Journal of Urban Economics 137 (2023) 103591

Contents lists available at ScienceDirect

Journal of Urban Economics

journal homepage: www.elsevier.com/locate/jue

Congestion and incentives in the age of driverless fleets☆

Federico Boffa[a,b], Alessandro Fedele[a], Alberto Iozzi[c,d,*]

[a] Free University of Bolzano, Italy
[b] Collegio Carlo Alberto, Italy
[c] Università di Roma 'Tor Vergata', Italy
[d] SOAS University of London, United Kingdom

ARTICLE INFO

JEL classification:
D4
H2
L1
R4

Keywords:
Congestion
Congestion externalities
Fleets
Autonomous vehicles
Congestion charges

ABSTRACT

The diffusion of autonomous vehicles (AVs) will expand the tools to manage congestion. Differently than fleets of traditional vehicles, operators of fleets of AVs will be able to assign different travelers to different routes, potentially inducing different congestion levels (and speed). We look at the effects of the technological transition from traditional to autonomous vehicles. Our model exhibits a unit mass of heterogeneous individuals. Some of them use the services of a fleet, while others do not, and travel independently. With few fleet users, the fleet technology (traditional vs automated vehicles) is immaterial to welfare. On the contrary, when there are many fleet users, we show that, if fleets do not price any individuals out of the market, the differentiation in congestion across routes under the automated fleet is welfare-reducing. When, instead, fleets price some individuals out of the market, the welfare effects of the transition are ambiguous and depend on the interplay between the extent of rationing by both types of fleets and the extent of differentiation by the AVs fleet. Finally, we characterize the tax restoring the first best with AVs. It involves charging different taxes across lanes, starkly different between independent travelers and the fleet. While independent travelers should be charged lane-specific congestion charges, the fleet should be imposed a scheme involving a congestion-based tax and a subsidy.

1. Introduction

Technological innovation is drastically affecting urban mobility. Understanding its effects on congestion is of utmost importance because congestion costs, while hidden and hard to measure due to their nature of opportunity costs, represent a significant component of traveling costs. Congestion costs, which are heterogeneous across individuals (reflecting heterogeneity in individuals' value of time, see Small, 2012, and Small et al. 2005), have been estimated to amount, on aggregate, to more than one hundred billion dollars yearly in the U.S. and to be steadily increasing over time (Schrank et al., 2011).

One of the biggest changes in urban mobility is the rise of fleets, triggered by the diffusion of smartphones and geo-localization systems. An increasing share of passenger travel is provided by the fleets of ridesharing companies, such as Uber and Lyft.[1] This trend will likely only be magnified when fully autonomous vehicles (AVs), software-driven cars that no longer require a human driver (thus allowing companies to save on the driver's cost), will be deployed on a large scale. Indeed, many predicts that consumers' investment in private cars

☆ We are very grateful to the Editor, Kristian Behrens, and two anonymous reviewers for their careful reading and extremely useful comments. We also thank Toomas Hinnosaar, Jakub Kastl, Ryan Kellogg, Ignacio Monzon, Giacomo Ponzetto, Salvatore Piccolo, Aleksey Tetenov, Matthew Turner, and Tommaso Valletti for useful discussions and helpful comments. We thank participants in the NBER Workshop on "The economics of autonomous and electric vehicles" at Stanford, in seminars at Collegio Carlo Alberto, Free U. of Bolzano, Hong Kong Polytechnic, IMT Lucca, SOAS U. of London, Tor Vergata U. of Rome, U. of the Balearic Islands, U. of Siena, U. of Naples Federico II, in workshops at Católica Porto Business School, Sapienza U. of Rome, U. of Sassari (IBEO), U. of Bergamo, UCLouvain (LED Research Group) and in conference sessions at AISRE, ASSET, EARIE, ERSA, IAERE, ITEA (Best paper award in Paris 2019), JEI, SIE, SIEPI, and SIET. All errors are ours. The authors acknowledge financial support from the Italian Ministry of University and Scientific Research (PRIN, grant 201797Z7NP), Boffa and Fedele from the Free University of Bolzano (grant WW201E).

* Corresponding author at: Università di Roma 'Tor Vergata', Italy.
E-mail addresses: federico.boffa@unibz.it (F. Boffa), alessandro.fedele@unibz.it (A. Fedele), alberto.iozzi@uniroma2.it (A. Iozzi).

[1] In 2016, fleets accounted for 15% of all intra San Francisco vehicle trips. In New York, the share of vehicle trips covered by fleets has increased dramatically in recent years and has doubled annually between 2014 and 2016 (Erhardt et al., 2019; Mangrum and Molnar, 2020).

https://doi.org/10.1016/j.jue.2023.103591
Received 1 December 2021; Received in revised form 24 July 2023
Available online 1 September 2023
0094-1190/© 2023 The Authors. Published by Elsevier Inc. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

is bound to shrink and that urban traffic will be organized around fleets (Fagnant and Kockelman, 2015; Ward et al., 2019).[2]

The diffusion of fleets will affect the organization of urban travel and the management of congestion. Congestion not only derives from transport infrastructures being inadequate relative to demand, but it is also the result of a standard externality. In a conventional setting, with atomistic drivers and traditional vehicles, drivers do not factor in their travel decisions the external effect in terms of congestion they impose on fellow travelers since. By contrast, fleets, because of their ability to centralize decisions, have the incentive to do so — although possibly in a way that deviates from welfare maximization.

Crucially, the way fleets may deal with congestion depends on the nature of their vehicles, traditional or autonomous. Fleets of AVs will allow their operators to assign different travelers to different routes. They can therefore affect the aggregate congestion costs at any point in time both by affecting the total number of travelers and by assigning travelers to more or less congested (slower or faster) routes based on their (heterogeneous) willingness to pay for time savings. Instead, the assignment of travelers to routes with different congestion levels by fleets of traditional vehicles is very costly and difficult to implement, hence not observed in practice. Fleets of traditional vehicles can then affect congestion through the total number of vehicles at any point in time only.[3]

In this paper, we look at the effects of the technology (traditional versus autonomous vehicles) on congestion and welfare, in an urban environment where a share of the transport services is provided by a monopolistic fleet. We specifically focus on the ability of an AVs fleet to assign travelers to routes with different congestion (and speed), which differentiates it from fleets of traditional vehicles.[4] Since individuals are heterogeneous in their benefit from traveling and in their disutility from congestion, we first show that welfare maximization requires differentiating the congestion level across routes. One may thus be tempted to argue that a fleet of AVs is welfare superior to a conventional fleet, due to its ability to assign travelers to different routes. We show that this intuition is incomplete. A fleet of AVs with market power uses the combination of prices and congestion differentiation as tools to price discriminate across different types of travelers. It thereby induces a level of differentiation across routes, as well as a level of travelers' rationing, that deviate from welfare maximization, and may even reduce welfare vis-à-vis a corresponding fleet of traditional vehicles.

In our model, individuals travel on a road segmented into two separate parallel and congested lanes. Alternatively, one can think of two different routes in a more complex network. The lanes are ex-ante identical, but can ex-post differ in the level of congestion. Some individuals use the service of a fleet (either traditional or autonomous) operated by a monopolist, while others travel independently. The congestion cost an individual suffers depends on her type, a congestion disutility parameter, common across individuals, and the number of vehicles traveling on the same lane, which determines its congestion and speed. Consistent with evidence pointing to a positive relationship between income and the value of time, we assume that individuals with a larger utility from the trip suffer from a larger congestion disutility.[5] We look at the equilibrium assignment of individuals to one of the two lanes or to not traveling in the cases of both traditional and AVs fleets, and we compare welfare.

We show that, when the fleet is sufficiently small, the outcome in terms of welfare is the same for the automated and the traditional monopolist. While the traditional monopolist does not differentiate across lanes because of the technological constraint, the automated monopolist is forced to charge the same price across the two lanes, since independent travelers can freely switch across them, up to the point where the level of congestion in the two lanes is identical. The ability to differentiate congestion level (and, as a result, speed) associated with AVs matters, instead, when the fleet is sufficiently large. When the common congestion disutility parameter is high, so that both types of monopolists prefer not to price some individuals out of the market, the differentiation in congestion induced by the automated monopolist turns out to be welfare-inferior compared to the no differentiation operated by the traditional monopolist. In the case in which all individuals travel with the fleet (so there are no independent individuals), this is due to the excess differentiation operated by the automated monopolist vis-à-vis the welfare-maximizing level, which is welfare inferior to the lack of differentiation under the traditional monopolist. On the contrary, when the common congestion disutility parameter is low, the welfare ranking across the two monopolists, traditional and automated, depends on the combination between the extent of congestion differentiation across lanes by the automated monopolist (which reduces welfare) and the extent of rationing, which we find to be weakly higher under the traditional monopolist. When rationing is welfare-reducing, then AVs may increase welfare. When rationing, instead, is welfare increasing, welfare is higher under a traditional fleet.

Part of our results parallels those obtained in the airline economics literature, under air carriers with market power, in terms of the congestion levels: see, for instance, Brueckner (2002), Basso (2008), and Verhoef and Silva (2017), and the empirical counterparts estimating the relation between airport concentration and congestion (Mayer and Sinai, 2003; Rupp, 2009; Daniel and Harback, 2008; Molnar, 2013). However, we crucially add individuals' heterogeneity in the congestion disutility, and, as a result, we examine how congestion management through travelers' sorting across lanes affects welfare. Our findings also relate to Czerny and Zhang (2015), who study price discrimination in the presence of congestion externalities, and to Lindsey et al. (2018), who analyze internalization by road users with significant shares of traffic flows in the context of a bottleneck model.

We conclude our analysis by investigating the case of a tax authority that imposes taxes to restore social optimality. We argue that the first best cannot be restored with traditional vehicles since technological constraints prevent congestion differentiation across lanes. With AVs, we show that, when all travelers are independent, lane-specific congestion charges, reflecting the marginal (external) cost imposed on the other vehicles, restore optimality. When, instead, all individuals travel with the fleet, restoring first best requires a very different tax/subsidy scheme, which, on aggregate, may involve a transfer to the monopolist. Our result aligns with findings obtained in the literature on airports when carriers have market power (Daniel, 1995; Brueckner, 2002; Pels and Verhoef, 2004; Brueckner, 2005; Basso and Zhang, 2007; Silva and

[2] Fleets of robotaxis, i.e., taxis operated by ride service companies through AVs, are already operating in selected urban areas. Examples include Baidu's autonomous ride-hailing platform Apollo Go which covers more than ten cities in China, Cruise - a GM subsidiary - operating in San Francisco, Austin and Phoenix, and Waymo - an Alphabet subsidiary - providing services in Phoenix and San Francisco.

[3] In principle, the route choice could be forced on drivers of traditional vehicles (and the GPS technology could allow for monitoring). However, monitoring costs would be prohibitively high, in part because of the costs of precisely identifying vehicles' location, in part because of agency problems (moral hazard on the side of drivers), and of legal issues (drivers are often currently acting as independent contractors, as opposed to employees). In fact, to the best of our knowledge, there is no fleet of conventional vehicles imposing such routing constraints on their drivers.

[4] The AVs technology brings about changes in a variety of other dimensions, including labor cost, individuals' disutility from congestion, contribution of each vehicle to congestion, risk of accidents and liability issues. We abstract away from them in the model, and we discuss some of those in Section 8.

[5] Estimates of the elasticity of time value to income range from about 0.5 to 1, and they are increasing at higher levels of income (Börjesson et al., 2012). In the context of a rising average income, the current trends of increasing inequality in many Western countries should only exacerbate such heterogeneity.

Verhoef, 2013). We innovate over that literature by characterizing a simple tax/subsidy scheme that restores the incentives to optimality not only in the level of rationing but also in the degree of differentiation across lanes.

To the best of our knowledge, only a few papers consider congestion with reference to AVs. Lamotte et al. (2017) develop a bottleneck model to investigate the commuters' choice between conventional and autonomous vehicles, while van den Berg and Verhoef (2016) focus on the impact of AVs on road capacity, studying the deployment of infrastructures resulting from the transition to AVs.[6] Finally, our paper is close to Ostrovsky and Schwarz (2018), who investigate the interplay between autonomous transportation, carpooling, and road pricing to achieve socially efficient outcomes.

The rest of the paper is organized as follows. Section 2 describes the model. Section 3 illustrates the first best. Section 4 characterizes the equilibrium in a fleet of traditional vehicles. Section 5 characterizes the equilibrium in an AV fleet. Section 6 provides the welfare analysis. Section 7 examines taxation to restore social optimality. Section 8 discusses the robustness of the results to some possible extensions of our model, including non-linear congestion costs, competition between fleets, change in the market fundamentals as a result of the transition to AVs, and the introduction of a dynamic setting. Section 9 concludes. Derivations and proofs of all Propositions are relegated to Appendix A. Appendix B describes the numerical simulations used to derive some welfare results when some individuals are priced out of the market. Appendix C replicates our results in the case of quadratic congestion costs.

2. The model

Lanes and individuals' utility. There is a unit mass of individuals, each with a unit demand for a trip from a common origin to a common destination. Trips occur along a single road connecting the origin and the destination. The road is divided into two lanes. As mentioned, one can also think of two different routes. The two lanes/routes are congested at any positive mass of travelers. They are ex-ante identical but may differ ex-post because of a different mass of travelers, leading to different congestion levels. We refer to the (weakly) more congested lane as the *slow* lane and denote its mass of travelers by s. Similarly, the (weakly) less congested lane is referred to as the *fast* lane, and its mass of travelers is denoted by $f \leq s$.

Individuals are heterogeneous. Their type, denoted by θ, is assumed to be uniformly distributed in the $[0,1]$ interval.[7] A type-θ individual's utility function is

$$U(\theta) = \begin{cases} 0 & \text{when not traveling;} \\ B(\theta) - \theta g s & \text{when traveling in the slow lane;} \\ B(\theta) - \theta g f & \text{when traveling in the fast lane.} \end{cases} \tag{1}$$

The term $B(\theta)$, with $B'(\theta) > 0$ and $B''(\theta) \leq 0$, is the gross benefit from traveling. The terms $\theta g s$ and $\theta g f$ denote the disutility from congestion. They depend on: (*i*) the actual level of congestion experienced by the individual, equal to the mass of travelers in the same lane, either s or f; (*ii*) a type-independent parameter $g > 0$, representing the common component of travel time disutility due to congestion; (*iii*) the type θ, representing the idiosyncratic component of travel time disutility due to congestion. We assume the congestion disutility to be linear in the traffic volume. Moreover, θ determines the travelers' value of travel while, at the same time, affecting the congestion cost they suffer. The assumption that both increase with θ is consistent with evidence that points to a positive relationship between the value of time and income (see, for instance, Small, 2012).

We assume travelers utility to be increasing in θ, i.e., $U'(\theta) > 0$. A sufficient condition for this to occur for any s and f is stated in the following:

Assumption 1. $g < g_{max} \equiv B'(1)$.

We further assume that every individual's utility from traveling is positive, which requires the following:

Assumption 2. $U(0) = B(0) > 0$.

Individuals' identity. We assume that the mass of individuals is composed of two different groups:

- *corporate individuals*: when traveling, they use vehicles that are managed by a fleet operator;
- *atomistic individuals*: when traveling, they use vehicles that are *not* managed by a fleet operator.

