral judgements (Leben, 2019, 100; Himmelreich, 2018). For example, it is a moral judgement that one should alter the route of an AV in order to lower the risk of a potential collision with a bicyclist and thereby increase the AV's risk of colliding with an object.

In order to develop a sound regulation of AD that is justifiable to all reasonable citizens, it is not only necessary to define the relevant risks and find a way to measure and balance them. It is also necessary to explicitly identify the (implicit) intuitions, arguments, and principles that underlie the moral deliberation of risk distribution in such mundane driving situations. Only then can they enter into a full reflective equilibrium and be part of an overlapping consensus. In the case of the above example, there could be several moral reasons for altering the behavior of an AV in favor of the bicyclist. For example, one reason could be summed up as a principle of reducing harm: the bicyclist is more vulnerable than the AV user and would likely sustain more serious injuries in case of a collision. Another reason could be an idea of fairness: since it is the AV user that increases the risk of severe injuries by driving a car with a high speed and mass, it would be unfair to burden primarily the bicyclist with this increased risk (see Baumann et al. 2019). These reasons would have to be balanced with other moral and non-moral considerations like an assumed moral difference between killing and letting die (as illuminated by the trolley problem), a precautionary principle or the need for an efficient mobility system, which can be at odds with optimizing AV for safety (Himmelreich, 2018, 680).

5 Substantial and Insubstantial Overlapping Consensus

In finding solutions to AD challenges, one has to rely on the overlapping consensus. By outlining the ethical challenges and the underlying values that seem to govern intuitive reactions, as shown in the section above, a liberal democracy may ascertain some potential candidates for the shared beliefs, values, or principles that are part of its overlapping consensus. But this is merely a first and tentative step. How this overlapping consensus is structured is an empirical question that requires further and thorough investigation from multiple perspectives. Some insight stems from the philosophical debate about trolley cases that uncover conflicts between widely shared intuitions. Other insights can include findings from other fields of philosophy, from sociology, or from empirical research in general. For example, the moral machine experiment by MIT, which gained popularity when it was first launched, can provide some limited data about parts of this overlapping consensus (Awad et al., 2018). In the case of this specific experiment, it is important not to mistake its empirical findings about the course of action a majority of the participants choose for a sufficient legitimation of a political decision. Instead, these findings have to be seen as forms of considered judgements and should be complemented by the attempt to find a full reflective equilibrium. Other empirical studies already explicitly aim to inform a reflective equilibrium process (Bergmann et al., 2018). These findings must be weighed against the background of values that are already agreed on, e.g., those that underlie much of the body of law in liberal democracies. Law itself is another important source for shared principles that can be applied to AD.[23]

Depending on how many shared beliefs, values, and principles exist in a given society, one might distinguish two different forms of overlapping consensus and their respective results regarding AD challenges: substantial and insubstantial overlapping consensus.

5.1 Substantial Overlapping Consensus

An overlapping consensus could be called substantial when there are many shared political principles (e.g., the duty to not objectify a person, principles of fairness), background theories (e.g., that there is a difference between killing and letting die), and other shared beliefs (e.g., knowledge about injury and death statistics for different forms of mobility). In this case, it is likely that one can also arrive at substantial solutions to AD challenges (e.g., that an AV in a real-world trolley case must retain its initial direction of movement if the risk for the least advantaged party cannot

[23] An interesting example offers de Sio (2017) who starts with existing traffic laws and proceeds to elaborate their underlying normative principles, with a special focus on the "doctrine of necessity" as a potential solution to dilemmatic emergency situations/trolley cases faced by AVs (Santoni de Sio 2017). As will become clearer in Sect. 5.1, this approach seems to be a promising way of how a substantial overlapping consensus concerning an ethical regulation of AD could be established in practice: Here, specific traffic laws and their underlying norms constitute an already established form of consensus. Applying them to the case of AD is then a good starting point to formulate regulations that have a high likelihood of passing the test of a full reflective equilibrium.

be significantly reduced). There could also be an acceptable justification for precise regulations on how AVs ought to behave in real-world collision situations with risk and uncertainty as well as in mundane driving situations, so that it is made clear in every possible case which risks can be imposed by AVs on their own passengers and on fellow road users.

