Equilibrium Neglect and Political Feasibility

click here for latest version

preliminary and incomplete PLEASE DO NOT DISTRIBUTE

Bnaya Dreyfuss*

September 16, 2025

Abstract

When voters underappreciate the equilibrium effects of policies, efficient reforms can become politically infeasible. I study the policy implications of *equilibrium neglect* and propose a potential remedy. I first develop a general framework of equilibrium neglect, and then use this framework to construct a portable and generally applicable potential remedy: using off-path contingent rebates, a social planner can generate a policy that (i) implements the social optimum, (ii) does not increase the deficit—on and off the equilibrium path—and (iii) is politically feasible even in the presence of equilibrium-neglectful voters. This solution can be applied in a large class of problems. Empirically, I first present new evidence suggesting that real-world surges in public support for congestion pricing are driven by frequent commuters who learn about its equilibrium benefits. I then administer an opinion poll on congestion pricing with a large representative sample from six US metros. Respondents are overly pessimistic about traffic-reduction effects of congestion pricing, which correlate with strong opposition to the policy. Adding contingent compensation that pays if traffic remains high significantly increases public support, especially among potential recipients. Turning to to the supply side of policymaking, I survey civil servants and US state legislators. Both groups recognize my proposed remedy's public appeal, and legislators' own support for congestion pricing increases; civil servants, however, show a preference for more conventional tools.

*Harvard University. Email: bdreyfuss@g.harvard.edu. I thank Raj Chetty, Katherine Coffman, Justin de Benedictis-Kessner, James Druckman, Benjamin Enke, Ed Glaeser, Yannai Gonczarowski, Ori Heffetz, Gabriel Kreindler, David Laibson, Shengwu Li, Matthew Rabin, Alex Rees-Jones, Joshua Schwartzstein, Jesse Shapiro, David Yang and participants at various Harvard workshops, BEAM 2025, and SITE 2025 for helpful feedback and suggestions. I thank Siena Polls and Benjamin and Joshua Moshes for generously sharing their data, and the Policymakers Lab for administering the civil servants survey. This project is supported by the Chae Family Fund, the Lab for Economic Applications and Policy (LEAP), Kenneth C. Griffin Research Fund, Institute for Quantitative Social Science (IQSS), the Center for American Political Studies (CAPS) and Time-sharing Experiments for the Social Sciences (TESS).

1 Introduction

Economists and the general public often disagree about policies. Despite years of policy advocacy, policies economists typically oppose, such as rent control and tariffs, are viewed positively in the eyes of much of the public, whereas policies economists tend to support, such as carbon taxes and congestion pricing, are viewed negatively (Caplan, 2001; Sapienza and Zingales, 2013).

There are many legitimate reasons for such disagreements: the models economists use to generate policy recommendations can be misspecified, oversimplified, or ignore aspects that people view as important. However, in many cases, there are good reasons to believe that the way people intuit about the costs and benefits of policies may lead them to the wrong conclusions. In particular, when evaluating a new policy, people often underestimate or overlook a crucial element: others' reaction to it, or its *equilibrium effects* (Dal Bó et al., 2018). This tendency can distort perceptions of a policy's indirect costs or benefits, leading to wrong assessment of its overall impact. I call this bias *equilibrium neglect*. Indeed, there is a large body of work, reviewed below, showing that in many contexts people fail to perform contingent reasoning (Niederle and Vespa, 2023), specifically when thinking about others (Eyster and Rabin, 2005).

Taking this body of work as the starting point, I ask: what are the implications for policy design? In particular, can policymakers design welfare-improving policies that can win support from equilibrium-neglecting voters? I first model what it means for agents to neglect equilibrium when evaluating a new policy, and then use this model to identify a portable and generally applicable potential remedy. The remedy adds *off-path* contingent rebates to an existing policy, ensuring that if the true benefits of the policy are positive, then the agent's *perceived* benefits of the policy are also positive, even when they neglect its equilibrium effects. I then empirically test the effectiveness of this remedy in increasing public support for congestion pricing in an opinion poll with a representative sample. Consistent with the theory, I find that respondents' beliefs about the effect of congestion pricing are overly pessimistic, and that pessimism correlates with opposition to the policy. The proposed remedy—contingent compensation mechanism—significantly increases support, especially among targeted recipients.

While equilibrium neglect can play a role in the political feasibility of various policies (e.g., taxation, price controls and industrial policy), I use congestion pricing for my theoretical running example and empirical application. Congestion pricing is particularly suitable for studying equilibrium neglect for a few reasons. First, it offers a straightforward distinction between direct effects (toll increase) and equilibrium effects (congestion reduction), both of which are unidimensional, easily observed, and precisely measured. Second, while there is consensus among economists and transportation engineers that it can increase welfare, congestion pricing is in practice notoriously difficult to implement

due to political feasibility constraints (Gu et al., 2018). Third, there is evidence (reviewed below) suggesting that EN is particularly relevant in the context of congestion pricing: support for congestion pricing after its implementation (i.e., after observing the benefits) is typically 10-20 percentage points higher compared to pre-implementation support (i.e., before observing the effects on traffic). In addition, there is evidence showing that people underappreciate the congestion-reducing effects of the policy ex ante, but strongly revise their beliefs having observed its effects.

Section 2 describes the contribution of this paper contributes to the literature on behavioral mechanism design and behavioral public policy, failures in contingent reasoning, and the political feasibility of environmental policies.

Section 3 presents the theoretical framework. A running example throughout the section is a model of road congestion, used to illustrate the main ideas and connect them to the empirical application. A social planner chooses a policy from a policy space. Policies shape payoffs and incentives and therefore each policy induces a different game. I model how agents perceive the costs and benefits of a policy change from the status, i.e., the perceived benefit of moving between games within a family of games. The effect of any policy change can be decomposed into a *direct effect*—the effect of a policy change holding others' behavior fixed—and an *equilibrium effect*—the effect of converging to the new equilibrium, holding the policy fixed. I allow agents to underweight or cognitively discount the latter effect, which I call *equilibrium neglect* (EN). EN can come in the form of explicit equilibrium skepticism—believing that others may not change their behavior in response to the policy change—or in the form of *naïveté*, an attenuated perception of equilibrium effects, which are typically complex, less salient and harder to predict than direct effects. In the context road congestion, I show that equilibrium neglect may result in inflated WTP for investing in lane expansions, by ignoring their effect on increased demand for driving, partially offsetting the improvement in commute times.

Section 3.4 quantifies the implications of EN to political feasibility. The *political feasibility* of a given policy depends on *perceived gains* (as opposed to real gains) relative to the status quo through a *political feasibility rule* that respects standard monotonicity and consensus properties. When agents are sufficiently neglectful, unboundedly beneficial policies can be politically infeasible, and unboundedly harmful policies politically feasible. Moving beyond mere existence, I characterize the conditions under which EN is likely to constrain political feasibility. Intuitively, when high switching costs are associated with large equilibrium benefits, there is a set of agents for whom—under sufficient equilibrium neglect—perceived gains (which only consider the taxes and switching costs) are negative, yet actual gains (which also take into account equilibrium benefits) are positive. I call such agents *spurious opposition*. Such cases naturally arise in generic settings where agents' types determine both equilibrium benefit and switching costs. In the context of road congestion, the value of time savings is typically associated with how much the agent value

their ability to drive, making congestion pricing particularly prone to spurious opposition composed of equilibrium-neglecting infra-marginal drivers: these drivers perceive the high cost of either paying the toll or switching their behavior, but not the equilibrium benefits from shorter commute times.

Section 3.5 describes my proposed remedy. Intuitively, spurious opposition arises from over-weighting a policy's direct effect, which holds others' behavior fixed. The remedy adds compensation that pays agents when others do not change their behavior—the very contingency they worry about. A sufficiently high contingent payment can therefore make the direct effect positive, closing the gap between perceived and actual gains, for any degree of equilibrium neglect. I call such policies *EN-robust*. Since compensation depends only on others' actions, individual incentives remain unchanged and the original equilibrium is preserved—the policies are *outcome equivalent*. Since payment is triggered only if status-quo behavior persists, which does not occur in equilibrium, no compensation is actually paid and the policies are *deficit equivalent*. Theorem 1 formalizes this intuition, showing that starting from any status quo, and for any constrained optimum, we can construct an EN-robust counterpart that preserves outcome- and deficit-equivalence.

Theorem 2 shows that if we additionally allow the compensation to be conditioned on an individual agent's past behavior, we can ensure that even off the equilibrium path, the deficit of the EN-robust policy never exceeds the status-quo deficit, even off the equilibrium path. The key insight is that compensation is triggered when agents do not respond to the policy. This non-responsiveness generates revenue surplus that used to fund that compensation. For instance, in congestion pricing, if congestion remains high despite the toll, toll revenue is also high. Knowing agents' status-quo actions allows us to only target potential members of the spurious opposition. In congestion pricing, the spurious opposition are inframarginal drivers who under-appreciate the benefits of shorter commute times, making status-quo drivers the natural targets for compensation. This enables us to construct an off-path compensation scheme that appeases precisely this spurious opposition, funded by the policy's own revenue, keeping budget constraints off path, while achieving the desired social outcome in equilibrium.

In section 4 I turn to empirically investigate equilibrium neglect and its potential remedy in the context of congestion pricing. As motivation, I first analyze existing polls from seven cities that implemented congestion pricing, showing that support consistently increases post-implementation. I analyze individual responses from two Siena polls conducted in New York state—before (December 2024, N=834) and after (March 2025, N=806) NYC's congestion pricing implementation—and classify respondents by geographical area based on county-level Manhattan commute intensity derived from ACS traffic flows. Relative to the low-commute group, the increase in net support among high-commute respondents is 29.8 higher (SE = 8.48). This finding is consistent with equilibrium neglect generating spurious opposition that, after implementation, dissipates

in geographical areas where benefits are visible.

Section 4.2 describes the design of the poll I administered. The poll has two goals: first, documenting equilibrium neglect and mapping it to policy attitudes in a real policy context. Second, testing whether the remedy implied by the theory—contingent compensation—is effective in increasing support in practice. Ex-ante it is far from obvious that contingent compensation will have a positive effect on support: understanding off-path rebates requires agents to engage in contingent reasoning, which is the very skill they may lack due to EN. In addition, contingent compensation can be complex, may feel unnatural or seen as non-credible, which could negatively affect public support.

I survey a representative sample drawn from NORC's AmeriSpeak probability panel. The sample consists of 2,056 adults residing in one of six major U.S. metropolitan areas: Boston, Chicago, Dallas, Los Angeles, San Francisco, and Washington D.C. The survey has a two-stage, between-subject design. In the first stage, all respondents see a standard congestion pricing policy modeled after the original $15 toll proposed by the MTA in NYC in 2024. Respondents then answer questions about their beliefs regarding congestion reduction and commute habits.

In the second stage, they are asked to state their support for one of three randomly assigned modified policies: the *EN-robust* treatment adds a "Monitoring & Assurance" component: respondents are told that traffic will be monitored and data will be publicly available, and that if after the implementation traffic decreases by less than 8% in a given month (compared to that month's baseline level), workers who commute to work in the downtown area would receive a $120 monthly compensation. The *Control* treatment includes only the monitoring component, without compensation. The third treatment, *Toll-Reduction*, lowers the toll from $15 to $9 and serves as benchmark to measure the impact of EN-robustness against a more conventional adjustment margin.

I present my findings in Section 4.4. First, there is strong opposition to standard congestion pricing, which in turn translates to strong opposition. About 63% of the respondents oppose the policy while only 21% support it, resulting in -42 percentage points *net support*. Respondents substantially underestimate the benefits from congestion pricing. Half the respondents are asked to predict changes to their own commute times. Around 39% indicate they expect no change, and another 20% that traffic conditions will *worsen*. The other half were asked to estimate the effect of NYC's congestion pricing, which provides a concrete, easily verifiable benchmark where actual performance can be observed. I find similarly high level of under-appreciation: 65% of respondents underestimate the actual reduction (which was approximately 12 minutes), and 33% estimate little to no improvement from the 23-minute baseline.

This over-pessimism in turn translates to opposition to the policy. For the NYC question, there is a 17.4 point difference in net support between those expecting meaningful improvements versus those expecting little change (-57 vs. -39.6; $p = 0.001$). Beliefs

about own-time commute predict support even more strongly: net support among those who predict any amount of time savings in their commute is 36.7 points higher than those who predict no savings (-62.4 vs. -25.7; $p < 0.001$). Translating beliefs into dollar units, I combine perceived time savings and respondents' value of time (inferred from income) into a *perceived benefit* index for each subject. An additional dollar of perceived benefit corresponds to about 2 point increase in net support (SE = 0.33).

To test my remedy, I compare 2nd-stage support across treatments. I find a significant effect of contingent compensation on support, concentrated among those it targets. Among respondents who work in the downtown area ($N = 184$)—those who would receive the compensation if congestion remains high—net support in EN-robust increases from baseline by 48 percentage points more than in control (SE = 12.21). The treatment effect on the whole sample ($N = 2,056$) is a more moderate yet highly significant 11 point increase (SE = 3.85). Therefore, contingent compensation increases support among the group it targets. By alleviating the concerns of those who are negatively affected, it can be an effective tool of increasing policy support. In contrast, Toll-Reduction shows no meaningful improvement relative to control, both among downtown workers and in the whole sample.

Consistent with the model, the treatment effect is driven by respondents whose beliefs about traffic reduction are overly pessimistic: the effect on those who predict no improvement is 14.41–17.85 points higher than those who predict some improvement (for whom the treatment effect is small and indistinguishable from zero). Furthermore, treatment is also concentrated among respondents who spend more time on the question: the effect on those spending above the median duration (31 seconds) is 17.24–18.18 points higher than those below the median (for whom the treatment effect is statistically zero). Taken together, these findings suggest that treatment effects in my data reflect a genuine attitude reaction to the actual content of the policy—the addition of contingent compensation—rather than surveyor demand effects and other arbitrary design artifacts.

Finally, to understand how these dynamics translate into political feasibility, Section 5 turns to the supply side of policymaking, surveying two key groups: state legislators and civil servants. I elicit policymakers' own beliefs and attitudes about congestion pricing as well as their second-order beliefs, i.e., their perceptions of the public's beliefs and attitudes. State legislators exhibit equilibrium neglect at rates similar to the general public, and as with the public, these beliefs strongly predict their policy support. Accordingly, their support for congestion pricing increases substantially for the EN-robust version, as does their estimate of public support for it. In contrast, civil servants largely demonstrate a textbook understanding of the policy's effects, and support standard congestion pricing. However, their support for EN-robust congestion pricing is *lower*, even as they correctly predict it will be more popular with the public. These findings highlight a potential disconnect, where the experts who design policies may be skeptical of the very remedies

that could make those policies politically viable with elected officials and the public they are meant to serve.

I conclude in Section 6.

2 Related Literature

My main contribution is to construct and test a theory-driven practical remedy for a feasibility challenge driven by cognitive limitations. The core insight contributes to behavioral mechanism design, and its practical implications contribute to the literature on behavioral public economics.

I start with the former. The paper adds to a growing literature on how cognitive limitations shape the design of mechanisms. The strategic simplicity program (Li, 2017; Bó and Hakimov, 2023; Pycia and Troyan, 2023; Nagel and Saitto, 2025) characterizes strategy-proof mechanisms by the amount of contingent or forward-looking reasoning agents need in order to identify their dominant strategy. A related strand studies which social outcomes and rules a planner can implement with cognitively limited participants: de Clippel (2014); de Clippel et al. (2019) provide implementation results when choice correspondences may depart from the classical benchmark and under level-k equilibrium, respectively. Gavan and Penta (2025) extend Nash implementation to admit deviations from equilibrium behavior while restricting outcomes to an acceptable set. Closest in spirit to my remedy, but different in context, Gonczarowski et al. (2023) and Dreyfuss et al. (2025), both modify existing matching mechanisms in outcome-equivalent ways to handle departures from the classical model. The key distinction between this body of work and my paper is focus: rather than studying incentive compatibility, i.e., behavior *within a mechanism,* I study how cognitive limitations affect individual rationality, i.e., preferences *over mechanisms*, which received much less attention.

Moving from abstract mechanisms to applications, this paper studies practical policy design questions in the presence of cognitively limited or biased agents. In doing so it joins the behavioral public economics literature, reviewed in Bernheim and Taubinsky (2018). For example, Chetty et al. (2009), Rees-Jones and Taubinsky (2020), and Allcott et al. (2019) document systematic biases and analyze their potential implications for the design of wellfare maximizing taxation. The same distinction as above applies: much of this work analyzes the behavior of cognitively limited agents *under* a given policy, whereas I study the implications of the constraint these limitations impose by distorting agents' preferences *over* policies.

This paper also contributes to the literature on the economic implications of failures in contingent thinking, reviewed in Niederle and Vespa (2023). A central case is misunderstanding how actions map to incentives, which matters both theoretically (Eyster and Rabin, 2005; Cohen and Li, 2023) and empirically (e.g., Esponda and Vespa, 2014).

Most relevant here is Dal Bó et al. (2018), who document systematic underappreciation of equilibrium effects in the lab, showing that subjects fail to recognize that the other will adjust their behavior in response to incentive changes, leading them to favor a socially dominated game. I take the core insight of this paper as my starting point: I generalize their framework to an abstract policy setting, derive implications for political feasibility, and use it to construct a remedy. Empirically, I document equilibrium underappreciation in a high-stakes policy setting and test the remedy in that same setting. Another relevant paper is Elmendorf et al. (2025), who document systematic underappreciation of housing price responses to supply shocks. More broadly, the paper relates to the growing literature on people's misspecified subjective models of the economy (Andre et al., 2022; Stantcheva, 2022; Chinoy et al., 2023; Binetti et al., 2024).

Finally, I contribute to the large literature on public support to environmental policies, reviewed in Carattini et al. (2018b), and to congestion pricing in particular, reviewed in Gu et al. (2018). This literature consistently finds that support for congestion pricing and other environmental policies is higher after implementation, and that people are often skeptical of effectiveness beforehand but typically revise beliefs and attitudes after experiencing the effects (see, for example, Schuitema et al. 2010 and Andersson and Nässén 2016 in the context of congestion pricing, Carattini et al. 2018a in the context of a garbage tax, and Murray and Rivers 2015 in the context of carbon taxes). Both findings are consistent with equilibrium neglect that dissipates once people are exposed to the true benefits of the policy. My first contribution to this literature is to offer a general framework that can explain and organize these findings. My second and main contribution is to propose a potential remedy to these frictions. Comparing my proposed remedy to the tools discussed in the literature, perceived effectiveness strongly predicts support, but evidence on the persistent effects of information-provision treatments is mixed (for example, Douenne and Fabre 2022 show small effects for carbon taxes; Baranzini et al. 2021 show null effects for congestion pricing, and Allcott and Knittel 2019 in the context of fuel economy; however, see for example Semken 2025 and Dechezleprêtre et al. 2025 who find positive effects in the context of climate change). In contrast, changes to the actual content of a policy have been shown to shape support. In particular, reducing the size of the tax (Baranzini et al., 2021) and earmarking revenues for specific purposes (Jaensirisak et al., 2005) are generally associated with higher support for congestion pricing. My proposal combines these approaches: it addresses over-pessimism not by attempting to change it directly, but by leveraging it into a policy change that increases perceived benefits, especially for over-pessimistic voters. The contingent compensation mechanism can be thought of as a specific form of earmarking: explicit revenue recycling that pays more when outcomes are worse, potentially persuading those most concerned about them.

