Equilibrium Neglect and Political Feasibility

Pre-results draft

Preliminary. Please Do not Distribute

Bnaya Dreyfuss*

March 21, 2025

Abstract

When voters underappreciate the equilibrium effects of policies, they may rank them incorrectly, which can often make socially optimal policies politically infeasible. I study the policy implications of *equilibrium neglect* (EN) and propose a potential remedy. I first develop a general framework of EN, 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 is portable and can be applied in a large class of problems. I design empirical tests for this remedy both in incentivized experiments and in opinion polls about congestion pricing. Data collection to be completed over the next few weeks.

*Harvard University. Email: bdreyfuss@g.harvard.edu. I thank Raj Chetty, Katherine Coffman, 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 seminar participants at Harvard for helpful feedback and suggestions. This project is supported by the Griffin Economics Research Fund.

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 the public, whereas policies economists tend to support, such as carbon taxes and congestion pricing, are viewed negatively. 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, in other words, its *equilibrium effects*. This tendency can distort perceptions of a policy's indirect costs or benefits, potentially leading to inaccurate assessment of its overall impact. We term 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; Dal Bó et al., 2018).

Taking this body of work as our starting point, we ask: what are the policy implications? In particular, can policymakers design welfare-improving policies that can win support from equilibrium-neglecting voters? We 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. Our remedy adds *off-path* contingent rebates to an existing policy, thus ensuring that the agent's perceived benefits of the policy are positive even if they neglect its equilibrium effects. We then design empirical tests for the effectiveness of this remedy in increasing public support in two different contexts: an incentivized experiment replicating and extending Dal Bó et al. (2018), and an opinion poll about a congestion pricing policy.

While equilibrium neglect can play a role in the political feasibility of various policy contexts (including taxation, price controls and industrial policy), we use congestion pricing both as our theoretical running example and as our main empirical application.

Congestion pricing is particularly suitable for studying equilibrium neglect for a few reasons. First, it offers a clean distinction between direct effects (price change) and equilibrium effect (congestion reduction), both of which are unidimensional, easily observed, and precisely measured. Second, while there is consensus among economists and transportation engineers that it is typically welfare increasing, congestion pricing is in practice notoriously difficult to implement due to political feasibility constraints (for a recent example from NYC, see 5). Third, there is evidence (reviewed below) suggesting that EN is particularly relevant in the context of congestion pricing: people have been consistently shown to underappreciate the congestion-reducing effects of the policy ex ante, but strongly revise their beliefs having observed its effects. In addition, support for congestion pricing is typically 10-20 percentage point lower before its implementation (i.e., before observing its effects on traffic) compared to after (i.e., having observed the effects).

This paper contributes to the literature on misspecified subjective models of the economy, failures in contingent reasoning, behavioral mechanism design, and the political feasibility of congestion pricing, all reviewed in Section 2.

We present our theoretical framework in Section 3. Based on the main insight from Dal Bó et al. (2018), we propose a general model of equilibrium neglect (EN) in policy evaluation. We construct a general setting where a policy space induces a class of games, and model how agents perceive costs and benefits of a new policy relative to the status quo. Decomposing the effect of a policy change 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—an agent is said to partly neglect equilibrium if they underweight the latter effect. This effect can be driven by explicit pessimism—believing that others may not change their behavior in response to the policy change—or by cognitively discounting equilibrium effects, which are typically ex-ante less salient, and harder to predict.

In Section 3.2 we apply this framework in the context of congestion pricing. We first construct a simple example where implementing congestion pricing generates a Pareto

improvement, and then show that even in this extreme case, if agents are sufficiently equilibrium neglectful, it may not generate a sufficiently high level of support. We show that this problem can be completely solved by using *explicit* redistribution of the collected revenue. With explicit redistribution compensation incresaes with congestion, thus effectively providing agents with insurance against bad outcomes: high congestion implies more toll-paying drivers, which increases compensation. Conversely, low compensation implies fewer toll-paying drivers and hence low congestion. As we show, this ensures positive perceived benefits even for equilibrium-neglectful agents, while reducing congestion and without increasing the deficit.

Theorem 1 generalizes this result, showing that the same logic can be applied in any setting with quasi-linear utility and taxation. In particular, starting from any status quo, and given any constrained social optimum, we can generate a policy that (i) achieves the same social outcome, (ii) does not increase the deficit and (iii) generates the level of support that would have been achieved absent equilibrium neglect. The policy achieves these goals by introducing an off-path component to the social optimum that, while not affecting the outcome and hence the deficit, increases its perceived benefits for equilibrium neglectful agents.

Theorem 2 shows that this off-path contingent rebate can be constructed in a way that guarantees deficit equivalence not just on the equilibrium path but even off path. In particular, we show that even off-path, the deficit is not higher than the status quo deficit. This alleviates potential credibility concerns, as voters might treat policies involving excessively large deficits (even off path) with suspicion. It also makes the remedy effective in the presence of equilibrium-neglecting policy makers who may themselves be pessimistic about equilibrium effects, and may be hesitant to commit to large off-path deficits.

Does the proposed remedy to equilibrium neglect using off-path rebates prove effective in practice? Ex-ante, there are reasons for skepticism: people might resist EN-robust policies, perceiving them as complex, unnatural, or non-credible. In addition, off-path rebates demand agents to engage in contingent reasoning—the very skill they may lack due to EN. We empirically test the effectiveness of our remedy in generating support in

two different settings: an incentivized lab experiment, and an opinion poll about a real congestion pricing policy.

Section 4 presents our experimental design and findings. Our design replicates and extends Dal Bó et al. (2018). Subjects begin by playing five rounds of a Prisoner's Dilemma. They then vote on a policy that changes incentives so that mutual cooperation becomes the new dominant-strategy equilibrium in the new game, called the Harmony Game. Dal Bó et al. (2018) find low support for the Harmony Game even though it is socially optimal, consistent with equilibrium neglect. Our Baseline treatment replicates the same payments as Dal Bó et al. (2018). We add to this design two EN-robust treatments that include an off-path component offering players a payment in the event that the other player keeps defecting (in contrast to the Nash prediction). Both EN-robust treatments are budget balanced on path, but only one of them is also budget balanced off-path.

Data collection for this experiment will be completed over the next few weeks. The main outcome will compare the level of support for the social optimum, i.e., the share of subjects voting for the Harmony game in each treatment, and the social outcome, i.e., the rate of cooperation in each treatment. According to our theory, applying the remedy will result in an increase in support for the social optimum, without changing the social outcome.