However, it is no simple task to reveal and then work with a substantial overlapping consensus: How can common beliefs, values, and principles that find widespread agreement on a societal level be identified and then brought into an equilibrium? While the sources outlined above can be of service here, it is important to acknowledge that it is not easy to move from Rawls's ideal-world overlapping consensus and the ideal method of full reflective equilibrium to real-world practice. Here, concepts explored within the discipline and practice of technology assessment (TA), such as the concept of responsible research and innovation (RRI) (Owen et al., 2012; Schomberg, 2013) or a value-led design approach (van de Poel, 2016; van de Poel & Zwart, 2010), could be used as tools for helping to make explicit the values, beliefs, and principles that are already shared in the context of AD and informing the political process for AD regulations in line with such a substantial overlapping consensus. There is also promising research available on how to use the method of wide reflective equilibrium in the context of TA practice (Doorn, 2010, 2012; Doorn & Taebi, 2018; van de Poel & Zwart, 2010; Taebi, 2017). This use of the method to identify a set of justified proposals based on a substantial overlapping consensus would already include relevant stakeholders in the process and could reflectively inform and structure the deliberation of a wider political public, thus facilitating the achievement of a full reflective equilibrium.

5.2 Insubstantial Overlapping Consensus

On the other hand, an overlapping consensus might be very limited, i.e., there might be very few and rather vague shared political principles, background theories, and beliefs. In this case, it is less likely that there will be a justification for solutions to AD challenges that is appealing to all reasonable citizens. If AD were nonetheless to be introduced in the transport system, a stopgap solution to the challenges of the automated decision-making process could be to impose on AVs approximately the same regulation as on human drivers. In this case (excluding issues of responsibility), no moral distinctions would be made between the decisions of an AV and those of a human driver, despite potential differences in ability — AVs would be allowed the same behavior in real-world trolley cases as is permitted for human drivers and could impose the same risks on fellow road users as are acceptable for human drivers according to legislation and court decisions. There would be no decisive moral solutions to these AD challenges based on a shared justification but rather workarounds based on the insubstantial consensus established in existing law and the democratic decision-making process that can enact regulation by majority vote. Additional regulation would then only be needed to specify an adequate testing procedure prior to the authorization of AVs in order to determine that the implementation of existing traffic regulations is successful, and to settle questions of accountability

for accidents (e.g., clarifying what cases fall under product liability or determining compensation for harm when accidents happen within the bounds of existing regulations and in the absence of technical malfunction).

While such a stopgap measure might be seen as a good compromise in the absence of a substantial overlapping consensus, it also seems to be somewhat problematic and unsatisfying. It is problematic for practical reasons: While the stopgap measure is a commonly proposed solution to AD challenges, it neglects the complex nature of traffic regulations for humans and the social interaction with which these regulations are balanced. It is unclear how well existing legal regulations for human drivers can be translated into an algorithmic form in order to be implementable in an autonomous system. In other words, the stopgap solution is especially vulnerable to the "challenge of specificity" (Himmelreich, 2018). This becomes apparent in regulations relating to certain abilities that a human can cultivate but a machine lacks or merely imitates (e.g., "anticipatory driving") and whose specifics are notoriously difficult to pinpoint and translate. While this problem is not isolated to the stopgap, it is exacerbated here by the fact that legal traffic regulations are full of this kind of difficult-to-define terminology. In addition, the stopgap solution seems to be somewhat unsatisfying for normative reasons: AD offers an opportunity to secure a safer transport system within which certain social and ethical values consistently inform decisions made in traffic. The benefit of automated decision-making in AD over human decision-making lies in its speed of action, its incorruptibility and its overall consistency in applying whatever rules have been implemented in it. For the stopgap solution, this means that in the best case, AD will follow traffic regulations more consistently than human drivers and thereby contribute to safer traffic overall. But why should we stop there, when we could instead implement better rules for decision-making in an AV — rules that a human cannot follow but a sophisticated machine can? Compared to the possibilities offered by a substantial overlapping consensus that identifies specific social and moral values for decision-making, an insubstantial overlapping consensus and its proposed stopgap solution appear as a squandered opportunity.