3 Theory

3.1 Preliminaries

There is a set N of agents.[1] Each agent $n \in N$ chooses an action $a_n \in A_n$. An action profile is denoted by $a = (a_n)_{n \in N}$, and the set of action profiles is denoted by $A \equiv \prod_{n \in N} A_n$. A social planner chooses a *policy profile* $T \in \mathbb{T}$. A policy profile is a collection of agent-specific policies, $T = (T_n)_{n \in N}$. Each agent n has preferences over action profiles and their own policy component, represented by $v_n : A \times \mathbb{T}_n \to \mathbb{R}$, denoted $v_n(a; T_n)$, where $\mathbb{T}_n$ is the set of possible policies for agent n. The set of utility functions is denoted $v = \{v_n\}_{n \in N}$.

It is useful view policies as operating through *consequences*. In this representation, each agent n has a consequence space C_n, and a policy is a function $T_n : A \to C_n$ mapping action profiles to consequences. With a sufficiently rich consequence space, there exists $u_n : A \times C_n \to \mathbb{R}$ such that

$$v_n(a; T_n) = u_n(a, T_n(a)).$$

Consequences can capture transfers, public goods, infrastructure, or resource allocations; policies specify how actions map to these (monetary or non-monetary) outcomes. A policy T_n is *unilateral* if its induced consequences depend only on n's own behavior: $T_n(a_n, a_{-n}) = T'_n(a_n)$, and *multilateral* otherwise.

An *economy* $\mathcal{E} = \langle N, A, v, \mathbb{T} \rangle$ is a policy-parametrized family of normal-form games.[2] This framework is aimed at capturing how agents perceive the effect of a policy change, i.e., moving across games within the family. For simplicity, I restrict attention to settings with a pure Nash Equilibrium for any $T \in \mathbb{T}$, as is the case in all congestion games and, more generally, all potential games. I further assume that a unique pure Nash equilibrium is selected under each T via some selection criterion. The (unique) Nash equilibrium under the policy profile T is denoted by $a^*(T)$. Agent n's best response to a profile a_{-n} and a policy T_n is denoted $a_n^*(a_{-n}, T_n)$.

To illustrate the different results in this section, I develop a simple congestion model. I introduce here its main components and how they map to the general setup.

Example (Road Congestion and Transportation Policies: Setup.) A transportation network connects where people live and where they work. In all but one application, I abstract away from route choice and consider a single road. There is a unit continuum of commuters N. Each commuter $n \in N$ makes a binary choice between driving ($a_n = 1$) and taking public transit ($a_n = 0$). The government determines the level of infrastructure

[1] When needed for integration (e.g., when N is uncountable), we endow N with a measure space structure $(N, \mathcal{N}, \mu)$ and assume relevant functions are measurable.

[2] For any policy profile $T \in \mathbb{T}$, the induced game is $G(T) \equiv \langle N, (A_n)_{n \in N}, (v_n(\cdot; T_n))_{n \in N} \rangle$.

I, a toll p drivers must pay, and a lump-sum transfer k. Let $\alpha(a)$ denote the share of commuters who drive in profile a. Utility from driving to agent n is given by

$$v_n(a;T) = \theta_n Q(\alpha(a);I) - (b+p) + k,$$

where $\theta_n \geq 0$ is agent n's value of time, $Q(\alpha(a);I)$ is time saving from driving given a level of congestion α and a level of infrastructure investment I. It is monotonically increasing in I, monotonically decreasing and continuously differentiable in α and $Q(0;I) > 0$ for all I. $b > 0$ captures driving costs (e.g., gas, depreciation) net of public transit ticket price. We assume that value of time θ is distributed according to F which is continuously differentiable with a strictly positive density over $[0,\infty)$. Utility from public transit is normalized to the lump-sum transfer k.

The policy space is $\mathbb{T} = (I,p,k) \subset \mathbb{R}^3$,[3] and is constrained by the *deficit* D—defined as government spending minus tax and (equilibrium) toll revenue—which cannot exceed $\overline{D}$.[4] Formally:

$$\mathbb{T} = \left\{(k,p,I) : k + I - p \cdot \alpha(a^*(T)) \leq \overline{D}\right\}.$$

The unique equilibrium in this game is a cutoff equilibrium:

Result 1. *For any $T \in \mathbb{T}$ there is a unique Nash equilibrium $a^*(T)$ characterized by the cutoff $\theta^*(T)$ where $a_n = 1$ if and only if $\theta_n \geq \theta^*(T)$.*

All proofs are in the Online Appendix.

3.2 Equilibrium Neglect

To understand how equilibrium neglect shapes political feasibility, we first need to model how agents evaluate policy changes. A policy change affects agents through two channels: first, directly, by changing their payoffs and second, indirectly, through its effect on others' behavior. The key insight is that agents assessing a proposed policy may fail to fully account for the latter.

Consider a status-quo policy T^0 with the corresponding equilibrium $a^0 = a^*(T^0)$. When a new policy T is proposed, the *total gain* for agent n is given by

$$\delta_n(T,T^0) = v_n\left(a^*(T);T_n\right) - v_n\left(a^*(T^0);T_n^0\right),$$

I will often refer to the total gain as the *actual* gain. A profile of total gains from T given T^0 is denoted $\delta(T,T^0) = (\delta_n(T,T^0))_{n\in N}$, and a generic profile of total gains is denoted δ.

[3] In this case, a policy T_n can be modeled as operating through consequences if the consequence space contains time saving and wealth, i.e., $C_n \subseteq \mathbb{R}^2$, where $c = (Q,w)$ represents time savings Q and wealth w and $u_n(a,c) = \theta_n Q + w$.

[4] If the government needs to raise a minimal revenue, we let $\overline{D} < 0$.

Let $\hat{a}_n = a_n^*\left(a_{-n}^0, T_n\right)$ denote agent n's best response to policy T, fixing others' behavior at the status quo. Any total gain can be decomposed into direct and equilibrium effects:

$$\delta_n(T, T^0) = \underbrace{v_n\left((\hat{a}_n,\, a_{-n}^0); T_n\right) - v_n\left(a^0; T_n^0\right)}_{\text{direct effect}} + \underbrace{v_n\left(a^*(T); T_n\right) - v_n\left((\hat{a}_n,\, a_{-n}^0); T_n\right)}_{\text{equilibrium effect}}. \quad (1)$$

The decomposition adds and subtracts the counterfactual $v_n\left((\hat{a}_n,\, a_{-n}^0); T_n\right)$—the agent's utility under the proposed policy T, with others' behavior held at the status quo. The *direct effect* captures the impact of moving from T^0 to T absent any equilibrium adjustments (i.e., fixing others' actions at a_{-n}^0). The *equilibrium effect* captures how others' behavioral responses to the policy change affect the agent's welfare, holding the policy fixed at T.

Agents suffer from *equilibrium neglect* when their *perceived* gain from a proposed policy systematically underweights its equilibrium effects:

Definition 1 (Perceived Gain)**.** *Let $\eta_n \in [0, 1]$. Agent n's perceived gain from proposed policy T under the status quo T^0 is given by:*

$$\tilde{\delta}_n(T, T^0; \eta_n) = v_n\left((\hat{a}_n,\, a_{-n}^0); T_n\right) - v_n\left(a^0; T_n^0\right) + \eta_n\left(v_n\left(a^*(T); T_n\right) - v_n\left((\hat{a}_n,\, a_{-n}^0); T_n\right)\right).$$

The *perceived* gain $\tilde{\delta}_n$ deviates from the real gain δ_n as a function of equilibrium neglect, parameterized by η (I index η_n by agent n but often generically refer to the parameter as η). I call agents with $\delta_n > 0$ and $\tilde{\delta}_n < 0$ the *spurious opposition*. Finally, a profile of equilibrium-neglect parameters is denoted $\boldsymbol{\eta} = (\eta_n)_{n \in N}$. For any T, T', denote the profile of perceived gains by $\tilde{\delta}(T, T'; \boldsymbol{\eta}) = \left(\tilde{\delta}_n(T, T'; \eta_n)\right)_{n \in N}$, and a generic profile of perceived gains is denoted $\tilde{\delta}$.

3.2.1 Discussion

The equilibrium neglect parameter, η, can be interpreted in a number of ways, which fall into one of two main categories. The first category is *equilibrium naïveté* where agents partly overlook equilibrium effects due to its complexity. Direct effects are typically straightforward to understand and compute.[5] In contrast, fixed-point reasoning is difficult and often requires perceiving and aggregating over a population distribution, adding another layer of complexity. Such difficult-to-calculate objects may be under-appreciated and discounted by η. Alternatively, the perceived gain, $\tilde{\delta}_n$ could be interpreted probabilistically, in which case $1 - \eta$ is the probability that the agent overlooks the equilibrium

[5]Notice that the agent's own optimization is included in the direct effect. The model and the main results could be extended to accommodate agents neglecting their own optimization, but this would significantly complicate notation without adding much insight.

effect when evaluating a proposed policy. A related interpretation is that η captures noisy cognition (Enke and Graeber, 2023): agents predict the total gain by relying on "cognitive signals" generated through sequential cognitive sampling or deliberation process, which attenuates the more complex equilibrium component towards zero.[6]

The second category is *equilibrium skepticism*, which reflects a more explicit belief. For example, η can capture the perceived probability of others' responsiveness: agents may assign a non-zero probability $(1-\eta)$ to the event that others' behavior will remain fixed at the status quo, thereby neglecting others' elasticity. This may reflect skepticism about others' willingness or ability to adjust behavior in response to incentives, or may arise from attributing others' current behavior to fixed dispositions rather than to their response to a given situation, resulting in systematic under-appreciation of others' elasticity.

As none of my results hinge on the particular psychological mechanism driving equilibrium neglect, in what follows I remain mostly agnostic to the interpretation of η.

The formulation has the standard structure nesting the classical model as a special case: when $\eta = 1$, the *perfect foresight* case, equilibrium effects are correctly perceived $(\tilde{\delta}_n(T, T^0; 1) = \delta_n(T, T^0))$. The other extreme is *full neglect*, $\eta = 0$ where agents completely disregard changes in others' behavior and only perceive the direct effect. The intermediate range $\eta \in (0, 1)$ captures *partial neglect*.[7]

Example (Over-investment in roads.) Let $T^0 = (0, 0, 0)$ and denote $\alpha^0 = \alpha(a^0)$. Consider a road-investment policy funded by a lump-sum tax: $T = (h, 0, -h)$. Focusing on the marginal dollar (i.e., small values of h), for status-quo PT riders (agents with $\theta_n < \theta^*(T^0)$), the actual gain equals perceived gain: for them, the result of a unit increase in investment is simply a tax increase. Therefore, $\tilde{\delta}_n(T, T^0; \eta) = -h$. The perceived gain of status-quo drivers $(\theta_n \geq \theta^*(T^0))$, is given by

$$\tilde{\delta}_n(T, T^0; \eta_n) = h \left(\underbrace{-1 + \theta_n \frac{\Delta Q(\alpha(T^0); 0)}{\Delta I}}_{\text{direct effect}} + \eta_n \underbrace{\theta_n \frac{\partial Q(\alpha^0; 0)}{\partial \alpha} \cdot \frac{\Delta \alpha(a^*(T^0))}{\Delta I}}_{\text{equilibrium effect}} \right).$$

The first part of the expression is the benefit from shorter commute time when infrastructure improves, and the second part captures the equilibrium increase in road usage, which partly offsets it. Since the equilibrium effect (slower commute) is negative, agents with $\eta < 1$ over-value the benefits from road investments. For example, fixing $\eta < 1$

[6] The expression in definition 1 is the mean posterior of this process if (i) predicting the (typically simple) direct effect involves minimal noise and (ii) predicting the (typically complex) equilibrium effect involves a non-trivial amount of cognitive noise, and the prior is centered around zero (a natural cognitive default when equilibrium effects average zero).

[7] The framework can also capture *over*-weighting of equilibrium effects when $\eta > 1$. While this form of bias can be important in policy discourse (and is perhaps common even, or maybe especially, among economists), I leave it for future work.

for all agents, there is always a non-empty set of agents who perceive a positive gain ($\tilde{\delta}_n(T, T^0; \eta) > 0$) but whose real gain is negative ($\delta_n(T, T^0) < 0$), i.e., a *spurious coalition* creating a gap between the normative and the perceived willingness to accept a tax increase, potentially leading to over-investment in roads.[8]

3.3 Political Feasibility

How do perceived gains translate to political feasibility?

Definition 2. *Given a status-quo policy T^0 and a proposed policy T, and associated profile of perceived gains $\tilde{\delta} \in \mathbb{R}^N$, a political feasibility rule is a function $f : \mathbb{R}^N \to \{T^0, T\}$. A policy T is politically feasible under status quo T^0 if $f(\tilde{\delta}) = T$.*

I impose two restrictions on the political feasibility rule:

Assumption 1. *The political feasibility rule f satisfies the following properties:*

1. *Monotonicity: if $f(\tilde{\delta}') = T$ and $\tilde{\delta}_n \geq \tilde{\delta}'_n$ for all n, then $f(\tilde{\delta}) = T$.*
2. *Consensus: if $\tilde{\delta}_n < 0$ for all $n \in N$ then $f\left(\tilde{\delta}\right) = T^0$, and if $\tilde{\delta}_n \geq 0$ for all $n \in N$ then $f\left(\tilde{\delta}\right) = T$.*

A common and intuitive choice for a political feasibility rule which I will often use is a minimal required share of weighted support:

$$f(\delta) = T \iff \sum_{n \in N} \omega_n \mathbf{1}_{\{\tilde{\delta}_n \geq 0\}} \equiv S_\omega(\tilde{\delta}) \geq \beta,$$

where the ω_n's are weakly positive influence weights with $\sum_{n \in N} \omega_n = 1$, and $\beta \in (0, 1]$ is a feasibility threshold.[9] This rule requires that a policy T has a sufficient (influence-weighted) support to overcome any opposition. Two focal cases are $\beta = 1$, where only Pareto improvements are politically feasible (e.g., because a vocal minority can maintain the status quo), and $\beta = \frac{1}{2}$ with equal influence weights—the simple majority or median voter case.

Alternatively, a rule might implement a weighted Kaldor-Hicks test, requiring that the weighted sum of welfare changes across all individuals exceeds some threshold: $\sum_{n \in N} \omega_n \tilde{\delta}_n \geq$

[8] This idea is captured by Goodwin (1997): "From the late 1950s onwards the transport planning orthodoxy was 'predict and provide'... first we forecast how much traffic there will be, and then we build enough road space to accommodate it. [...] 'Predict-and-provide' actually meant, inevitably, 'predict-and-*under*provide' [since] supply of road space will not—because it cannot—be increased to match demand. [Eventually, a transport committee] reported that road construction in conditions of congestion normally results in an increase in the total volume of traffic, hence a shorter period of relief from congestion, thereby consigning to history the untenable assumption that the total volume of traffic is unaffected by traveling conditions."

[9] When N is uncountable, interpret $\sum_{n \in N}$ as integration with respect to the measure on N, and assume the perceived gains δ_n and weights ω_n are measurable functions of n.

b. For some set of weights $\omega \in \mathbb{R}^N$. A natural choice for the minimum required aggregate welfare improvement is $b = 0$, with $b > 0$ capturing any standard frictions or status-quo bias.

3.4 Implications

Having defined the two main objects of this paper—equilibrium neglect and political feasibility—we want to quantify the degree to which the former imposes a constraint on the latter.

3.4.1 Cyclical Preferences and Unbounded Reversals

The first question one might ask is whether an equilibrium-neglecting agent has coherent policy preferences. Proposition 1 shows that the answer is generally negative:

Proposition 1 (Cyclical Preferences)**.** *Let $\succ_n$ be a preference relation over the space of policies $\mathbb{T}$, defined as $T \succ_n T' \iff \tilde{\delta}_n(T, T'; \eta_n) > 0$. Then for any $\eta_n < 1$, there exists an economy $\mathcal{E}$ such that $\succ_n$ is not asymmetric, i.e., $T \succ_n T' \succ_n T$ for some $T, T' \in \mathbb{T}$.*

Note that asymmetry violations are more severe than transitivity violations: they admit *pairwise* contradictions (not just chains). The proof can be shown in the context of the running congestion example.

Example (Cyclical transportation preferences.) Consider the case where *two* routes, R_1 and R_1, connect where people live and work. Commuters choose between public transit ($a_n = 0$) and driving on route R_i ($a_n = i$, $i \in \{1, 2\}$). Time savings on R_i is

$$Q_i(\alpha_i; T) = s_i - \alpha_i(a),$$

where $\alpha_i(a)$ is the share of commuters choosing R_i in profile a, and s_i is chosen by the planner. In particular, let $\mathbb{T} = \{T_1, T_2\} = \{(s_1, s_2) \in \{(\overline{s}, \underline{s}), (\underline{s}, \overline{s})\}\}$, interpreted as choosing which route gets traffic management (e.g., traffic light) priority, where T_i prioritizes route R_i. Equilibrium time savings on both routes must be equal, i.e.,

$$s_1 - \alpha_1(a) = s_2 - \alpha_2(a),^{10}$$

implying actual gains from a policy change are always zero: $\delta_n(T_i, T_{-i}) = 0$ for $i \in \{1, 2\}$ and all $n \in N$. However, perceived gains are given by:

$$\tilde{\delta}_n(T_i, T_{-i}; \eta_n) = \theta_n(\overline{s} - \underline{s}) + \eta_n \theta_n(\alpha_i(a^*(T_{-i})) - \alpha_i(a^*(T_i))) = \theta_n(1 - \eta_n)(\overline{s} - \underline{s}),$$

[10] To maintain equilibrium uniqueness, we can impose an arbitrary selection criterion for which drivers choose which route.

where the second equality follows from the equilibrium condition. The direct gain from T_i under T_{-i} is the time savings increase $(\overline{s} - \underline{s})$ on the newly prioritized R_i. The equilibrium effect completely offsets the direct gain as drivers re-optimize and congestion on R_i increases until gains are completely competed away: $-(\overline{s} - \underline{s})$. Since $\overline{s} > \underline{s}$, for any $\eta_n < 1$, perceived gains are strictly positive: under T_1, agents perceive benefits from T_2, and under T_2, they perceive gains from T_1, generating a 2-cycle. Intuitively, under T_1, agents look at R_2 and see an emptier road. Neglecting equilibrium, they believe that giving this emptier road traffic-management priority will make it faster for them, neglecting the increase in drivers that will follow.

As the example highlights, EN is not a status quo bias: here, EN generates a "grass is greener" cycle, making a change appear beneficial from the perspective of the current equilibrium. Therefore, while under EN policy preferences are biased *by* the status quo, they are not necessarily biased *towards* it.

Another question that might arise is whether EN-driven distortions on political feasibility are bounded. The following proposition shows that the answer is generally negative: with sufficient neglect, agents may oppose arbitrarily beneficial policy, and may support arbitrarily harmful one.

Proposition 2 (Welfare Reversals)**.** *Fix any feasibility rule f satisfying monotonicity and consensus, and any $K > 0$. There exists an economy $\mathcal{E}$ and $\boldsymbol{\eta} \in \mathbb{R}^N$ where:*

1. *(Infeasible unbounded gains) There exist $T^0, T^+ \in \mathbb{T}$ such that $\delta_n(T^+, T^0) \geq K$ for all $n \in N$, yet we have $f(\tilde{\delta}(T^+, T^0; \boldsymbol{\eta})) = T^0$.*

2. *(Feasible unbounded losses) There exist $T^0, T^- \in \mathbb{T}$ such that $\delta_n(T^-, T^0) \leq -K$ for all $n \in N$, yet we have $f(\tilde{\delta}(T^-, T^0; \boldsymbol{\eta})) = T^-$.*

3.4.2 Spurious Opposition

Propositions 1 and 2 show the existence of settings where EN distorts personal and societal policy rankings. How relevant are these to real-world economic settings? For EN to impose a constraint on feasibility, there needs to be a spurious opposition: a share of agents whose actual gains are positive, but perceived gains are negative. The following proposition characterizes the conditions under which achieving a social outcome with budget-balanced taxation generates spurious opposition.

I focus here on aggregate games with quasi-linear utilities and unidimensional types. While not a necessary condition for a spurious opposition to exist, it is a natural and analytically tractable benchmark that transparently captures the key mechanism.