We let the mass of corporate individuals be equal to $\mu \in [0,1]$, and the remaining mass $1-\mu$ be composed of atomistic individuals. These two proportions are exogenously given. Belonging to either group may depend on individual preferences and on long-run decisions, such as, for instance, the choice to buy a vehicle, which are not modeled here. Moreover, we let the distribution of the two groups of individuals be independent of the type θ: in any subinterval of the unit line, there is a fraction μ of corporate individuals and a fraction $1-\mu$ of atomistic individuals. While this makes our analysis more tractable, it also reflects the lack of clear evidence on how the value of travel/congestion disutility θ is related to the choice of owning private vehicles vis-à-vis using fleet services (see, for instance, Alemi et al., 2019; Tirachini, 2020).[8]

The fleet. The fleet of vehicles providing transportation services to corporate individuals is managed by a monopolistic firm. We assume that the monopolist incurs zero fixed and marginal costs, it can charge a uniform price per lane, but it cannot set different prices to different customers using the same lane, possibly because of asymmetric information or a privacy protection regulation (Montes et al., 2018).[9]

We explore two different technological scenarios:

- *traditional vehicles* (sometimes referred to as non-AVs): the monopolist does not possess the technology to affect the allocation of corporate travelers across lanes. This is meant to reflect the status before the adoption of AVs when the technology does not allow to cheaply and effectively monitor travelers' itineraries;
- *autonomous vehicles*: the monopolist has the technology to allocate corporate travelers across lanes. This captures the situation in which the fleet operator uses AVs and centrally determines itineraries by programming the autonomous driving software.

Our simple two-lanes/routes setting can be interpreted as a stylized illustration of a more complex route network, where the AVs technology, with its ability to monitor travelers' itineraries accurately, appears to

[6] Silva et al. (2016) and Lindsey et al. (2019) also analyze equilibria when drivers are non-atomistic; Simoni et al. (2019) use agent-based simulations to evaluate the impact of different congestion pricing and tolling strategies in the presence of AVs.

[7] For a uniform distribution assumption of travelers' type, see, e.g., Brueckner (2002).

[8] In his comprehensive survey on ride hailing, Tirachini (2020) reports that the choice of using the services of a fleet depends on several factors, including age, availability of public transport, residential and workplace density, and flexibility of job hours, but that the relationship with income is not clear.

[9] An additional reason that may prevent a firm from charging personalized pricing is consumers' aversion to these pricing practices: see, for instance, Leibbrandt (2020) and http://news.bbc.co.uk/2/hi/business/914691.stm (last accessed, March 2, 2023) for the case of Amazon.

be a necessary tool to allocate travelers to one of many routes, each associated with a price and congestion level.

The game. We initially look at a two-stage game. In the first stage, the monopolist sets the fares to be paid by corporate individuals, $p \geq 0$ for the (weakly) slow lane and $P \geq p$ for the (weakly) fast lane. In the second stage, all corporate and atomistic individuals simultaneously make their travel decisions. All players have full information on the entire game and the equilibrium concept is subgame perfection.

In Section 7, we add an initial stage to this game, in which a tax authority chooses a tax scheme. After this initial stage, the rest of the game unfolds as described before.

Individuals' incentives. In the second stage of the game, individuals, both atomistic and corporate, choose whether or not to travel and, if they travel, in which lane to do so.

To illustrate the individual incentives, note that, for a type θ, the decision depends on her utility $U(\theta)$ and on the fares (or taxes) she pays when traveling in the slow or the fast lane. A type-θ individual travels in the slow lane if and only if her individual rationality (IR) constraint holds,

$$B(\theta) - \theta g s - p \geq 0, \tag{2}$$

and her incentive compatibility (IC) constraint holds,

$$B(\theta) - \theta g s - p \geq B(\theta) - \theta g f - P. \tag{3}$$

Similarly, she travels in the fast lane if and only if

$$B(\theta) - \theta g f - P \geq 0, \tag{4}$$

$$B(\theta) - \theta g f - P \geq B(\theta) - \theta g s - p. \tag{5}$$

3. First best

We start by considering the benchmark of a utilitarian welfare-maximizing social planner, who perfectly knows the economy. It can decide which individuals travel and directly allocate travelers to the two lanes. The social welfare function is simply given by

$$\int_0^1 U(\theta) d\theta. \tag{6}$$

Welfare maximization requires partitioning travelers into at most three groups. Some low θ-types may not travel, while all the other θ's are assigned to the two lanes, with the highest θ's traveling in the fast lane.[10] Then, (6) may be rearranged as

$$W \equiv \int_{1-s-f}^{1-f} (B(\theta) - \theta g s)\, d\theta + \int_{1-f}^{1} (B(\theta) - \theta g f)\, d\theta, \tag{7}$$

where the first integral represents the aggregate utility of types traveling in the slow lane, and the second is the aggregate utility of types traveling in the fast lane. The planner's problem is

$$\max_{s \geq 0, f \geq 0} W \quad \text{s.t. } s + f \leq 1. \tag{8}$$

The solution to the planner's problem is characterized in the next Proposition.

Proposition 1. *Let s_{FB} and f_{FB} denote the interior solutions to problem (8), i.e., $s_{FB} + f_{FB} < 1$.[11] Also, let $\bar{s}_{FB}$ and $\bar{f}_{FB}$ denote the corner solutions to problem (8), i.e., $\bar{s}_{FB} + \bar{f}_{FB} = 1$. Then, at the social optimum,*

- *when $g > g_{FB}$, rationing occurs and the lanes have different congestion levels, $s_{FB} > f_{FB}$. The individuals with $\theta \in [0, 1 - s_{FB} - f_{FB}]$ do not travel, those with $\theta \in [1 - s_{FB} - f_{FB}, 1 - f_{FB}]$ travel in the slow lane and those with $\theta \in [1 - f_{FB}, 1]$ travel in the fast lane. s_{FB} and f_{FB} satisfy*

$$f_{FB}(s_{FB}) = \frac{1}{3}\left(2\left(1 + s_{FB}\right) - \sqrt{7 s_{FB}^2 - 4 s_{FB} + 4}\right); \tag{9}$$

- *when $g \leq g_{FB}$, no rationing occurs, and the lanes have different congestion levels, $\bar{s}_{FB} > \bar{f}_{FB}$. The individuals with $\theta \in [0, \bar{s}_{FB}]$ travel in the slow lane, and those with $\theta \in [\bar{s}_{FB}, 1]$ travel in the fast lane. $\bar{s}_{FB}$ and $\bar{f}_{FB}$ are*

$$\bar{s}_{FB} = \frac{1}{2} + \frac{\sqrt{7} - 2}{6} \approx 0.6076 \quad \text{and} \quad \bar{f}_{FB} = \frac{1}{2} - \frac{\sqrt{7} - 2}{6} \approx 0.3924; \tag{10}$$

where

$$g_{FB} \equiv \frac{36\, B(0)}{4 + \sqrt{7}} \approx 5.4179 \times B(0). \tag{11}$$

This Proposition characterizes the socially optimal allocation of individuals. The planner may exclude the individuals with the lowest benefit from traveling so that the market is not fully covered. This occurs when the utility from traveling in the slow lane enjoyed by the type-0 individual, $B(0)$, is lower than the increase in the aggregate congestion costs this individual imposes on all fellow travelers in the slow lane.[12] All the remaining travelers are sorted in the two lanes. A mass of travelers equal to s_{FB} (or $\bar{s}_{FB}$, in the case of full coverage) is allocated to the slow lane, and a mass of travelers f_{FB} (or $\bar{f}_{FB}$, in the case of full coverage) to the fast lane. Intuitively, travelers allocated to the fast lane are those with the highest θ.

When rationing occurs, s_{FB} and f_{FB} cannot be characterized explicitly, but their relation is given in (9). This expression shows that the planner optimally differentiates across lanes: travelers with relatively high θ, who suffer the most from congestion, are assigned to the fast, less congested, lane. In addition, comparative statics results show that $\frac{\partial s_{FB}}{\partial g} < 0$ and $\frac{\partial f_{FB}}{\partial g} < 0$. This reflects the intuition that a larger common component of the congestion disutility, g, is associated with a lower mass of travelers in each lane and, therefore, lower market coverage in the first best.

4. Traditional vehicles: equilibrium analysis

In this section, we study the scenario with traditional non-autonomous vehicles, in which the monopolist cannot monitor the route taken by travelers. This will be shown to imply, straightforwardly, that the monopolist cannot differentiate the congestion level between the lanes, which, in equilibrium, are equally congested.

We start by solving backward our two-stage game. In the second stage, atomistic and corporate individuals simultaneously make their travel choices after observing the fares set by the monopolist. We first investigate the behavior of the mass $1 - \mu \in [0, 1]$ of atomistic individuals. They rely on traditional vehicles not managed by a fleet and pay no fares. We denote with s^a and f^a the mass of atomistic travelers in the slow and the fast lane, respectively. Using Assumptions 1 and 2, it is easy to show that, when no fare is paid, the IR constraint (4) and the IC constraint (5) hold for any atomistic traveler. As a result, all atomistic individuals decide to travel, so that

$$s^a + f^a = 1 - \mu, \tag{12}$$

and choose the (weakly) less congested lane.

[10] See Appendix A for the derivation of this result.

[11] The subscript FB is a mnemonic for First Best.

[12] Condition $g > g_{FB}$ can be rearranged as $B(0) < g \frac{(\bar{s}_{FB})^2}{2}$, where the RHS is indeed the aggregate marginal congestion cost for those traveling in the slow lane.

We now turn to corporate individuals. We denote with s^c and f^c the mass of those traveling in the slow and fast lane, respectively. Because of the impossibility of monitoring travelers' itineraries using a fleet of traditional vehicles, the monopolist cannot allocate a traveler to the lane she paid for. It follows that, for any pair of prices $p < P$ set by the monopolist for the two lanes, all corporate travelers will pay p and still choose the lane that gives her the higher utility. The IC constraint (5) becomes $B(\theta) - \theta g f - p \geq B(\theta) - \theta g s - p$, which always holds.

Given the behavior of corporate and atomistic individuals, any second-stage equilibrium allocation of travelers is such that the two lanes are equally congested. Indeed, if a faster lane existed, the travelers in the slow lane would prefer to switch lanes to enjoy lower congestion. This holds true up to the point where the fast lane becomes as congested as the slow lane. Using (12), we can write

$$s = f = \frac{s^c + f^c + 1 - \mu}{2}. \tag{13}$$

To fully characterize the equilibrium of this second stage, we need to determine the total mass of travelers in each lane, given by the IR constraints (2) and (4). Because of the equal mass of travelers in the two lanes, the constraints collapse to a single one that, using (13), can be written as

$$B\left(1 - \frac{s^c + f^c}{\mu}\right) - \left(1 - \frac{s^c + f^c}{\mu}\right) g \, \frac{s^c + f^c + 1 - \mu}{2} - p \geq 0, \tag{14}$$

where $\theta = 1 - \frac{s^c + f^c}{\mu}$ is the type of the corporate individual indifferent between not traveling and traveling in either lane.

In the first stage of the game, denoting by c the total mass of corporate travelers ($c \equiv s^c + f^c$), the maximization problem faced by the monopolist may be written as

$$\max_{c \geq 0} \left[B\left(1 - \frac{c}{\mu}\right) - \left(1 - \frac{c}{\mu}\right) g \, \frac{c + 1 - \mu}{2} \right] c, \tag{15}$$
$$\text{s.t.} \quad c \leq \mu,$$

where we take the IR constraint (14) to be binding, as it must be in equilibrium.

The following Proposition characterizes the equilibrium allocation of travelers using traditional vehicles.

Proposition 2. *Let c_N denote the interior solution to problem (15), i.e., $c_N < \mu$.*[13] *Then, in equilibrium, lanes are always equally congested and*

- *when $g < g_N$, rationing occurs. Atomistic and corporate individuals with $\theta \in \left[1 - \frac{c_N}{\mu}, 1\right]$ travel, while corporate individuals with $\theta \in \left[0, 1 - \frac{c_N}{\mu}\right]$ do not travel. Any allocation of travelers is such that*

$$\begin{aligned} s = f &= \tfrac{c_N + 1 - \mu}{2} < \tfrac{1}{2}, \\ s^a + f^a &= 1 - \mu, \\ s^c + f^c &= c_N < \mu; \end{aligned} \tag{16}$$

- *when $g \geq g_N$, no rationing occurs, so all individuals travel. Any allocation of travelers is such that*

$$\begin{aligned} s = f &= \tfrac{1}{2}, \\ s^a + f^a &= 1 - \mu, \\ s^c + f^c &= \mu; \end{aligned} \tag{17}$$

where

$$g_N \equiv 2\left[B'(0) - B(0)\right]. \tag{18}$$

The inability to monitor travelers' itineraries forces the monopolist to charge the same price in both lanes. This determines an identical congestion level across lanes.

The monopolist optimally screens low-θ corporate individuals out of the market when the common component of congestion disutility g is sufficiently low: $g < g_N$. This rationing condition starkly contrasts the corresponding condition in the first-best, which, instead, commands to ration when the parameter when g is sufficiently large, $g > g_{FB}$. To get an intuition for this result, consider that, in our framework, the individual congestion disutility increases with g. As g gets smaller, the individuals' utility function $U(\theta)$ turns more sensitive to θ, as the now lower congestion disutility only partially compensates the higher willingness to pay associated with a higher θ.[14]

Two additional aspects are worth discussing. First, there is a multiplicity of travelers allocation that satisfy conditions (16) and (17). They all are an equilibrium of the game, and, given that the two lanes share the same congestion level, they yield the same utility level to each market participant.

Second, the monopolist's equilibrium fare is determined by the binding IR constraint of the lowest-θ corporate traveler: either $p = B\left(1 - \frac{c_N}{\mu}\right) - \left(1 - \frac{c_N}{\mu}\right) g \, \frac{c_N + 1 - \mu}{2}$ in the case of rationing, or $p = B(0)$ when the market is fully covered.

5. Autonomous vehicles: equilibrium analysis

In this section, we study the scenario where the monopolist operates a fleet of AVs. The automated monopolist can monitor the route chosen by each traveler, thereby preventing them to travel in a lane different than the one they paid for. We will show that, as a result, the monopolist charges two different prices in the two lanes, as long as the share of corporate travelers is large enough. This determines different levels of congestion in the two lanes and induces travelers to sort across them based on their heterogeneous value of time. Differentiation across lanes increases the monopolist's profits compared to the non-AVs case.

Again, we start by solving backward our two-stage game, in which all individuals (both atomistic and corporate) simultaneously make their travel decisions. We denote with s^c and f^c the mass of corporate travelers in the slow and fast lane. Similarly, s^a and f^a are the mass of atomistic travelers in the two lanes.

We first focus on choices by atomistic individuals. As in the case of traditional vehicles, they all travel, as in (12), and choose the (weakly) less congested lane.

We next move to corporate individuals. Unlike the case of the traditional monopoly, using an AVs fleet allows the monopolist to monitor travelers' itineraries and, thus, to ensure that the lane used by the traveler is consistent with the price she paid for. Each traveler's choice depends on the prices charged in each lane, p ad P, which are at this stage taken as given. It also depends on the expected congestion level prevailing in the two lanes — itself determined by the choices of fellow travelers, atomistic and corporate. In particular, two classes of second-stage equilibria can emerge.