In section 5, we present a design for an opinion poll to test our proposed remedy for equilibrium neglect. Following the theoretical framework in Section 3, we apply our EN-robust mechanism in the context of congestion pricing policy. To ensure ecological validity and generalizability across diverse urban settings, we employ a representative sample of 2,000 respondents from six U.S. metropolitan areas with known traffic problems (Boston, Washington D.C., Chicago, Dallas, San Francisco, and Los Angeles), recruited through NORC's AmeriSpeak Panel, a probability-based panel designed to be representative of the U.S. household population. We use a two-stage, between-subject design comparing support for a standard congestion pricing policy to support for its EN-robust counterpart and a reduced-toll alternative. For the baseline policy, we model our design after the orig-

inal $15 toll proposed by the MTA in NYC in 2024. To generate EN-robustness without increasing the deficit on- and off-path, we add a "Monitoring & Assurance" component, offering workers who commute to the downtown area a $120 monthly compensation if traffic decreases by less than 8% after introducing the toll. We randomly assign respondents to one of three conditions ($N = 800$ for both the treatment and control conditions, and $N = 400$ for the toll reduction condition), keeping other parts of the policy, including information about its expected traffic reduction, constant across treatments, while comparing support against both a simple monitoring policy and a reduced $9 toll policy (mirroring NYC's actual policy compromise).

We conclude in Section 6.

2 Related Literature

It is widely acknowledged that people's models of the economy are often misspecified. Most relevant to our paper, Dal Bó et al. (2018) experimentally shows that people underappreciate equilibrium effects, which causally makes them support the suboptimal Prisoner's Dilemma over Harmony Game. Other work modeling and estimating people's potentially misspecified subjective models of the economy includes Andre et al. (2022); Stantcheva (2022); Chinoy et al. (2023); and Binetti et al. (2024). This paper contributes to this literature by constructively using a particular formal model of misperception to generate and empirically test a potential policy remedy.

More generally, our paper is related to the large body of work on failures in contingent reasoning, reviewed in detail in Niederle and Vespa (2023). In particular, failing to understand how actions and incentives are related has been shown to be particularly important, both theoretically (Eyster and Rabin, 2005; Cohen and Li, 2023) and empirically (e.g., Esponda and Vespa, 2014).

Similar to how these documented empirical regularities have been modeled and then applied to characterize different mechanisms by their strategic simplicity (e.g., Li, 2017; Bó and Hakimov, 2023; and Pycia and Troyan, 2023), we use a general framework to derive

a generally applicable remedy for equilibrium neglect and test it empirically. While these papers are focused on people's behavior *within* a mechanism, we are interested in people's preferences *over* different mechanisms or policies.

Last, by using congestion pricing as our main application, we also contribute to the large literature on public acceptability of congestion pricing, reviewed in Gu et al. (2018). This literature consistently finds support for congestion pricing to be higher after its implementation, and that people are overly pessimistic about the congestion-reducing effects of such policies, strongly revising their beliefs post-implementation (Schuitema et al., 2010). Both findings are consistent with equilibrium neglect.

In terms of remedies, while the perceived effectiveness of the policy in reducing congestion has been shown to strongly predict support, informational treatments are found ineffective (Baranzini et al., 2021). In contrast, real changes to the policy have been shown to generate support. In particular, a reduction in the size of the toll (Baranzini et al., 2021), as well as earmarking revenues for specific purposes (Jaensirisak et al., 2005), are both generally associated with higher levels of support. Our proposed remedy addresses equilibrium neglect not by attempting to change people's beliefs, but by introducing a real change to the policy. However, rather than decreasing the toll, which would affect its effectiveness, our theory-driven remedy involves a particular earmarking by transparently re-investing the revenue in order to increase the perceived benefits for equilibrium-neglecting voters.

3 Theory

3.1 Framework

In this section, we present our framework capturing equilibrium neglect. This formulation generalizes and extends the Dal Bó et al. (2018) framework, which only applied in their setup, which had two players, two games (policies) and two actions.

3.1.1 Preliminaries

There is a set of N of agents $n \in N$. Each agent $n \in N$ chooses an action a_n from an action space A_n. We denote the action profile by $a = \{a_n\}_{n \in N}$, and the set of possible action profiles by $A \equiv \prod_{n \in N} A_n$. For each player $n \in N$ there is a *policy function* $T_n \in \mathbb{T}_n$ that maps action profiles into a consequence for the player: $T_n : A \to C_n$. We may think of T_n as capturing taxes, subsidies, public goods and infrastructure, or any regulation a government might impose on an agent's actions. We denote profiles of policy functions, which we simply call "policy," by $T = \{T_n\}_{n \in N}$. The set of possible policies is denoted by $\mathbb{T}$. Agents have preferences over action profiles and their own consequences (but not directly over policies), and so player n's utility is given by $u_n : A \times C_n \to \mathbb{R}$.

A set of players N, action profiles A, utility functions $u = \{u_n\}_{n \in N}$ combined with a set of possible policies $\mathbb{T}$ jointly define a class of games governed by T. This framework is aimed at capturing the perceived effect of a policy change, i.e., moving across games within a given class. It will be convenient to denote $v_n(\cdot \mid T) \equiv u_n(\cdot, T_n(\cdot))$, i.e., the agent's utility function within the game induced by the policy T. For simplicity, we restrict our attention to settings with a unique pure Nash Equilibrium for any $T \in \mathbb{T}$ and denote the Nash equilibrium under policy T by $a^*(T)$. Last, we denote the best response for agent n to others' action profile a_{-n} and a policy function T_n by $a_n^*(a_{-n}, T_n)$.

We say that a policy T is *anonymous* if $T_n(a) = T_{n'}(a)$ for all $n, n' \in N$ and for all $a \in A$. In what follows we largely focus on anonymous functions. When there is no ambiguity, we will often suppress the index and denote anonymous policies by the (unique) policy function they contain i.e., $T(a) = T_n(a)$. We say that a policy function is *action-based* if $T_n(a)$ is only a function of a_n, i.e., if the payoff depends only the agent's own action. In contrast, a policy function is *profile-based* if T_n also depends on a_{-n}. The distinction between action- and profile-based policies will become important in our main result on feasibility under equilibrium neglect.