Let N be a unit continuum. Each agent $n \in N$ has type $\theta \in \mathbb{R}_+$. All agents share the action space $A_n = \{0, 1, \ldots, M\}$. Others' actions affect agents through the aggregator

$\alpha(a) = \int_{n \in N} a_n \, dn$. The policy space is the set of anonymous, unilateral, on-path budget-balanced linear tax schedules, i.e., $T_n(a_n) = -pa_n + p\,\alpha(a^*(p))$. We index a policy by its linear tax p. Utilities are type-dependent, quasi-linear in money, and depend on the aggregator:

$$v_n(a; T) = w(a_n, \alpha(a); \theta_n) + T_n(a_n).$$

I impose the following monotonicity and single-crossing conditions:

Assumption 2 (Monotonicity)**.** *$w(a_n, \alpha)$ is monotonic in a_n and conditionally monotonic in α:*

1. *For any θ and α, $w(a_n, \alpha; \theta)$ is strictly increasing in a_n.*
2. *For any θ and a_n, $w(a_n, \alpha; \theta)$ is weakly decreasing in α, and there exists $h < M$ such that if $a_n > h$ it is strictly decreasing.*

Assumption 3 (Strong Single Crossing)**.** *w has strictly bounded increasing differences in (θ, a_n) and (θ, α); i.e., there exists $\epsilon > 0$ such that:*

1. *For any $a_n > a_n'$, α, and $\theta > \theta'$,*

$$[w(a_n, \alpha; \theta) - w(a_n', \alpha; \theta)] - [w(a_n, \alpha; \theta') - w(a_n', \alpha; \theta')] \geq \epsilon(\theta - \theta').$$

2. *For any $\alpha < \alpha'$, $a_n > h$, and $\theta > \theta'$,*

$$[w(a_n, \alpha'; \theta) - w(a_n, \alpha; \theta)] - [w(a_n, \alpha'; \theta') - w(a_n, \alpha; \theta')] \geq \epsilon(\theta - \theta').$$

Assumption 3 is a standard single-crossing/increasing-differences condition, strengthened so the increase is bounded away from zero, ensuring preferences are sufficiently responsive to types. In particular, its first part ensures that for any tax t, high actions are chosen by some sufficiently high types, and the second part ensures that a first-order increase in type translates to a first-order increase in the WTP for externality reduction. Models with preferences linear in type satisfy this assumption by construction.

The next result shows that, under these assumptions, any budget-balanced linear tax that improves on the status quo generates spurious opposition:

Proposition 3 (Spurious opposition)**.** *Suppose Assumptions 2–3 hold and fix a status quo policy $T^0 = p^0$ with associated equilibrium α^{p^0}. Any budget-balanced linear tax p with associated equilibrium $\alpha^p < \alpha^{p^0}$ generates spurious opposition: there exist θ^{EN} and $\bar{\eta}(\theta) \in (0, 1)$ such that for all agents with $\theta_n > \theta^{EN}$ and all $\eta_n \in [0, \bar{\eta}(\theta)]$,*

$$\delta_n(T, T^0) > 0 \quad \textit{but} \quad \tilde{\delta}_n(T, T^0; \eta_n) < 0.$$

It is also instructive to express this result in terms of its implications to weighted-support based feasibility rules:

Corollary 1. *Let f be a weighted-support rule with thershold β, and assume $\eta_n = \eta$. Under the assumptions of Proposition 3, there exist a nonempty set $B \subset (0, 1]$ and a threshold $\overline{\eta} < 1$ such that for all $\beta \in B$:*

$$f(\tilde{\delta}(T, T^0; \boldsymbol{\eta})) = T \iff \eta > \bar{\eta}$$

Therefore, for some feasibility thresholds, the socially beneficial taxation p is feasible only if neglect is not too large.

The intuition for the result is as follows: budget balanced taxation always creates net payers. When preferences are sufficiently rich, if switching costs and externality exposure lie on the same dimension (or equivalently, if there is a monotonic mapping from costs to externality exposure; this is guranteed by Assumption 3), then, despite being net payers, high types' actual gains remain positive. Provided their EN parameter is sufficiently low, this generates spurious opposition.

Proposition 3 delivers three main takeaways on the prevalence of EN, on the types it affects, and on the economic conditions for it to constrain feasibility. First, it shows that the constraint EN imposes on feasibility extends beyond the contrived examples discussed earlier in this section, and can arise naturally in common economic settings.

Second, the proposition identifies high types as the set of agents most likely to spuriously oppose a policy—typically interpreted as those with higher skill or wealth. To the extent that skill and wealth correlate with greater influence weight, this makes spurious opposition more likely to prove pivotal. Furthermore, if perceived elasticity η is partly determined by the agent's own elasticity, higher types, whose switching costs are higher, are expected to exhibit lower η, which may exacerbate the constraint still further.

Last, the proposition provides an intuitive sufficient condition on the economic settings where EN is more likely to constrain feasibility: spurious opposition emerges when high switching costs—generating negative perceived gains—coincide with large equilibrium benefits—generating positive actual gains. Put differently, settings particularly prone to spurious oppositions are those where high usage is associated with greater exposure to equilibrium effects.

For example, EN is unlikely to pose a feasibility challenge when a few agents (e.g., factory owners) value a harmful activity (e.g., polluting a river), but the benefit from its reduction is concentrated among a different group (e.g., downstream residents). Here, while taxing pollution may generate opposition from factory owners, it need not be spurious, but driven by real economic interest.

Examples where EN is likely to constrain feasibility include dynamic electricity pricing, where high-usage consumers face the steepest rate increases but gain most from

grid stability; government-provided deposit insurance financed by a risk-based fee, where banks with the riskiest portfolios pay the highest premiums but, as the institutions most vulnerable to bank runs, also gain the most from preventing them; and, finally, as my running example illustrates, in the context of congestion pricing, where those who place the highest value on driving are also those who place the highest value on the shorter commute times that result from congestion pricing.

Example (Politically Infeasible Congestion Pricing.) From this point on, I fix I and suppress it from the notation. First of all, notice that since value of time is linear in type, the example satisfies Assumptions 2 (with $h = 0$) and 3 by construction.[11]

As is common in settings with externality when $p = 0$, congestion is inefficiently high: the threshold for driving in the socially optimal allocation is $\theta^* > \theta^0$.

Result 2. *Without tolls ($p = 0$), the equilibrium driving rate exceeds the social optimum:* $\alpha^0 > \alpha^*$.

Going back to Pigou (1920) and later Vickrey (1969), the canonical solution for this inefficiency is congestion pricing. A budget-balanced congestion pricing is a toll that implements the socially optimum allocation, and a lump-sum rebate k redistributing the revenue. In the proof of Result 2, I derive the optimal toll as

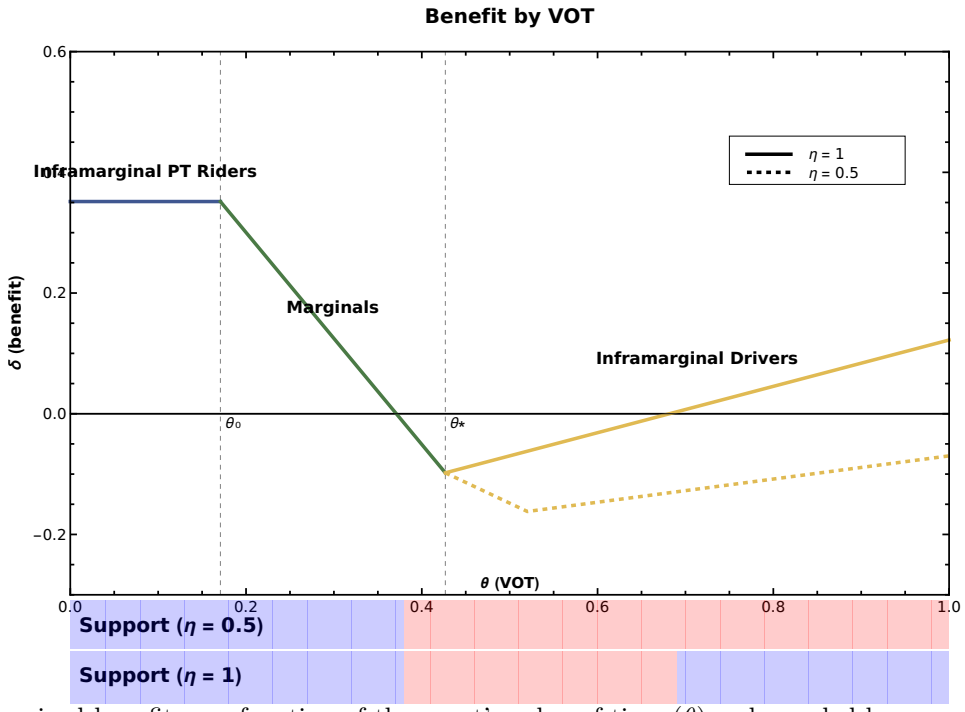
$$p^* = -Q'(1 - F(\theta^*)) \cdot \mathbb{E}[\theta|\theta \geq \theta^*] \cdot (1 - F(\theta^*)),$$

where θ^* is the driving cutoff at the optimum. As is typical in these settings, the optimal toll is the externality: the effect of adding an additional driver factored by the mean value of time among drivers and the size of the driving population. The budget-balanced transfer is therefore $\alpha^* p$. Let $Q(\alpha^0) = Q^0$ and $Q(\alpha^*) = Q^*$.

Who are the actual winners and losers from congestion pricing? The solid line in Figure 1 plots gains from congestion pricing as a function of type. Three classes of commuters emerge: the leftmost flat segment in blue contains infra-marginal public-transit riders ($\theta \leq \theta^0$); agents in this group pay no toll and do not adjust their behavior and their gain is simply the flat transfer $p^*\alpha^*$. The middle, downward-sloping segment in green contains the marginals ($\theta \in (\theta^0, \theta^*]$)—those who switch to public transit in response to the toll. The gain for these commuters is $p\alpha^* - \theta Q^*$—the size of the transfer minus the switching cost, which increases with their value of time. Finally, the rightmost, upward-sloping segment in yellow represents infra-marginal drivers who pay the toll and continue driving ($\theta > \theta^*$). For this group, the gain is $\theta(Q^* - Q^0) - p^*(1 - \alpha^*)$—the increase in time savings scaled by their value of time minus the net toll after the flat transfer, which is always negative under budget-balanced policy.

[11] Strictly speaking, for the example to be a special case of the setting of Proposition 3, b would need to be a part of the status-quo policy (i.e., $p^0 = b$.)

Figure 1: Congestion Pricing: Perceived vs. Actual Benefit



Note: Perceived benefit as a function of the agent's value of time (θ), color coded by group (marginal vs. inframarginal). The solid line represent perfect foresight ($\eta = 1$), and the dashed line represents EN ($\eta = 0.5$). The vertical dashed lines mark the cutoff values θ_0 (between inframarginal PT riders and marginals) and θ^* (between marginals and inframarginal drivers). Solid bars at the bottom indicate positive (blue) and negative (red) perceived benefit (i.e., support./opposition) under perfect foresight (bottom row) and EN (top row). The status quo has no toll and no transfer. The proposed policy is a budget-balanced congestion pricing with a toll p^* implementing the social optimum (α^*) and rebates equilibrium toll revenue ($p^*\alpha^*$) as a flat transfer. Parameters: $Q(\alpha) = 1.5\alpha$, $b = 0.3$ and $\theta_n \sim U(0, 1)$.

The figure shows that real gains from congestion pricing are typically non-monotonic in θ: for low-θ agents (all infra-marginal PT riders and lower-θ marginals) driving has low enough value to be fully offset by the transfer. Medium-θ agents lose: for marginals, driving is valuable enough that the transfer cannot offset the switching cost, while for lower-θ infra-marginal drivers, the value of time savings is too small to cover the toll. Finally, agents with high θ value time savings enough to come out ahead after paying the toll (net of the transfer).

For *perceived gains*, the dashed curve in Figure 1 shows perceived gains for $\eta = 0.5$. Marginal PT riders and infra-marginals do not drive in equilibrium, and therefore their gains are independent of congestion, making perceived and actual gains coincide. For infra-marginal drivers, perceived benefits are:

$$\tilde{\delta}_n(\tilde{T}, T^0; \eta_n) = \underbrace{p^*(\alpha^* - \hat{a}_n)}_{T^*(\hat{a}_n)} - \underbrace{\big(1 - \hat{a}_n\big)(\theta_n Q^0 - b)}_{w_n(\hat{a}_n, a^0_{-n}) - w_n(a^0)} + \eta_n \big[\underbrace{\theta_n(Q^* - \hat{a}_n Q^0)}_{w_n(a^*) - w_n(a^0)} - \underbrace{(1 - \hat{a}_n)(p^* + b)}_{T^*(a^*_n) - T^*(\hat{a}_n)} \big],$$

where $\hat{a}_n = \mathbf{1}_{\theta_n Q^0 \geq p^* + b}$ is the optimal action under the new toll and the old congestion

level. The direct effect is the transfer minus toll or the switching cost (depending on $\hat{a}_n$), and the equilibrium effect is the new time saving net the old time savings or the toll-inclusive driving cost (also depending on $\hat{a}_n$).

The yellow dashed line in Figure 1 shows, how the attenuation of congestion-reduction effect initially makes infra-marginal drivers perceive themselves as potential marginals ($\hat{a}_n = 0$) causing perceived benefits to *decrease* with θ (as for marginals), before increasing again (when $\hat{a}_n = 1$).

The spurious opposition includes all agents to the right of where the solid line meets the x-axis denoted θ^{EN} (from proposition 3)—agents with positive real gains, whose perceived gains are negative under sufficient neglect. For this group, perceived gains are positive iff

$$\eta \geq \frac{p^*(1-\alpha^*)}{\theta_n(Q^* - Q^0)} \equiv \overline{\eta}(\theta_n)$$

i.e., when the equilibrium-neglect parameter is higher than the toll-to-equilibrium-benefit ratio. Consequently, when infra-marginal drivers are pivotal, EN can cause congestion pricing to become politically infeasible. For example, with equal weights, whenever $0 < S(\delta) - \beta < 1 - F(\theta^{EN})$, congestion pricing is feasible if and only if η is sufficiently high.

3.5 Remedy

With standard tools—anonymous, unilateral and budget-balanced taxation—the political feasibility of any constrained social optimum may be constrained by EN. Can alternative tools help overcome these constraints? For the remainder of this section, I focus on economies with quasi linear utilities where the policy space is the set of transfer schemes: $\mathbb{T} \subseteq (A \to \mathbb{R})^{\mathbb{N}}$, where $v_n(a; T_n) = w_n(a) + T_n(a)$. A policy's *deficit at* a is defined as $D_T(a) = \sum_{n \in N} T_n(a)$, and the *equilibrium deficit* as $D_T^* = \sum_{n \in N} T_n(a^*(T)$.

Our objective is to construct a politically feasible policy that implements a under fiscal constraints. Formally, given a status quo T^0, a feasibility rule f, a target action profile $a \in A$ and a maximal deficit level $\overline{D}$, our goal is to find a transfer scheme T that satisfies the following criteria:

1. *Outcome implementation:* $a^*(T) = a$.
2. *Budget constraint:* $D_T^* \leq \overline{D}$.
3. *Political feasibility:* $f(\tilde{\delta}(T, T^0; \boldsymbol{\eta})) = T$.

Notice that achieving these goals may be impossible for reasons entirely unrelated to EN. Indeed, there is a substantial body of work on the challenges in achieving each criterion: mechanism design, public finance and social choice theory are dedicated to the study of implementation, fiscal constraints, and preference aggregation—presenting

a rich set of challenges even absent any cognitive constraints. To isolate the problem, it is instructive to start with a policy that achieves some constrained optimum a under the budget constraint, while being politically feasible under perfect foresight, i.e., $f(\delta(T, T^0)) = T$.

Consider a policy T with associated equilibrium a^T and deficit D. A spurious opposition arises when $v_n((\hat{a}_n, a^0_{-n}); T)$, i.e., the agent's utility from T holding others' behavior fixed at a^0, is significantly lower than $v_n(a^T; T)$, i.e., the agents actual utility under T, is large. We therefore seek a modified policy $\tilde{T}$ that closes this gap while inducing the same equilibrium and satisfying the same budget constraint.

As it turns out, a multilateral policy $\tilde{T}$ that supplements T with a compensation k contingent on a_{-n}, which pays whenever $a_{-n} = a^0_{-n}$, can achieve exactly these three goals. To see why, observe that this compensation depends only on others' actions, and therefore with quasi-linear utilities has no effect on each agent's choice of a_n. It must therefore result in the same equilibrium as T. This also implies that the two policies have the same deficit, as the compensation is never paid in equilibrium. Finally, since $v_n((\hat{a}_n, a^0_{-n}); \tilde{T}) = v_n((\hat{a}_n, a^0_{-n}); T) + k$, a sufficiently large k can close the gap between $v_n((\hat{a}_n, a^0_{-n}); T)$ and $v_n(a^T; T)$.

The theorem below formalizes this intuition. I call two policies that induce the same Nash equilibrium *outcome equivalent*, and two policies that have the same equilibrium deficit *deficit equivalent*. Finally, we say that action profiles a and a' are *multilaterally distinct* if they differ for more than one player, i.e., $|\{n \in N : a_n \neq a'_n\}| \geq 2$ (when N is a continuum, replace "more than one player" with "a set of positive measure").[12]

Theorem 1. *Let T^0 be a status quo policy, let f be a monotonic feasibility rule, and let T^* be an anonymous unilateral policy. Suppose that $a^*(T^*)$ and $a^*(T^0)$ are multilaterally distinct. If T^* is politically feasible under T^0 with respect to f when $\boldsymbol{\eta} = \mathbf{1}$, then there exists a multilateral policy $\tilde{T}$ that is outcome and deficit equivalent to T^*, and is politically feasible with respect to f for all $\boldsymbol{\eta} \in [0,1]^N$.*

We call a policy *EN-robust* if it is politically feasible with respect to f for all $\boldsymbol{\eta} \in [0,1]^N$. When discussing policies that implement constrained optima, "EN-robust" refers specifically to those that also satisfy the outcome and deficit requirements.

Theorem 1 shows that provided that a policy is politically feasible without equilibrium neglect, we can construct its EN-robust counterpart that maintains feasibility even in the presence of equilibrium-neglecting voters. It imposes no structure on heterogeneity, requires no knowledge about the player other than what *others* did in the status quo,[13] and

[12] Requiring the two profiles are multilaterally distinct ensures $a^*_{-n} \neq a^0_{-n}$ for all $n \in N$.

[13] Strictly speaking, in non-symmetric games $\tilde{T}$ may not anonymous since, while the payment is the same for all players, the quantity it is conditioned on, a^0_{-n}, may differ across players. However, the same result holds approximately for an anonymous $\tilde{T}$ when there are sufficiently many players (and exactly when N is a continuum) by conditioning the payment on the distribution of others' actions or on an aggregate quantity.

applies quite generally: given any monotonic feasibility rule, starting from any status quo, and for any constrained optimum achieved via transfers. Crucially, the policy generates the same equilibrium, without increasing the deficit on the equilibrium path.

While contingent compensations maintain the same deficit in equilibrium, they can increase it off the equilibrium path. A potential worry is that policies promising large contingent compensations may be met with skepticism by voters—especially equilibrium-neglecting voters, who believe these will actually be triggered. Since the goal is precisely to persuade such neglectful voters, off-path deficit may counteract the persuasive effect of the compensation. Moreover, policymakers, who may themselves suffer from equilibrium neglect, may reject policies with large off-path compensations to avoid potential large deficits. This raises a natural question: can contingent compensations be constructed such that the deficit is constrained not just in equilibrium, but at every action profile?