If $1 - \mu < s^c - f^c$, then s^c and f^c may be part of an equilibrium of the second stage where

$$\begin{aligned} s &= s^c, \\ f &= f^c + (1 - \mu) < s. \end{aligned} \tag{19}$$

This is an equilibrium provided that, at s^c and f^c, the IR and IC constraints (2)–(5) are satisfied. In this class of equilibria, the mass of atomistic travelers is relatively low. There is a difference in congestion between lanes, and all atomistic individuals travel in the fast lane.[15]

If instead $1 - \mu \geq s^c - f^c$, an equilibrium may exist only with no differentiation across lanes, so that (13) holds. Such equilibrium is only compatible with $p = P$. For this to be an equilibrium, s^c and

[13] The subscript N is a mnemonic for Non-AVs.

[14] Observe that, by a similar argument, the monopolist rations also when $B'(0)$ is relatively large. When $B'(0)$ is large, $U(\theta)$ gets steeper in θ, which makes quantity less sensitive to fares.

[15] See Appendix A for the derivation of (19).

f^c must satisfy the IR constraint in (14). In this class of equilibria, no differentiation across lanes emerges. This result is due to the relatively high mass of atomistic travelers vis-à-vis the difference in the congestion level across the two lanes. If a faster lane existed, the atomistic travelers in the slow lane would benefit from moving there to enjoy lower congestion.[16]

We turn to the analysis of the first stage of the game, in which the monopolist sets the fares to maximize profits. We separately consider the two classes of second-stage equilibrium allocations of travelers. We note that the monopolist can perfectly anticipate the travelers' choices for any pair of prices p and P. As a result, by choosing the prices, the monopolist induces the congestion levels in the two lanes. Since congestion is an important quality dimension, congestion differentiation could be interpreted as quality differentiation. Then, we could interpret the monopolist offer as a menu of (one per lane) price-congestion (or, equivalently, price-quality) combinations: (p, s) and (P, f), where $p \le P$ and $s \ge f$.[17] The monopolist then provides two vertically differentiated service levels to extract surplus from high θ types (i.e., with a high willingness to pay for low congestion).[18] As previously discussed, the monopolist cannot set different prices for different customers using the same lane, possibly because of informational or regulatory constraints. Hence, quality differentiation, through differentiating congestion across lanes, is the only instrument in the monopolist's hand to price discriminate.

When μ is small, no congestion differentiation emerges across lanes in the second stage, and the monopolist proposes the same fare ($p = P$) to all corporate travelers. In this case, the equilibrium outcome is identical to that described in Proposition 2.

The monopolist problem is significantly different when μ is relatively large, and the second-stage equilibrium allocation of travelers exhibits different congestion levels across lanes, resulting from two different fares in the two lanes. The set of IR and IC constraints (2)–(5) faced by the monopolist are, by a standard argument, reduced to two: first, the IR constraint (2) for the corporate traveler indifferent between not traveling and traveling in the slow lane, that is, for the type $\theta = 1 - \frac{s^c+f^c}{\mu}$; second, the IC constraint (5) for the corporate traveler indifferent between the slow and the fast lane, that is, for the type $\theta = 1 - \frac{f^c}{\mu}$. Using (19), these constraints can be written as

$$B\left(1 - \frac{s^c + f^c}{\mu}\right) - \left(1 - \frac{s^c + f^c}{\mu}\right) g s^c - p \ge 0, \tag{20}$$

$$B\left(1 - \frac{f^c}{\mu}\right) - \left(1 - \frac{f^c}{\mu}\right) g(f^c + 1 - \mu) - P \ge B\left(1 - \frac{f^c}{\mu}\right) - \left(1 - \frac{f^c}{\mu}\right) g s^c - p. \tag{21}$$

Solving (20) and (21) for p and P when they hold as equalities and substituting them into the monopolist profit yields the following maximization problem:

$$\max_{s^c \ge 0, f^c \ge 0} \left[B\left(1 - \frac{s^c + f^c}{\mu}\right) - \left(1 - \frac{s^c + f^c}{\mu}\right) g s^c\right](s^c + f^c) + \left(1 - \frac{f^c}{\mu}\right) g \left[s^c - (f^c + 1 - \mu)\right] f^c, \tag{22}$$
$$\text{s.t. } s^c + f^c \le \mu,$$

The next Proposition illustrates the allocation of travelers.

[16] This implies that, under full coverage, differentiation across lanes cannot emerge when $\mu \le \frac{1}{2}$.

[17] In this menu of contracts, the congestion/quality component of each offer, rather than being set by the monopolist, is endogenously determined by the individuals' choices, based on the monopolist prices (see, for instance, Reitman, 1991).

[18] Our problem is in line with the price discrimination literature. It indeed satisfies the standard assumptions of this literature, that is $\frac{\partial U(.)}{\partial \theta} > 0$ (see Assumption 1), as well as the Spence-Mirrlees condition $\frac{\partial U(\theta,s)}{\partial s \partial \theta} = \frac{\partial U(\theta,f)}{\partial f \partial \theta} = -g < 0$.

Proposition 3. *Let s^c_A and f^c_A be the interior solutions to problem (22), i.e., $s^c_A + f^c_A < \mu$.[19] Also, let $\bar{s}^c_A$ and $\bar{f}^c_A$ be the corner solutions to problem (22), i.e., $\bar{s}^c_A + \bar{f}^c_A = \mu$. Then,*

- *when $\mu \le \mu_A$, the equilibria are as in Proposition 2;*
- *when $\mu > \mu_A$ and $g < g_A$, in equilibrium rationing occurs and the lanes have different congestion levels: corporate individuals with $\theta \in \left[0, 1 - \frac{s^c_A + f^c_A}{\mu}\right]$ do not travel, corporate travelers with $\theta \in \left[1 - \frac{s^c_A + f^c_A}{\mu}, 1 - \frac{f^c_A}{\mu}\right]$ travel in the slow lane, and corporate travelers with $\theta \in \left[1 - \frac{f^c_A}{\mu}, 1\right]$ travel in the fast lane, along with all atomistic travelers. s^c_A, f^c_A, s^a_A and f^a_A satisfy*

$$s^a_A = 0, \tag{23}$$
$$f^a_A = 1 - \mu, \tag{24}$$
$$s = s^c_A > f = f^c_A + 1 - \mu, \tag{25}$$
$$f^c_A(s^c_A) = \frac{1}{3}\left(2\mu - 1 + 2s^c_A - \sqrt{7\left(s^c_A\right)^2 - 2s^c_A(2-\mu) + 1 - \mu + \mu^2}\right); \tag{26}$$

- *when $\mu > \mu_A$ and $g \ge g_A$, in equilibrium no rationing occurs and the lanes have different congestion levels: corporate travelers with $\theta \in \left[0, 1 - \frac{\bar{f}^c_A}{\mu}\right]$ travel in the slow lane and corporate travelers with $\theta \in \left[\frac{\bar{f}^c_A}{\mu}, 1\right]$ travel in the fast lane, along with all atomistic travelers. $\bar{s}^c_A$, $\bar{f}^c_A$, s^a_A and $\bar{f}^c_A$ satisfy*

$$\bar{s}^a_A = 0, \tag{27}$$
$$\bar{f}^a_A = 1 - \mu, \tag{28}$$
$$s = \bar{s}^c_A > f = \bar{f}^c_A + 1 - \mu, \tag{29}$$
$$\bar{s}^c_A = \frac{1}{2} + \frac{\sqrt{4\mu^2 - 2\mu + 1} - 2(1-\mu)}{6}, \tag{30}$$
$$\bar{f}^c_A = \frac{1}{2} - \frac{\sqrt{4\mu^2 - 2\mu + 1} + 4(1-\mu)}{6}; \tag{31}$$

where

$$g_A \equiv \begin{cases} 2\left[B'(0) - B(0)\right] & \text{if } \mu \in \left[0, \frac{1}{2}\right], \\ K(\mu)\left[B'(0) - B(0)\right] & \text{if } \mu \in \left(\frac{1}{2}, 1\right], \end{cases} \tag{32}$$

with $K(\mu) \equiv \frac{18\mu}{8\mu^2 + 5\mu - 1 + (4\mu - 1)\sqrt{4\mu^2 - 2\mu + 1}}$, and

$$\mu_A \equiv \begin{cases} \frac{1}{2} & \text{if } g \ge g_A, \\ \tilde{\mu} & \text{if } g < g_A. \end{cases} \tag{33}$$

When the mass of corporate travelers is sufficiently large, the Proposition shows that the monopolist sets two prices, p and P, which induce two different congestion levels in the two lanes. High θ individuals travel in the least congested lane, along with all the atomistic individuals, while low θ individuals travel in the more congested lane. When, instead, the mass of corporate travelers is small (i.e., $\mu \le \mu_A$), the monopolist with AVs is not able to offer two differentiated services. For any level of μ, at the equilibrium, rationing of low-θ corporate individuals occurs if and only if g is sufficiently low.

Several features of the above results deserve a discussion, which, for expositional reasons, we present separately for the case of full and partial coverage.

Full coverage: $g \ge g_A$. The monopolist fully covers the market when g is sufficiently large. In other words, rationing does not occur only if the parameter for congestion disutility g is relatively large, with the same logic as in Proposition 2.

[19] The subscript A is a mnemonic for AVs.

No differentiation across lanes occurs in equilibrium when μ is sufficiently small ($\mu \leq \frac{1}{2}$), where the critical value of $\mu = \frac{1}{2}$ solves $1 - \mu = \bar{s}_A^c - \bar{f}_A^c$. The presence of at least as many atomistic as corporate individuals prevents the monopolist from inducing a different congestion level across lanes. In this situation, the mass of travelers in the two lanes is identical, and the resulting multiplicity of equilibria is equivalent to that described in Proposition 2, but for a different reason. While the traditional monopolist cannot differentiate across lanes due to technological constraints, uniform congestion across lanes arises endogenously with AVs. Indeed, the automated monopolist possesses the technology to monitor corporate travelers' itineraries and, therefore, to induce their allocation in the lanes. However, it cannot affect atomistic individuals' lane choices. Hence, the presence of a relatively high proportion of atomistic individuals willing to travel in the fast lane undoes any possible congestion gap across lanes. The monopolist charges the same prices for both lanes, equal to $p = B(0)$.

When instead μ is larger than $\frac{1}{2}$, the mass of atomistic travelers is relatively low and cannot bridge the congestion gap across lanes resulting from monopolist choices. Anticipating that all the atomistic travelers use the fast lane, the monopolist charges a sufficiently low price in the slow lane, so that a large mass of corporate travelers is attracted to this lane, and a relatively high price in the fast lane, while few corporate travelers then choose. Unlike the no differentiation case, here the equilibrium allocation of travelers is unique. The equilibrium fares are given by $p = B(0)$ and $P = B(0) + g\left(1 - \frac{\bar{f}_A^c}{\mu}\right)[\bar{s}_A^c - \bar{f}_A^c - (1-\mu)]$.

An interesting feature of this equilibrium is that full coverage may occur with a monopolistic fleet of AVs but not at the social optimum.[20] This implies that the monopolist may dispatch more travelers than the social planner. This finding is at odds with the standard outcome that a monopolist reduces total output. Intuitively, this may happen when g is relatively large so that the demand function faced by the monopolist is sensitive to prices. The monopolist then prefers to set low fares, with the effect that everybody travels. At the same time, because of the relatively large g, rationing occurs at the social optimum. This result depends on the fact that the monopolist prices are set based on the valuation of the marginal consumers, being determined by the binding IR constraint for the θ-type indifferent between not traveling and traveling in the slow lane and the binding IC constraint for the θ-type indifferent between traveling in the slow and in the fast lane. This is clearly in contrast with the objective of the social planner, aimed at maximizing the benefits of all individuals traveling. More generally, the misalignment between monopolist's and planner's incentives depends on the fact that an increase in the number of travelers adversely affects the utility of all the travelers due to the external congestion effect. If we interpret congestion as a quality dimension, our result can be directly related to the seminal work of Spence (1975) on the endogenous quality choice by a monopolist.[21]

Additionally, when μ is sufficiently high, the automated monopolist choices induce an overdifferentiation of congestion across lanes, considering both corporate and atomistic travelers: too few travelers travel in the fast lane as compared to the socially optimal level, $\bar{f}_A^c + 1 - \mu < \bar{f}_{FB}$, and too many in the slow lane, $\bar{s}_A^c > \bar{s}_{FB}$.[22] Again, interpreting congestion as a quality measure allows us to directly relate our findings to Mussa and Rosen (1978), who show that a multi-product monopolist differentiates the quality levels of its products more than a planner would.

The automated monopolist's incentives to differentiate congestion across lanes are magnified by the fact that the high θ travelers' willingness to pay for the fast lane depends on the difference in the extent of congestion across lanes.[23] By inducing a larger difference in congestion across lanes, ceteris paribus, the automated monopolist is able to extract a higher surplus from high θ types traveling in the fast lane.[24]

Partial coverage: $g < g_A$. When g is sufficiently small, the monopolist chooses to price some individuals out of the market.

When the market is partially covered, the monopolist does not differentiate congestion across lanes when $\mu \leq \tilde{\mu}$. $\tilde{\mu}$ is obtained by solving $1 - \mu = s_A^c - f_A^c$ and it is higher than $\frac{1}{2}$. It is above the corresponding value identified under full coverage. When the monopolist optimally prices some corporate travelers out of the market, fewer atomistic travelers than under full coverage are needed to prevent the monopolist from inducing differentiation. With no differentiation, the equilibrium is the same as in Proposition 2 with rationing, and the monopolist charges $p = P = B\left(1 - \frac{c_A}{\mu}\right) - \left(1 - \frac{c_A}{\mu}\right) g \frac{c_A + 1 - \mu}{2}$, where c_A is the equilibrium total mass of corporate travelers, identical to c_N from Proposition 2.

When, instead, $\mu \geq \tilde{\mu}$, the mass of atomistic travelers is relatively low, so the monopolist optimally differentiates congestion across lanes. Again, low θ corporate types will travel in the slow lane, while all the atomistic travelers travel in the fast lane, along with high θ corporate types. The presence of atomistic travelers hurts the profit that the automated monopolist makes on corporate individuals because it increases the congestion level in the most profitable lane for the monopolist, thereby decreasing the price that high θ corporate types are willing to pay for that lane.

In this context, an increase in the congestion cost g has two effects. First, as already discussed, the demand functions faced by the monopolist for each of the two lanes turn more sensitive to its own fares, so the monopolist prefers to set (weakly) low fares, to induce (weakly) more travelers in equilibrium. To see this, consider two individuals traveling in the slow lane and denote θ_s and $\theta_s' > \theta_s$ their types. Since $\frac{\partial U(\theta)}{\partial \theta \partial g} < 0$, a higher g is associated to a lower $U(\theta_s') - U(\theta_s)$. As a result, a higher congestion cost decreases the heterogeneity in the willingness to pay across travelers in the same lane and moves the monopolist's tradeoff towards low fares. Second, the total congestion costs increase more rapidly with θ as g rises. As g increases, the individuals' willingness to pay to avoid congestion, equal to the congestion cost they suffer, increases more rapidly with θ. Thus, individuals in the fast lane, holding prices and congestion levels in the two lanes constant, are less willing to substitute with the slow lane. This establishes a parallel between our notion of substitution across lanes in an environment of vertical differentiation, and the notion of elasticity of substitution in an environment of monopolistic competition, where increasing (decreasing) elasticity of substitution means that varieties are less (more) differentiated (see Parenti et al. 2017).[25]

[20] In Appendix A, we derive the sufficient conditions for this result.