6 A Veto Against AD?

A further question is whether it is desirable for a given society to implement AD in its transport system at all. This question must also be answered within an overlapping consensus. A decisive answer pro or contra AD implementation most likely depends on the quality of the overlapping consensus itself and the solutions to various AD challenges. In our view, both an insubstantial and a substantial overlapping consensus could lead to a veto of AD implementation for a number of different reasons.

We see three ways an insubstantial overlapping consensus might lead, not to a stopgap solution, but to an outright prohibition of AD:

First, there could be a significant disagreement about the stopgap measure itself. If according to the stopgap an AV could be licensed upon proving that it drives at least as safely as a normal (or good) human driver, then it is necessary to articulate precisely what the political standard entails that specifies who counts as an adequately responsible human driver. If there were no minimal agreement on such a standard or its respective tests for AVs, then it would be impossible to license AVs.

Second, in an insubstantial overlapping consensus, there may be disagreement or uncertainty about whether AD is morally permissible or not. For precautionary reasons, it might then be impermissible to license AVs.

And third, an insubstantial overlapping consensus could also be the result of a fundamental and unresolvable disagreement concerning the values that should be implemented in an AV. With respect to the fact of reasonable pluralism, different reasonable moral doctrines could arrive at different and incommensurable answers to the questions raised by AD. This worry seems to be inherent in the broad media attention given to trolley cases in reporting on AD, because it focuses exactly on this point: there seem to be different and perhaps even incommensurable solutions to dilemma situations, each justified by a specific comprehensive doctrine and without the prospect of ultimately proving what the *true* or *correct* moral solution is. However, some shared solution is necessary if AD is to be implemented, because AD must be accessible to every reasonable citizen. To demonstrate this point, imagine the following somewhat eccentric scenario: Under the assumption that utilitarians and Kantians would come to different conclusions regarding AD challenges, if AVs were simply implemented with utilitarian values, a Kantian might not be able to use such AVs. As the technology became more embedded in society, this would ever more severely affect their life. If they could de facto no longer participate in the transportation system, this is, prima facie, something that cannot be justified to them. The fear is that with an insubstantial overlapping consensus, reasonable pluralism might lead to an impasse in the case of AD, rendering an adequate regulation and a subsequent legitimate implementation of the technology impossible.

Yet even if there is a substantial overlapping consensus, it could be the case that AD should not be implemented if (1) there are shared values that prohibit AD or (2) it is technologically impossible to implement shared values in an AV.

In the first case, there may be a substantial overlapping consensus in the sense that certain values are agreed upon, such as a shared belief that people have an absolute (political) duty not to objectify persons. This could be a belief that is justifiable to all citizens based on their finding reasons for adhering to this principle within their own comprehensive doctrines.[24] If it then turns out that it is impossible to implement AD without violating the (political) duty of not objectifying persons, e.g., because the shared understanding is that automated decisions about risk-distribution always take the form of killing and never of letting die, there would in turn be an absolute duty not to implement AD.

[24] The reasoning behind rule-utilitarianism is an example of this: a utilitarian could agree to such a principle on a political level, because they believe that including it in a specific political system will maximize the utility overall, even if it lowers utility in individual cases.