The theorem below shows that $\tilde{T}$ can be constructed so that it maintains off-path budget constraint in the following sense: at the status quo profile, the deficit equals the status quo deficit and at any other action profile, it equals the deficit of the new policy. However, constructing such policies requires additional information: the compensation size must be conditioned on each player's status quo action.

Formally, let T^0 be a status quo policy with associated equilibrium $a^0 = a^*(T^0)$. A policy T is past-action based (PA-based) if transfers to each player depend not only on current actions but also on that player's action under the status quo. Formally, $T_n(a) = T'_n(a; a^0_n)$ where a^0_n is player n's action under T^0 treated as a fixed parameter.[14]

The following theorem says that we can construct EN-robust PA-based $\tilde{T}$ that has the same deficit as T^0 at a^0, and as T^* at all other profiles:

Theorem 2. *Let T^0 be a status quo policy, let f be a weighted-support based rule, and let T^* be an anonymous unilateral policy. Suppose that $a^*(T^*)$ and $a^*(T^0)$ are multilaterally distinct. If T^* is politically feasible under T^0 with respect to f when $\boldsymbol{\eta} = \mathbf{1}$, then there exists a PA-based multilateral policy $\tilde{T}$ that is outcome and deficit equivalent to T^*, and is politically feasible with respect to f for all $\boldsymbol{\eta} \in [0,1]^N$. Furthermore, the deficit under $\tilde{T}$ equals that of the status quo at a^0 and that of T^* otherwise: $D_{\tilde{T}}(a^0) = D_{T^0}(a^0)$ and $D_{\tilde{T}}(a) = D_{T^*}(a)$ for all $a \in A \setminus a^0$.*

To maintain the status quo deficit at the status quo profile, the compensation must depend on past actions and is given by

$$T^0_n(a^0) - T^*_n(a^0_n), \tag{2}$$

[14]In relation to the discussion in footnote 13, when N is a continuum, a^0_{-n} is the same for all players, and we can write $T_n(a) = T'(a; a^0_n)$, i.e., the same function for all players given their status quo behavior. With finitely many players, an anonymous PA-based policy $T'_n(a; a^0_n)$ achieves deficit equivalence only approximately.

the difference between what agent n pays under the status quo and what they would pay under the new policy if they maintained their status quo behavior. Looking at 2, it is straightforward see that summing over all players, we get

$$D_{\tilde{T}}(a^0) = D_{T^*}(a^0) + D_{T^0}(a^0) - D_{T^*}(a^0) = D^*_{T^0}.$$

Crucially, this expression requires only knowledge of the agent's past actions.[15] While seemingly challenging in practice, past actions can often be well proxied by readily available observables. For example, in my empirical application I show that workplace location can be used to predict driving frequency.

The intuition has two main parts. First, when the population is non-responsive to taxation, revenue is higher, funding the off-path compensation. Second, knowing an agent's status quo action provides an upper bound on their direct harm, enabling more precise targeting of potential members of the spurious opposition. In congestion pricing, for example, high congestion generates more toll revenue, and an agent who drives in the status quo is harmed at most by the toll size (an agent taking public transit isn't harmed at all). The policy thus adds compensation that targets those directly harmed, funded by the revenue surplus, precisely offsetting the direct effect which reproduces the status-quo deficit. As an immediate implication, when policy T^* is not intended to raise revenue (i.e., $D(T^*(a^*)) \leq D(T^0(a^0))$), the PA-based $\tilde{T}$ maintains the same budget constraint as T^* in all contingencies.

One implication of past-action based targeting is that when $\eta < 1$, some agents perceive lower benefits under $\tilde{T}$ than under T^*. To see why, observe that the contingent compensation in $\tilde{T}$, given by 2, can be negative. Therefore, while the contingent compensation ensures that all agents with positive actual gains from T^* support $\tilde{T}$ over the status quo, it may lead some of them to support $\tilde{T}$ less intensively than they would support T^*. This is precisely why Theorem 2 holds only for weighted-support based feasibility rules, which depend on the sign of $\tilde{\delta}_n$, but not on its intensity. Consequently, the theory predicts that relative to its standard counterpart, the increase in the support for PA-based EN-robust policy should be concentrated among potential recipients, i.e., those whose past actions imply higher contingent compensation.

Example (Congestion-contingent compensation.) To see how the two types of policies differ, we can apply Theorems 1 and 2 in the congestion model.

Pseudo Rebates Consider the anonymous multilateral congestion pricing policy $\tilde{T}$ that augments T^* with contingent compensation $\tilde{k}(a)$ that pays when congestion remains

[15] In fact, the social planner need not observe the player's actual status quo action: the compensation depends only on what the agent paid previously and what they would pay under the new policy with unchanged behavior, which may be a strictly weaker informational requirement.

high:

$$\tilde{T}(a) = \underbrace{p^*(\alpha^* - a_n)}_{T^*(a_n)} + \underbrace{\mathbf{1}_{\alpha(a)=\alpha^0} p^*}_{\tilde{k}(a)},$$

In addition to the toll p^*, and the flat transfer $k^* = \alpha^* p^*$ (both identical to T^*), this policy compensates all agents p^* when congestion remains high ($\alpha(a) = \alpha^0$), effectively rebating the toll. This rebate completely eliminates the spurious opposition. To see why, recall that only infra-marginal drivers are in the spurious opposition. For these drivers:

$$\begin{aligned}\tilde{\delta}_n(\tilde{T}, T^0; \eta_n) = \\ &\underbrace{p^*(\alpha^* - \hat{a}_n)}_{T^*(\hat{a}_n)} - \underbrace{(1-\hat{a}_n)(\theta_n Q^0 - b)}_{w_n(\hat{a}_n, a^0_{-n}) - w_n(a^0)} + \underbrace{p^*}_{\tilde{k}(a^0)} + \eta_n \big[\underbrace{\theta_n(Q^* - \hat{a}_n Q^0)}_{w_n(a^*) - w_n(a^0)} - \underbrace{(1-\hat{a}_n)(p^* + b)}_{T^*(a^*_n) - T^*(\hat{a}_n)} - \underbrace{p^*}_{\tilde{k}(a^0)} \big] \\ &= \eta \left[\theta_n(Q^* - Q^0) - (1-\alpha^*)p^* \right] + (1-\eta) \left[p^* \alpha^* + (1-\hat{a}_n)(p^* + b - \theta_n Q^0) \right].\end{aligned}$$

The first expression is positive for all drivers with positive real gains, while the second is always positive (since $\hat{a}_n = 0 \iff p^* + b \geq \theta_n Q^0$). Intuitively, the direct harm is bounded by the size of the toll, and the rebate completely offsets it. Notice that the rebate only depends on the aggregate level of congestion, making it a *pseudo rebate* since it is independent of the commuter's own driving choice. Therefore, the effective toll under $\tilde{T}$ remains p^*, the equilibrium remains α^* and the compensation is not triggered in equilibrium. Since T^* is budget-balanced in equilibrium, so is $\tilde{T}$: $D^*_{\tilde{T}} = D^*_{T^*} = 0$. This demonstrates Theorem 1.

However, at a^0, the deficit is

$$D_{\tilde{T}}(a^0) = \underbrace{(1+\alpha^*)p^*}_{\text{transfers}} - \underbrace{\alpha^0 p^*}_{\text{revenue}} > 0,$$

so the policy runs a deficit off equilibrium.[16]

Commuter Compensation Now consider the PA-based multilateral policy $\tilde{T}$:

$$\tilde{T}_n(a) = \underbrace{p^*(\alpha^* - a_n)}_{T^*(a_n)} + \mathbf{1}_{\alpha(a)=\alpha^0} \underbrace{p^*(a^0_n - \alpha^*)}_{T^0(a^0_n) - T^*(a^0_n)},$$

This *commuter compensation* policy compensates status-quo drivers when congestion remains high despite having a toll in place. Following the same derivation, for infra-marginal drivers (all status-quo drivers who receive compensation), if $\delta_n(T^*, T^0) \geq 0$,

[16] Setting $\tilde{k}(a^0) = (1-\alpha^*)p^*$, i.e., making the *total* transfer at a^0 equal the toll p^*, still yields a deficit $p^*(1-\alpha^0) > 0$ which is positive whenever infra-marginal PT riders exist.

then $\tilde{\delta}_n(\tilde{T}, T^0; \eta) \geq 0$ for all $\eta \in [0,1]$.[17] The same outcome- and deficit-equivalence arguments apply. Since α^0 is the share of commuters with $a_n^0 = 1$, the deficit at a^0 is:

$$D_{\tilde{T}}(a^0) = \underbrace{\underbrace{\alpha^* p^*}_{k^*} + \underbrace{(1-\alpha^0)}_{\text{share } a_n^0=0}(0-\alpha^*)p^* + \underbrace{\alpha^0}_{\text{share } a_n^0=1}(1-\alpha^*)p^*}_{\text{transfers}} - \underbrace{\alpha^0 p^*}_{\text{revenue}} = 0.$$

This demonstrates Theorem 2.

3.5.1 Discussion

Transfers can help increase political support. The challenge is in designing a transfer that is both credible to agents and costless to the planner. A large payment contingent on an event all parties deem unlikely (e.g., snow in the Sahara) has little persuasive power, while one contingent on a likely outcome (bad weather day in Boston) may be prohibitively expensive. The mechanism I propose operates precisely on the wedge generated by equilibrium neglect: it identifies the contingency—the persistence of status-quo behavior—that EN agents find plausible (and/or salient), but the planner knows lies out the new equilibrium path. Similar to how a warranty functions, the planner leverages their information to bridge a potential perception friction by offering insurance against the very outcome voters (mistakenly) fear.

It is instructive to compare this approach to more conventional remedies for political opposition, viewed through the lens of congestion pricing. Decreasing the toll below its socially optimal level (as was recently done in New York City), may increase support, but would likely violate both outcome equivalence (by inducing a higher level of driving compared to the social optimum) and generate lower revenue. A planner could use subsidies instead of tolls to achieve the same goal (e.g., by paying drivers to drive outside of peak hours). However, this would violate the budget constraint. An alternative policy, seemingly similar to contingent compensation, is refunding drivers the tolls they paid if congestion remains high. However, this approach effectively rewards driving, which distorts incentives and violates outcome equivalence. Finally, trials and pilot programs, such as the one used to introduce congestion pricing in Stockholm (Schuitema et al., 2010), offer another alternative by making equilibrium benefits tangible and directly observable, allowing direct experience to correct the perceptions that fuel spurious opposition (the next section presents global evidence consistent with such corrections). However, ultimately this approach faces the same fundamental feasibility challenge: if a policy is politically infeasible to implement permanently, it may also be infeasible to trial, as agents who expect to lose from it have no incentive to agree to a temporary version. The

[17] Infra-marginal PT riders and marginals have $\tilde{\delta}_n(\tilde{T}, T^0; \eta) \leq \tilde{\delta}_n(T^*, T^0; \eta)$ with strict inequality for $\eta < 1$. However, $\delta_n(T^*, T^0) \geq 0 \Rightarrow \tilde{\delta}_n(\tilde{T}, T^0; \eta) \geq 0$ still holds, ensuring EN-robustness under any weighted-support feasibility rule.

off-path compensation mechanism avoids this by offering a compensations that undo the payments when the equilibrium prediction does not pan out.

A potential worry is that state-contingent policies are themselves too complex. If agents neglect equilibrium effects due to cognitive constraints, they may also struggle to process the contingent nature of the proposed remedy. This raises an important empirical question: can EN-robust policies increase support in practice? I address this question in section 4.

4 Congestion Pricing: Public Opinion

I test the theory in a policy opinion poll on congestion pricing. As discussed in the introduction, congestion pricing is particularly suitable for studying equilibrium neglect and testing EN-robustness, for a few reasons. First, it has a natural and clean distinction between the direct effect (toll increase) and equilibrium effect (congestion reduction). Second, convergence to the new equilibrium is typically fast, significantly reducing timing-related complications.[18] Third, the main outcome, traffic volume/travel time, has a single dimension, further reducing potential complications in the definition of the action space A. Fourth, adjustment speed and lower dimensionality both make predictions quite reliable. For example, in its final Enviromental Assessment document, the Metropolitan Transportation Authority (MTA) predicted a 13% decrease in daily entrances to the Central Business in Manhattan under the 9$ tolling scenario. As of September 7th, the average weekly decrease is 12%.[19] Last, outcomes can be easily and transparently measured, making contingent payments simpler and easier to credibly commit to. I start by reviewing motivating evidence from past polls.

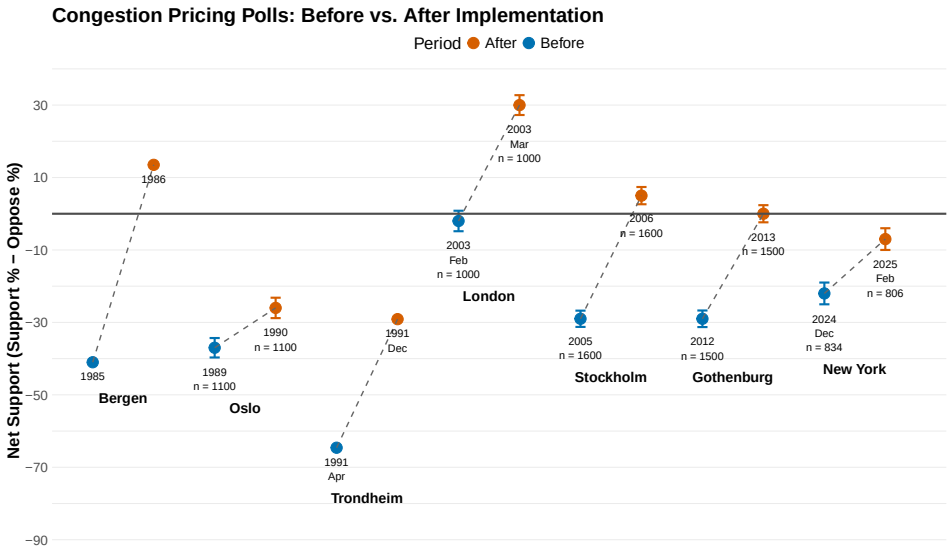
4.1 Motivating Evidence from Existing Polls

As motivation, I start by analyzing existing polls. Figure 2 plots *net support*—the percentage supporting minus the percentage opposing the policy—before versus after implementation of congestion pricing in seven different cities (to the best of my knowledge, these are all the cities for which there is evidence). Support consistently increases post-implementation. To the extent that pre-implementation responses reflect perceived gains and post-implementation responses reflect respondents' actual experience (i.e., actual gains), these results are consistent with equilibrium neglect generating spurious opposition that dissipates post-implementation, once benefits realize.

[18] Theoretically, congestion games have a potential function which guarantees convergence to a Nash equilibrium in pure strategies via best-response dynamics.

[19] For the MTA prediction, see: https://www.mta.info/document/158191. For current data on entries, see: https://metrics.mta.info/?cbdtp/vehiclereductions

Figure 2: Support for Congestion Pricing Across Cities



Note: Support for congestion pricing pre- (blue) and post- (orange) implementation across cities. Y-axis: net support (% support - % oppose). Aggregate data assembled from various publicly available sources. Error bars represent standard errors, and are added when the number of observations was provided.

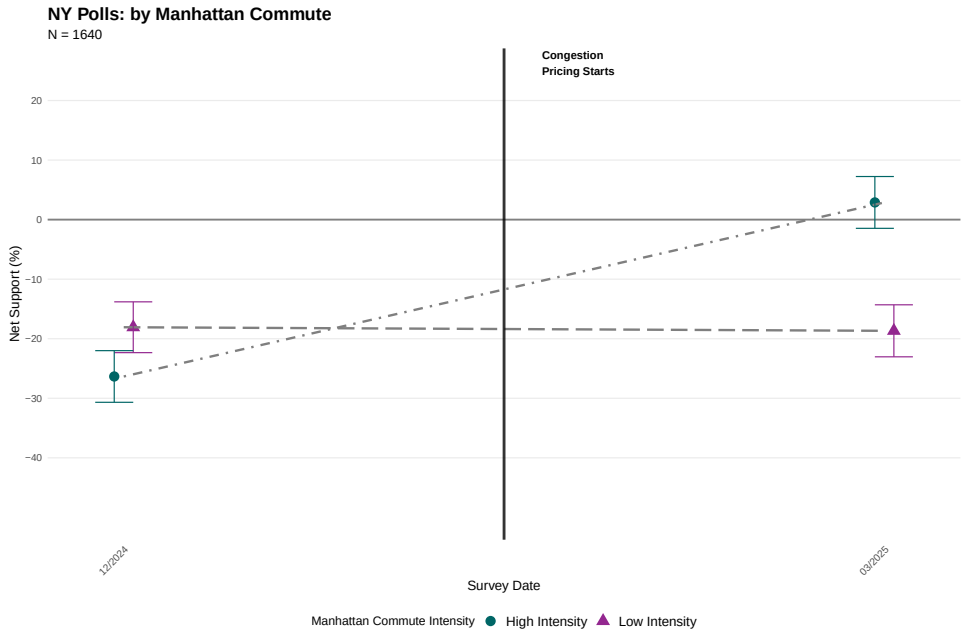
To show this channel more directly, I analyze individual responses from New York ($N = 1,640$), which I obtained from Siena Polls. I analyze two polls: one conducted one month before congestion pricing began in NYC (December 2–5, 2024; $N = 834$) and another about two months after (March 2–6, 2025; $N = 806$). Both polls sampled registered voters from New York state using identical polling methodology. For each respondent, I observe their support response and a rich set of demographics, including county of residence. To the extent that attitude shifts are driven by policy exposure causing belief updating, respondents less likely to commute to Manhattan over the two-month period between implementation and the poll are also less likely to update their beliefs and attitudes. Using ACS county-level traffic flow data, I classify respondents as high-commute (above median Manhattan traffic flow) or low-commute (below median).

Figure 3 plots net support before and after implementation separately for each commute intensity group. It shows that the increase in support in NYC is driven by high-commute respondents. This group started with slightly lower level of support, but increased by 29.8 percentage points more (SE = 8.48) after implementation. The effect remains large when controlling for party-specific trends and a rich set of demographics.[20]

[20]One might worry that differential partisan dynamics could confound this result, particularly since the policy became more salient after President Trump took office in January 2025. Controlling for party-specific trends and a rich set of demographics shows that indeed Republicans became more negative post implementation (-20.5 points; SE = 10.09). However, the differential increase among high-exposure participants becomes only slightly smaller (25.34 points; SE = 8.42) and remains highly statistically significant.

While question wording differs across waves,[21] the differential increase suggests policy exposure plays an important role in driving attitude shifts, consistent with respondents holding overly pessimistic beliefs that change once they experience the gains. In the next section, I provide more direct evidence that over-pessimism shapes opposition using my own survey.

Figure 3: Support for Congestion Pricing in NY by Manhattan Commute Intensity



Note: Support for congestion pricing in NYC among NY registered voters pre- and post-implementation, split by commute intensity to Manhattan. Total $N = 1,640$. Sampling weights applied. High intensity (green circles): respondents residing in counties above median Manhattan commute share. Low intensity (purple triangles): below median. Commute intensity: commuters traveling to Manhattan (NY county) as share of all commuters, based on 2020 ACS commuting flows data. Error bars represent standard errors. Diff-in-diff estimate: 29.83 (SE = 8.48).

4.2 Survey Design

The survey has two main purposes, testing two main components of the theory. The first is descriptive, testing for the presence of equilibrium neglect (in the form of congestion-reduction skepticism) in an ecologically important high-stakes policy setting, and then showing how it maps to policy stances in a representative sample. Second, testing whether

[21]The pre-implementation question reads: "In June, Governor Hochul put a $15 congestion pricing toll plan for Manhattan on hold. Last month, after the election, Governor Hochul reinstituted a $9 congestion pricing plan, set to begin in January. Do you support or oppose the Governor's decision to implement the $9 congestion pricing toll for Manhattan?" (Support/Oppose/In the middle), whereas the post-implementation question reads: "Governor Hochul says that congestion pricing in Manhattan is working, reducing traffic and raising revenue for the MTA, and it should remain. President Trump says that congestion pricing is nothing more than an unfair tax on working commuters, and it should be eliminated. Do you think congestion pricing:" (should be eliminated/should remain/in the middle).

remedy suggested by the theory, contingent compensation, is indeed an effective tool for increasing public support.