[21] Spence (1975) indeed shows that, as a result of the misalignment between monopolist interest in the value of quality of the marginal unit and the planner's interest in the quality of all the units, the monopolist chooses a quality level that may be either lower or higher than the socially optimally one.

[22] This occurs when $\mu > \frac{1}{9} + \frac{2\sqrt{7}}{9} \approx 0.6991$.

[23] Indeed, using (20) and (21) when holding as equality, we have $P = \left[B\left(1 - \frac{s^c+f^c}{\mu}\right) - \left(1 - \frac{s^c+f^c}{\mu}\right) g s^c\right] + \left(1 - \frac{f^c}{\mu}\right) g\,[s^c - (f^c + 1 - \mu)]$.

[24] If we assumed that the automated monopolist could implement a group pricing scheme, where she observes types, can condition the price offers to the type and is forced to charge a uniform price within each lane, then the automated monopolist incentives to overdifferentiate vis-à-vis the social planner would be attenuated. In this case, prices would not need to satisfy the IC constraints. The price in the fast lane would then be equal to the IR constraint of the marginal traveler in that lane, and would no longer depend on the difference in the extent of congestion across lanes, thereby reducing the monopolist's incentive to increase it.

[25] Our results also confirm the importance of allowing for a variable elasticity of substitution (which has been shown to be a crucial determinant of several outcomes) in models of monopolistic competition; see Parenti et al. (2017).

Finally, we look at the effects of the introduction of the AVs technology on monopolist profits. By allowing the monopolist to discriminate across lanes, the AVs technology is equivalent to the lift of a uniform price constraint on the monopolist maximization problem. When this technological opportunity can be exploited (i.e., when the share of corporate individuals is large enough, that is, when $\mu > \mu_A$), there is a clear benefit for the monopolist. Formally, this can be seen when comparing the monopolist maximization problems in (15) and (22). For the automated monopolist, it is feasible to set identical fares to those set by its traditional counterpart. The outcome, and profits, would then be identical to those under the traditional monopolist. If the automated monopolist chooses a different pair of fares (such as those in Proposition 3, which induce differentiation across lanes), then they must be associated with a (weakly) higher profit. The higher profitability of AVs fleets might contribute to explaining Tesla's prospect mentioned above to stop selling private vehicles and sell robotaxis instead.

6. Welfare analysis

In this section, we investigate the welfare effects of a technological transition of vehicles from traditional to autonomous. To this aim, we contrast the equilibrium welfare under the AVs and non-AVs scenarios. The welfare comparison turns out to be driven by the interplay between the effects of the two margins in which the monopolist's choice may differ across the two scenarios: the extent of congestion differentiation across lanes and the extent of rationing. In turn, such margins are crucially affected by the magnitude of the congestion disutility parameter g and the relative share of corporate travelers μ.

To better illustrate this interplay, we start by analyzing the case in which, given parameter values, full coverage emerges in equilibrium with both traditional and automated vehicles. This happens when the congestion disutility parameter is sufficiently large (i.e., $g \geq g_N$). In this case, the only margin in which the monopolist choice may differ across the two scenarios is the extent of differentiation across lanes.

Full coverage. When $\mu \leq \frac{1}{2}$, we showed that neither monopolist differentiate the congestion level across lanes. Hence, the choices by the two monopolists are identical and lead, in both cases, to social welfare given by

$$W = \int_0^1 B(\theta)d\theta - \int_0^1 \frac{1}{2}\theta g d\theta. \tag{34}$$

Consider now the case of $\mu > \frac{1}{2}$. Since a traditional monopolist does not differentiate congestion across lanes, Eq. (34) gives social welfare. Instead, social welfare under the automated monopolist is

$$W = \int_0^1 B(\theta)d\theta - \mu\left[\int_0^{1-\frac{f^c}{\mu}} \theta\, g\, s^c\, d\theta + \int_{1-\frac{f^c}{\mu}}^1 \theta\, g\,(f^c + 1 - \mu)\, d\theta\right] - (1-\mu)\int_0^1 \theta\, g\,(f^c + 1 - \mu)\, d\theta. \tag{35}$$

Congestion differentiation across lanes, therefore, occurs both in the first best and, as long as $\mu > \frac{1}{2}$, under the automated monopoly, while it does not under the traditional monopolist. One could thus expect that differentiation, performed by the automated monopolist, increases welfare. In fact, we find it always hurts welfare. This results from the interplay of two effects, which play out differently for different levels of μ: one effect relates to the identity of travelers in the two lanes and the other to the extent of differentiation.

We first focus on the effect of the travelers' identity. The automated monopolist's differentiation in the congestion level across lanes leads all the atomistic travelers, including the low θ types, to travel in the fast lane. Since those with a low θ suffer little from congestion, this misallocation generates a welfare loss relative to the traditional monopolist. This welfare-reducing effect is particularly strong when the proportion of atomistic travelers is relatively high.

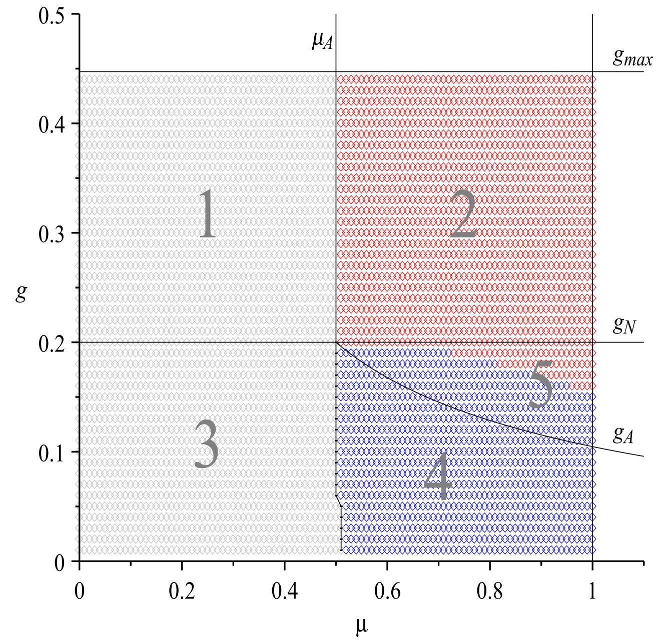

Fig. 1. Welfare comparisons when $B(\theta) = b_0 + b\sqrt{\theta + \frac{1}{4}}$, $b_0 = 0.4$ and $b = 1$.
The figure spans all admissible values for μ (horizontal axis) and g (vertical axis), illustrating for each of these combinations the welfare comparison with and without AVs. The blue area illustrates combinations of parameters such that welfare is higher with AVs (i.e., $W_N < W_A$), the red area illustrates combinations of parameters such that welfare is higher without AVs (i.e., $W_A < W_N$), the gray area illustrates combinations of parameters such that welfare is identical with and without AVs (i.e., $W_A = W_N$). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

The second effect relates to the extent of differentiation, whose impact on welfare depends on the level of μ. The differentiation level is monotonically increasing in μ, small when μ is close to $\frac{1}{2}$, and more than the welfare-maximizing level when μ is close to 1. For sufficiently low μ, the differentiation under the automated monopolist increases welfare over the no-differentiation by the traditional monopolist. As μ becomes large enough, the large extent of differentiation turns welfare reducing relative to no-differentiation (see footnote 22).

Overall, we find that welfare is higher under the traditional monopolist for any level of μ. When μ is sufficiently close to 1, the two effects of travelers' identity and the degree of differentiation go in the same direction. In contrast, when μ is sufficiently close to $\frac{1}{2}$, the adverse effect of differentiation on welfare due to travelers' identity turns out to outweigh the positive impact of the extent of differentiation.

These results are described in the following:

Proposition 4. *Assume $g \geq g_N$, so that full coverage emerges both with the traditional and the automated monopolist. Let W_N and W_A denote social welfare under the traditional and automated monopolist, respectively. Then, in equilibrium,*

- *when $\mu \in \left[0, \frac{1}{2}\right]$, social welfare is identical in the two regimes, $W_N = W_A$, and does not vary with μ;*
- *when $\mu \in \left(\frac{1}{2}, 1\right]$, social welfare does not vary with μ under the traditional monopolist, while it is strictly decreasing in μ under the automated monopolist, so that $W_A < W_N$ and $\frac{d(W_N - W_A)}{d\mu} > 0$.*

These results are illustrated in Fig. 1, which provides the result of a welfare comparison between the two scenarios for different combinations of μ on the horizontal axis, and g on the vertical axis. The parametric combinations of interest for Proposition 4 span Regions 1

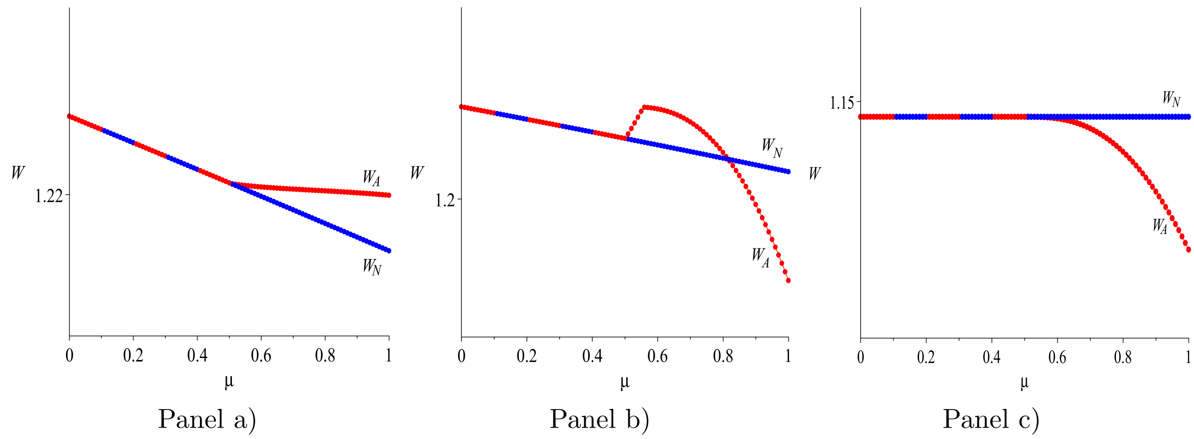

Fig. 2. Equilibrium welfare as a function of μ.

Equilibrium welfare in the case of the traditional monopolist (W_N, blue line) and automated monopolist (W_A, red line) as a function of μ, when $B(\theta) = b_0 + b\sqrt{\frac{1}{4} + \theta}$, $b_0 = 0.4$, $b = 1$ and for different values of g. In panel (a), $g = 0.08$; in panel (b), $g = 0.18$; and in panel (c), $g = 0.40$. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

and 2. Similarly, panel (c) of Fig. 2 reports the value of W_N and W_A as a function of μ, for given values of the other parameters of the model.[26]

Partial coverage. We now move to consider sufficiently low levels of the congestion disutility parameter such that partial coverage occurs in equilibrium in at least one of the scenarios. This arises when $g < g_N$. Indeed, we showed that the minimum level of g such that the traditional monopolist fully covers the market, g_N, weakly exceeds the corresponding level for the automated monopolist, g_A, so that $g_N \geq g_A$. The inequality is strict when the share of corporate travelers is sufficiently high so that the automated monopolist differentiates congestion across lanes in equilibrium.

When μ is sufficiently small, $\mu \leq \mu_A$, both the traditional and the automated monopolist do not differentiate congestion across lanes. Despite the different motivations behind the absence of discrimination, their incentives as to the extent of rationing are fully aligned. Hence, welfare under the two scenarios is identical. Region 3 in Fig. 1 illustrates the point. The comparisons between welfare under the traditional and the automated monopolist turns richer when μ is larger than μ_A so that the automated monopolist finds it optimal to differentiate congestion across lanes. The two monopolists' strategies depend on the values of g and μ, When g is sufficiently small, $g < g_A$, both the traditional and the automated monopolist cover the market only partially. When g is larger, $g_A \leq g < g_N$, the automated monopolist fully covers the market, while the traditional monopolist covers the market only partially. Regions 4 and 5 in Fig. 1 illustrate these two cases, respectively.

Our results are illustrated in the following:

Proposition 5. *Assume $g < g_N$, so that partial coverage emerges under either the traditional or the automated monopolist (or both). Let W_N and W_A denote social welfare under the traditional and automated monopolist, respectively. Then, in equilibrium*

- *when $\mu \in [0, \mu_A]$, social welfare is identical in the two cases, $W_N = W_A$;*
- *when $\mu \in (\mu_A, 1]$, the welfare comparison between the two cases depends on how the traditional and automated monopolists' choices in terms of rationing and differentiation impact welfare. Welfare is always higher under the traditional monopolist when μ is sufficiently large, and g is sufficiently close to g_N.*

The Proposition illustrates that the introduction of AVs does not change social welfare when the mass of corporate travelers is such that no differentiation across lanes arises for both the traditional and the automated monopolist. Instead, when the mass of corporate travelers is sufficiently large (i.e., $\mu > \mu_A$, as in regions 4 and 5 in Fig. 1), social welfare is different across the two scenarios. The lack of analytical solutions for monopolist choices hampers the welfare comparison between the two scenarios. We, therefore, illustrate our results through numerical simulations, using the following explicit functional form for the gross benefit function: $B(\theta) = b_0 + \sqrt{\frac{1}{4} + \theta}$. Full details of the methodology are provided in Appendix B.

One preliminary note is in order. The welfare effects of rationing under monopoly depend on the comparison with the planner's rationing strategy. When the planner does not ration, while the monopolist does, rationing is welfare-reducing, and the reduction in welfare is clearly larger, the larger the extent of rationing. On the contrary, if welfare maximization commands rationing, full coverage under monopolist is not welfare maximizing. We perform our numerical analysis for combinations of parameters such that a social planner would never ration.[27] Hence, rationing is welfare-reducing for any combination of parameters used in our simulations. We will return to this aspect at the end of this section and argue that this does not impose any severe limitations on our results.

We first consider the case in which the automated monopolist does not ration, while its traditional counterpart does. This case is illustrated in Region 5 of Fig. 1, and it is also displayed in panel (b) of Fig. 2. Here, the automated monopolist differentiates congestion across lanes ($\mu > \mu_A$), while the traditional monopolist does not. A trade-off emerges between welfare-reducing differentiation by the automated monopolist and welfare-reducing rationing by the traditional monopolist. As discussed above, the welfare cost of differentiation increases with the share of corporate travelers μ. Also, the welfare cost of rationing depends on the extent of rationing, which in turn rests on the congestion disutility parameter, and in particular, for the traditional monopolist, on the distance between g and g_N. In line with these intuitions, our numerical simulations show that in the north-east portion of Region 5, where the welfare cost of the automated monopolist's differentiation is relatively high (as μ is relatively high), and the welfare cost of the traditional monopolist's rationing is relatively low (as g is relatively

[26] Notice that the amount of coverage in the first best (and in particular whether the planner fully covers the market or not) is immaterial to Proposition 4.