The second case refers to an argument presented by Himmelreich (2018, 675; see also Keeling, 2020, 301): Imagine there is a shared agreement about the values that should inform the rules and criteria for decision-making to be implemented in AVs. To ensure that AVs actually adhere to these values, they would have to be implemented using a top-down approach where specific rules for decision-making are implemented in an AV by the programmer. In actual engineering practice, however, automated decision-making is implemented with a bottom-up approach where methods of machine learning are used. While this approach is very effective, the drawback for the political and moral issues of AD is that it is opaque: it often remains unclear and nearly impossible to trace in retrospect what inferences machine learning algorithms draw from the available data and subsequently what kind of "reasons," rules or criteria they apply in their decision-making processes. So even if there is a shared agreement on the values that should inform the rules of the decision-making process of an AV, it is not certain that machine-learning algorithms would infer these rules from the training data. But if it is not feasible with a top-down approach and technically impossible with a bottom-up approach to ensure that the decisions of an AV truly express a society's shared values, then one possible conclusion could be to prohibit AD.

In view of the history of technological progress, it seems unlikely that a society could significantly hinder the development of technology towards increasing automation even in the face of such fundamental ethical challenges.[25] On the political stage of a non-ideal world, the purported benefits of AD could outweigh these concerns, even if a prohibition would be the conclusion of a substantial overlapping consensus. Because of this, even with a justified veto against AD, it might be sensible to at least try to shape the budding technology as much as possible in conformity with the shared values of society. In any case, we do not seek to propose an elaborate solution to the issue of authorization of AD, but simply wish to point out that it is not adequately addressed in the current debate.

7 Conclusion

We have shown that the problem of reasonable pluralism is important for the debate on AD challenges and that against this background a Rawlsian justificatory framework deserves due attention. According to this framework, solutions to AD challenges have to be identified within public reason by forming a full reflective equilibrium. We elaborated on this by taking a closer look at two exemplary AD challenges, which are central in the current debate: real-world trolley cases and rules for risk distribution. With respect to these challenges, a substantial overlapping consensus might lead to a specific regulation for AD while an insubstantial overlapping

[25] See also JafariNaimi (2018) who cautions against a sense of inevitability where AD is concerned and instead urges us to "restore the deep sense of uncertainty accompanied with this technology" (JafariNaimi 2018, 14) and to rethink mobility in general in order to increase safety especially for vulnerable groups.

consensus might result in a stopgap solution that imposes (approximately) the same regulation on AD as on human drivers. A further but related question that needs more attention is the issue of whether it is desirable to authorize AD at all.

Acknowledgements We would like to thank our colleagues at the Karlsruhe Institute of Technology (KIT) for many discussions and their valuable input to the paper: especially our colleagues at the Institute for Technology Assessment and Systems Analyses (ITAS) and at the Department of Philosophy. We would also like to thank the organizers and participants of the 2021 Rabb Symposium "Embedding AI in Society" and the CEPE/IACAP Joint Conference 2021 "The Philosophy and Ethics of Artificial Intelligence" for the opportunity to present and discuss our work. Last but not least, we would like to thank the two anonymous reviewers who contributed greatly to the improvement of the paper with their detailed comments and critical inquiries.

Author Contribution Both authors contributed equally to this paper.

Funding Open Access funding enabled and organized by Projekt DEAL.

Data Availability Not applicable.

Code Availability Not applicable.

Declarations

Conflict of Interest None.

References

Awad, E., Dsouza, S., Kim, R., Schulz, J., Henrich, J., Shariff, A., Bonnefon, J.-F., & Rahwan, I. (2018). The moral machine experiment. *Nature, 563*(7729), 59–64. https://doi.org/10.1038/s41586-018-0637-6

Baumann, M. F., Brändle, C., Coenen, C., & Zimmer-Merkle, S. (2019). Taking responsibility: A responsible research and innovation (RRI) perspective on insurance issues of semi-autonomous driving. *Transportation Research Part a: Policy and Practice, 124*(June), 557–572. https://doi.org/10.1016/j.tra.2018.05.004

Bergmann, L. T., Schlicht, L., Meixner, C., König, P., Pipa, G., Boshammer, S., & Stephan, A. (2018). Autonomous vehicles require socio-political acceptance—An empirical and philosophical perspective on the problem of moral decision making. *Frontiers in Behavioral Neuroscience, 12*. https://doi.org/10.3389/fnbeh.2018.00031

BonJour, L. (1985). *The structure of empirical knowledge*. Harvard University Press.