It follows a simple two-stage, between-subject design: subjects first indicate support for a standard congestion pricing policy, answer beliefs and habits questions, and then in the second stage indicate support for a randomly assigned modified congestion pricing policy. These two stages correspond to the two goals stated above: directly documenting equilibrium neglect, and then testing the effectiveness of EN robustness. The experimental materials are available in the Online Appendix.

4.2.1 Baseline support

At the beginning of the survey, all respondents are presented with a standard congestion pricing policy in the major downtown area of their metro area. According the policy, modeled after the originally proposed NYC Central Business District Tolling Program, vehicles entering the downtown area during peak hours would be charged a $15 daily toll, and the revenue would be used for transportation improvements.

Respondents are presented with the arguments made by proponents (citing research from other cities suggesting it can cut traffic by about 15–20% and reduce travel times by 10–15 minutes) and opponents (saying that the policy is an unfair burden that will not meaningfully reduce congestion). Adding supporters arguments has two main reasons. First, it is meant to eliminate trivial information or attention frictions which can presumably be corrected by simple informational treatments (rather than the involved policy change I am proposing). Second, it provides respondents with a natural traffic-reduction benchmark which is exploited in the EN-robust condition in the second stage. Opponents' arguments are presented to avoid one-sided presentation, which can bias the results.[22] Respondents indicate their level of support for the policy on a five-point scale: Strongly Support, Somewhat Support, Neither Support nor Oppose, Somewhat Oppose, and Strongly Oppose.

4.2.2 Commute Habits & Beliefs

Respondents then answer questions in two order-randomized blocks. Block 1 contains two questions about commute habits: (i) the location of their workplace (downtown, other part of town, suburbs, rural, and home), and (ii) how frequently they drive into the downtown area (from "Never" to "5 or more times per week"). The first question is meant to identify downtown workers, the recipient population of the contingent compensation

[22] A balanced presentation in policy questions is standard practice (American Association for Public Opinion Research, 2023), meant to prevent framing/experimenter demand effects which have been shown to persist under one-sided presentation of arguments (Schuman and Presser, 1996).

in the EN-robust condition. The second question is used in constructing the perceived benefit index described below.

Block 2 elicits beliefs about the travel time reduction as a result of congestion pricing and the level of certainty in the prediction. The question asked in this block is randomized: half of the respondents are asked to predict how congestion pricing would affect their own travel time ("If congestion pricing were implemented according to the proposal you just saw, how do you think it would affect driving time from your home to downtown [city]?"). Respondents select from nine options organized in 5-minute increments, ranging from "Driving time would decrease by more than 15 minutes" to "Driving time would not change" to 'Driving time would decrease by more than 15 minutes." The other half of the sample is asked to predict the effect congestion pricing had in NYC. This question asks respondents to estimate post-implementation travel times through the Holland Tunnel during rush hour. Respondents are told the pre-toll average commute was 23 minutes (compared to 8 minutes with no traffic), and select from seven 5-minute increments ranging from "10 minutes or less" to "36 minutes or more." All respondents then report their level of certainty (0% to 100%) in their prediction following the same format as (Enke and Graeber, 2023).

4.2.3 Second-Stage Support

All respondents then see a modified congestion pricing policy and are asked to state their support on the same scale. The policy they see is randomly assigned. It is identical to the policy presented at baseline, with one of the following modifications, where the modification is highlighted.

EN-robust (40% of the sample) The policy adds a "Monitoring & Assurance" clause: "Traffic levels would be monitored and published monthly. If monthly traffic decreases by less than 8% (about half the typical 15–20% reduction) all workers who commute to work in the downtown area would receive $120 tax credit for that month, regardless of how they commute." Since understanding the compensation mechanism relies on contingent reasoning—which equilibrium-neglecting respondents may struggle with—the treatment adds an explanation meant to explain the logic and reduce confusion: "This is intended to ensure affected workers either benefit from shorter commute times or receive compensation if traffic does not decline as expected."

Control (40% of the sample) The policy only includes the "monitoring" part without the contingent compensation ("Traffic levels would be monitored and published monthly"). To make the policy descriptions comparable, this treatment also included a similar explanation: "This is intended to provide data about the program's effects on traffic patterns."

Toll Reduction (20% of the sample): The toll is reduced from \$15 to \$9.[23]

Discussion Theorem 2 from the theory section identifies status-quo commuters, those who currently drive into the tolled area, as the recipients of the contingent compensation in the budget balanced EN-robust policies. Since conditioning on past behavior may be impractical and may seem unnatural, the contingent compensation in the EN-robust condition targets downtown workers, an approximation meant to identify those who would be most directly affected by the policy. While not all downtown workers drive, and not all frequent drivers work downtown, this population segment has many regular commuters into the area and represents a natural, clearly defined, administratively feasible group for implementing the compensation.[24]

To avoid knife-edge cases, rather than conditioning the compensation on zero improvement in traffic conditions, I condition it on half the predicted reduction. One may think of this version as an approximate budget-balanced version of the exact budget-balanced version proposed in the theory section.

Comparing 2nd-stage support between treatment and control conditions (controlling for baseline support) isolates the effect of contingent compensation itself, since the control condition absorbs any potential confounding factors including framing effects, repeated elicitation effects, and the effect of increasing perceived information/transparency from monitoring. This comparison therefore provides a lower bound on the overall effect of introducing compensation into an existing congestion pricing policy, which would likely be amplified by these additional forces.

Finally, the toll-reduction condition is meant to serve as a benchmark. In the face of strong opposition, a natural reaction by a politician is to lower the toll to appease the opposition. This condition is therefore included to test off-path compensation against a more conventional adjustment margin. Notice that unlike off-path compensation, starting from any constrained optimal toll, this solution leads to worse equilibrium outcomes *and* lower revenue and/or higher deficit.

Parameter Selection The key parameters are based on the NYC tolling program:

Peak-Hour Tolls The \$15 baseline toll reflects the original proposal for New York City's congestion pricing program. Toll reduction to \$9 was selected based on NYC's

[23]The original toll level, \$15, is striked through, and the new toll level, \$9 is highlighted.

[24]The mean monthly trips for respondents who work downtown ($N = 184$) is 8.825 (median = 6) compared to the rest of the sample ($N = 1863$) whose mean monthly trips is 2.055 (median = 0.5), p-value < 0.000. Not only is workplace location strongly predictive of driving to downtown, conditioning on other demographics adds very explanatory power: regressing monthly trips on workplace location alone yields adjusted $R^2 = 0.178$; adding a rich set of demographics (income, education, employment, political affiliation, marital status, religion etc.) yields adjusted $R^2 = 0.1931$.

revised proposal announced in November 2024, which was adjusted in response to opposition. These provide two focal benchmarks that respondents can evaluate and allow us to compare our contingent compensation approach against a recently policy compromise that was actually implemented.

Projected Traffic Reduction The expected traffic reduction with a $15 toll is taken from the MTA's environmental assessment of the congestion pricing program.[25] The 8% threshold for compensation represents approximately half the projected reduction.

Monthly Compensation The $120 monthly compensation amount was calculated to ensure approximate budget balance when the compensation is triggered. The calculation takes into account average daily revenue from toll; administrative costs and exemptions (estimated at 50% of gross revenue), and the ratio between total commuters and the number of workers who commute to work in the downtown area. Further details are in Online Appendix B.

4.2.4 Second Stage (belief)

After stating support for the modified policy, respondents answer a second question about their congestion reduction beliefs, allowing us to identify a treatment effect on beliefs. Those who answered a question about NYC in the first stage now answer about their own own commute, and those who answer about their own commute at baseline answer about NYC. In other words, each respondent answers both questions (own commute and NYC traffic reduction), in randomized order (before vs. after exposure to treatment).

4.3 Predictions

The model makes three main predictions. First, it predicts a non-trivial amount of equilibrium neglect, manifested in overly pessimistic beliefs about the effect of congestion pricing on commute time. Second, those pessimistic beliefs in turn should be associated with lower support for the policy. The primary measure I will be using is net support. Finally, contingent compensation is predicted to increase public support. Notice that absent of altruistic motives, contingent compensation should only affect the support of potential recipients, i.e., those who work downtown. Therefore, we should expect the treatment effect to be concentrated in this sub-population. As discussed before, the comparison to toll-reduction condition is a more conventional tool that serves as a benchmark. Finally, while the theory does not predict that the treatment will change beliefs, this is a plausible psychological mechanism: Providing off-path compensation may signal confidence in

[25]The June 2024 Environmental Assesment of the MTA for NYC predicted a 17% reduction in traffic volume. See https://www.mta.info/document/142711.

the effectiveness of the policy and therefore may make people more optimistic about its congestion-reducing effect. My design allows for testing of belief shifts as a mechanism for increase in support.

4.4 Results

4.4.1 Survey Administration and Sample

The poll was fielded by NORC's AmeriSpeak® Panel in April 2025.[26] The sample consists of a total of 2,056 adults residing in one the following Metropolitan Statistical Areas: Boston–Cambridge–Newton, MA-NH ($N = 161$); Chicago–Naperville–Elgin, IL-IN ($N = 590$); Los Angeles-Long Beach-Anaheim, CA ($N = 613$); Washington-Arlington-Alexandria, DC-VA-MD-WV ($N = 209$); Dallas-Fort Worth-Arlington, TX ($N = 211$); and San Francisco-Oakland-Berkeley, CA ($N = 272$).[27]

4.4.2 First Stage Support and Beliefs

In this section, I first establish baseline support for congestion pricing, then quantify the degree of equilibrium neglect—specifically, skepticism about travel time reduction—and finally show that such skepticism strongly predicts opposition.

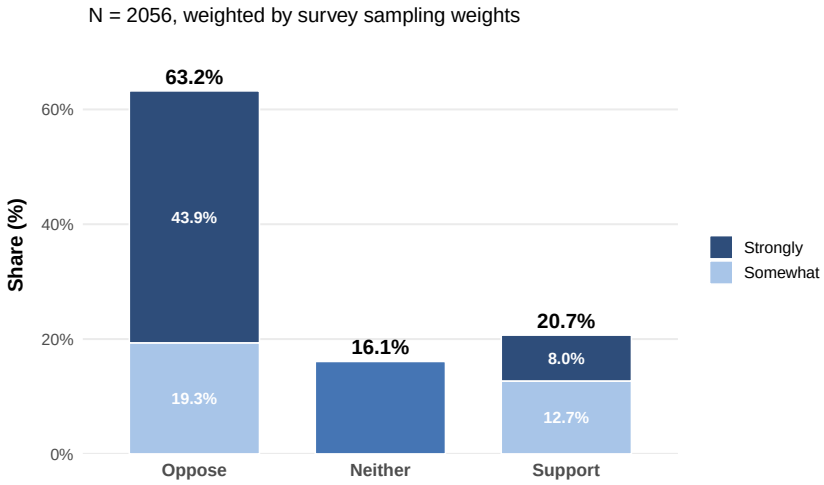
I start with support. Figure 4 plots the share of each support stance, applying sampling weights. An overwhelming majority opposes congestion pricing policy in my sample: 63.2% of the population oppose congestion pricing (43.9% strongly oppose), while only 20.7% support it; the remaining 13% are neutral (choosing "neither support nor oppose"). This translates to a net support of -42.5 percentage points.

From this point on I focus on *net support* as a unified single-number, easy-to-interpret measure of support in the population. Figure 5 plots net support in each of the six metro areas in the sample. It shows that while there is substantial variation, from -57.1% in Dallas to -20.8% in San Francisco, congestion pricing remains largely unpopular across these different regions. To simplify the interpretation, for the rest of the analysis, results are reported without applying sampling weights. Applying these weights does not meaningfully change any of the findings.

[26] AmeriSpeak® is a probability-based, nationally representative household panel operated by NORC at the University of Chicago. Households are randomly selected using area probability and address-based sampling from the NORC National Sample Frame, covering approximately 97% of the U.S. household population, excluding only P.O. Box addresses, some unlisted USPS addresses, and certain new constructions. The weighted AAPOR RR31 response rate for this panel is 26.1%.

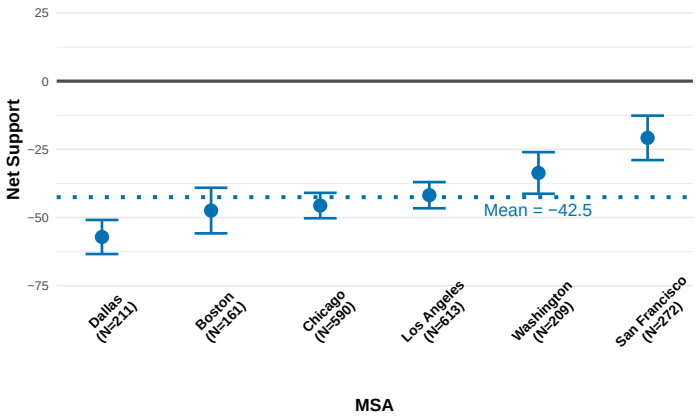
[27] These six MSAs were selected based on NORC's feasibility constraints, and in order to represent reasonably diverse geographic regions across the United States that suffer from high congestion. The 2024 INRIX Global Traffic Scorecard ranks Chicago, Los Angeles, and Boston as 2nd, 3rd, and 4th respectively for U.S. traffic congestion, with Washington DC (9th), San Francisco (15th), and Dallas (20th) also appearing in the top 20.(INRIX and Pishue, 2024).

Figure 4: Distribution of Support Stance



Note: Distribution of support for $15 congestion toll.

Figure 5: Net Support by Metropolitan Statistical Level



Note: Net support across the six Metropolitan Statistical Areas (MSAs) in the sample, weighted by sampling weight. Net support is defined as (%support − %oppose), where each category includes both "strongly" and "somewhat." Error bars: standard errors.

Next, I quantify the degree of equilibrium neglect—in the form of over-pessimism about congestion reduction—in the data by comparing beliefs to an objective measure of time savings. I start with predictions about changes to own commute time. For each zip code z in my sample, I model post-congestion pricing time savings as:

$$\text{time saving}_z = \alpha_z(\text{current rush-hour time}_z - \text{current off-peak time}_z) \tag{3}$$

where $\alpha_z = 1$ represents complete congestion elimination (rush-hour time equals off-peak time) and $\alpha_z = 0$ represents no improvement. This formulation allows natural extrapolation across locations accounting for the fact that tolling can only improve travel times to the extent congestion was slowing traffic initially.

Assuming $\alpha_z = \alpha$ across locations, I estimate α from NYC data by plugging pre-congestion pricing travel times from Moshes and Moshes (2025) into the formula above.[28] Defining rush hours as the tolling period in NYC (5 am – 9 pm) and off-peak as 1 am – 3 am, I obtain $\hat{\alpha} = 0.566$, i.e., on average, 56.6% of rush-hour congestion delay was eliminated. Since the routes leading to downtown in Moshes and Moshes (2025) only include the entrances, extrapolating $\hat{\alpha}$ from this data to my sample implicitly assumes that much of the delay people experience happens in entering the city. If commute times on other parts of the route improve by less, this approach may overstate time savings for people's typical commutes. On the other hand, it also likely *understates* predicted time savings under a \$15 toll, since the toll in NYC is \$9. Either way, it provides a useful benchmark for measuring the degree to which respondents' commute times could improve under congestion pricing.

I estimate time savings for each respondent by using Google Maps Distance Matrix API to calculate commute times from their zip codes to downtown during both rush and off-peak hours. Multiplying the difference by $\hat{\alpha}$ yields *estimated* time savings for each respondent, which I contrast with their *perceived* time savings from survey responses.

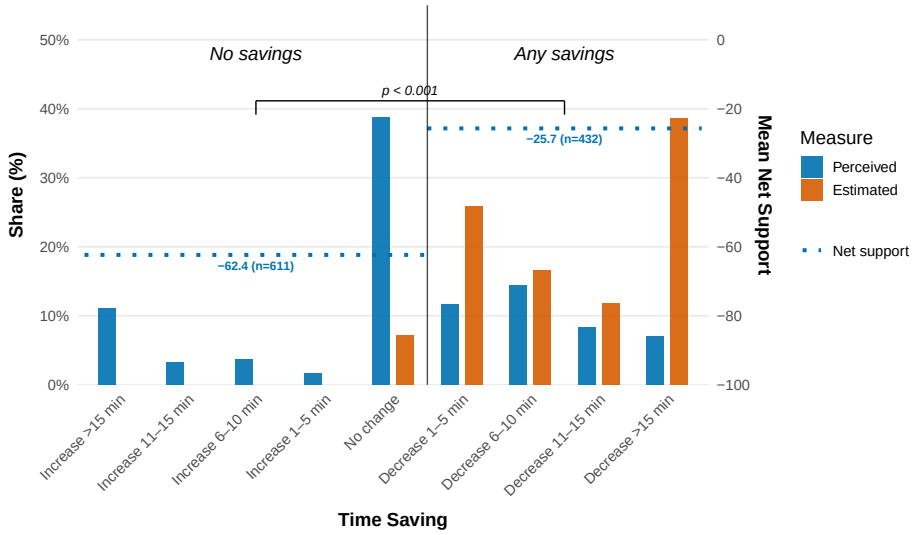
The distribution of calculated time savings (Figure 6; orange bars) shows that most respondents are predicted to experience significant time reductions, with the largest share (39%) estimated to save more than 15 minutes per trip. Only about 7% of respondents are predicted to experience a change in travel time that is less than one minute. In contrast, perceived time savings (blue bars) is consistent with strong equilibrium neglect among a solid majority of respondents: About 39% believe congestion pricing would result in no change at all to travel times, and another 20% believe travel times would *increase*. Less than 30% believe that their travel time would decrease by 6 minutes or more.

Notably, the 39% share of respondents selecting the option "Driving time would not change" is unlikely to be driven by the choice of the scale (15-minute increase to 15-minute decrease) or misunderstanding (e.g., mistaking "increase" for "decrease"). Interestingly,

[28] I am grateful to Benjamin and Joshua Moshes for giving me access to their raw data.

these beliefs appear to reflect substantive views rather than cognitive noise and confusion: limiting the sample to respondents with below-median cognitive uncertainty (i.e., those reporting certainty in their prediction $> 50\%$; $N = 460$), beliefs are *more* pessimistic, with about 51% of respondents predicting no change, and another 20% predicting an increase in travel time.[29]

Figure 6: Estimated Time vs. Perceived Time Saving and Net Support



Note: Perceived vs. estimated time savings from congestion pricing ($N = 1{,}043$). Orange bars show estimated savings, calculated as $\hat{\alpha} \times$ (rush-hour time − off-peak time) with $\hat{\alpha} = 0.566$ extrapolated from Moshes and Moshes (2025); ZIP-code level travel times are obtained from Google Maps Distance Matrix API. Blue bars show perceived savings from survey responses. Dotted lines indicate mean net support by perceived savings group ("No savings" vs. "Any savings"). p-value: test of the difference in mean net support between the two groups.