[27] Given the functional form of choice for $B(.)$, Assumption 1 becomes $g < \frac{1}{\sqrt{5}} \approx 0.4472$ and the no rationing condition under the first best, $g \leq g_{FB}$, becomes $g \leq \frac{72}{5} - \frac{18}{5}\sqrt{7} \approx 4.8753$, which is implied by Assumption 1.

close to g_N), welfare is higher with traditional vehicles: $W_A < W_N$. On the contrary, in the rest of the region, the rationing distortion is more harmful than the congestion differentiation distortion, so $W_A > W_N$.

Consider now the case when the automated monopolist differentiates congestion across lanes ($\mu > \mu_A$), and both monopolists ration some travelers ($g < g_A$). In Fig. 1, this case is illustrated in Region 4. This is also displayed in panel (a) of Fig. 2. A similar tradeoff as in Region 5 emerges between the welfare cost of differentiation under AVs and the welfare cost of rationing, which is higher under traditional vehicles. Our numerical simulations reveal that the traditional monopolist rations more travelers than the traditional monopolist and that this excess rationing is more harmful from the welfare standpoint, so that $W_A > W_N$.

We now discuss parameter values for which the planner rations some individuals. If both monopolists fully cover the market, the planner's rationing does not affect our above illustrated welfare comparison. If, instead, at least one type of monopolist does not fully cover the market (i.e., $g_{FB} < g_A$ for at least some values of μ), the welfare comparison is affected, as the welfare cost of rationing is reduced (and may even flip to a welfare benefit). In terms of welfare, this generally increases the appeal of the more intense rationing by the traditional monopolist. However, even in the most extreme cases in which the planner rations some individuals even for very low values of the congestion disutility parameters (so g_{FB} is close to 0), social welfare under the automated monopolist is likely to be higher than under the traditional one. This is because the traditional monopolist's excess rationing with respect to the automated monopolist may outplay the welfare benefits obtained by the automated monopolist's choice to differentiate speed for low enough g and for μ sufficiently close to $\frac{1}{2}$. Hence, the qualitative features of Fig. 1 remain unchanged even in cases where the first best requires rationing for a large set of congestion disutility parameters.

7. First-best restoring taxes

Up to now, we have excluded government intervention. This laissez-faire approach is often observed in practice. While economists advocate road pricing and congestion taxes as tools to improve upon market outcomes, these are rarely implemented in practice (notable exceptions include London, Stockholm, and Singapore). This is likely for political economy reasons (Oberholzer-Gee and Weck-Hannemann, 2002), such as the fear of penalizing low-income travelers that are inflexible in their arrival time (Hall, 2018, 2021).

This section characterizes the tax/subsidy schemes to restore the first best. We start by noting that a tax scheme able to achieve the first best does not exist with traditional vehicles. Indeed, restoring the first best requires having both an efficient level of rationing and an efficient level of differentiation across the lanes (with the corresponding optimal allocation of travelers in the lanes). When vehicles are traditional, the combination of technological and economic constraints does not allow effective monitoring of drivers' route choices, thereby hindering differentiation. AVs technology overcomes this obstacle.

To derive the socially optimal tax scheme with AVs, we move to a three-stage game in which the tax authority sets the taxes in the first stage, while the two subsequent stages are identical to those analyzed in the previous sections. We restrict to the unit (per travel) taxes, differentiated by lane but not by the travelers' identity. We denote by t and T the unit tax for the slow and fast lanes, respectively. We are interested in showing that the socially optimal tax scheme implies that the tax to be levied on corporate travelers is different than that to be applied on atomistic travelers. To this aim, we first briefly discuss the optimal tax scheme in case the market is populated by atomistic individuals only, so $\mu = 0$. We then turn to characterize the optimal road tax/subsidy scheme for corporate travelers. For simplicity, we look at the other polar case of all individuals being corporate, so $\mu = 1$.

In our framework, optimal road taxes have two goals: (*i*) they need to induce to travel those and only those individuals whose private benefit from a trip is larger than the social cost they impose on fellow travelers; (*ii*) they need to restore the efficient allocation of travelers across lanes. These conditions are to be met both for atomistic and corporate travelers. As extensively discussed in the previous sections, both atomistic and corporate travelers are allocated suboptimally, but for different reasons. The nature of the tax system restoring the social optimum must then be different for the two types of travelers. Atomistic individuals do not internalize the effect of their choice on other travelers. The tax imposed on them must then correct for this distortion. We show that, as intuitive, it has a Pigouvian nature. On the other hand, corporate travelers are misallocated due to the monopoly profit-maximizing incentives: when $\mu = 1$, the monopolist overdifferentiates to extract a larger surplus from individuals with a high θ. The resulting distortion shows that a monopolistic fleet with AVs factors in its decisions the effects of its choice on other individuals but does so in a way that deviates from welfare maximization.

We first derive the unit taxes on atomistic travelers when all individuals are atomistic and use AVs not managed by a fleet. Let t_A^a and T_A^a respectively denote the unit taxes levied on atomistic individuals traveling in the slow and fast lane that restore social optimum. These taxes are described in the following Proposition.

Proposition 6. *Assume all individuals are atomistic. The pairs of taxes that replicate the social optimum are*

- *when $g \le g_{FB}$ (i.e., when no rationing occurs in the first best),*

$$\begin{aligned} t_A^a &\le B(0), \\ T_A^a &= t_A^a + g\frac{5-\sqrt{7}}{18}; \end{aligned} \tag{36}$$

- *when $g > g_{FB}$ (i.e., when rationing occurs in the first best),*

$$\begin{aligned} t_A^a &= g\, s_{FB}\left(1 - f_{FB} - \frac{s_{FB}}{2}\right), \\ T_A^a &= t_A^a + g\left(1 - f_{FB}\right)\left(s_{FB} - f_{FB}\right). \end{aligned} \tag{37}$$

Taxes modify the after-tax net utility that atomistic individuals enjoy from traveling, thereby affecting their choices as to whether to travel and in which lane. They are incentive compatible congestion charges, reflecting the external cost imposed by the marginal travelers on fellow travelers. As a result, they align the incentives of the marginal travelers to social optimality.

In particular, when $g \le g_{FB}$, misallocation of travelers across lanes is the only distortion to be solved since there is no rationing both with atomistic travelers and in the first best. Hence, the optimal pair of taxes – as in (36) – should not restrict market coverage. A multiplicity of low enough t_A^a, including $t_A^a = 0$, delivers this. On the other hand, optimal differentiation across lanes is obtained through an appropriate difference $T_A^a - t_A^a$, which ensures that the identity of the marginal traveler is the same as in the first best. When instead g is larger than g_{FB}, a social planner excludes atomistic individuals with a low θ value. Hence, both the total number of travelers and their allocation across lanes have to be corrected. The tax in the slow lane, t_A^a in (37), is then uniquely determined and ensures that the level of market coverage replicates the first best. On the other hand, the fast lane tax, T_A^a, induces an optimal degree of differentiation across lanes.

Next, we characterize the tax/subsidy scheme on the automated monopolist ensuring that, when $\mu = 1$, the equilibrium allocation of travelers is as in the social optimum. Let t_A^c and T_A^c denote the unit taxes that restore the social optimum: they are levied on the monopolist managing the AVs fleet for each travel in the slow and fast lane, respectively. We derive a scheme that is additively composed of two terms, a tax component and a subsidy component. Both components are possibly differentiated by lanes and do not depend on the type θ.

The characterization of this scheme is provided in the following Proposition.

Proposition 7. *Assume all individuals are corporate. The tax/subsidy scheme on the AVs monopolist that replicates the social optimum is as follows:*

$$t_A^c = g\,s - z_A, \tag{38}$$

$$T_A^c = g\,f - Z_A, \tag{39}$$

where

$$\begin{array}{ll} z_A = Z_A = 0 & \text{if } g < g_1, \\ z_A = Z_A = B'(0) - B(0) + g\frac{4+\sqrt{7}}{18} & \text{if } g_1 \le g \le g_2, \\ z_A = \tilde{z}_A \text{ and } Z_A = \tilde{Z}_A & \text{if } g_2 < g < g_{max}, \end{array} \tag{40}$$

and

$$g_1 \equiv \max\left\{0, \frac{18\left[B(0) - B'(0)\right]}{4+\sqrt{7}}\right\}, \tag{41}$$

$$g_2 \equiv \min\left\{g_{FB}, g_{max}\right\}, \tag{42}$$

$$\tilde{z}_A \equiv \left(s_{FB} + f_{FB}\right) B'\left(1 - s_{FB} - f_{FB}\right) + 2g s_{FB}\left(1 - \frac{3}{4}s_{FB} - f_{FB}\right), \tag{43}$$

$$\tilde{Z}_A \equiv \left(s_{FB} + f_{FB}\right) B'\left(1 - s_{FB} - f_{FB}\right) - \frac{3}{2}g\left(\frac{2}{3}s_{FB}^2 + f_{FB}^2 - \frac{4}{3}f_{FB}\right). \tag{44}$$

The Proposition illustrates that social optimality is restored by imposing on the automated monopolist the pair t_A^c and T_A^c consisting of a tax component, $g\,s^c$ and $g\,f^c$, based on the mass of travelers in that lane and increasing in the common component of the congestion disutility, and a subsidy component, z_A and Z_A, exogenously determined based on the socially optimal number of travelers.

The tax component has two effects. First, it reverses the effect of the congestion parameter on the monopolist's incentive to cover the market. Since the tax component is increasing in g, the monopolist is now induced to ration travelers when g is sufficiently large, as in the case of the social planner. Second, it causes the monopolist to shift passengers from the slow to the fast lane, correcting the monopolist's excess differentiation across lanes. This is because the tax component is higher in the slow lane, as it is based on the mass of travelers in each lane. The logic of this tax is similar, for instance, to that of the tax on quality in Cremer and Thisse (1994) and Lambertini and Mosca (1999), in the context of a vertically differentiated oligopoly, as long as in our model congestion is interpreted as a quality level.

The subsidy component is instead designed to induce the AVs monopolist to choose the first best level of rationing. In particular, it takes different values depending on the relative level of the congestion parameter g. When g is sufficiently small ($g < g_1$), the monopolist always prefers to cover the entire market, as the social planner would do. In this case, the subsidy component is not needed. Under a reasonable goal of government expenditure minimization, it is then optimal to set it equal to zero. When g takes on an intermediate level ($g_1 \le g \le g_2$), the monopolist prefers to exclude some travelers, while the social optimum would instead imply a fully covered market. The subsidy component then provides the monopolist with the incentive to avoid rationing. When g is large ($g_2 < g < g_{max}$), both the monopolist and the social planner choose to ration. In this case, the subsidy component induces the monopolist to provide the same level of rationing as the social planner would do.[28]

Overall, the two components of the tax/subsidy scheme serve very different purposes and have remarkably different features. The levels of the tax components only depend on the monopolist's choices, being contingent on s^c and f^c. In this respect, this tax is simple to set, as it places a relatively low informational requirement (the value of g only) on the tax authority. On the other hand, the subsidy component does not depend on the monopolist's choices. However, it requires perfect knowledge of the travelers' benefit function and of the solution to the first-best problem, putting a much heavier informational load on the tax setter.

While individuals do not directly pay the tax (nor directly receive the subsidy), they are affected by the introduction of the scheme through the change in the monopolist pricing policy. Indeed, after the introduction of a tax, the monopolist has the incentive to charge a pair of fares such that individuals, given their IR and IC in (20) and (21), find it optimal to choose an allocation across lanes that reproduces the social optimum. The effects of the scheme on each individual depend on the change in price and in travelers' allocation lanes before and after its implementation.

We start by illustrating this in case full coverage occurs both at the social optimum and under a monopoly without taxes. The introduction of the scheme turns out to increase the net utility of all types except for type 0 (who is indifferent). In particular, it does not change the price of the slow lane, while it reduces the price of the fast lane. At the same time, the slow lane becomes relatively less congested, while the fast lane becomes more congested. As a result, the low θ types, who travel in the slow lane both before and after the introduction of the scheme, enjoy a net benefit in the form of lower congestion while paying the same price.[29] Intermediate θ types, who move to the fast lane after the introduction of the scheme, pay a higher price and enjoy lower congestion. It turns out that the beneficial effect of lower congestion outplays the adverse effect of a higher price for all θ types in this interval. High θ types, who travel in the fast lane both before and after, enjoy a lower fare but suffer from more congestion. Even in this case, the aggregate effect of the scheme is positive for all θ types in the interval. A lower fare's benefits outweigh the adverse effects of higher congestion.

The welfare comparisons are less clear-cut when we abandon the case of full coverage. In these cases, the prices of both lanes, as well as their extent of congestion, change before and after the introduction of the tax. It is possible to find instances where both variations benefit (or harm) some types of travelers, indicating a clear effect on their net utility. For instance, consider the case in which some low θ travelers are not traveling before the introduction of the scheme while they are traveling after. Then, their utility after the scheme increases, except for the marginal traveler who is paying a fare equal to her willingness to pay (and is hence indifferent). On the contrary, travelers who are no longer traveling as a result of the introduction of the scheme see a decline in their net utility.

When we look at the effect of the scheme on the automated monopolist, we find that, in equilibrium, the subsidy component may exceed the tax component so that the monopolist receives a net subsidy from the tax authority. However, a net subsidy does not necessarily imply that the monopolist overall benefits from the application of the scheme. Indeed, the monopolist is affected by the scheme not only, directly, through the additional revenue/cost implied by the subsidy and tax components, but also, indirectly, through the reduction in profit resulting from the change in prices and the mass of travelers in each lane.

To overcome the distributional concerns and the related political difficulties associated with a positive transfer to the AVs monopolist (emphasized, for instance, by Pels and Verhoef (2004) in the context of airline companies), a potential solution consists in coupling it with an upfront fixed license that preserves budget neutrality, while ensuring non-negative profit to the monopolist.

Overall, our findings suggest that, when the prevalence of AVs will make the differentiation of the extent of congestion across routes a reality, the optimal tax should exhibit two features. First, the tax on atomistic travelers should differ from that on corporate travelers. Second, for each type of traveler, the amount of the tax will depend on the congestion level (speed of service) of the selected route.

[28] Notice that, depending on the relative values of $B(0)$, $B'(0)$ and g, some of the intervals in (40) may be empty.

[29] Type 0, who is not bothered by congestion, is instead indifferent.

8. Discussion and extensions

Our analysis could be possibly extended in several directions, including non-linear congestion costs, competition between fleets, changes in the market fundamentals following the development of AVs and travelers' choice between owning a car and using a fleet. This section discusses why we decided not to incorporate these extensions and the robustness of our results to them.

Non-linear congestion costs. We obtain our results under the assumption that an individual's disutility from congestion increases linearly with the mass of travelers in the same lane. Our results are however robust to a different specification of the relationship between the disutility of congestion and the mass of travelers: in Appendix C, we analyze a model in which the congestion disutility is quadratic in the mass of travelers and show that our results are qualitatively unaffected.