3.1.2 Perceived Benefits

To formalize equilibrium neglect and its effects on political feasibility, we need to formalize how agents perceive the costs and benefits of a policy change, allowing agents to partially neglect the effect of a policy change on others' behavior:

Definition 1 (Perceived benefits)**.** *Let $\eta \in [0,1]$. The perceived benefit of policy T' under a status quo T for agent n is given by*

$$\delta_n(T', T; \eta) = v_n\left((\hat{a}_n, a^*_{-n}(T)) \mid T'\right) - v_n\left(a^*(T) \mid T\right) + \eta\left[v_n\left(a^*(T') \mid T'\right) - v_n\left((\hat{a}_n, a^*_{-n}(T)) \mid T'\right)\right]$$

where $\hat{a}_n = a^*_n(a^*_{-n}(T), T'_n)$ is the agent's best response to the policy change holding others' behavior fixed.

The definition of perceived benefits relies on decomposing the effect of moving from T to T' into two parts: the first, captured by the first two terms, is the *direct effect*: the effect of a policy change on the agent's personal consequences holding the behavior of others fixed.[1] The second effect is the *equilibrium effect* of moving from the status-quo equilibrium to the new equilibrium—i.e., the effect of others' behavior change on the agent's utility—holding the policy fixed.

The perceived equilibrium effect is cognitively discounted by $\eta \leq 1$. This discount factor has a few potential interpretations. First, while the direct effect is typically straightforward to calculate, predicting the equilibrium effect is much more difficult. For example, it typically requires the agent to hold an explicit belief about the distribution of preferences in the population. It is therefore discounted by η.

A slightly different interpretation of η is probabilistic: we can think of it as representing the probability that the agent assigns to the economy converging to the new equilibrium as opposed to the event where the policy changes and yet others' behavior remains fixed.

In the most extreme case, $\eta = 0$ the voter only perceives (or believes in) the direct

[1] We include in this effect the player's own behavioral adjustment in reaction to the policy. In the Online Appendix we show that allowing agents to neglect their own individual reaction does not meaningfully change our main results.

effect, and completely neglects (or misbelieves in-) changes in utility coming from others' behavioral reactions. Conversely, the $\eta = 1$ case embeds standard models in which voters perfectly perceive equilibrium effects.[2]

3.1.3 Political Feasibility

To the extent that $\eta < 1$ causes voters to rank policies incorrectly, EN can have direct policy implications by making socially desirable policies politically infeasible.

The following definition formalizes political feasibility as a function of perceived benefits for voters:

Definition 2. *Let T^0 be a status-quo policy, let $\{\eta_n\}_{n\in N}$, and let $\beta \in [0,1]$. A policy T is politically feasible under the status quo T^0 with respect to the threshold β given η if the share of players for whom $\delta_n(T', T; \eta_n) \geq 0$ is at least β.*

Two particularly interesting cases for feasibility are $\beta = 1$, i.e., the case where only Pareto improvements are politically feasible (e.g., because a vocal minority has enough power to maintain the status quo), and $\beta = \frac{1}{2}$, i.e., the median voter case.

3.1.4 Quasilinear Utility, Taxation and Deficit

While our framework can in principle deal with a wide range of policy tools such as infrastructure investment, zoning laws, and bans,[3] in what follows we focus on monetary consequences and quasilinear utility, i.e., we assume $\mathbb{T} = \{T : A \times N \to \mathbb{R}\}$ and $v_n(a \mid T) = w_n(a) + T_n(a)$.

This allows us to use the size of the deficit as a natural way to constrain and rank policies within $\mathbb{T}$. The *deficit* of policy T is given by $D_T(a^*(T)) = \sum_n T_n(a^*(T))$ or the equivalent integral in a continuum setting. Notice that $D_T(a^*(T^*)$ is the *on-path* deficit. Later in this section we discuss *off-path* deficits. Last, we say that T is (on-path) budget-balanced if $D_T(a^*(T)) \leq 0$.

[2]Notice that this framework can also capture over-weighting of equilibrium effects when $\eta > 1$. While this form of bias can certainly be important in policy discourse (and is perhaps common even, or maybe especially, among economists), it is not our focus, and we leave it for future work.

[3]While the action space for each agent does not depend on the policy, our framework can capture a ban on some action a_n as choosing for all agents a policy function $T_n((a_n, a_{-n})) = c_n$ for all a_{-n} such that $u(\cdot, c_n)$ is arbitrarily low.

3.2 Application: Congestion pricing

3.2.1 Setup

In this section, we outline our main application of EN: congestion pricing. There is a unit continuum of agents and a single road. All agents decide between driving ($a_n = 1$) and riding the train ($a_n = 0$).

Riding the train always takes l minutes; in contrast, driving time depends on the share of agents who choose to drive. We assume that the cost of congestion is given by a simple step function: the road has some capacity $\alpha < 1$, and driving takes $l-(b-c\cdot\mathbf{1}(\gamma(a) > \alpha))$ where b is the time saving from uncongested driving, $\gamma(a)$ is the share of road users and c is the cost of congestion. Slightly abusing notation, we denote $c(a) = c \cdot \mathbf{1}(\gamma(a) > \alpha)$.

A *standard* congestion pricing policy is action-based and anonymous, and can be written as $T_n(a) = k - p \cdot a_n$, i.e., an unconditional lump-sum transfer and a price for using the road. We denote a standard congestion pricing policy by $(p, k) \in \mathbb{R}^2$.

Agents have quasi linear utility from time and money:

$$v_n(a \mid T) = I_n - \theta_n(l - a_n(b - c(a))) + T_n(a),$$

where I_n is income and θ_n captures the value of time for agent n. For simplicity, we assume $\theta_n \sim U(0, 1)$ and index agents by their type.

To simplify the presentation but WLOG we assume $I_n = 0$ for all n and $l = 0$. To avoid equilibrium-multiplicity issues, we assume $c < b$.[4] It is therefore straightforward to see that absent of congestion pricing, i.e., if $(p, k) = (0, 0) \equiv T^0$, all agents drive (i.e., $\gamma(a^*(T^0) = 1)$ and the road is congested. We will treat this policy as the status quo.

3.2.2 Political Feasibility of Congestion Pricing

Congestion is socially suboptimal if total welfare is lower under the status quo, unpriced equilibrium than in the case where the road is used exactly at capacity and only the top α types drive, i.e., if $\int_0^1 \tilde{\theta}(b - c)d\tilde{\theta} \leq \int_{1-\alpha}^1 \tilde{\theta}bd\tilde{\theta}$. A sufficient condition for this to hold is

[4] If $c > b$, the equilibrium determines only the share of drivers riding the train but not their type.

$\frac{b-c}{b} \leq \alpha$. The sufficient condition holds if the road capacity is high enough (so that the social cost of eliminating congestion is not too high) or if the cost of congestion is high enough (so that eliminating congestion is sufficiently socially beneficial).