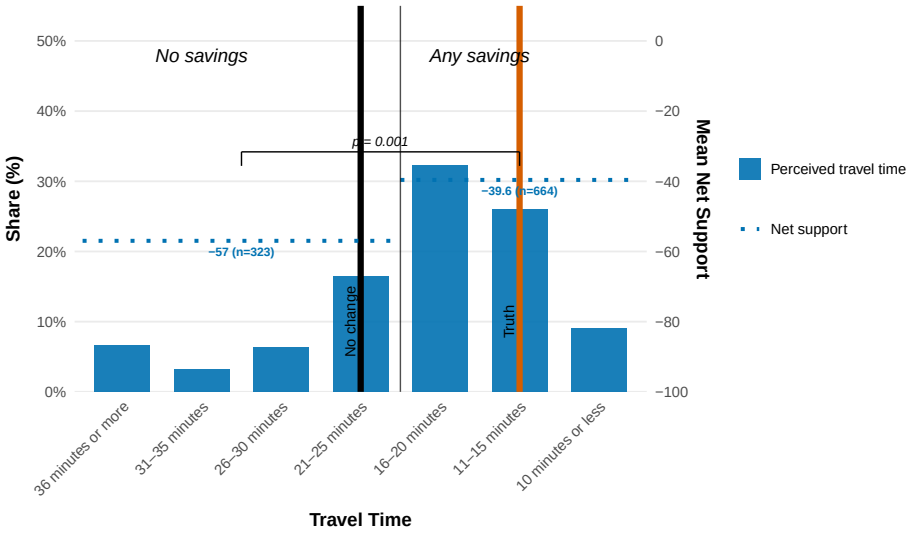
Predictions about one's own time savings are likely the most relevant measure for determining attitudes toward congestion pricing. However, they are subjective, subject to interpretation, and hypothetical. In contrast, beliefs about traffic reduction in NYC have a concrete interpretation and can be easily compared to objective and verifiable ground truth. However, the two main goals of this exercise, namely measuring the prevalence of equilibrium neglect and correlating it with policy support—make using these beliefs a particularly strong test of the theory. First, information about "other cities" is already embedded in the baseline elicitation as part of supporters' arguments ("Supporters cite research from other cities suggesting it can cut traffic by about 15–20% and reduce travel times by 10–15 minutes"), potentially mechanically shrinking the prevalence of equilibrium neglect among respondents. Second, if respondents view NYC as fundamentally different from their own area, this should significantly weaken the correlation between

[29] In the context of the discussion in section 3.2.1, this also suggests that in the context of congestion pricing, equilibrium skepticism, rather than equilibrium naïveté, is the more prevalent form of equilibrium neglect.

their beliefs and their attitude. Both of these make using these beliefs a stronger test of the theory.

Figure 7 plots the distribution of responses to the question about NYC. While less prevalent than own-time predictions, these beliefs still show strong skepticism towards congestion-reducing effects of tolls: while the actual post-congestion pricing travel time through Holland tunnel is around 11 minutes, 65% of respondents estimate it to be 16 minutes or more. Notably, 33% of respondents estimate it at 21 minutes or more, indicating little to no improvement, or even a worsening in traffic conditions compared to the 23 minutes baseline.

Figure 7: Perceived Travel Time in NYC and Net Support



Note: Perceived travel times through the Holland Tunnel under congestion pricing ($N = 987$). Blue bars show the distribution of beliefs about the new travel time. The solid black line ("no change") marks no meaningful change relative to pre-congestion pricing travel time. The solid orange line ("Truth") indicates the estimated travel time based on Moshes and Moshes (2025). Dotted lines indicate mean net support by perceived savings group ("No savings" vs. "Any savings"). p-value: test of the difference in mean net support between the two groups.

Responses to both questions show a meaningful share of respondents under-appreciating equilibrium effects, with a large share estimating them at approximately zero. Do these pessimistic beliefs map to stronger opposition to the policy? Starting with beliefs about own travel time, the dotted lines in figure 6 indicate net support splitting the sample into a group of respondents that believe in any time-saving effect of congestion pricing ($N = 432$), and those who believe in no improvement at all ($N = 611$), showing a 36.7 point difference in net support in this group ($p < 0.001$). As expected, beliefs about NYC are less predictive of support for congestion pricing in one's own city: Figure 7 shows that the difference in net support between respondents indicating a belief in any meaningful improvements ($N = 664$) and those who believe that travel times remained roughly the same or worse ($N = 323$) is 16.4 points ($p = 0.001$).

To map these results back to the theory, I construct a measure of actual and perceived benefit and correlate it with support. The measure converts perceived time savings into dollar terms. I construct a per-trip perceived benefit index for each respondent:

$$\tilde{\delta}_{trip} = \theta_n \times \text{perceived time savings}_n - \$15, \tag{4}$$

where θ_n represents the value of time (VOT), set at 50% of hourly household income following US DOT guidelines Belenky (2011), and perceived time saving is the belief about respondents' own commute.[30] To take into account trip frequency, I construct $\tilde{\delta}_{month}$ which multiplies per-trip benefit by the self-reported number of monthly driving trips.

Figure 8 plots net support against different benefit measures. Panels a and c plot perceived per-trip and monthly benefit. An additional perceived per-trip dollar benefit is associated with a 2.033 points increase in net support (SE = 0.326). At the monthly level, additional \$10 increase in perceived benefit are associated with an increase of 1.3 points in net support (SE = 0.044). Both relationships are highly statistically significant. Panels b and d plot *estimated* benefits, replacing perceived with estimated time savings. They show that, unsurprisingly, estimated benefit measures are essentially uncorrelated with support, with significantly smaller coefficients, both not different from zero. Therefore, living in congestion-prone areas that would typically benefit from the time saving associated with congestion pricing is not associated with higher support congestion pricing.

To summarize, I find substantial equilibrium neglect in the population, reflected in pessimistic beliefs about congestion reduction—both in hypothetical estimates about respondents' own commute time reductions, and in their estimate of its actual effect in NYC. Consistent with my model, these pessimistic beliefs are turn strongly associated with increased opposition to the policy.

4.4.3 Second Stage: Main Results

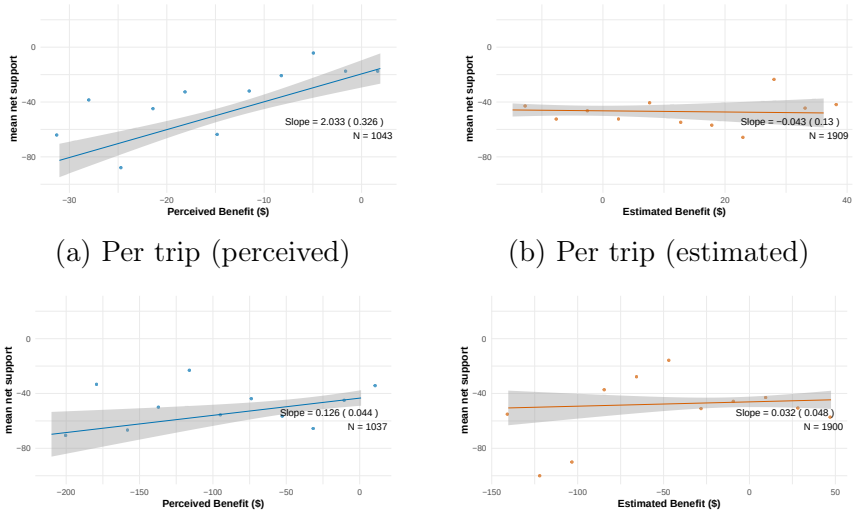
Having shown that pessimism is prevalent and predicts opposition to congestion pricing, I now turn to the remedy: are policies that include contingent compensation indeed robust to equilibrium neglect? We should expect the effect of EN-robustness to be concentrated in downtown workers, who are the potential recipients of the compensation. Other respondents should only respond to the treatment to the extent that they are altruistic towards people who work in the downtown area or have other fairness concerns.

Panel a in Figure 9 compares net support among downtown workers in EN-robust and Control across stages, using a diff-in-diff design.[31] In EN-robust, second-stage net

[30] This index maps to $\delta_n(T, T^0; \eta)$ from the theory section with the subtle difference that rather than plugging in the true effect and discounting by η, I directly plug in a potentially attenuated effect.

[31] My intention was to stratify and block-randomize treatment based on baseline support to ensure balanced baseline support across conditions. Unfortunately, due to technical limitations, the vendor

Figure 8: Benefit and Net Support



(a) Per trip (perceived)

(b) Per trip (estimated)

(c) Monthly (perceived)

(d) Monthly (estimated)

Note: Estimated and perceived benefits from congestion pricing are compared to net support. *Perceived benefit* (panels a and c) is calculated as $\theta_n \times$ perceived time savings – \$15, where θ_n is the value of time, estimated at 50% of hourly household income, and perceived time savings are taken from survey responses. *Estimated benefit* (panels b and d) replaces perceived time savings with estimated time savings (see Figure 4). Monthly benefit (panels c and d) multiplies the relevant benefit by the self-reported number of driving trips to downtown. Bins are equally spaced, and the fitted line is from a linear regression of net support on the relevant benefit measure. Data are winsorized at the 5th and 95th percentiles.

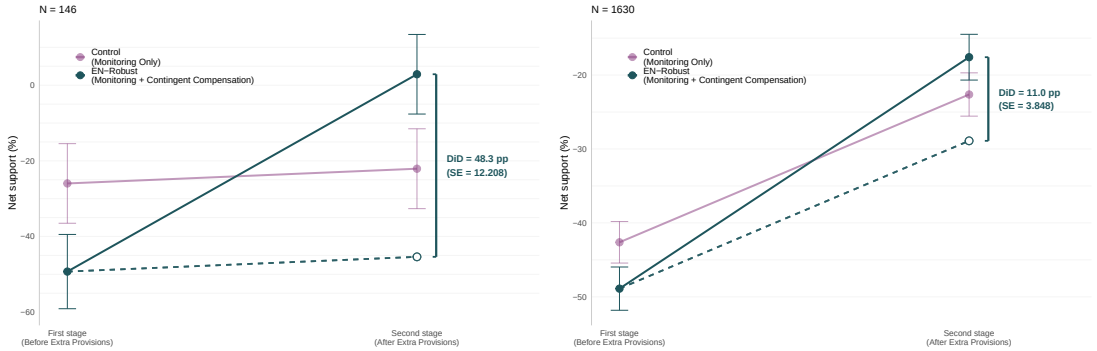
support increases 52.2 points (from −49.3% to +2.9 effectively neutralizing the initially widespread opposition. In Control there is a much smaller 3.9 points increase (from −26 to −22), resulting in a diff-in-diff estimate of 48.3 (SE = 12.208).

Looking at the whole sample (Figure 9, panel (b)), support increases from baseline to second stage in both treatments, but this increase is 11 points higher in EN-robust than in Control (SE = 3.848). The increase in Control likely reflects the effect of monitoring alone combined with framing effects. The first row in Table 1 shows similar results from a regression of second-stage net support on treatment, controlling for baseline support, which remain stable with the inclusion a rich set of controls. Taken together, these results show that contingent compensation indeed increases support for congestion pricing in the population, and, as expected, its effects are concentrated, and are particularly large, within the group it targets.

Toll Reduction To benchmark the effect, I included a treatment reducing the toll from \$15 to \$9. The second row in Table 1 shows that its effects are significantly smaller. It

was unable to implement this stratification. This resulted in an imbalanced treatment allocation where baseline net support in Control is 6.3 points higher than in EN-robust and Toll Reduction. While not statistically implausible (p-value testing difference in means = 0.08), it requires correcting for it when comparing across treatments, either by differencing out baseline support, or by controlling for it (a regression controlling for baseline support was the main specification I pre-registered).

Figure 9: Effect of Treatment on Support for Congestion Pricing



(a) Downtown Workers ("ToT") (b) All Respondents (ATE)

Note: Net support across stages in EN-robust and Control. Panel (a) plots the results for downtown workers only ($N = 146$). Panel (b) reports results for the entire sample ($N = 1630$). Each figure plots mean net support in the first and second stage. The solid teal line represents EN-robust, and the solid purple line represents Control. Error bars indicate standard errors. The dashed teal line shows the counterfactual of EN-robust under parallel trends, based on the observed change in Control. The bracket indicates the difference-in-differences estimate, reported in net support percentage points.

Table 1: Treatment Effects on Net Support

	Downtown Workers			Whole Sample		
	(1)	(2)	(3)	(4)	(5)	(6)
EN-robust	40.908	38.545	42.351	9.244	9.127	10.154
	(11.255)	(11.773)	(14.028)	(3.379)	(3.378)	(3.411)
Toll Reduction	4.783	1.309	3.044	-8.564	-8.210	-6.651
	(13.380)	(13.732)	(15.234)	(4.040)	(4.039)	(4.050)
Baseline Support	✓	✓	✓	✓	✓	✓
Basic Demographics		✓	✓		✓	✓
All Demographics			✓			✓
R^2	0.458	0.473	0.638	0.375	0.380	0.404
Observations	184	184	184	2043	2043	2043

Note: The table reports regressions of net support for the policy on treatment, controlling for baseline support; the omitted category is Control. Columns 1–3 restrict the sample to respondents who work in the downtown area, and are eligible for contingent compensation. Columns 4–6 include the whole sample. "Basic demographics" control for income, sex, age, and education. "All demographics" additionally include ethnicity, metropolitan area, housing type, employment status, marital status, partisanship, ideology, trust in media, religion and religious attendance, household size, and phone ownership. Standard errors are in parentheses.

is positive but statistically indistinguishable from zero among downtown workers, and if anything, has a *negative* effect on the whole population. As treatment effect is measured relative to Control, this suggests that the effect of reducing the toll is similar, or smaller in magnitude to a simple transparency treatment. Ex-post, I speculate that from a theoretical point of view, this finding may not be surprising: an equilibrium-neglecting agent still perceives negative benefits from a $9 toll, and should still oppose it. At the same time, agents with perfect foresight should only change their mind if they believe the reduction is toll is worth more than the potential reduction in time savings it will cause. Therefore, the overall effect of a toll reduction is ambiguous.

4.4.4 Additional Analyses

Differential Effects by Response Time and Prior Beliefs Apparent opinion shifts in polls may be driven by methodological artifacts rather than genuine reaction to the substance. Finding that contingent compensation increases support especially among its intended recipients is consistent with treatment effects that are indeed driven by reaction to the actual content of the policy rather than other survey design artifacts. Table 2 presents two additional pieces of evidence supporting this interpretation.

The treatment effect on respondents who spend above the median time (31 seconds) on the second-stage support question is 17.24–18.18 points higher (SE = 6.94–6.99) than those below the median. The main treatment effect (i.e., the effect on the below-median group) is statistically indistinguishable from zero. This suggests that the contingent compensation mechanism affects only respondents who spent enough time to understand its content.[32]

Second, consistent with the theory, the treatment effect is 14.41–17.85 (SE = 8.37–8.53) points higher among respondents whose first-stage beliefs indicate no perceived improvement in travel time—those who predict no change or worsening conditions on their own commute, or who estimated no improvement through the Holland Tunnel when asked about NYC congestion pricing—compared to those who believe in some improvement. Also consistent with the theory, the treatment effect on this latter group is indistinguishable from zero. Note that this specification does not control for baseline support: doing so would partial out and absorb the component of elicited beliefs that drives attitudes, leaving only residual, attitude-irrelevant part of the variation in beliefs (e.g., noise and measurement error).

[32]Splitting the sample by whether respondents spent more than 18 seconds on the question—the minimal time required for average human reading speed (238 words/minute) to process the 71 words in the added EN-robust clause—yields similar results: interaction effects of 18.19–18.6 (SE = 8.24–8.27) and main effects indistinguishable from zero (-1.08 to -5.63; SE = 6.94–7.31). In the absence of direct attention checks, one can also think of these (non-preregistered) specifications as a form of screening less attentive respondents.

Table 2: Treatment Effects on Net Support, by Beliefs and Duration

	By Skepticism			By Duration		
	(1)	(2)	(3)	(4)	(5)	(6)
A. Main Effects						
EN-robust	-2.891	-2.583	1.080	-1.423	-1.590	-0.382
	(5.808)	(5.787)	(5.678)	(5.139)	(5.142)	(5.170)
Toll Reduction	-11.872	-10.923	-8.298	-9.877	-9.252	-8.010
	(6.831)	(6.811)	(6.649)	(5.100)	(5.103)	(5.118)
B. Interaction Effects						
EN-robust x No Perceived Savings	17.849	17.356	14.407			
	(8.532)	(8.509)	(8.373)			
Toll Reduction x No Perceived Savings	-3.915	-5.381	-1.943			
	(10.216)	(10.185)	(9.989)			
EN-robust x Above Median Duration				18.176	17.669	17.238
				(6.937)	(6.958)	(6.991)
Toll Reduction x Above Median Duration				2.310	2.215	3.502
				(8.416)	(8.411)	(8.430)
Baseline Support				✓	✓	✓
Basic Demographics		✓	✓		✓	✓
All Demographics			✓			✓
R^2	0.023	0.032	0.118	0.378	0.380	0.406
Observations	2023	2023	2023	2043	2043	2043

Note: The table reports regressions of net support for the policy on treatment, controlling for baseline support; the omitted category is Control. Columns 1–3 interact treatment effects with an indicator for congestion-reduction skepticism, defined as reporting no perceived time savings either in one's own commute or in New York City generally. Columns 4–6 interact treatment effects with an indicator for whether the respondent's response time on the second-stage support question was above or below the median of 31 seconds. "Basic demographics" control for income, sex, age, and education. "All demographics" additionally include ethnicity, metropolitan area, housing type, employment status, marital status, partisanship, ideology, trust in media, religion and religious attendance, household size, and phone ownership. Standard errors are in parentheses.

Treatment Effect on Beliefs While the model makes no clear prediction about treatment effects on beliefs, contingent compensation might plausibly work through a signaling mechanism—policymakers committing to large payments if traffic does not fall could increase citizens' confidence in the equilibrium traffic-reduction prediction. Finding that effects are concentrated among targeted recipients suggests this is unlikely the case. Indeed, regressing second-stage beliefs—either about the respondent's own commute or about effects in NYC—on treatment yields small coefficients that are all statistically indistinguishable from zero (-0.69 to -0.02; SE = 0.59–0.61).

5 Congestion Pricing: Policymakers

The previous section documents equilibrium neglect among the general public in the context of congestion pricing, showing that it is prevalent and predicts policy opposition, while EN-robust policies with compensation mechanisms gain higher support. How do these findings translate to political feasibility? In the absence of referenda, political feasibility is ultimately driven by the willingness of policymakers—both civil servants and elected officials—to engage with it and implement it. This willingness may depend on their own assessment of policy effectiveness and their perception of public support.

This raises several questions: Are policymakers themselves prone to equilibrium neglect? Do they perceive the public to be EN? How do these beliefs map to their stance on congestion pricing? This section presents results from surveys eliciting both first- and second-order beliefs among two groups of policymakers: US state legislators and civil servants from the US, UK, and Australia.

5.1 Sample

State Legislators To survey politicians, I rented a booth at the 2025 National Convention of State Legislators Summit and invited attendees to participate in a short in-person survey. State legislators had a special badge colors, allowing me to distinguish them from staff and other attendees. Only those with legislator badges were invited to participate. 130 participants started the survey, yielding 117-118 responses to relevant questions after attrition. Participants were offered compensation ($5-25 Starbucks gift cards, randomly drawn) with the option to donate it to a food bank in their state.

Civil Servants I surveyed 51 federal-level civil servants from the US, UK, and Australia recruited through the Policymakers Lab. Participants received $15 Amazon gift cards.

5.2 Design

The state legislators survey was conducted in collaboration with another researcher studying housing development support. The order of the parts of the survey was randomized. Both the state legislators and civil servants surveys included additional elicitation of EN and perceived public EN on other topics, to be discussed in future iterations of this paper. Here I focus on congestion pricing questions that paralleled the opinion poll.