Competition between fleets. Our analysis assumes a monopolistic fleet. It is well possible that the AVs fleets market, despite network effects, will evolve into an oligopolistic industry (see, for instance, Zhang et al. 2019). However, we have not extended the analysis to competition. The results under competition are likely sensitive to the model of competition we consider. We believe the most natural way of introducing competition would involve considering an oligopoly with two lanes, in which firms compete in quantity, and each can dispatch its vehicles in each of the two lanes, and the congestion in each lane then depends on the sum of the vehicles dispatched by all companies in that lane. In this setting, the market outcome would be affected by the externality a vehicle dispatched by a firm imposes on the rivals' vehicles. However, the results would qualitatively align with those in our monopoly model. Alternative oligopolistic settings include price competition, dynamic competition, or environments in which each oligopolist dispatches its vehicles in a dedicated lane. This latter model would likely share some properties with those obtained in the literature on rivalry with congestible facilities (see, among others, Van Dender, 2005; Basso and Zhang, 2007; van den Berg and Verhoef, 2012). An extension analyzing the interaction between profit-maximizing fleets and a public transit company is provided in Boffa et al. (2023).

Change in the market fundamentals with the transition to AVs. In our setting, the technological transition to AVs does not affect the market fundamentals. In fact, we might think that AVs reduce congestion disutility g since travelers can spend their time in their vehicles more productively. Our analysis can accommodate this, by considering a lower level of g with AVs compared to that with non-AVs in each scenario. However, the welfare effect of the technological transition would have to be modified accordingly. AVs would entail an additional direct beneficial welfare effect due to the increase in individual utility resulting from the lower g. Furthermore, the lower g might induce the automated monopolist to ration a larger share of individuals. The welfare effect of this potentially more intense rationing depends on whether rationing is welfare-improving or welfare-reducing.[30]

In principle, the technological transition to AVs might produce additional changes. More effective fleet management may affect both the demand and the supply of fleet services. It may reduce waiting times and search costs for corporate travelers while at the same time inducing a more efficient use of the available capacity when this is given. Overall, this may lead to an expansion in the use of the service. In our model, this could be reflected in an increase in the proportion of corporate travelers μ. This may allow the automated monopolist to increase the extent of differentiation across lanes. The welfare effect of this strategy is unambiguously negative if all individuals travel. At the same time, it is subtler, depending on the interplay between the extent of rationing and differentiation, if the market is not fully covered.

Importantly, the above discussed effects of the technological transition on both congestion disutility and the use of fleets service are at this stage still uncertain: this is why we do not incorporate them in our model. For example, while AVs may reduce congestion disutility by increasing the value of time spent on a vehicle, they might also affect overall travel time through several channels pointing in different directions, with an unclear combined effect. On the one hand, AVs may be better driven than traditional vehicles, thereby contributing to a reduction in travel time (Cummins et al., 2021; Lopez et al., 2018; Mirzaeian et al., 2021; Overtoom et al., 2020; Stern et al., 2018). On the other hand, they might also, in particular in the initial stage of their rollout, be driven very cautiously to minimize accidents and increase the perceived safety of the technology and its acceptability by public opinion.[31] Also, AVs might induce a different pattern of vehicle usage, thereby possibly increasing congestion for a constant number of vehicles. In addition, the adoption of the AVs technology may affect the supply side as well, particularly the size of the fleet and, therefore, in combination with changes in demand, the potential capacity constraints faced by the operator. The direction and the extent of the changes are currently hard to forecast, because of the uncertain magnitude of the investment required for the transition, and the resulting effect on the structure of the operating cost.

Dynamics. The model ignores dynamics. Adding dynamics to our model would entail several possible consequences. First, firms learn over time about individual characteristics, which would provide fleets with more information on individual types. This potentially increases the tailorization of its pricing policy, with the consequences discussed in footnote 3.

If we add a more complex network and demand structure to the model, a tension between static and dynamic optimization may emerge. If the fleet is capacity constrained, the profit maximizing allocation of vehicles reflects spatial consideration as well. Maximizing profits over time also requires solving a fleet allocation problem over space. For example, a vehicle could be used to carry a traveler with low willingness to pay heading to an area with high demand at the arrival time, rather than a traveler with higher willingness to pay but going to an area with low demand at the arrival time (Brancaccio et al., 2023; Buchholz, 2022; Buchholz et al., 2020; Castillo et al., 2023; Fréchette et al., 2019; Rosaia, 2020). Additionally, in this context, even the congestion level of the lane to which a traveler is allocated may be based on the need to satisfy demand at the arrival point timely.

This misalignment between static and dynamic profit maximizing solutions would occur for both the traditional and the automated monopolist. However, AVs technology expands the opportunities to address dynamic issues efficiently. Indeed, in fleets using traditional vehicles, a dispatching strategy aimed at improving the dynamic outcome requires modifying drivers' incentives through the price system. Due to frictions (associated with informational constraints and, potentially, market power), this generally cannot be done efficiently. An example of a practice aimed at modifying drivers' incentives is the so-called surge pricing, applied, for instance, by Uber. There, a surcharge is applied to the final price and to the remuneration of drivers to provide drivers with incentives to converge towards areas where demand is expected to be larger than usual, e.g., after an event that attracts a large crowd (Guda and Subramanian, 2019; and Castillo, forthcoming). The adoption of the AVs technology, while not solving the misalignment of static and dynamic incentives, provides the fleet operator with a powerful and efficient instrument to consider all the dynamic incentives in its strategy. Since our model is not able to accommodate dynamic

[30] The reduction in g, by increasing individuals' willingness to pay (for all θ, except for $\theta = 0$), would also (at least weakly) increase the automated monopolist profit.

[31] A smooth transition, without accidents, towards AVs might help overcome the documented tendency for humans to show a larger aversion to delegating tasks to machines rather than to other humans (Gogoll and Uhl, 2018).

considerations, we are not able to speculate on how such considerations could affect its predictions, in particular, on the extent of congestion. However, it appears reasonable to imagine that its superior ability to deal with dynamic incentives represents an additional advantage of the AV technology.

9. Conclusions

Our paper analyzes the effects of the technological transition from traditional to autonomous vehicles in an environment in which some individuals use the services of a (monopolistic) fleet while others travel independently. The AVs technology enables the fleet operator to monitor the itineraries of the vehicles she dispatches. Hence, a fleet composed of AVs, different from one of traditional vehicles, can assign travelers to routes with a different level of congestion, resulting in different speeds and prices.

Differentiation across routes is a fundamental tool to address congestion, since heterogeneity in congestion disutility, which reflects primarily heterogeneity in individuals' value of time, is known to be substantial. Accordingly, in our model, we show that welfare maximization requires differentiating congestion across routes, assigning individuals with a higher value of travel time to a less congested (faster) route, and those with a lower value of travel time to a more congested (slower) route.

Besides differentiation of congestion across routes, the other, perhaps more intuitive, tool to address congestion, is rationing, consisting in screening some individuals out of the market. In our model, welfare maximization commands screening individuals with a low value from traveling out of the market when the negative externality they impose on fellow travelers is sufficiently high.

We compare the market outcome and the associated welfare level in the two scenarios of automated and traditional fleets. We first find that the technological transition to AVs makes a difference, in terms of outcome and welfare, only when the share of fleet users is sufficiently large (and, correspondingly, when the share of independent travelers is sufficiently low).

In the cases of a sufficiently large fleet, we show that, despite the potential welfare benefits from differentiation, a fleet using AVs does not necessarily improve welfare. When there is no rationing, and all the individuals use the fleet, the automated monopolist overdifferentiates the congestion level across routes. This turns out to be welfare inferior to the uniform level of congestion operated by the traditional monopolist. More generally, as long as there is no rationing and the fleet is sufficiently large so that the market outcome is different in the two scenarios, the outcome under the traditional monopolist is welfare superior.

When the fleets (traditional and/or autonomous) ration some individuals, the welfare effect of the transition is subtler and depends on the interplay between the extent of rationing and the extent of differentiation. In general, we find that rationing is more intense under the traditional monopolist, which tends to make the AV fleet more appealing, from the welfare standpoint, when welfare maximization does not prescribe rationing.

We finally study how to restore first best with a tax system. We show that, with AVs, the tax differs depending on the traveler's identity, whether she is independent or using the fleet service. Independent travelers are charged lane-specific congestion charges, while the fleet is charged a more complex scheme involving a tax and a subsidy component.

Our analysis shows that the ability to differentiate the extent of congestion by the AVs fleets will dramatically change the tools to manage congestion and the incentives by fleet operators. A full understanding of the consequences of the rollout of the AVs fleets is then of utmost importance to devise efficient urban transport policies.

CRediT authorship contribution statement

Federico Boffa: Conceptualization, Methodology, Software, Formal analysis, Writing – original draft, Writing – review & editing, Funding acquisition. **Alessandro Fedele:** Conceptualization, Methodology, Software, Formal analysis, Writing – original draft, Writing – review & editing, Funding acquisition. **Alberto Iozzi:** Conceptualization, Methodology, Software, Formal analysis, Writing – original draft, Review & editing, Funding acquisition.

Appendix A. Derivations and proofs

This Appendix contains the proofs of all Propositions, together with the derivations of additional results contained in the paper.

Derivation of (7). We show here that the maximization of (6) implies the maximization of (7). First, we show that, if the planner optimally excludes some travelers, it would exclude travelers with the lowest θ's. By contradiction, suppose that it dispatches a θ'-type, and it does not dispatch a θ''-type, with $\theta' < \theta''$. Then, a switch between the θ''-type and the θ'-type would leave congestion unaffected while increasing the aggregate net benefit from traveling because $B'(\theta) > 0$: as a result, the social welfare would increase.

Second, we show that it is never optimal to assign to the same lane travelers in non-contiguous partitions of the unit line. Let types θ' and θ'' be traveling in a lane with a mass of travelers equal to $l' + l''$. Assume, by contradiction, that $\theta' \in [\bar{\theta}', \bar{\theta}' + l']$ and $\theta'' \in [\bar{\theta}'', \bar{\theta}'' + l'']$, with $\bar{\theta}' + l' < \bar{\theta}''$. Let also type θ''' be traveling in a lane with a mass of travelers equal to l''' and that $\theta''' \in [\bar{\theta}' + l', \bar{\theta}'']$. Next, assume that $l' + l'' < l'''$ so that the lane where types θ' and θ'' travel is less congested than the one where type θ''' travels. A switch between types θ' and θ''' would leave congestion in both lanes unaltered while increasing the aggregate net benefit from traveling because $\theta' < \theta'''$ and $B'(\theta) > 0$. Similarly, assume that $l''' < l' + l''$ so that the lane where types θ' and θ'' travel is more congested than the one where type θ''' travels. A switch between types θ'' and θ''' would leave congestion in both lanes unaltered while increasing the aggregate net benefit from traveling because $\theta'' > \theta'''$ and $B'(\theta) > 0$.

As a result, the welfare function can be rewritten as follows

$$W = \int_{1-l_1-l_2}^{1-l_2} \left(B(\theta) - \theta g l_1\right) d\theta + \int_{1-l_2}^{1} \left(B(\theta) - \theta g l_2\right) d\theta,$$

where l_1 is the mass of travelers in lane 1 and l_2 is the mass of travelers in lane 2. It is easy to prove that $l_1 > l_2$, so that travelers with lower θ's are placed in lane 1, the more congested lane. Consider two travelers, θ' and θ'', with $\theta' < \theta''$ and first suppose θ' uses lane 1, while θ'' uses lane 2. In equilibrium, the aggregate net benefit for the two travelers is $w \equiv B\left(\theta'\right) - \theta' g l_1 + B\left(\theta''\right) - \theta'' g l_2$. Suppose now that θ'' uses lane 1 and θ' uses lane 2. The aggregate net benefit for the two travelers is $w' \equiv B\left(\theta''\right) - \theta'' g l_1 + B\left(\theta'\right) - \theta' g l_2$. Then, $w - w' = B\left(\theta'\right) - \theta' g l_1 + B\left(\theta''\right) - \theta'' g l_2 - (B\left(\theta''\right) - \theta'' g l_1 + B\left(\theta'\right) - \theta' g l_2) = g(\theta'' - \theta')(l_1 + l_2) > 0$, which shows that social welfare is higher in the first case. The planner problem may then be written as in (7). □

Proof of Proposition 1. The Lagrangian of problem (8) is

$$\mathcal{L}_{FB} \equiv \int_{1-s-f}^{1-f} (B(\theta) - \theta g s)\, d\theta + \int_{1-f}^{1} (B(\theta) - \theta g f)\, d\theta - \lambda(s + f - 1).$$

At the solutions to this problem, denoted by s_{FB}, f_{FB} and λ_{FB}, Kuhn–Tucker conditions require

$$\frac{\partial \mathcal{L}_{FB}}{\partial s} = B\left(1 - s_{FB} - f_{FB}\right) - 2 g s_{FB}\left(1 - f_{FB} - \frac{3}{4} s_{FB}\right) - \lambda_{FB} = 0, \tag{A-1}$$

$$\frac{\partial \mathcal{L}_{FB}}{\partial f} = B\left(1 - s_{FB} - f_{FB}\right) + g\left(s_{FB}^2 + \frac{3}{2} f_{FB}^2 - 2 f_{FB}\right) - \lambda_{FB} = 0, \tag{A-2}$$

$$\frac{\partial \mathcal{L}_{FB}}{\partial \lambda} = s_{FB} + f_{FB} - 1 \le 0, \qquad \lambda_{FB} \ge 0 \text{ and } \qquad \frac{\partial \mathcal{L}_{FB}}{\partial \lambda}\lambda_{FB} = 0. \tag{A-3}$$

Assume that $\lambda_{FB} = 0$ and $s_{FB} + f_{FB} - 1 < 0$. Substitute $\lambda_{FB} = 0$ in (A-1) and (A-2), equate them and solve w.r.to f_{FB} to obtain (9).

Assume that $\lambda_{FB} \ge 0$ and $s_{FB} + f_{FB} = 1$. Substitute $s_{FB} = 1 - f_{FB}$ in (A-1) and (A-2), equate them and solve for f_{FB} to obtain $\bar{f}_{FB}$ as in (10). Use $s_{FB} = 1 - f_{FB}$ to get $\bar{s}_{FB}$ as in (10). Plug $\bar{f}_{FB}$ and $\bar{s}_{FB}$ thus obtained into (A-1) or (A-2) and solve for λ_{FB} to obtain $\lambda_{FB} = B(0) - g\frac{4+\sqrt{7}}{36}$. Solve $\lambda_{FB} \ge 0$ for g to get (11). ■

Derivation of the comparative static results in the first best. We derive here the comparative statics results mentioned in Section 3, i.e., $\frac{ds_{FB}}{dg} < 0$ and $\frac{df_{FB}}{dg} < 0$.