We first establish the baseline result: in this simplified setting, congestion pricing combined with lump-sum transfers maintaining budget balancedness can generate a Pareto improvement:

Proposition 1. *There exists a congestion pricing policy $T^* = (p^*, k^*)$ that satisfies the following:*

1. *T^* eliminates congestion*
2. *T^* is budget-balanced*
3. *T^* Pareto dominates T^0*

The fact that T^* Pareto dominates T^0 implies $\delta_\theta(T^*, T^0; 1) \geq 0$ for all θ. In other words, absent of equilibrium neglect, T^* is always politically feasible under T^0:

Corollary 1. *Let $\eta = 1$. Then T^* is politically feasible for any $\beta \in [0, 1]$.*

What happens when $\eta < 1$? Since much of the gains (and all of the surplus) from congestion pricing come from reduction in equilibrium congestion, lowering the value of η can narrow the set of feasibility thresholds such that the policy is politically feasible:

Proposition 2. *T^* is politically feasible under T^0 iff $\beta \leq \overline{\beta}(\eta)$, where*

1. *$\overline{\beta}(\eta)$ is weakly increasing in η*
2. *$\overline{\beta}(1) = 1$, $\overline{\beta}(0) = \min\{\frac{\alpha(1-\alpha)b}{b-c}, 1\}$*

Notice that this result implies that in some cases even Pareto improvements are politically infeasible. For example, if $\frac{b-c}{b} > \frac{1}{2}$ we have $\overline{\beta}(0) < \frac{1}{2}$—a Pareto improvement that cannot achieve a simple majority.

3.3 EN-robust Optimal Congestion Pricing

The fundamental challenge for achieving political feasibility is to make agents who neglect equilibrium effects support a policy whose benefits come from equilibrium effects. Indeed, in our application, equilibrium neglect can make a Pareto improvement infeasible. Moreover, since k^* is the highest possible transfer that keeps the budget balanced, there is no other standard pricing policy that achieves the no-congestion optimum and is politically feasible without running a deficit.

A natural candidate for a remedy is the use of subsidies. For example the social planner can subsidize the use of public transportation instead of charging a price for driving. However, subsidies are costly, and depending on the social cost of congestion and the degree of equilibrium neglect, the necessary expansion in the program's budget may induce its own set of feasibility issues.

Our objective is therefore to find a policy satisfying the following desiderata: (i) eliminates congestion, (ii) keeps the budget balanced and (iii) is politically feasible in the presence of equilibrium-neglecting agents. More generally, given a (possibly constrained) social optimum T^* that induces an equilibrium a^*, we are looking for a policy $\tilde{T}$ that satisfies: (i) *outcome equivalence*, i.e., $a^*(\tilde{T}) = a^*$, (ii) *deficit equivalence*, i.e., $D_{\tilde{T}}(a^*) = D_{T^*}(a^*)$, and (iii) *EN-robustness*, i.e., if T^* is feasible with a threshold β absent of equilibrium neglect, $\tilde{T}$ is feasible with the same threshold β under any degree of equilibrium neglect. When there is no ambiguity, we will often refer to EN-robust policies satisfying outcome and budget equivalence simply as "EN-robust policies," while referring to the baseline standard policy as "EN-sensitive."

As the analysis above shows, in the context of congestion pricing, a standard—i.e., anonymous and action-based—congestion pricing policy satisfying these conditions may not exist. What about other types of policies?

Relaxing anonymity can typically help increase political feasibility by allowing transfers targeted at agents whose direct losses from the policy change are the highest. However, the practical use of such policies may be limited, since they require the planner to condition the policies on types, which are typically assumed to be private information.

Instead, our remedy relies on the use of profile-based policies. Intuitively, EN constrains feasibility through agents' mistaken beliefs on others' actions. Profile-based policies allow us to condition payoffs on *others'* actions, and therefore can guarantee the agents benefits even under their mistaken perceptions about others'. Unlike relaxing anonymity, profile-based policies do not require any additional information.

The social planner can propose a profile-based anonymous policy conditioning price on congestion: $T_n(a) = k^* - a_n \cdot p^* \cdot \mathbf{1}(\gamma(a) \leq \alpha)$. Such policy may be intuitively appealing as it effectively ensures toll-paying agents a refund in the event that congestion persists—precisely the event that equilibrium-neglecting voters perceive as likely. Notice that this policy also satisfies deficit equivalence, since without congestion, the deficit is still zero. However, crucially, this policy does *not* satisfy outcome equivalence. In particular, it is easy to show that if congestion is an equilibrium under T^0, it is also an equilibrium under T. In other words, relative to T^*, T satisfies deficit equivalence and EN-robustness, but not outcome equivalence.

Offering a compensation for congestion for agents who drive does not satisfy outcome equivalence because it changes driving incentives. Therefore, in order to achieve political feasibility while maintaining outcome and deficit equivalence, we need to compensate for congestion while keeping driving incentives unchanged. This implies conditioning the *transfer*, rather than the price, on congestion, thus providing compensation for equilibrium-neglecting voters while keeping driving incentives constant:

Proposition 3. *Let $T^* = (p^*, k^*)$ be the standard optimal congestion pricing policy. Then the policy $\tilde{T}_n(a) = p^* \cdot \gamma(a) - a_n \cdot p^*$ is outcome and deficit equivalent to T^*, and is politically feasible under T^0 for any $\beta \in [0, 1]$ and any $\eta > 0$.*

This proposition shows that the social planner can guarantee the trifecta of outcome equivalence, deficit equivalence (relative to T^*), and EN robustness using a non-standard albeit quite natural and intuitive policy tool: *explicit* redistribution that conditions the payout on the share of agents driving. Redistribution achieves political feasibility by effectively providing insurance for voters: when congestion is high, so is compensation, guaranteeing benefits to drivers and non-drivers in all contingencies.

Notice that by construction, this policy keeps the budget balanced both on- and off path: since the size of the compensation always equals the size of the revenue, the government is never giving away more than it collects, independent of the action profile. This makes explicit redistribution not only natural, but also credible, since it does not increase the deficit relative to the status quo, in any contingency.

3.4 EN Robustness: a General Remedy

The previous section outlines how EN can make a socially desirable policy politically infeasible even when it generates benefits for all voters. It then shows how the social planner can generate an alternative policy that achieves political feasibility even under EN, without chaining the outcome or deficit. Of course, the EN feasibility constraint extends beyond the simple example we analyzed. Does the remedy we proposed generalize as well?