All participants first see a congestion pricing policy identical to Control from the public opinion poll. Since the relevant measure from a policymaker's perspective is the effect on the overall population, rather than asking about their own commute time, they are asked to predict the change in commute time for a driver who currently spends 30 minutes on his commute every morning, compared to the 20 minutes this same route takes without congestion. The possible answers had a similar symmetric structure as the opinion poll: more than 10 minute decrease, 3–10 minute decrease, no meaningful change, 3–10 minute increase, more than 10 minute increase. Among the possible answers, 3–10 is the only economically reasonable one, with the rest representing an under- or over-appreciation of the time-saving effects of congestion pricing.[33]

Participants are then asked to predict the share of the population from the general public who would mention any time reduction when asked to predict the consequences of the policy. This question captures policymakers' perceived level of equilibrium neglect among the public, focusing on what I view as the most ecologically relevant aspect: whether the equilibrium effect is top of mind for citizens when faced with such a policy. Next, they indicate their own level of support[34] followed by a question asking them to predict the level of support among the public. Finally, all participants then see the EN-robust version of the policy, with the added contingent compensation highlighted, and indicate again their own level of support and then estimate the level of public support for that version of the policy.

5.3 Results

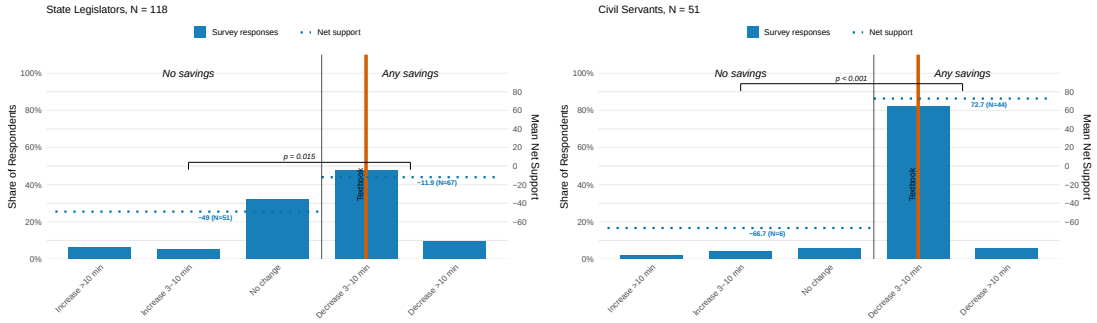
Beliefs I start with first-order beliefs. State legislators exhibit an under-appreciation of time savings from congestion pricing similar to that of the general public. As shown in Figure 10, 43% predict no improvement in commute times. This suggests that EN

[33]Predicting a decrease larger than 10 minutes likely over-estimates the reduction, as the question mentions 20 minutes as the free-flow travel time. I interpret choosing "no meaningful change" as representing under-appreciation of equilibrium effects. While it could also reflect a genuine belief in a meaningful time reduction smaller than 3 minutes, results cited above from NYC suggest this would likely underestimate the effect on travel times, given NYC implemented a 40% lower toll and say 60% improvement relative to free-flow traffic.

[34]To emphasize that the question asks about their attitude in their capacity as policymakers, the question reads: "How much would you recommend or support a similar policy in your current role?" with the same five-point scale as in the opinion poll.

may extend from voters to elected officials and political competition does not necessarily select for individuals with a better grasp of equilibrium effects. In contrast, only 12% of civil servants predict no improvement; a large majority correctly anticipate time savings, with the largest share (82%) choosing the "textbook" 3-10 minute decrease. These beliefs strongly predict support: among state legislators, net support is 37 points higher for those who predict time savings than for those who do not. The difference is larger among civil servants, at 140 points, though this should be interpreted with caution, as the no-time-savings group in this sample contains only six respondents.

Figure 10: Policymakers' Perceived Time Savings



(a) State Legislators (b) Civil Servants

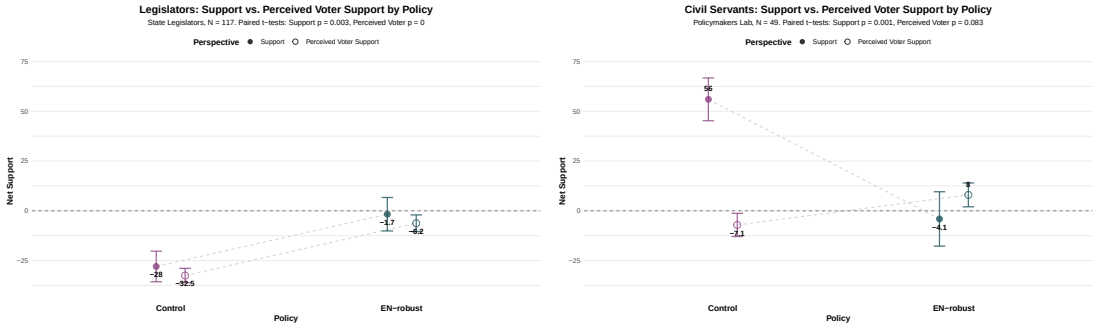
Note: Perceived time savings from congestion pricing for a hypothetical driver. The baseline time in the question is 30, and the free-flow time is 20 minutes. Blue bars show the distribution of beliefs about time savings. The solid orange line ("Textbook") indicates meaningful traffic reduction up to the free-flow speed. Dotted lines indicate mean net support by perceived savings group ("No savings" vs. "Any savings"). p-value: test of the difference in mean net support between the two groups.

Second-order beliefs also strongly predict policy support (see Online Appendix for full regression table). A one percentage point increase in a respondent's estimate of the share of the public who mention time savings is associated with a 0.93 (SE = 0.28) point increase in a state legislator's net support and a 1.76 (SE = 0.47) point increase for a civil servant. These results hold when controlling for the respondent's own beliefs, suggesting that a policymaker's perception of public understanding is an input in their support, independent of their own assessment of the policy.

Support for EN-robust Policies Finally, I look at support for Control and EN-robust congestion pricing, and perceived public support for each. Figure 11 plots net support and perceived public support (converted to net support points) in both groups. State legislators' own support for EN robust is more than 26 points higher than Control, moving from -28 to -1.7. Their perception of public support follows a similar pattern, with estimated public net support rising from -32.5 to -6.2 between Control and EN-robust.

For civil servants, the pattern is more complex. While support for the standard congestion pricing policy (Control) is overwhelmingly high (+56 net support), it drops

Figure 11: Policymakers' Net Support and Perceived Net Support for Control and EN-robust



(a) State Legislators (b) Civil Servants

Note: Net support and perceived public net support for Control and EN-robust. The full circles represent actual net support for each policy, and the hollow circles represent perceived public net support. Purple: Control. Teal: EN-robust. Panel (a) plots results for state legislators ($N = 117$), and panel (b) plots the results for civil servants ($N = 49$).

significantly for the EN-robust version to -4.1. This drop may reflect concerns that the contingent compensation could compromise traffic reduction, increase the deficit, suggesting that budget and outcome equivalence may not be immediately obvious without explicit explanation. It may also reflect worries about logistical complexity. At the same time, civil servants on average seem to predict that public support the EN-robust congestion pricing is higher, with estimated mean public net support increasing 15 points from -7.1 to $+8$. Interestingly, despite their different starting points, both elected officials and civil servants converge on a similar, roughly neutral level of support for the EN-robust congestion pricing policy.

5.4 Discussion

Taken together, these findings can offer two insights. First, they may help explain why contingent compensation mechanisms are not currently common in public policy: those tasked with designing policies—civil servants—seem to prefer conventional tools. This preference, however, may lead them to champion policies whose political feasibility is questionable. Second, the results highlight the potential of EN-robust policies for easing opposition from both elected officials and the public, even if it comes at the cost of less enthusiastic support from administrators. I speculate that explaining outcome and deficit equivalence will likely be effective in convincing civil servants, especially when combined with evidence on increased public support, but this remains an open question for future research.

6 Conclusion

This paper describes a problem and proposes a solution. At its core, the problem is one any economist who has picked up a newspaper will recognize: the bread and butter of the discipline—thinking through the implications of best response—is a mode of reasoning that does not always come naturally to the general public. Similarly, the remedy should seem equally familiar and employs one of the oldest tools in a policymaker's toolkit: if you want to make people happier with a policy, simply offer them more money. The key, however, lies not in *that* transfers are offered, but in *how* they are designed, recognizing that the same deviation that can limit the set of policies we are able to choose from also creates scope for "policy arbitrage" allowing us to generate perceived gains at no fiscal cost and align public support with the policy's actual benefits.

While this paper focuses on congestion pricing, equilibrium neglect is a general phenomenon that likely constrains the political feasibility of a wide range of socially beneficial policies, from carbon taxation to financial regulation. While off-path contingent compensation is designed to be similarly portable, tailoring it to a particular context requires carefully thinking about who the potential spurious opposition is, and what to condition on in order to appease it.

Future research could explore several avenues. Applying and testing this framework in other policy domains is a natural next first step. Second, the disconnect identified between civil servants and elected officials merits further investigation: what makes policy experts prefer conventional tools, and how can they be persuaded of their utility in bridging the gap between optimal policy and public opinion? Finally, further work could explore the boundaries of the remedy and how it interacts with examining how the framing and complexity of contingent payment mechanisms themselves interact with voters' cognitive limitations.

References

Hunt Allcott and Christopher Knittel. Are Consumers Poorly Informed about Fuel Economy? Evidence from Two Experiments. *American Economic Journal: Economic Policy*, 11(1):1–37, February 2019. ISSN 1945-7731. doi: 10.1257/pol.20170019. URL `https://www.aeaweb.org/articles?id=10.1257/pol.20170019`.

Hunt Allcott, Benjamin B Lockwood, and Dmitry Taubinsky. Regressive Sin Taxes, with an Application to the Optimal Soda Tax*. *The Quarterly Journal of Economics*, 134(3):1557–1626, August 2019. ISSN 0033-5533. doi: 10.1093/qje/qjz017. URL `https://doi.org/10.1093/qje/qjz017`.

American Association for Public Opinion Research. Question Wording. Technical report,

American Association for Public Opinion Research, 2023. URL https://aapor.org/wp-content/uploads/2023/01/Question-Wording.pdf.

David Andersson and Jonas Nässén. The Gothenburg congestion charge scheme: A pre–post analysis of commuting behavior and travel satisfaction. *Journal of Transport Geography*, 52:82–89, April 2016. ISSN 0966-6923. doi: 10.1016/j.jtrangeo.2016.02.014. URL https://www.sciencedirect.com/science/article/pii/S0966692316000430.

Peter Andre, Carlo Pizzinelli, Christopher Roth, and Johannes Wohlfart. Subjective Models of the Macroeconomy: Evidence From Experts and Representative Samples. *The Review of Economic Studies*, 89(6):2958–2991, November 2022. ISSN 0034-6527. doi: 10.1093/restud/rdac008. URL https://doi.org/10.1093/restud/rdac008.

Andrea Baranzini, Stefano Carattini, and Linda Tesauro. Designing Effective and Acceptable Road Pricing Schemes: Evidence from the Geneva Congestion Charge. *Environmental and Resource Economics*, 79(3):417–482, July 2021. ISSN 1573-1502. doi: 10.1007/s10640-021-00564-y. URL https://doi.org/10.1007/s10640-021-00564-y.

Peter Belenky. Revised departmental guidance on valuation of travel time in economic analysis. Technical report, US Department of Transportation, Washington, DC, 2011.

B. Douglas Bernheim and Dmitry Taubinsky. Behavioral Public Economics. In B. Douglas Bernheim, Stefano DellaVigna, and David Laibson, editors, *Handbook of Behavioral Economics: Applications and Foundations 1*, volume 1 of *Handbook of Behavioral Economics - Foundations and Applications 1*, pages 381–516. North-Holland, January 2018. doi: 10.1016/bs.hesbe.2018.07.002. URL https://www.sciencedirect.com/science/article/pii/S2352239918300022.

Alberto Binetti, Francesco Nuzzi, and Stefanie Stantcheva. People's Understanding of Inflation, May 2024. URL https://www-nber-org.ezp-prod1.hul.harvard.edu/papers/w32497.

Inácio Bó and Rustamdjan Hakimov. Pick-an-Object Mechanisms. *Management Science*, September 2023. ISSN 0025-1909. doi: 10.1287/mnsc.2023.4908. URL https://pubsonline.informs.org/doi/abs/10.1287/mnsc.2023.4908. Publisher: INFORMS.

Bryan Caplan. What Makes People Think Like Economists? Evidence on Economic Cognition from the "Survey of Americans and Economists on the Economy". *The Journal of Law and Economics*, 44(2):395–426, October 2001. ISSN 0022-2186. doi: 10.1086/322812. URL https://www.journals.uchicago.edu/doi/10.1086/322812. Publisher: The University of Chicago Press.

Stefano Carattini, Andrea Baranzini, and Rafael Lalive. Is Taxing Waste a Waste of Time? Evidence from a Supreme Court Decision. *Ecological Economics*, 148:131–151, June 2018a. ISSN 0921-8009. doi: 10.1016/j.ecolecon.2018.02.001. URL https://www.sciencedirect.com/science/article/pii/S0921800917300915.

Stefano Carattini, Maria Carvalho, and Sam Fankhauser. Overcoming public resistance to carbon taxes. *WIREs Climate Change*, 9(5):e531, 2018b. ISSN 1757-7799. doi: 10.1002/wcc.531. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/wcc.531. _eprint: https://wires.onlinelibrary.wiley.com/doi/pdf/10.1002/wcc.531.

Raj Chetty, Adam Looney, and Kory Kroft. Salience and Taxation: Theory and Evidence. *American Economic Review*, 99(4):1145–1177, September 2009. ISSN 0002-8282. doi: 10.1257/aer.99.4.1145. URL https://www.aeaweb.org/articles?id=10.1257/aer.99.4.1145.

Sahil Chinoy, Nathan Nunn, Sandra Sequeira, and Stefanie Stantcheva. Zero-Sum Thinking and the Roots of U.S. Political Divides, September 2023. URL https://www.nber.org/papers/w31688.

Shani Cohen and Shengwu Li. Sequential Cursed Equilibrium. Paper, arXiv.org, August 2023. URL https://econpapers.repec.org/paper/arxpapers/2212.06025.htm.

Ernesto Dal Bó, Pedro Dal Bó, and Erik Eyster. The Demand for Bad Policy when Voters Underappreciate Equilibrium Effects. *The Review of Economic Studies*, 85(2):964–998, April 2018. ISSN 0034-6527. doi: 10.1093/restud/rdx031. URL https://doi.org/10.1093/restud/rdx031.

Geoffroy de Clippel. Behavioral Implementation. *American Economic Review*, 104(10):2975–3002, October 2014. ISSN 0002-8282. doi: 10.1257/aer.104.10.2975. URL https://www.aeaweb.org/articles?id=10.1257/aer.104.10.2975.

Geoffroy de Clippel, Rene Saran, and Roberto Serrano. Level-k Mechanism Design. *The Review of Economic Studies*, 86(3):1207–1227, May 2019. ISSN 0034-6527. doi: 10.1093/restud/rdy031. URL https://doi.org/10.1093/restud/rdy031.

Antoine Dechezleprêtre, Adrien Fabre, Tobias Kruse, Bluebery Planterose, Ana Sanchez Chico, and Stefanie Stantcheva. Fighting Climate Change: International Attitudes toward Climate Policies. *American Economic Review*, 115(4):1258–1300, April 2025. ISSN 0002-8282. doi: 10.1257/aer.20230501. URL https://www.aeaweb.org/articles?id=10.1257/aer.20230501.

Thomas Douenne and Adrien Fabre. Yellow Vests, Pessimistic Beliefs, and Carbon Tax Aversion. *American Economic Journal: Economic Policy*, 14(1):81–110, Febru-

ary 2022. ISSN 1945-7731. doi: 10.1257/pol.20200092. URL https://www.aeaweb.org/articles?id=10.1257/pol.20200092.

Bnaya Dreyfuss, Ofer Glicksohn, Ori Heffetz, and Assaf Romm. Deferred Acceptance with News Utility. *Management Science*, July 2025. ISSN 0025-1909. doi: 10.1287/mnsc.2024.05446. URL https://pubsonline.informs.org/doi/10.1287/mnsc.2024.05446. Publisher: INFORMS.

Christopher S. Elmendorf, Clayton Nall, and Stan Oklobdzija. The Folk Economics of Housing. *Journal of Economic Perspectives*, 39(3):45–66, August 2025. ISSN 0895-3309. doi: 10.1257/jep.20241428. URL https://www.aeaweb.org/articles?id=10.1257/jep.20241428.

Benjamin Enke and Thomas Graeber. Cognitive Uncertainty*. *The Quarterly Journal of Economics*, 138(4):2021–2067, November 2023. ISSN 0033-5533. doi: 10.1093/qje/qjad025. URL https://doi.org/10.1093/qje/qjad025.

Ignacio Esponda and Emanuel Vespa. Hypothetical Thinking and Information Extraction in the Laboratory. *American Economic Journal: Microeconomics*, 6(4):180–202, November 2014. ISSN 1945-7669. doi: 10.1257/mic.6.4.180. URL https://www.aeaweb.org/articles?id=10.1257/mic.6.4.180.

Erik Eyster and Matthew Rabin. Cursed Equilibrium. *Econometrica*, 73 (5):1623–1672, 2005. ISSN 1468-0262. doi: 10.1111/j.1468-0262.2005.00631.x. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1468-0262.2005.00631.x. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1468-0262.2005.00631.x.

Malachy James Gavan and Antonio Penta. Safe Implementation. *Theoretical Economics*, 2025.

Yannai A. Gonczarowski, Ori Heffetz, and Clayton Thomas. Strategyproofness-Exposing Mechanism Descriptions, July 2023. URL http://arxiv.org/abs/2209.13148. arXiv:2209.13148 [econ].

P. B. Goodwin. Solving congestion (when we must not build roads, increase spending, lose votes, damage the economy or harm the environment, and will never find equilibrium). Report 97/60, ESRC Transport Studies Unit, University College London, London, 1997. URL https://discovery.ucl.ac.uk/id/eprint/1244/. Publication Title: ESRC Transport Studies Unit, University College London: London.

Ziyuan Gu, Zhiyuan Liu, Qixiu Cheng, and Meead Saberi. Congestion pricing practices and public acceptance: A review of evidence. *Case Studies on Transport Policy*, 6

(1):94–101, March 2018. ISSN 2213-624X. doi: 10.1016/j.cstp.2018.01.004. URL https://www.sciencedirect.com/science/article/pii/S2213624X16300608.

INRIX and Bob Pishue. INRIX 2024 Global Traffic Scorecard, 2024. URL https://inrix.com/scorecard/.

S. Jaensirisak, M. Wardman, and A. D. May. Explaining Variations in Public Acceptability of Road Pricing Schemes. *Journal of Transport Economics and Policy (JTEP)*, 39(2):127–154, May 2005.

Shengwu Li. Obviously Strategy-Proof Mechanisms. *American Economic Review*, 107 (11):3257–87, November 2017. doi: 10.1257/aer.20160425. URL http://www.aeaweb.org/articles?id=10.1257/aer.20160425.

Benjamin Moshes and Joshua Moshes. Congestion Pricing Tracker, 2025. URL https://www.congestion-pricing-tracker.com/.

Brian Murray and Nicholas Rivers. British Columbia's revenue-neutral carbon tax: A review of the latest "grand experiment" in environmental policy. *Energy Policy*, 86: 674–683, November 2015. ISSN 0301-4215. doi: 10.1016/j.enpol.2015.08.011. URL https://www.sciencedirect.com/science/article/pii/S0301421515300550.

Lea Nagel and Roberto Saitto. A Measure of Complexity for Strategy-Proof Mechanisms. 2025.

Muriel Niederle and Emanuel Vespa. Cognitive Limitations: Failures of Contingent Thinking. *Annual Review of Economics*, 15(Volume 15, 2023):307–328, September 2023. ISSN 1941-1383, 1941-1391. doi: 10.1146/annurev-economics-091622-124733. URL https://www.annualreviews.org/content/journals/10.1146/annurev-economics-091622-124733. Publisher: Annual Reviews.