Bonnefon, J.-F., Shariff, A., & Rahwan, I. (2016). The social dilemma of autonomous vehicles. *Science, 352*(6293), 1573–1576. https://doi.org/10.1126/science.aaf2654

Brimont, L., Saujot, M., & Sartor, O. (2017). Accelerating sustainable mobility with autonomous vehicles. Field Actions Science Reports. *The Journal of Field Actions,* no. Special Issue 17 (December): 22–25.

Chaturvedi, A. (2020). Implications of data privacy once autonomous vehicles hit the roads. Geospatial World (blog). January 14, 2020. https://www.geospatialworld.net/blogs/implications-of-data-privacy-once-autonomous-vehicles-hit-the-roads/. Accessed 28 Dec 2020.

Daniels, N. (1979). Wide reflective equilibrium and theory acceptance in ethics. *The Journal of Philosophy, 76*(5), 256. https://doi.org/10.2307/2025881

Daniels, N. (1996). *Justice and justification: Reflective equilibrium in theory and practice. Cambridge Studies in Philosophy and Public Policy*. Cambridge University Press.

Daniels, N. (2020). "Reflective equilibrium." In *The Stanford Encyclopedia of Philosophy, edited by E. N. Zalta, Summer 2020*. Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/sum2020/entries/reflective-equilibrium/.

Davnall, R. (2020). Solving the single-vehicle self-driving car trolley problem using risk theory and vehicle dynamics. *Science and Engineering Ethics, 26*(1), 431–449. https://doi.org/10.1007/s11948-019-00102-6

Dhar, V. (2016). Equity, safety, and privacy in the autonomous vehicle era. *Computer, 49*(11), 80–83. https://doi.org/10.1109/MC.2016.326

Doorn, N. (2010). Applying Rawlsian approaches to resolve ethical issues: Inventory and setting of a research agenda. *Journal of Business Ethics, 91*(1), 127–143. https://doi.org/10.1007/s10551-009-0073-5

Doorn, N. (2012). Exploring responsibility rationales in research and development (R&D). *Science, Technology, & Human Values, 37*(3), 180–209. https://doi.org/10.1177/0162243911405344

Doorn, N., & Taebi, B. (2018). Rawls's wide reflective equilibrium as a method for engaged interdisciplinary collaboration: Potentials and limitations for the context of technological risks. *Science, Technology, & Human Values, 43*(3), 487–517. https://doi.org/10.1177/0162243917723153

Elgin, C. (2005). Non-foundationalist epistemology: Holism, coherence, and tenability. In M. Steup & E. Sosa (Eds.), *Contemporary Debates in Epistemology* (pp. 156–167). Blackwell.

Erman, E., & Möller, N. (2015). Political legitimacy in the real normative world: The priority of morality and the autonomy of the political. *British Journal of Political Science, 45*(1), 215–233. https://doi.org/10.1017/S0007123413000148

Faber, K., & van Lierop, D. (2020). How will older adults use automated vehicles? Assessing the role of AVs in overcoming perceived mobility barriers. *Transportation Research Part a: Policy and Practice, 133*(March), 353–363. https://doi.org/10.1016/j.tra.2020.01.022

Foot, P. (1967). The problem of abortion and the doctrine of the double effect. *Oxford Review, 5*, 5–15.

Gledhill, J. (2012). Rawls and realism. *Social Theory and Practice, 38*(1), 55–82. https://doi.org/10.5840/soctheorpract20123813

Goodall, N. (2019). More than trolleys: Plausible, ethically ambiguous scenarios likely to be encountered by automated vehicles. *Transfers, 9*(2), 45–58. https://doi.org/10.3167/TRANS.2019.090204

Goodall, N. J. (2014a). Machine ethics and automated vehicles. In G. Meyer & S. Beiker (Eds.), *Road vehicle automation* (pp. 93–102). Springer International Publishing. https://doi.org/10.1007/978-3-319-05990-7_9

Goodall, N. J. (2014b). Ethical decision making during automated vehicle crashes. *Transportation Research Record: Journal of the Transportation Research Board, 2424*(1), 58–65. https://doi.org/10.3141/2424-07

Grunwald, A. (2016). Societal risk constellations for autonomous driving. Analysis, historical context and assessment. In M. Maurer, J. C. Gerdes, B. Lenz & H. Winner (Eds.), *Autonomous driving: Technical, legal and social aspects* (pp. 641–63). Springer. https://doi.org/10.1007/978-3-662-48847-8_30.