The following theorem shows that the same logic generalizes to *any* situation captured by our framework:

Theorem 1. *Let T^0 be a status quo policy, let $\beta \in [0,1]$, and let T^* be an anonymous action-based policy. If T^* is politically feasible under T^0 with respect to the threshold β assuming $\eta_n = 1$ for all n, then there exists an anonymous (profile-based) policy $\tilde{T}$ that is outcome and deficit equivalent to T^*, and is politically feasible with respect to β for all $\eta \in [0,1]^{|N|}$.*

The theorem shows that, starting from any status quo, for any action-based policy that is politically feasible assuming $\eta = 1$, there exists an EN-robust policy that achieves the same social outcome and deficit. In other words, given any constrained optimum that achieves standard political feasibility, we can generate at no additional cost an equivalent policy that completely eliminates potential EN constraints that may have prevented its implementation.

The proof of the theorem is constructive and therefore provides a recipe for constructing an EN-robust implementation. In essence, it takes the original policy and adds an

off-path rebate component: a contingent compensation paid only when others' behavior remains unchanged. Since the additional component only pays off the equilibrium path and since the policies are outcome equivalent, the deficit remains unchanged.

3.5 EN Robustness and Off-path Deficit

While relying on off-path rebates never increases the deficit *on path*, it may increase the deficit off path, which has potential downsides. For example, voters might view a policy promising excessively large compensation as less credible and may cause them to suspect whether the social planner will follow through. Moreover, even when the social planner can credibly commit to such policies, they themselves might suffer from equilibrium neglect, or may be skeptic of the equilibrium predictions. Therefore, the off-path rebates would be constructed in a way that prevents running large deficits, even in off-path contingencies.

We define an *off path* deficit as $D_T(a)$ where $a \neq a^*(T)$ (notice that there are many such deficits). We will now show that we can generate a policy that is deficit equivalent to either the new policy or the status quo, but doing so requires knowledge of the status quo action.

Let T^0 be some status-quo policy and let $a^0 = a^*(T^0)$. We say that a policy T is an anonymous past-action based (PA based) if T_n depends on a^0_n. As it turns out, knowing the agent's status quo action is a way to guarantee a form of off-path deficit equivalence:

Theorem 2. *Let T^0 be a status quo policy and let $a^0 = a^*(T^0)$, let $\beta \in [0,1]$, and let T^* be an anonymous action-based policy. If T^* is politically feasible under T^0 with respect to the threshold β assuming $\eta_n = 1$ for all n, then there exists an anonymous PA (profile-based) policy $\tilde{T}$ that is outcome and deficit equivalent to T^* and is politically feasible with respect to β for any $\eta \in [0,1]$. Moreover, $\tilde{T}$ satisfies $D_{\tilde{T}}(a) = D_{T^*}(a)$ for all $a \in A \setminus a^0$, and $D_{\tilde{T}}(a^0) = D_{T^0}(a^0)$.*

In other words, allowing the policy to depend on agents' status-quo actions, we can restrict the off-path deficit in the contingency where players do not change their behavior– the only contingency where $\tilde{T}$ and T^* diverge—to be the same as the status quo deficit.

In other words, the EN-robust policy achieves the desired social outcome while ensuring that even off path, the deficit does not increase relative to the status quo.

3.5.1 Discussion

Every EN-sensitive socially optimal action-based policy has an outcome and deficit equivalent EN-robust counterpart. As shown in the proof of Theorems 1 and 2, to construct such policies, the social planner adds a sufficiently large contingent compensation that pays when other players' behavior is the same as the status quo. In order to make the policy deficit-equivalent to the status quo policy under the status-quo profile, this compensation needs to depend on past actions. In particular, it is given by

$$T_n^0(a^0) - T^*(a_n^0),$$

i.e., the difference between what the agent pays under the status quo, and what the agent pays under the new policy if their behavior remains the same. Notice that this sum only depends on the agents' actions, not their type.[5] Intuitively, this difference serves as an upper bound on the losses from the direct effect, but aggregating it over all agents results in the same deficit as in the status quo.

Off-path payments can operate through different psychological channels. First, they might simply change perceived utility: if equilibrium neglect is driven by an explicit belief that with some probability the status quo does not change, then off-path payments ensure that the agent utility is higher even in this contingency. Second, they can operate through salience: when agents mistakenly overweight direct effects (relative to equilibrium effects), off-path payments ensure that this effect is sufficiently positive. Last, while we do not explicitly model this channel, off path payments can operate through persuasion: on top of increasing utility from the direct effect, promising an off-path payment can credibly signal to potentially skeptic agents that the social planner is confident in their

[5]In fact, the social planner does not need to know the player's status-quo action, since the payment only depends on what the agent paid in the status quo, and what they would pay if they did not change their behavior, which in some cases (such as in our congestion-pricing application) may be a strictly weaker requirement.

equilibrium prediction. Therefore, we conjecture that adding the contingent component to the policy can increase support by directly affecting η, making agents less pessimistic about potential equilibrium effects.

4 Experiment

The previous section proposes a remedy to equilibrium neglect by using off-path rebates. Is this solution effective when implemented in practice? Ex ante, there are reasons for skepticism. Perhaps the most important one is the fact that in order to gain support, EN-robust policies require agents to think through different benefits across different contingencies. However, the very reason agents suffer from EN may be due to failure in exactly that type of contingent reasoning. Agents may also oppose EN-robust policies because they seem complex, unnatural, or non-credible.

The next two sections test its effectiveness empirically. Our first context is a controlled incentivized lab setting that has been previously shown to generate equilibrium neglect (Dal Bó et al., 2018). We treat this test as a clean proof of concept, showing that EN-robust policies can indeed increase support in a setting where EN has been pinned down as an important driver of opposition to the social optimum. In the next section, we move to a more ecological setting: opinion polls on congestion pricing among residents of the NYC metro area, where, at the time of writing this paper, congestion pricing is frequently debated. While neither application proves that the above-mentioned reasons for skepticism are universally invalid, they do show that in some contexts, EN-robust policies can be highly effective in generating support, at no cost to the social planner.

4.1 Analysis and Predictions

In this section, we analyze behavior in the experiment using the equilibrium neglect framework. The first part of the analysis follows a similar logic to the analysis in Dal Bó et al. (2018), and the second part adds off-path payments to generate EN-robustness.

Consider a two-player two-action game, where agents choose between cooperate (C)

and defect (D). There is a baseline payoff I, over which the cost of cooperation is given by c, and the positive externality of cooperation (i.e., the benefit to the other player) is given by b, where $b > c > 0$. Consider first the zero-tax status quo policy, $T^0 = 0$. While the social optimum is (C, C), which results in $I + b - c$ for each player, since $c > 0$, the dominant strategy equilibrium is $a^*(T^0) = (D, D)$, which results in I for each player, and is socially suboptimal.