Denote the FOCs of the maximization problem for the planner in (8), given in (A-1) and (A-2), as $h_s(s_{FB}, f_{FB}, g) = 0$ and $h_f(s_{FB}, f_{FB}, g) = 0$, after substituting $\lambda_{FB} = 0$. Implicit differentiation of the FOCs w.r.to g gives

$$\frac{ds_{FB}}{dg} = \frac{\frac{\partial h_f}{\partial g}\frac{\partial h_s}{\partial f_{FB}} - \frac{\partial h_f}{\partial f_{FB}}\frac{\partial h_s}{\partial g}}{\frac{\partial h_f}{\partial f_{FB}}\frac{\partial h_s}{\partial s_{FB}} - \frac{\partial h_f}{\partial s_{FB}}\frac{\partial h_s}{\partial f_{FB}}};$$

$$\frac{df_{FB}}{dg} = \frac{\frac{\partial h_f}{\partial g}\frac{\partial h_s}{\partial s_{FB}} - \frac{\partial h_f}{\partial s_{FB}}\frac{\partial h_s}{\partial g}}{\frac{\partial h_f}{\partial f_{FB}}\frac{\partial h_s}{\partial s_{FB}} - \frac{\partial h_f}{\partial s_{FB}}\frac{\partial h_s}{\partial f_{FB}}}.$$

In problem (8), SOCs require $\frac{\partial h_f}{\partial f_{FB}}\frac{\partial h_s}{\partial s_{FB}} - \frac{\partial h_f}{\partial s_{FB}}\frac{\partial h_s}{\partial f_{FB}} > 0$. As a result,

$$\text{sgn}\left(\frac{ds_{FB}}{dg}\right) = \text{sgn}\left(\frac{\partial h_f}{\partial g}\frac{\partial h_s}{\partial f_{FB}} - \frac{\partial h_f}{\partial f_{FB}}\frac{\partial h_s}{\partial g}\right)$$

and

$$\text{sgn}\left(\frac{df_{FB}}{dg}\right) = \text{sgn}\left(\frac{\partial h_f}{\partial g}\frac{\partial h_s}{\partial s_{FB}} - \frac{\partial h_f}{\partial s_{FB}}\frac{\partial h_s}{\partial g}\right).$$

Using (A-1) and (A-2) yields

$$\frac{\partial h_s}{\partial s_{FB}} = -B'(1 - s_{FB} - f_{FB}) - 2g\left(1 - f_{FB} - \frac{3}{2}s_{FB}\right),$$

$$\frac{\partial h_s}{\partial f_{FB}} = -B'(1 - s_{FB} - f_{FB}) + 2gs_{FB},$$

$$\frac{\partial h_f}{\partial s_{FB}} = -B'(1 - s_{FB} - f_{FB}) + 2gs_{FB},$$

$$\frac{\partial h_f}{\partial f_{FB}} = -B'(1 - s_{FB} - f_{FB}) + g\left(3f_{FB} - 2\right),$$

$$\frac{\partial h_s}{\partial g} = -2s_{FB}\left(1 - f_{FB} - \frac{3}{4}s_{FB}\right),$$

$$\frac{\partial h_f}{\partial g} = (s_{FB})^2 + \frac{3}{2}(f_{FB})^2 - 2f_{FB}.$$

When evaluated at the equilibrium relationship (9), the numerator of $\frac{ds_{FB}}{dg}$ is negative for any $s_{FB} \le \bar{s}_{FB}$. Hence, $\frac{ds_{FB}}{dg} < 0$. Also, from (9), $\frac{df_{FB}}{ds_{FB}} > 0$, hence $\frac{df_{FB}}{dg} = \frac{df_{FB}}{ds_{FB}}\frac{ds_{FB}}{dg} < 0$. □

Derivation of (13). For all θ-type atomistic travelers in the slow lane, it must be the case that

$$B(\theta) - \theta g(s^a + s^c) \ge B(\theta) - \theta g(f^a + f^c),$$

which reduces to $s^a + s^c \le f^a + f^c$ or, using $s = s^a + s^c$ and $f = f^a + f^c$, to $s \le f$. For all θ-type corporate travelers in the slow lane, it must be the case that

$$B(\theta) - \theta g(s^a + s^c) - p \ge B(\theta) - \theta g(f^a + f^c) - p,$$

which reduces to $s \le f$ as well. Since s cannot be strictly lower than f, it must be the case that s^a, s^c, f^a and f^c are such that $s = f$. Using (12) and the fact that $s + f = s^a + s^c + f^a + f^c$, we get (13).

Proof of Proposition 2. The Lagrangian of the monopolist problem in (15) is

$$\mathcal{L}_N \equiv \left[B\left(1 - \frac{c}{\mu}\right) - \left(1 - \frac{c}{\mu}\right)g\frac{c + 1 - \mu}{2}\right]c - \lambda(c - \mu).$$

At the solutions, denoted by c_N and λ_N, Kuhn–Tucker conditions require

$$\frac{\partial \mathcal{L}_N}{\partial c} = -B'\left(1 - \frac{c_N}{\mu}\right)\frac{c_N}{\mu} + B\left(1 - \mu - \frac{c_N}{\mu}\right) - \frac{g}{2}\left(1 - \frac{2c_N}{\mu} - \frac{3c_N^2}{\mu} + 4c_N\right) - \lambda_N = 0. \tag{A-4}$$

$$\frac{\partial \mathcal{L}_N}{\partial \lambda} = c_N - \mu \le 0, \qquad \lambda_N \ge 0 \qquad \frac{\partial \mathcal{L}_N}{\partial \lambda}\lambda_N = 0. \tag{A-5}$$

Assume that $\lambda_N \ge 0$ and $c_N = \mu$. Substitute $c_N = \mu$ into (A-4) and solve for λ_N to obtain $\lambda_N = B(0) - B'(0) + \frac{g}{2}$. Solve $\lambda_N \ge 0$ for g to obtain $g \ge g_N$, where g_N is given by (18).

Assume that $\lambda_N = 0$ and $c_N < \mu$; the solution c_N is implicitly given by (A-4) after plugging $\lambda_N = 0$.

Derivation of (19). We aim to prove that $s^a = 0$ when $1 - \mu < s^c - f^c$. Assume, by contradiction, that $s^a > 0$. For all θ-type atomistic travelers in the slow lane, it must be the case that

$$B(\theta) - \theta g(s^a + s^c) \ge B(\theta) - \theta g(f^a + f^c),$$

which reduces to $s^a + s^c \le f^a + f^c$ or, using (12), to $1 - \mu \ge s^c - f^c + 2s^a$. This contradicts $1 - \mu < s^c - f^c$. It follows that $s^a = 0$ and, substituting $s^a = 0$ into (12), $f^a = 1 - \mu$ as in (19). □

Proof of Proposition 3. The Lagrangian of the monopolist's problem in (22) is

$$\mathcal{L}_A \equiv \left[B\left(1 - \frac{s^c + f^c}{\mu}\right) - \left(1 - \frac{s^c + f^c}{\mu}\right)gs^c\right](s^c + f^c) + \left(1 - \frac{f^c}{\mu}\right)g\left[s^c - (f^c + 1 - \mu)\right]f^c - \lambda(s^c + f^c - \mu).$$

At the solutions, denoted by s_A^c, f_A^c and λ_A, Kuhn–Tucker conditions require

$$\frac{\partial \mathcal{L}_A}{\partial s^c} = -B'\left(1 - \frac{s_A^c + f_A^c}{\mu}\right)\frac{s_A^c + f_A^c}{\mu} + B\left(1 - \frac{s_A^c + f_A^c}{\mu}\right) + \frac{gs_A^c\left(3s_A^c + 4f_A^c - 2\mu\right)}{\mu} - \lambda_A = 0 \tag{A-6}$$

$$\frac{\partial \mathcal{L}_A}{\partial f^c} = -B'\left(1 - \frac{s_A^c + f_A^c}{\mu}\right)\frac{s_A^c + f_A^c}{\mu} + B\left(1 - \frac{s_A^c + f_A^c}{\mu}\right) + g\frac{2\left(s_A^c\right)^2 + f_A^c\left(4(1-\mu) + 3f_A^c - 2\right) - (1-\mu)\mu}{\mu} - \lambda_A = 0. \tag{A-7}$$

$$\frac{\partial \mathcal{L}_A}{\partial \lambda} = s_A^c + f_A^c - \mu \le 0, \qquad \lambda_A \ge 0 \qquad \frac{\partial \mathcal{L}_A}{\partial \lambda}\lambda_A = 0. \tag{A-8}$$

Assume that $\lambda_A \ge 0$ and $s_A^c + f_A^c = \mu$. Substitute $s_A^c = \mu - f_A^c$ in (A-6) and (A-7), equate them and solve for f_A^c to obtain $\bar{f}_A^c$ as in (31). Use $\bar{s}_A^c = \mu - \bar{f}_A^c$ to obtain $\bar{s}_A^c$ as in (30). Plug $\bar{s}_A^c$ and $\bar{f}_A^c$ thus obtained into (A-6) and (A-7) to get $\lambda_A = B(0) - B'(0) + g\frac{(4\mu-1)\sqrt{4\mu^2-2\mu+1}+8\mu^2+5\mu-1}{18\mu}$. Solve $\lambda_A \ge 0$ for g to get $g \ge K(\mu)\left[B'(0) - B(0)\right]$.

Assume $\lambda_A = 0$ and $s_A^c + f_A^c < \mu$. Substitute $\lambda_A = 0$ in (A-6) and (A-7), equate them and solve for f_A^c to get (26).

We now derive the threshold $\tilde{\mu}$, that is, we check under which conditions $1 - \mu < (\ge)\, s_A^c - f_A^c$ in equilibrium under partial coverage. Using (26), we find that, for any admissible s_A^c and f_A^c, $1 - \mu > s_A^c - f_A^c$ at $\mu = \frac{1}{2}$ and $1 - \mu < s_A^c - f_A^c$ at $\mu = 1$. By implicitly differentiating s_A^c and f_A^c w.r.to μ it is possible to establish $\frac{\partial\left(s_A^c - f_A^c\right)}{\partial\mu} > 0$. It follows there exists a $\tilde{\mu} \in \left(\frac{1}{2}, 1\right)$ such that $1 - \mu \ge s_A^c - f_A^c$ for any $\mu \in \left[\frac{1}{2}, \tilde{\mu}\right]$ and $1 - \mu < s_{AC}^c - f_A^c$ for any $\mu \in (\tilde{\mu}, 1]$. ■

Proof of Proposition 4. Let $\mathcal{B} \equiv \int_0^1 B(\theta)d\theta$. Denote by $\bar{W}_N$ and $\bar{W}_A$ the equilibrium welfare in full coverage with and without AVs, respectively. When $\mu \le \frac{1}{2}$, both the traditional and the automated monopolist are unable to differentiate across lanes, and welfare is given by Eq. (34). At the equilibrium as in (17), they are equal to

$$\bar{W}_N = \bar{W}_A = \mathcal{B} - \frac{1}{4}g. \tag{A-9}$$

When, instead, $\mu > \frac{1}{2}$, the traditional monopolist does not differentiate across lanes, so that social welfare is as in (A-9). On the contrary, the automated monopolist differentiates across lanes and social welfare is given by Eq. (34). When evaluated at the equilibrium $\bar{s}_A^c$ and $\bar{f}_A^c$ as in (30) and (31), it is equal to

$$\bar{W}_A = \mathcal{B} - \frac{16\mu^3 - 24\mu^2 + 45\mu - 1 + \left(8\mu^2 - 10\mu - 1\right)\sqrt{4\mu^2 - 2\mu + 1}}{108\mu} g. \tag{A-10}$$

The result follows immediately. ∎

Proof of Proposition 5. When $\mu \in [0, \mu_A]$, the result follows immediately from Propositions 2 and 3. Focus now on the case $\mu \in (\mu_A, 1]$. Denote $W_N(g,\mu)$ and $W_A(g,\mu)$ the equilibrium welfare without and with AVs, respectively. From (A-9) and (A-10), the difference $W_N(g_N, 1) - W_A(g_N, 1)$ is positive and finite. Since $W_N(.)$ and $W_A(.)$ are continuous both in g and μ, then there always exists ϵ positive and sufficiently close to 0 such that $W_N(g_N - \epsilon, 1-\epsilon) - W_A(g_N - \epsilon, 1-\epsilon) > 0$. ∎

Derivation of sufficient conditions for the automated monopolist overproviding coverage. We show conditions under which more travelers travel under the automated monopolists than in the first best, as discussed in Section 5. In particular, we provide sufficient conditions under which the automated monopolist fully covers the market, while the social planner does not ($\bar{s}_A + \bar{f}_A > s_{FB} + f_{FB}$).

A sufficient condition for the automated monopolist to fully cover the market when the social planner rations some low-θ travelers is given by $\max\{g_A, g_{FB}\} < g < g_{max}$. The interval $\max\{g_A, g_{FB}\} < g_{max}$ is empty iff

$$B'(0) \ge \frac{2\left(8\mu^2 + 5\mu - 1 + (4\mu - 1)\sqrt{4\mu^2 - 2\mu + 1}\right) + 4\sqrt{7}\mu}{36\mu} B'(1).$$

This inequality holds iff

$$\mu \le \mu^+ \equiv \left(\frac{6}{19}\sqrt{7} + \frac{17}{76}\right)\sqrt{\frac{175\,237}{179\,776} - \frac{10\,013}{44\,944}\sqrt{7}} - \frac{89}{848}\sqrt{7} + \frac{577}{1696} \approx 0.72.$$

As a result, the automated monopolist fully covers the market when the planner does not when $\mu > \mu^+$ and $\max\{g_A, g_{FB}\} < g < g_{\max}$.

Proof of Proposition 6. Assume $g \le g_{FB}$, so that full coverage occurs in the first best. Since the type-0 atomistic traveler gets utility $B(0) - t$ from traveling in the slow lane, Assumptions 1 and 2 imply that full coverage occurs as in the social optimum when $t = t_A$. Given t_A, substitute $\bar{s}_{FB}$ as in (10) into the IC constraint (5) to write

$$B\left(\frac{1}{2} + \frac{\sqrt{7}-2}{6}\right) - \left(\frac{1}{2} + \frac{\sqrt{7}-2}{6}\right) g \left(\frac{1}{2} - \frac{\sqrt{7}-2}{6}\right) - T \ge B\left(\frac{1}{2} + \frac{\sqrt{7}-2}{6}\right) - g\left(\frac{1}{2} + \frac{\sqrt{7}-2}{6}\right)^2 - t_A$$

and solve it w.r. to T when holding as an equality to obtain T_A.