Hansson, S. O. (2013). *The ethics of risk*. Palgrave Macmillan UK. https://doi.org/10.1057/9781137333650

Hevelke, A., & Nida-Rümelin, J. (2015). Responsibility for crashes of autonomous vehicles: An ethical analysis. *Science and Engineering Ethics, 21*(3), 619–630. https://doi.org/10.1007/s11948-014-9565-5

Hillerbrand, R., Milchram, C., & Schippl, J. (2019). "The capability approach as a normative framework for technology assessment." *TATuP - Zeitschrift Für Technikfolgenabschätzung in Theorie Und Praxis, 28*(1), 52–57. https://doi.org/10.14512/tatup.28.1.52

Himmelreich, J. (2018). Never mind the trolley: The ethics of autonomous vehicles in mundane situations. *Ethical Theory and Moral Practice, 21*(3), 669–684. https://doi.org/10.1007/s10677-018-9896-4

JafariNaimi, N. (2018). Our bodies in the trolley's path, or why self-driving cars must *not* be programmed to kill. *Science, Technology, & Human Values, 43*(2), 302–323. https://doi.org/10.1177/0162243917718942

Jubb, R. (2015). Playing Kant at the court of King Arthur. *Political Studies, 63*(4), 919–934. https://doi.org/10.1111/1467-9248.12132

Keeling, G. (2020). Why trolley problems matter for the ethics of automated vehicles. *Science and Engineering Ethics, 26*(1), 293–307. https://doi.org/10.1007/s11948-019-00096-1

Larmore, C. (2013). What is political philosophy? *Journal of Moral Philosophy, 10*(3), 276–306. https://doi.org/10.1163/174552412X628896

Leben, D. (2017). A Rawlsian algorithm for autonomous vehicles. *Ethics and Information Technology, 19*(2), 107–115. https://doi.org/10.1007/s10676-017-9419-3

Leben, D. (2019). *Ethics for robots: How to design a moral algorithm*. Routledge/Taylor & Francis Group.

Lin, P. (2016). Why ethics matters for autonomous cars. In M. Maurer, J. C. Gerdes, B. Lenz & H. Winner (Eds.), *Autonomous driving: Technical, legal and social aspects* (pp. 69–85). Springer. https://doi.org/10.1007/978-3-662-48847-8

Nyholm, S., & Smids, J. (2016). The ethics of accident-algorithms for self-driving cars: An applied trolley problem? *Ethical Theory and Moral Practice, 19*(5), 1275–1289. https://doi.org/10.1007/s10677-016-9745-2

Owen, R., Macnaghten, P., & Stilgoe, J. (2012). Responsible research and innovation: From science in society to science for society, with society. *Science and Public Policy, 39*(6), 751–760. https://doi.org/10.1093/scipol/scs093

van de Poel, I. (2016). An ethical framework for evaluating experimental technology. *Science and Engineering Ethics, 22*(3), 667–686. https://doi.org/10.1007/s11948-015-9724-3

van de Poel, I., & Zwart, S. D. (2010). Reflective equilibrium in R & D networks. *Science, Technology, & Human Values, 35*(2), 174–199. https://doi.org/10.1177/0162243909340272

Pype, P., Daalderop, G., Schulz-Kamm, E., Walters, E. & von Grafenstein, M. (2017). Privacy and security in autonomous vehicles. In D. Watzenig & M. Horn (Eds.), *Automated driving: Safer and more efficient future driving* (pp. 17–27). Springer International Publishing. https://doi.org/10.1007/978-3-319-31895-0_2