Consider the anonymous action-based policy space as described in Table 2, where taxes can only depend on the player's own actions. We denote the tax/subsidy imposed on action $a_i \in \{C, D\}$ by $T(a_i)$. For policies satisfying $b > -T(D) > -T(C) + c$, the dominant strategy equilibrium becomes $a^*(T) = (C, C)$, which results in $I+b-c+T(C) > I$ for each player, a strict improvement over the status quo. However, the perceived benefit from imposing T is given by

$$\delta(T, T^0; \eta) = -c + T(C) + \eta b,$$

and is positive iff $\eta \geq \frac{c-T(C)}{b}$. Thus, sufficiently neglectful agents support the status quo.

Table 1: Class of games with action-based policy space

	C	D
C	$b - c + T(C)$, $b - c + T(C)$	$-c + T(C)$, $b + T(D)$
D	$b + T(D)$, $-c + T(C)$	$T(D)$, $T(D)$

Note: Implied class of games allowing action-based policies. Each cell represents the payoff above the I baseline.

To allow off-path payments, we now introduce profile-based policies. We denote the payment resulting from the other player's choice $a_{-i} \in \{C, D\}$ by $h(a_{-i})$. Let $\tilde{T}(a) = T(a_i) + h(a_{-i})$ and assume $h(C) = 0$. The perceived benefit from $\tilde{T}$ under the status quo T^0 is given by

$$\delta(\tilde{T}, T^0; \eta) = -c + T(C) + h(D) + \eta(b + h(D)).$$

Consistent with Theorem 1, $\tilde{T}$ is outcome and deficit equivalent to the status quo,

and for $h(D) \geq -c + T(C)$, $\tilde{T}$ is EN-robust: the perceived benefits are positive for any $\eta \geq 0$. Consistent with Theorem 2, setting the off-path compensation to $h(D) = -T(D)$ ensures that the off-path deficit and the status quo deficit are both equal to zero while still maintaining EN-robustness. Notice that since the status quo equilibrium is symmetric, a_n^0 (i.e., agent n's status-quo NE action) is known by construction.

Table 2: Class of games with profile-based policy space

	C	D
C	$b - c - T(C)$, $b - c - T(C)$	$-c - T(C) + h(D)$, $b - T(D)$
D	$b - T(D)$, $-c - T(C) + h(D)$	$-T(D) + h(D)$, $-T(D) + h(D)$

Note: Implied class of games allowing profile-based policies, assuming $h(C) = 0$. ach cell represents the payoff above the I baseline.

To summarize, in this setup the status quo equilibrium is socially suboptimal, and using action-based taxation can strictly improve both players' utility. However, sufficiently equilibrium neglectful players will not support the new policy. By introducing off-path compensation, the social planner can achieve the same outcome with the same deficit while ensuring that all players support the new policy. In the next section present the experimental design first replicating the main finding in Dal Bó et al. (2018) on the prevalence of equilibrium-neglectful agents, and then on the effectiveness of our proposed remedy.

4.2 Experimental Design

Our experimental design replicates and extends Dal Bó et al. (2018). The experiment has two main parts. In part 1, each subject plays five rounds of prisoner's dilemma against another player. Subject are (truthfully) told that the other player's action is drawn from a past sample of real players. Effectively, this means that the other player's action is a randomly iid drawn binary variable that equals C (cooperate) with probability s where s is determined by the share of cooperation from the previous sample. To determine s we behavior in Dal Bó et al. (2018) as our sample of "first movers."[6] The payment in PD

[6]Subjects in the main sample are only told that their payments is determined based on an action taken by a past player who faced the same payment matrix when making their choices, but not how the

is identical across all treatments, and is the same as in Dal Bó et al. (2018). Following Dal Bó et al. (2018), the payment is presented in terms of Experimental Points (EPs). At the end of each round, participants learn the other player's action and the outcome.

After completing five rounds of PD in Part 1, participants are presented with two options: the original game (PD) and an alternative game (HG). They are asked to submit a selection for which game should be played in Part 2, which consists of five more rounds. This selection is then implemented with a 50% chance. Otherwise, a game is chosen at random. This probabilistic implementation of choice mirrors the Random Dictator and Simple Majority treatments in Dal Bó et al. (2018). In these treatments, no player knows whether they are pivotal at the time of voting, yet each player has a non-zero chance of being pivotal. In addition, it also allows us to get more data on behavior in HG in Baseline and compare it to behavior in the EN-robust variants of the game.

Payments in HG vary across treatments. As shown in Table 3 there are three treatments. In the *Baseline* treatment, HG payments are identical to the (non-EN-robust) payments in Dal Bó et al. (2018). In the two *EN-robust* treatments, the payments contain an off-path component satisfying EN-robustness. The payments vary by the size of their off-path deficit. In one treatment, Off-Path Deficit (OPD), the off-path deficit is higher than the status quo deficit by 1 EP, whereas the other treatment, Off-Path Budget Balanced (OPBB), is off-path deficit equivalent to PD. This variation allows us to quantify the gains from increasing the off-path deficit, allowing us to test whether increasing the off-path deficit translates to increased support.

After choosing the game for Part 2, but before learning about the outcome, we elicit participants' beliefs about the probability that a randomly chosen player from the past sample will choose to cooperate in both PD and HG using an incentive compatible elicitation mechanism Karni (2009). We elicit beliefs after participants make their game selection to make sure beliefs do not affect the choice, but before learning about the selected game, to make sure the outcome does not affect elicited beliefs.

Participants then learn about the outcome and play five more rounds of the chosen

other player's payment was determined. In practice, to determine payment for first movers, they were matched with each other.

Table 3: Payments in Prisoners' Dilemma (PD) and Harmony Game (HG)

Treatment		**PD**		**HG**	
		C	D	C	D
Baseline (Dal Bó et al., 2018)	C	9,9	3,11	8,8	2,7
	D	11,3	5,5	7,2	1,1
EN Robust (OPBB)	C	9,9	3,11	8,8	6,7
	D	11,3	5,5	7,6	5,5
EN Robust (OPD)	C	9,9	3,11	8,8	7,7
	D	11,3	5,5	7,7	6,6

Note: Payoff matrices across treatments. Conversion rate: 20 points per $1.

game, followed by demographic questions.