Arthur Cecil Pigou. *The Economics of Welfare*. Macmillan and Co., Limited, London, 1 edition, 1920. ISBN 978-0-230-24931-8. URL https://link.springer.com/book/9780230249318.

Marek Pycia and Peter Troyan. A Theory of Simplicity in Games and Mechanism Design. *Econometrica*, 91(4):1495–1526, 2023. ISSN 1468-0262. doi: 10.3982/ECTA16310. URL https://onlinelibrary-wiley-com.ezp-prod1.hul.harvard.edu/doi/abs/10.3982/ECTA16310. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.3982/ECTA16310.

Alex Rees-Jones and Dmitry Taubinsky. Measuring "Schmeduling". *The Review of Economic Studies*, 87(5):2399–2438, October 2020. ISSN 0034-6527. doi: 10.1093/restud/rdz045. URL https://doi.org/10.1093/restud/rdz045.

Paola Sapienza and Luigi Zingales. Economic Experts versus Average Americans. *American Economic Review*, 103(3):636–642, May 2013. ISSN 0002-8282. doi: 10.1257/aer.103.3.636. URL https://www.aeaweb.org/articles?id=10.1257/aer.103.3.636.

Geertje Schuitema, Linda Steg, and Sonja Forward. Explaining differences in acceptability before and acceptance after the implementation of a congestion charge in Stockholm. *Transportation Research Part A: Policy and Practice*, 44(2):99–109, February 2010. ISSN 0965-8564. doi: 10.1016/j.tra.2009.11.005. URL https://www.sciencedirect.com/science/article/pii/S0965856409001256.

Howard Schuman and Stanley Presser. *Questions and Answers in Attitude Surveys: Experiments on Question Form, Wording, and Context.* SAGE, March 1996. ISBN 978-0-7619-0359-8. Google-Books-ID: Je640UKqNaYC.

Christoph Semken. The Marginal Impact of Emission Reductions: Estimates, Beliefs and Behavior. *Working Paper*, May 2025.

Stefanie Stantcheva. Understanding of Trade, May 2022. URL https://www.nber.org/papers/w30040.

William S. Vickrey. Congestion Theory and Transport Investment. *The American Economic Review*, 59(2):251–260, 1969. ISSN 0002-8282. URL http://www.jstor.org/stable/1823678. Publisher: American Economic Association.

A Proofs

This section contains the proofs for all Results, Propositions, and Theorems, following the order they appear in the paper.

A.1 Result 1

Proof. Any equilibrium must be a cutoff equilibrium. Let $\theta_1 > \theta_2$. If both types play best responses and type θ_1 weakly prefers public transit, then type θ_2 strictly prefers public transit. Similarly, if type θ_2 weakly prefers driving, then type θ_1 strictly prefers driving.

Any cutoff equilibrium with cutoff θ must satisfy the indifference condition:

$$\theta Q(1 - F(\theta); I) = b + p$$

Since the LHS is strictly increasing in θ (both θ and $Q(1 - F(\theta); I)$ are increasing in θ) and ranges from 0 to ∞, while the RHS equals $b + p > 0$, this equation has a unique solution $\theta^*(T)$.

Therefore, there exists a unique Nash equilibrium $a^*(T)$ where $a_n = 1$ iff $\theta_n \geq \theta^*(T)$. □

A.2 Proposition 2

Proof. An example of such structure is the experimental design in Dal Bó et al. (2018), where sufficiently neglectful participants are predicted to choose the Prisoners' Dilemma over Harmony Game (see their Proposition 1). The payoff matrices in their experiment can be transformed to generate unbounded gains from the latter and unbounded losses from the former, without changing this prediction. Here I provide a simpler example for completeness.

Let $N = \{1, 2\}$ with $A_n = \{0, 1\}$ and utilities $v_n(a; T) = a_n(1 - T) - (K + 1)a_{-n}$.

For part 1, set $T^0 = 0$ and $T^+ = 2$. The equilibrium under T^0 is $a^*(T^0) = (1, 1)$ with payoffs $-M$, while under T^+ it is $a^*(T^+) = (0, 0)$ with payoffs 0. Thus $\delta_n(T^+, T^0) = K$. The direct effect from T^0 to T^+ is $v_n((0, 1); T^+) - v_n((1, 1); T^0) = -(K+1) - (-K) = -1$, while the equilibrium effect is $K + 1$. For $\eta_n < \frac{1}{K+1}$, the perceived gain $\tilde{\delta}_n(T^+, T^0; \eta_n) = -1 + \eta_n(K + 1) < 0$, making T^+ infeasible by consensus.

For part 2, set $T^0 = 2$ and $T^- = 0$. Symmetric arguments show that for $\eta_n < \frac{1}{K+1}$, T^- is feasible by consensus. □

A.3 Proposition 3

Proof. WLOG let $p^0 = 0$, and let $p > 0$,[35] with associated Nash equilibrium $a^p = a^*(p)$ and $\alpha^t = \alpha(a^p)$ (notice that α^p is decreasing in p).

Lemma 1 (Unbounded Action)**.** *For any tax rate $p > 0$, any aggregator α, and any action $a_n \in \{0, \ldots, M\}$, there exists a threshold $\overline{\theta}$ such that if $\theta_n > \overline{\theta}$, $a_n^*(\alpha, p) \geq a_n$.*

Proof. It suffices to show that for any $p > 0$ and α, some type chooses action M. An agent prefers M to $M - 1$ if the marginal benefit exceeds p: $w(M, \alpha; \theta) - w(M - 1, \alpha; \theta) > p$.

By the Assumption 3, this marginal benefit grows at least linearly with type:

$$[w(M, \alpha; \theta) - w(M - 1, \alpha; \theta)] \geq [w(M, \alpha; 0) - w(M - 1, \alpha; 0)] + \epsilon\theta.$$

As $\theta \to \infty$, the right-hand side grows without bound. Therefore, for any finite t, there exists $\overline{\theta}$ such that for all $\theta > \overline{\theta}$, the marginal benefit exceeds p. For these types, the optimal action is M. □

Let $\hat{a}_n = \text{argmax}_{a_n \in A_n}\{w(a_n, \alpha^0; \theta_n) - pa_n\}$. Since $\alpha^p < M$, by Lemma 1, there exists $\overline{\theta^1}$ such that if $\theta > \overline{\theta^1}$, $\hat{a}_n > \alpha^p$.

[35] It is WLOG since we can always redefine $t' = t - t^0$.

The direct effect is given by

$$w(a_n^0, \alpha^0; \theta_n) - w(\hat{a}_n, \alpha^0; \theta_n) - p(\hat{a} - \alpha^p).$$

By best-response, the first part is weakly negative. The second part is strictly negative for $\hat{a}_n > \alpha^p$, i.e., if $\theta > \overline{\theta^1}$. Therefore, for $\theta_n > \overline{\theta}^1$, $\tilde{\delta}_n(T, T^0; 0) < 0$.

The actual gain for agent n is given by

$$w(a_n^p, \alpha^p; \theta_n) - w(a^0, \alpha^0; \theta_n) - p(a_n^p - \alpha^p)$$

We then have,

$$\begin{aligned} w(a_n^p, \alpha^p; \theta_n) - w(a^0, \alpha^0; \theta_n) - p(a_n^p - \alpha^p) &\geq w(a_n^0, \alpha^t; \theta_n) - w(a^0, \alpha^0; \theta_n) - p(a_n^0 - \alpha^p) \\ &\geq w(a_n^0, \alpha^p; \theta_n) - w(a^0, \alpha^0; \theta_n) - p(M - \alpha^p) \end{aligned}$$

Where the first inequality follows from best response, and the second from the definition of the action space. Using the same arguments as before, we can show that the RHS grows without bound, and in particular there exists $\overline{\theta^2}$ such that for all $\theta \geq \overline{\theta^2}$, the actual gain is positive i.e., $\delta(T, T^0) > 0$.

Therefore for all n with $\theta_n > \overline{\theta} \equiv \max\{\overline{\theta^1}, \overline{\theta^2}\}$, $\delta(T, T^0) > 0$ but $\tilde{\delta}_n(T, T^0; 0) < 0$.

To derive the bound on η, notice that

$$\tilde{\delta}_n(T, T^0; \eta) = \tilde{\delta}_n(T, T^0; 0) + \eta[\delta_n(T, T^0) - \tilde{\delta}_n(T, T^0; 0)]$$

therefore for $\eta < \frac{-\tilde{\delta}_n(T,T^0;0)}{\delta_n(T,T^0) - \tilde{\delta}_n(T,T^0;0)} \equiv \overline{\eta}(\theta_n)$ we have $\tilde{\delta}_n(T, T^0; \eta) < 0$, as needed. □

A.4 Result 2

Proof. From Result 1, the equilibrium cutoff θ^0 satisfies:

$$\theta^0 Q(1 - F(\theta^0)) = b$$

Clearly, any optimum must be cutoff based. Total welfare from an allocation where types $\theta > \theta'$ drive is given by:

$$W(\theta') = \int_{\theta'}^{\infty} [\theta Q(1 - F(\theta)) - b] dF(\theta)$$

The first-order condition yields:

$$\theta^* Q(1 - F(\theta^*); I) = b - Q'(1 - F(\theta^*)) \cdot \int_{\theta^*}^{\infty} \theta dF(\theta)$$

Since $Q' < 0$ and $\int_{\theta^*}^{\infty} \theta dF(\theta) > 0$:

$$\theta^* Q(1 - F(\theta^*)) > b = \theta^0 Q(1 - F(\theta^0))$$

Since $\theta \mapsto \theta Q(1 - F(\theta))$ is strictly increasing, we have $\theta^* > \theta^0$, which implies $\alpha^0 = 1 - F(\theta^0) > 1 - F(\theta^*) = \alpha^*$.

The gap $b - \theta^* Q(1 - F(\theta^*)) = Q'(1 - F(\theta^*)) \cdot \mathbb{E}[\theta | \theta \geq \theta^*] \cdot (1 - F(\theta^*))$ represents the marginal congestion externality, i.e., the social cost imposed on existing drivers by an additional driver. Setting the toll $p^* = -Q'(1 - F(\theta^*)) \cdot \mathbb{E}[\theta | \theta \geq \theta^*] \cdot (1 - F(\theta^*))$ internalizes this externality and achieves the social optimum. □

A.5 Theorem 1

Proof. Fix a status quo T^0 and a monotonic feasibility rule f. Let $v_n(a; T) = w_n(a) + T_n(a)$, and let T^* be a policy that is politically feasible with respect to f under T^0 and induces an equilibrium $a^*(T^*) = a^*$. Denote the multilaterally distinct status-quo equilibrium $a(T^0) = a^0$. Let

$$\tilde{T}_n(a) = T_n^*(a_n) + h_n(a_{-n})$$

where $h_n(a_{-n}) = k \cdot \mathbf{1}_{a_{-n} = a^0_{-n}}$.

Starting with outcome equivalence, by quasi-linearity and the definition of equilibrium we have:

$$\arg\max_{a_n \in A_n} v_n(a_n, a^*_{-n}; \tilde{T}) = \arg\max_{a_n \in A_n} v_n(a_n, a^*_{-n}; T^*) + h_n(a^*_{-n}) = \arg\max_{a_n \in A_n} v_n(a_n, a^*_{-n}; T^*) = a^*_n.$$

Therefore, if a^* is an equilibrium under T^*, it must also be an equilibrium under $\tilde{T}$. To show deficit equivalence, since a^* and a^0 are multilaterally distinct, we have $a^0_{-n} \neq a^*_{-n}$ for all $n \in N$. Therefore, $h_n(a^*_{-n}) = 0$ for all n, which implies $\tilde{T}_n(a^*) = T_n^*(a^*)$ for all n, and thus $D^*_{T^*} = D^*_{\tilde{T}}$.

Finally, we need to show that $\tilde{T}$ is EN-robust. By monotonicity of f, it suffices to show $\tilde{\delta}_n(\tilde{T}_n, T^0_n; \eta) \geq \delta_n(\tilde{T}_n, T^0_n)$ for all $n \in N$ and all $\boldsymbol{\eta} \in [0, 1]^N$.

For all n and all $\eta \in [0, 1]$, we have

$$\tilde{\delta}_n(\tilde{T}, T^0; \eta) = \delta_n(T^*, T^0; \eta) + (1 - \eta)k.$$

Therefore,

$$\begin{aligned}\tilde{\delta}_n(\tilde{T}_n, T^0_n; \eta) &= \eta \delta_n(\tilde{T}_n, T^0_n) + (1 - \eta)[\tilde{\delta}_n(\tilde{T}_n, T^0_n; 0)] \\ &= \eta \delta_n(T^*_n, T^0_n) + (1 - \eta)[\tilde{\delta}_n(T^*_n, T^0_n; 0) + k].\end{aligned}$$

By best response, the direct effect is bounded by the difference in transfers:

$$\tilde{\delta}_n(T_n^*, T_n^0; 0) = w_n(\hat{a}_n, a_{-n}^0) + T_n^*(\hat{a}_n) - w_n(a^0) - T_n^0(\hat{a}_n) \geq T_n^0(a^0) - T_n^*(a^0),$$

and therefore for all $k \geq \max\{T_n^*(a^0) - T_n^0(a^0), 0\}$, we have

$$\tilde{\delta}_n(\tilde{T}_n, T_n^0; \eta) \geq \delta_n(\tilde{T}_n, T_n^0)$$

for all $n \in N$ and all $\boldsymbol{\eta} \in [0,1]^N$, as needed. □

A.6 Theorem 2

Proof. Fix a status quo T^0 and a weighted-support based feasibility rule f. Let $v_n(a;T) = w_n(a) + T_n(a)$, and let T^* be a policy that is politically feasible with respect to f under T^0 and induces an equilibrium $a^*(T^*) = a^*$. Denote the status-quo equilibrium $a(T^0) = a^0$. Let $\tilde{T}_n(a) = \tilde{T}'_n(a; a_n^0) = T_n^*(a_n) + k_n \mathbf{1}_{a_{-n}=a_{-n}^0}$, where $k_n = T_n^*(a^0) - T_n^0(a^0)$. Note that relative to the previous case, constructing $\tilde{T}$ requires only one additional piece of information: the difference between what agent n pays under the status quo versus under T^*, holding behavior fixed.

Outcome and (on-path) deficit equivalence follow from the same steps as in the proof of Theorem 1. To establish EN-robustness with respect to f, we must show that if $\delta_n(T, T^0) \geq 0$, then $\tilde{\delta}_n(\tilde{T}, T^0; \eta) \geq 0$ for all $\eta \in [0,1]$.

Suppose $\delta_n(T^*, T^0) \geq 0$. From the proof of Theorem 1, we have

$$\begin{aligned}\tilde{\delta}_n(\tilde{T}_n, T_n^0; \eta) &= \eta\delta_n(T_n^*, T_n^0) + (1-\eta)[\tilde{\delta}_n(T_n^*, T_n^0; 0) + k_n] \\ &= \eta\delta_n(T_n^*, T_n^0) + (1-\eta)[\tilde{\delta}_n(T_n^*, T_n^0; 0) + T_n^*(a^0) - T_n^0(a^0)]\end{aligned}$$

and also

$$\tilde{\delta}_n(T_n^*, T_n^0; 0) \geq T_n^*(a^0) + T_n^0(a^0).$$

Therefore, if $\delta_n(T_n^*, T_n^0) \geq 0$, then $\tilde{\delta}_n(\tilde{T}_n, T_n^0; \eta) \geq 0$ for all $\eta \in [0,1]$, as required.

Finally, to characterize the deficit: for any $a \in A \setminus \{a^0\}$, we have $D_{\tilde{T}}(a) = D_{T^*}(a)$ by construction. For a^0, we have

$$D_{\tilde{T}}(a^0) = \sum_{n \in N}[T_n^*(a_n^0) + k_n] = \sum_{n \in N} T_n^0(a^0) = D_{T^0}(a^0)$$

as required. □

B Calculation of Budget-balanced Compensation

Revenue Calculation

For our calculations, we use the following notation:

- V: number of cars entering the city center on a typical weekday in the untolled scenario
- w: weekend traffic discount factor
- T': base toll (passenger vehicles)
- T: average toll (across vehicle types)
- $\tau(T)$: Traffic adjustment given toll (note: equals $1 - \rho(T)$ from the main text)
- r: average daily revenue
- $R = 365r$: annual revenue
- a: administrative costs and exemptions

Given a toll T, the new level of traffic is $\tau(T) \cdot V$. Taking into account around 30% of days in a given year being holidays and weekends, the average daily discount factor (i.e., averaging across weekends and weekdays) is given by:

$$w' = 0.7 + 0.3w$$

The average daily revenue is then given by:

$$r = a \cdot T \cdot \tau(T) \cdot w' \cdot V$$

We assume a holiday/weekend discount factor $w = 0.7$, which implies $w' = 0.91$. For NYC, we have $V = 700,000$ (**?**). The share of taxis entering Manhattan is approximately 0.35, and taxis pay on average $\frac{1}{8}$ of the toll,[36] which implies:

$$T = 0.65T' + 0.35 \cdot 0.18 \cdot T' \approx 0.69T'$$

Solving for a:

$$a = \frac{R}{T' \cdot \tau(T) \cdot 441,919}$$

[36] In NYC, Taxis pay $\frac{T}{12}$, and FHVs pay $\frac{T}{6}$, and we assume equal shares.

Imputed Deadweight Parameters

The table below imputes the deadweight parameter for different proposed tolling scenarios from the MTA's Environmental Assessment (**??**):

MTA EA Tolling Scenario	Base Toll (T')	MTA Projected Traffic ($\tau(T)$)	MTA Projected Revenue (R in billions)	Imputed Deadweight Parameter (a)
Adopted (original)	\$15	0.827	0.900	**0.450**
Adopted (revise phase 1)	\$9	0.876	0.500	**0.393**
Adopted (revise phase 2)	\$12	0.845	0.700	**0.428**
Adopted (revise phase 3)	\$15	0.827	0.900	**0.450**
A	\$9	0.846	1.060	**0.863**
B	\$10	0.843	0.830	**0.610**
C	\$14	0.827	1.100	**0.589**
D	\$19	0.813	1.340	**0.538**
E	\$23	0.801	1.480	**0.498**
F	\$23	0.817	1.020	**0.337**
G	\$12	0.840	1.100	**0.677**

Table 3: Imputed Deadweight Parameters for MTA Tolling Scenarios

Off-Path Compensation

Denote the number of workers who live outside the tolled area but are employed inside of it by E. The upper bound on total compensation per worker per day is therefore given by:

$$c = \tau(T) \cdot r \cdot E = T \cdot a \cdot w' \cdot V \cdot E$$

We denote the monthly compensation by $C = 31c$. According to ACS data, in NYC there are 1 million employees who commute to work in the CBD, which implies $VE = 0.7$. Plugging in our estimate from before and assuming $a = 0.5$ (approximately the average across all tolling scenarios), we get:

$$C(T) = 10T$$

For example, $C(15) = 150$. However, one may want to set aside a budget dedicated to public transit even off the equilibrium path. Setting this amount to $\frac{1}{5}$ of the revenue results in:

$$C(15) = 120 \quad \text{and (if I were to implement a \$9 toll + compensation)} \quad C(9) = 72.$$

C Policymakers: Additional Results

Table 4: Net Support and First- and Second-order Beliefs

	State Legislators		Civil Servants	
	(1)	(2)	(3)	(4)
Time savings (% public)	0.9308	0.772	1.76	0.891
	(0.284)	(0.328)	(0.471)	(0.471)
Time savings (own)		16.537		111.247
		(16.107)		(30.167)
R^2	0.085	0.092	0.229	0.405
Observations	118	118	49	49

Note: The table reports regressions of net support for congestion pricing on first- and second-order beliefs. Time savings (% public) is a respondent's estimate of the share of the general public mentioning time savings when asked to predict the consequences. Time savings (own) is an indicator that equals one when a respondent predicts any time savings (as opposed to an increase or "no meaningful change").