Rannenberg, K. (2016). Opportunities and risks associated with collecting and making usable additional data. In M. Maurer, J. C. Gerdes, B. Lenz & H. Winner (Eds.), *Autonomous driving: Technical, legal and social aspects* (pp. 497–517). Springer. https://doi.org/10.1007/978-3-662-48847-8_24

Rawls, J. (1974). The independence of moral theory. *Proceedings and Addresses of the American Philosophical Association, 48*, 5. https://doi.org/10.2307/3129858

Rawls, J. (1985). Justice as fairness: Political not metaphysical. *Philosophy & Public Affairs, 14*(3), 223–251.

Rawls, J. (1995). Political liberalism: Reply to Habermas. *The Journal of Philosophy, 92*(3), 132. https://doi.org/10.2307/2940843

Rawls, J. (1997). The idea of public reason revisited. *The University of Chicago Law Review, 64*(3), 765–807. https://doi.org/10.2307/1600311

Rawls, J. (1999). *A theory of justice* (Revised). Belknap Press.

Rawls, J. (2001). *Justice as fairness: A restatement*. Edited by E. Kelly. Harvard University Press.

Rawls, J. (2005). *Political liberalism. Expanded ed. Columbia Classics in Philosophy*. Columbia University Press.

Rodríguez-Alcázar, J. (2017). Beyond realism and moralism: A defense of political minimalism. *Metaphilosophy, 48*(5), 727.

Rodríguez-Alcázar, J., Bermejo-Luque, L. & Molina-Pérez, A. (2020). Do automated vehicles face moral dilemmas? A plea for a political approach. *Philosophy & Technology*. https://doi.org/10.1007/s13347-020-00432-5

Rossi, E. (2019). Being realistic and demanding the impossible. *Constellations, 26*(4), 638–652. https://doi.org/10.1111/1467-8675.12446

SAE International (2018). "J3016B: Taxonomy and definitions for terms related to driving automation systems for on-road motor vehicles - SAE International." SAE Interational. https://www.sae.org/standards/content/j3016_201806/

de SantoniSio, F. (2017). Killing by autonomous vehicles and the legal doctrine of necessity. *Ethical Theory and Moral Practice, 20*(2), 411–429. https://doi.org/10.1007/s10677-017-9780-7

Schmidt, M. W. (forthcoming). Das Überlegungsgleichgewicht als Lebensform: Versuch zu einem vertieften Verständnis der durch John Rawls bekannt gewordenen Rechtfertigungsmethode.

von Schomberg, R. (2013). "A vision of responsible research and innovation." *Responsible Innovation* (pp. 51–74). John Wiley & Sons Ltd. https://doi.org/10.1002/9781118551424.ch3

Schreurs, M. A., & Steuwer, S. D. (2016). Autonomous driving—Political, legal, social, and sustainability dimensions. In M. Maurer, J. C. Gerdes, B. Lenz & H. Winner (Eds.), *Autonomous driving: Technical, legal and social aspects* (pp. 149–71). Springer. https://doi.org/10.1007/978-3-662-48847-8_8.

Taebi, B. (2017). Bridging the gap between social acceptance and ethical acceptability. *Risk Analysis, 37*(10), 1817–1827. https://doi.org/10.1111/risa.12734

Thomas, A. (2017). Rawls and political realism: Realistic utopianism or judgement in bad faith? *European Journal of Political Theory, 16*(3), 304–324. https://doi.org/10.1177/1474885115578970

Thomson, J. J. (1976). Killing, letting die, and the trolley problem. *Monist, 59*(2), 204–17. https://doi.org/10.5840/monist197659224.

Wallach, W., & Colin, A. (2008). *Moral machines: Teaching robots right from wrong*. Oxford University Press.

Williams, B. (2005). *In the beginning was the deed: Realism and moralism in political argument*. Princeton University Press.