4.2.1 Comparison to Dal Bó et al. (2018)

The experiment replicates the main treatments from Dal Bó et al. (2018), using a similar UI (see Online Appendix) and experimental instructions. The main differences between the two experiments are related to the fact that we recruit our sample online, and use an online interface, whereas Dal Bó et al. (2018) use students in a physical lab. This forced us to make a few necessary changes and adapt the design to allow for different participants to play at different times. Last, due to budget constraints, the exchange rate from Experimental Points to dollars we used was $1 per 20 experimental points, which is lower than Dal Bó et al. (2018)'s $1 per three experimental points.

First, we match participants to players from past samples, rather than having the two players make their choices simultaneously. Second, the game for Part 2 is chosen either by the player or at random, and does not depend on other players' choices. We generally view these differences as minor: we are only interested in establishing a baseline result qualitatively similar to Dal Bó et al. (2018) on the level of support for PD in order to test our remedy in the same setting.

4.3 Results

The experiment is programmed and run using Qualtrics. Participants will be recruited from Prolific and randomly assigned to one of the three treatments: Baseline, EN-robust OPBB and EN-robust OPD.

4.3.1 In-game Behavior

We first test outcome equivalence: conditional on being implemented, do EN-robust variants induce more/less cooperation than the Baseline HG? This section will compare cooperation rates across conditions.

4.3.2 Political Feasibility

Are players more likely to support EN-robust policies? This section will compare the share of participants selecting HG in the selection stage in each treatment.

4.3.3 Beliefs

Do EN-robust policies make increase support through persuasion? As discussed in section 3.5.1, when the social planner signals confidence in behavioral responses it may reducing agents' equilibrium neglect (increasing η) and make them more optimistic about positive equilibrium effects. This section will compare beliefs in cooperation across treatments.

5 Survey Design

5.1 Motivation

The previous section demonstrates the effectiveness of EN-robustness in a highly controlled setting. While serving as a reasonable first step, and a convenient wind tunnel setting where equilibrium neglect plays an important role to empirically test our remedy, one can rightfully question how well it ports to other settings. In particular, in this setting, off-path payments can be effectively communicated simply by changing a number within a table cell. In contrast, since real policies are typically not communicated using

payoff matrices, contingent payments may feel unnatural, explaining them may be cumbersome, and understanding them requires thinking through contingencies. This worry looms particularly large when studying equilibrium neglect—a particular form of failure in contingent reasoning.

For our main ecological application, we choose congestion pricing. As discussed in the introduction, congestion pricing is particularly suitable for our purposes for a few reasons. First, congestion pricing has a clean distinction between the direct effect (price change) and equilibrium effect (congestion reduction). Second, convergence to the new equilibrium is essentially immediate, significantly reducing timing-related complications. Third, the main outcome has a single dimension (amount of cars/travel time), further reducing complications regarding the definition of the action space A. Last, the outcome can be easily and transparently measured, making it easier to credibly commit to contingent payments.

5.2 Design

To empirically test whether our proposed contingent compensation mechanism increases public support for congestion pricing policies, we employ a survey experiment with a representative sample from six U.S. metropolitan areas with known traffic problems: Boston, Washington D.C., Chicago, Dallas, San Francisco, and Los Angeles ($N = 2,000$).

We use a two-stage, between-subject design, where all respondents first evaluate a standard congestion pricing policy, then are randomly assigned to evaluate one of three modified policy versions.

5.2.1 Baseline Elicitation

All respondents are presented with a hypothetical standard congestion pricing policy in the downtown area of their city, charging vehicles a \$15 daily toll during peak hours. After indicating their level of support, respondents answer questions about their commute habits, personal valuation of time savings, and beliefs about the policy's effectiveness.

5.2.2 Treatment Conditions

Respondents are then randomly assigned to one of three conditions to evaluate a modified policy:

Treatment Condition ($N = 800$): 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."

We focus on workers as the recipient group for compensation because they approximately correspond to the "status quo drivers" identified in our theoretical framework—individuals who currently drive into the congested area and would be directly affected by the policy. While not all workers drive, the set of downtown workers encompasses most regular commuters into the area and represents a clearly defined, administratively feasible group for policy implementation.

Control Condition ($N = 800$): The policy includes only the "monitoring" part of the added clause "Traffic levels would be monitored and published monthly." This condition is meant to isolate the effect of the treatment, from the effect of repeated elicitation, and the pure monitoring effect.

Toll Reduction Condition ($N = 400$): The policy has a reduced toll from $15 to $9.

In all elicitations, respondents indicate their support using a five-point scale: Strongly support, Somewhat support, Neither support nor oppose, Somewhat oppose, or Strongly oppose.

5.2.3 Parameter Selection

Our choice of key parameters is informed by empirical evidence from existing congestion pricing schemes and proposals, particularly those in New York City:

Peak-Hour Toll The $15 baseline toll reflects the original proposal for New York City's congestion pricing program. We selected the $9 reduced toll based on NYC's revised proposal announced in 2024, which was adjusted in response to public and political opposition. These values provide realistic benchmarks that respondents can evaluate and allow us to compare our contingent compensation approach against an actual policy compromise implemented in response to equilibrium neglect.

Projected Traffic Reduction The expected 17% traffic reduction with a $15 toll is derived from the Metropolitan Transportation Authority's environmental assessment of the congestion pricing program. The 8% threshold for compensation represents approximately half the expected reduction, providing a clear benchmark that would indicate significant underperformance of the policy while still being achievable.

Monthly Compensation The $120 monthly compensation amount was calculated to ensure budget balance, even in scenarios where the congestion reduction fails to materialize. Our calculations incorporate several factors: average daily revenue from toll collection, accounting for both weekday and weekend traffic patterns; administrative costs and exemptions (estimated at 50% of gross revenue); the ratio between total commuters and the number of workers who commute to work in the downtown area (based on NYC data); and a portion of revenue (20%) reserved for transit improvements even in off-equilibrium scenarios.

Using these inputs, we derived a monthly compensation formula of approximately $C(T) = 8T$, where T is the toll amount. For a $15 toll, this yields a monthly compensation of $120. This amount is high enough to potentially influence policy preferences while remaining fiscally feasible through the toll revenue that would be collected in low-reduction scenarios.

5.3 Outcome Measures

Our primary outcome measure is *net support*—the percentage supporting minus the percentage opposing—compared across conditions. This allows us to quantify the treatment

effect as the increase in support resulting from the policy modification, with the control condition capturing potential increases driven merely by adding a clause and repeating the elicitation.

The comparison to the toll reduction condition serves as a benchmark, measuring the effectiveness of our contingent compensation approach relative to a more standard policy tool.