Assume $g \ge g_{FB}$, so that partial coverage occurs in the first best. To obtain t_A, consider that the marginal net effect on social welfare of a θ-type traveler deciding to travel in the slow lane (as opposed to not traveling) is given by the LHS of (A-1), while his private benefit is given by $B(\theta) - \theta g s$. The difference between this private benefit and the LHS of (A-1) is positive and therefore corresponds to the non-internalized component of the marginal social cost, which we set equal to t_A. Given t_A, use s_{FB} and f_{FB} in the IC constraint (5) to write

$$B\left(1 - f_{FB}\right) - \left(1 - f_{FB}\right) g f_{FB} - T = B\left(1 - f_{FB}\right) - \left(1 - f_{FB} - s_{FB}\right) g s_{FB} - t_A.$$

and solve it w.r. to T when holding as an equality to obtain T_A. ∎

Proof of Proposition 7. The monopolist profits are given by $\pi = [p - t(s,f)]\,s + [P - T(s,f)]\,f$. The monopolist maximizes these profits subject to the standard IR and IC constraints. Incorporating these constraints in its objective function and applying a per traveler tax/subsidy scheme given by $t = s^c g - z$ and $T = f^c g - Z$, the maximization problem for the monopolist is given by

$$\begin{aligned} \max_{\substack{s^c \ge 0,\\ f^c \ge 0}} \; & \left(B\left(1 - s^c - f^c\right) - \left(1 - s^c - f^c\right) g s^c\right)\left(s^c + f^c\right) \\ & + g\left(s^c - f^c\right)\left(1 - s^c - f^c\right) f^c \\ & - \left(s^c g - z\right) s^c - \left(f^c g - Z\right) f^c \\ \text{s.t.} \;\; & s^c + f^c \le 1. \end{aligned} \tag{A-11}$$

The Lagrangian of this problem is

$$\begin{aligned} \mathcal{L}_T \equiv & \left(B\left(1 - s^c - f^c\right) - \left(1 - s^c - f^c\right) g s^c\right)\left(s^c + f^c\right) \\ & + g\left(s^c - f^c\right)\left(1 - s^c - f^c\right) f^c - \left(g s^c - z\right) s^c \\ & - \left(f^c g - Z\right) f^c - \lambda\left(s^c + f^c - 1\right). \end{aligned}$$

At the solutions, denoted by s_T^c, f_T^c and λ_T, the Kuhn–Tucker conditions require

$$\begin{aligned} \frac{\partial \mathcal{L}_T}{\partial s} = & B\left(1 - s_T^c - f_T^c\right) - B'\left(1 - s_T^c - f_T^c\right)\left(s_T^c + f_T^c\right) \\ & - \left(4 - -3\, s_T^c - 4 f_T^c\right) g s_T^c + z - \lambda_T = 0; \end{aligned} \tag{A-12}$$

$$\begin{aligned} \frac{\partial \mathcal{L}_T}{\partial f} = & B\left(1 - s_T^c - f_T^c\right) - B'\left(1 - s_T^c - f_T^c\right)\left(s_T^c + f_T^c\right) \\ & - g\left(3\left(f_T^c\right)^2 + 2\left(s_T^c\right)^2 - 4 f_T^c\right) + Z - \lambda_T = 0 \end{aligned} \tag{A-13}$$

$$\frac{\partial \mathcal{L}_T}{\partial \lambda} = s_T^c + f_T^c - 1 \le 0, \qquad \lambda_T \ge 0 \text{ and } \qquad \frac{\partial \mathcal{L}_T}{\partial \lambda}\lambda_T = 0. \tag{A-14}$$

Assume that $\lambda_T = 0$ and $s_T^c + f_T^c \le 1$. Substituting $\lambda_T = 0$ and $z = Z$ in (A-12) and (A-13), equalize them and solve w.r.to f_T^c to get f_T^c expressed as a function of s_T^c that is identical to the one at the social optimum as in (9). This shows that, for any coverage, when $z = Z$ the tax component of the tax scheme provides the monopolist the incentives to differentiate as in the social optimum.

Assume instead that $\lambda_T \ge 0$ and $s_T^c + f_T^c = 1$. Substituting $s_T^c = 1 - f_T^c$ and $z = Z$ in (A-12) and (A-13), equating them and solving w.r.to f_T^c gives $f_T^c = f_{FB}$. Using again $s_T^c = 1 - f_T^c$ gives $s_T^c = s_{FB}$.

Plugging f_T^c and s_T^c thus obtained into (A-12) or (A-13) and solving w.r.to λ_T it gives $\lambda_T = B(0) - B'(0) + g\frac{4+\sqrt{7}}{18} + z$. Solving $\lambda_T \ge 0$ w.r.to g gives the condition on g such the monopolist wants to cover the entire market, that is $g \le g_T \equiv \frac{18(B(0) - B'(0) + z)}{4+\sqrt{7}}$. Let $g_T|_{z=0}$ be defined as the critical value of g when the monopolist receives a subsidy equal to zero in both lanes. Notice that $g_T|_{z=0} < g_{FB}$, so that when $g \le g_T|_{z=0}$, a monopolist receiving a subsidy equal to zero in both lanes would cover the entire market, in line with the social planner. This, together with the results shown in the first paragraph after Eq. (A-14) implies that, when $g \le g_T|_{z=0}$, social optimality is restored by a tax scheme such that $t_C = g s^c$ and $T_C = g f^c$.

When instead $g_T|_{z=0} < g \le \frac{36\,B(0)}{4+\sqrt{7}}$, the social planner would cover the market, while a positive subsidy is required for the monopolist to cover the entire market. The smallest subsidy to ensure the same coverage condition as at the social optimum is obtained by solving $g = g_T$ w.r.to z to get $z_C = Z_C = B(0) - B'(0) + \frac{4+\sqrt{7}}{18}$.

Focus now on $g > \frac{36\,B(0)}{4+\sqrt{7}}$. In this case, the social planner partially covers the market, hence the scheme should deliver the same result under the automated monopolist. The tax scheme which restores social optimality is then computed as follows. Equalize pairwise the FOCs w.r.to s^c or f^c in the monopolist problem in (A-11) to the FOCs w.r.to s or f in the social planner problem in (8), then solve for z and Z to obtain z_C and Z_C. ■

Appendix B. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.jue.2023.103591.

References

Alemi, F., Circella, G., Mokhtarian, P., Handy, S., 2019. What drives the use of ridehailing in California? Ordered probit models of the usage frequency of Uber and Lyft. Transp. Res. C 102, 233–248.

Basso, L., 2008. Airport deregulation: Effects on prices and capacity. Int. J. Ind. Organ. 26, 1015–1031.

Basso, L., Zhang, A., 2007. Congestible facility rivalry in vertical structures. J. Urban Econ. 61, 218–237.

Boffa, F., Ceesay, M., Iozzi, A., 2023. Fleets, Congestion and Consumers' Choice in Urban Transport. Mimeo.

Börjesson, M., Fosgerau, M., Algers, S., 2012. On the income elasticity of the value of travel time. Transp. Res. A 46, 368–377.

Brancaccio, G., Kalouptsidi, T., Papageorgiou, N., Rosaia, M., 2023. Search frictions and efficiency in decentralized transport markets. Q. J. Econ. http://dx.doi.org/10.1093/qje/qjad023.

Brueckner, J.K., 2002. Airport congestion when carriers have market power. Amer. Econ. Rev. 92 (5), 1357–1375.

Brueckner, J.K., 2005. Internalization of airport congestion: A network analysis. Int. J. Ind. Organ. 23 (7–8), 599–614.

Buchholz, N., 2022. Spatial equilibrium, search frictions, and dynamic efficiency in the taxi industry. Rev. Econom. Stud. 89 (2), 556–591.

Buchholz, N., Doval, L., Kastl, J., Matějka, F., Salz, T., 2020. The Value of Time: Evidence from Auctioned Cab Rides. National Bureau of Economic Research, Working Paper 27087.

Castillo, J.C., 2022. Who benefits from surge pricing? Econometrica (forthcoming).

Castillo, J.C., Knoepfle, D.T., Weyl, E.G., 2023. Matching and Pricing in Ride Hailing: Wild Goose Chases and How to Solve Them. Available at SSRN: https://ssrn.com/abstract=2890666 or http://dx.doi.org/10.2139/ssrn.2890666.

Cremer, H., Thisse, J.-F., 1994. Commodity taxation in a differentiated oligopoly. Internat. Econom. Rev. 35 (3), 613–633.

Cummins, L., Sun, Y., Reynolds, M., 2021. Simulating the effectiveness of wave dissipation by FollowerStopper autonomous vehicles. Transp. Res. C 123, 102954.

Czerny, A., Zhang, A., 2015. Third-degree price discrimination in the presence of congestion externality. Can. J. Econ. 48 (4), 1430–1455.

Daniel, J.I., 1995. Congestion pricing and capacity of large hub airports: A bottleneck model with stochastic queues. Econometrica 63 (2), 327–370.

Daniel, J.I., Harback, K.T., 2008. (When) do hub airlines internalize their self-imposed congestion delays? J. Urban Econ. 63 (2), 583–612.

Erhardt, G.D., Roy, S., Cooper, D., Sana, B., Chen, M., Castiglione, J., 2019. Do transportation network companies decrease or increase congestion? Sci. Adv. 5 (5).

Fagnant, D., Kockelman, K., 2015. Preparing a nation for autonomous vehicles. Transp. Res. A 77 (4), 167–181.

Fréchette, G.R., Lizzeri, A., Salz, T., 2019. Frictions in a competitive, regulated market: evidence from Taxis. Am. Econ. Rev. 109 (8), 2954–2992.

Gogoll, Jan, Uhl, Matthias, 2018. ‚Rage against the machine: automation in the moral domain. J. Behav. Exp. Econ. 74, 97–103. http://dx.doi.org/10.1016/j.socec.2018.04.003.

Guda, H., Subramanian, U., 2019. Your uber is arriving: Managing on-demand workers through surge pricing, forecast communication, and worker incentives. Manage. Sci. 65 (5), 1949–2443.

Hall, J.D., 2018. Pareto improvements from Lexus lanes: The effects of pricing a portion of the lanes on congested highways. J. Public Econ. 158, 113–125.

Hall, J.D., 2021. Can tolling help everyone? Estimating the aggregate and distributional consequences of congestion pricing. J. Eur. Econom. Assoc. 19 (1), 441—474.

Lambertini, L., Mosca, M., 1999. On the regulation of a vertically differentiated market. Aust. Econ. Pap. 38 (4), 354–366.

Lamotte, R., de Palma, A., Geroliminis, N., 2017. On the use of reservation-based autonomous vehicles for demand management. Transp. Res. B 99, 205–227.

Leibbrandt, A., 2020. Behavioral constraints on price discrimination: Experimental evidence on pricing and customer antagonism. Eur. Econ. Rev. 121, 103303.

Lindsey, R., de Palma, A., Silva, H.E., 2018. Equilibrium in a dynamic model of congestion with large and small users. hal-01760135.

Lindsey, R., de Palma, A., Silva, A., 2019. Equilibrium in the bottleneck model with atomic and non atomic users. Transp. Res. B 124, 82–107.

Lopez, P.A., Behrisch, M., Bieker-Walz, L., Erdmann, J., Flötteröd, Y.-P., Hilbrich, R., Lücken, L., Rummel, J., Wagner, P., Wiessner, E., 2018. Microscopic traffic simulation using SUMO. In: 21st International Conference on Intelligent Transportation Systems. ITSC, pp. 2575–2582.

Mangrum, D., Molnar, A., 2020. The Marginal Congestion of a Taxi in New York City. Mimeo.

Mayer, C., Sinai, T., 2003. Network effects, congestion externalities, and air traffic delays: Or why not all delays are evil. Amer. Econ. Rev. 93 (4), 1194–1215.

Mirzaeian, N., Cho, S.-H., Scheller-Wolf, A., 2021. A queueing model and analysis for autonomous vehicles on highways. Manage. Sci. 67 (5), 2904–2923.

Molnar, A., 2013. Congesting the Commons: A Test for Strategic Congestion Externalities in the Airline Industry. Mimeo.

Montes, R., Sand-Zantman, W., Valletti, T., 2018. The value of personal information in online markets with endogenous privacy. Manage. Sci. 65 (3), 1342–1362.

Mussa, M., Rosen, S., 1978. Monopoly and product quality. J. Econom. Theory 18, 301–317.

Oberholzer-Gee, F., Weck-Hannemann, H., 2002. Pricing road use: politico-economic and fairness considerations. Transp. Res. D 7 (2), 357–371.

Ostrovsky, M., Schwarz, M., 2018. Carpooling and the Economics of Self-Driving Cars. NBER Working Paper No. 24349.

Overtoom, I., Correia, G., Huang, Y., Verbraeck, A., 2020. Assessing the impacts of shared autonomous vehicles on congestion and curb use: A traffic simulation study in The Hague, Netherlands. Int. J. Transp. Sci. Technol. 9 (3), 195–206.

Pels, E., Verhoef, E., 2004. The economics of airport congestion pricing. J. Urban Econ. 55 (2), 257–277.

Reitman, D., 1991. Endogenous quality differentiation in congested markets. J. Ind. Econ. XXXIX (6), 621–647.

Rosaia, N., 2020. Competing Platforms and Transport Equilibrium: Evidence from New York City. Mimeo.

Rupp, N.G., 2009. Do carriers internalize congestion costs? Empirical evidence on the internalization question. J. Urban Econ. 65 (1), 24–37.

Schrank, D., Lomax, T., Eisele, B., 2011. TTI's 2011 Urban Mobility Report. Texas Transportation Institute, The Texas A & M University System.

Silva, H., Lindsey, R., de Palma, A., van den Berg, V., 2016. On the existence and uniqueness of equilibrium in the bottleneck model with atomic users. Transp. Sci. 51 (3), 863–881.

Silva, H., Verhoef, E., 2013. Optimal pricing of flights and passengers at congested airports and the efficiency of atomistic charges. J. Public Econ. 106, 1–13.

Simoni, M., Kockelmana, K., Gurumurthya, K., Bischoff, J., 2019. Congestion pricing in a world of self-driving vehicles: An analysis of different strategies in alternative future scenarios. Transp. Res. C 98, 167–185.

Small, K., 2012. Valuation of travel time. Econ. Transp. 1, 2–14.

Small, K., Winston, C., Yan, J., 2005. Uncovering the distribution of motorists' preferences for travel time and reliability. Econometrica 73 (4), 1367–1382.

Spence, A., 1975. Monopoly, quality, and regulation. Bell J. Econ. 6 (2), 417–429.

Stern, R.E., Cui, S., Delle Monache, M.L., Bhadani, R., Bunting, M., Churchill, M., Hamilton, N., Haulcy, R., Pohlmann, H., Wu, F., Piccoli, B., Seibold, B., Sprinkle, J., Workai, D.B., 2018. Dissipation of stop-and-go waves via control of autonomous vehicles: Field experiments. Transp. Res. C 89, 205–221.

Tirachini, A., 2020. Ride-hailing, travel behaviour and sustainable mobility: an international review. Transportation 47, 2011–2047.

van den Berg, V., Verhoef, E., 2012. Is the travel time of private roads too short, too long, or just right? Transp. Res. B 46, 971–983.

van den Berg, V., Verhoef, E., 2016. Autonomous cars and dynamic bottleneck congestion: The effects on capacity, value of time and preference heterogeneity. Transp. Res. B 94, 43–60.

Van Dender, K., 2005. Duopoly prices under congested access. J. Reg. Sci. 45 (2), 343–362.

Verhoef, E.T., Silva, H.E., 2017. Dynamic equilibrium at a congestible facility under market power. Transp. Res. B 105, 174–192.

Ward, J.W., Michalek, J.J., Azevedo, I.L., Samarasa, C., Ferreira, P., 2019. Effects of on-demand ridesourcing on vehicle ownership, fuel consumption, vehicle miles traveled, and emissions per capita in U.S. states. Transp. Res. C 108, 289–301.

Zhang, K., Chen, H., Yao, S., Xu, L., Ge, J., Liu, X., Nie, M., 2019. An efficiency paradox of uberization. Available at SSRN: https://ssrn.com/abstract=3462912 or http://dx.doi.org/10.2139/ssrn.3462912.