5.4 Supplementary Measures

Beyond the primary support question, we collect data on commuting patterns, driving habits, time valuation, and beliefs about policy effectiveness.

A key hypothesis is that pessimism about congestion reduction will more strongly predict opposition to the standard policy than to the compensation-enhanced version, since compensation provides insurance against ineffectiveness.

We measure perceived effectiveness twice (pre- and post-treatment) with randomization between beliefs about respondents' own city and beliefs about NYC. This allows us to assess both how beliefs affect support and whether treatment potentially changes beliefs, though our theory makes no specific prediction about treatment effects on perceived effectiveness.

5.5 Results

5.5.1 Sample

Our sample consists of 2,000 respondents recruited through NORC's AmeriSpeak Panel, a probability-based panel designed to be representative of the U.S. household population. AmeriSpeak employs a two-stage stratified sampling method with coverage of 97% of U.S. households. For our study, we restrict the sample to residents of six major metropolitan statistical areas (MSAs) with significant traffic congestion problems: Boston, Washington D.C., Chicago, Dallas, San Francisco, and Los Angeles.

We selected these six MSAs based on feasibility constraints, and represent diverse

geographic regions across the United States. They also consistently rank among cities with the highest traffic congestion in national studies, increasing the realism and relevance of our hypothetical scenarios.

5.5.2 Political Feasibility

We start with our main question: can EN-robustness increase support for congestion pricing? This section will compare net support in the second-stage elicitation across the three treatments.

5.5.3 Perceived Benefits and Support

In this section I will construct, for each subject, a "true" and "perceived" benefit index, and show that the latter is predictive of support for the standard policy, while the former is not.

5.5.4 Treatment Effect on Beliefs

In this section I will compare beliefs across treatments, focusing on the second stage, testing whether changing the treatment changes participants' beliefs about the policy.

6 Conclusion

TBD...

References

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.

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.

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.

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.

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.

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.

Edi Karni. A Mechanism for Eliciting Probabilities. *Econometrica*, 77(2):603–606, 2009. URL http://www.jstor.org/stable/40263877.

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.

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.

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.

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.

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

A Proofs

A.1 Proposition 1

Proof. Let $(p^*, k^*) = (b(1-\alpha), \alpha \cdot b(1-\alpha))$.

To see part (1), notice that without congestion, the share of agents who drive given a price p is given by $1 - \frac{p}{b}$. Therefore, plugging in p^* indeed implies no congestion. In other words, under p^* the equilibrium is such that only drivers with $\theta \in [1-\alpha, 1]$ drive.

Moving on to part (2), notice that since $\gamma(a^*((p^*, k^*))) = \alpha$, the revenue is given by $\alpha p^* = \alpha b(1-\alpha) = k^*$ which implies $D_{T^*}(a^*(T^*)) = 0$.

To show that T^* Pareto dominates T^0, we focus on the agent indifferent between driving and riding the train: $\tilde{\theta} \equiv 1 - \alpha$. We have

$$\begin{aligned} v_{\tilde{\theta}}(a^*(T^*) \mid T^*) &= \underbrace{b(1-\alpha)\alpha}_{\text{transfer}} + \underbrace{(1-\alpha)b}_{\text{time}} - \underbrace{b(1-\alpha)}_{\text{price}} \\ &= b(1-\alpha)\alpha \\ &\geq v_{\tilde{\theta}}(a^*(T^0) \mid T^0) = (1-\alpha)(b-c) \iff \alpha \geq \frac{b-c}{b}, \end{aligned}$$

which we assumed holds as a sufficient condition for congestion being suboptimal. For all higher types, $\theta \in [1-\alpha, 1]$, we have

$$\frac{\partial v_{\tilde{\theta}}(a^*(T^*) \mid T^*)}{\partial \theta} = b > \frac{\partial v_{\tilde{\theta}}(a^*(T^*) \mid T^*)}{\partial \theta} = b - c,$$

and therefore the inequality holds for them as well.

For all lower types, $\theta \in [0, 1-\alpha]$, we have $\frac{\partial v_{\tilde{\theta}}(a^*(T^*)|T^*)}{\partial \theta} = 0 < \frac{\partial v_{\theta}(a^*(T^0)|T^0)}{\partial \theta} = b - c$, and therefore they all strictly gain from moving from T^0 to T^*.

□

A.2 Proposition 2

Proof. First part: For a given policy T, $v_\theta(a \mid T)$ is always weakly decreasing in the share of road users (i.e., if $\gamma(a') > \gamma(a)$ then $v_\theta(a \mid T) \geq v_\theta(a' \mid T)$). Since $\gamma(a^*(T^*)) = \alpha <$

$\gamma(a^*(T^0)) = 1$, we have $v_\theta(a^*(T^*) \mid T^*) - v_\theta(a^*(T^0) \mid T^*) \geq 0$, which implies $\frac{\partial \delta_\theta(T^*, T^0; \eta)}{\partial \eta} \geq 0$. Therefore, the share of agents for which $\delta_\theta(T^*, T^0; \eta) \geq 0$ is increasing in η.

Moving on to the second part, when $\eta = 1$, Proposition 1 applies. When $\eta = 0$, we have two types: perceived non-drivers, and perceived drivers. These are the types that are (not) driving when the price is p^* and *the road is congested*. Solving for the perceived marginal (the type indifferent between driving and riding the train, when the price is p^* and the road is congested) yields $\hat{\theta} = \frac{b(1-\alpha)}{b-c}$.

For $\theta \in [0, \hat{\theta}]$ we have $\delta_\theta(T^*, T^0; 0) = \underbrace{b(1-\alpha)\alpha}_{k^*} - \underbrace{\theta(b-c)}_{\text{time loss}}$. Solving $\delta_\theta(T^*, T^0; 0) = 0$ for θ yields $\theta^* = \frac{b(1-\alpha)\alpha}{b-c}$, which implies that for $\theta \in [0, \min\{\theta^*, 1\}]$ we have $\delta_\theta(T^*, T^0; 0) \geq 0$, and therefore for $\eta = 0$, the feasibility threshold is given by $\min\{\frac{b(1-\alpha)\alpha}{b-c}, 1\}$, as needed.

Notice that for perceived drivers, the perceived benefit is given by $\underbrace{b(1-\alpha)\alpha}_{k^*} - \underbrace{b(1-\alpha)}_{p^*}$ which is always negative. □

A.3 Proposition 3

Proof. Notice that since $\tilde{T}$ and T^* only differ in the component that is independent of a_n, by quasi linearity we can write $v_\theta(a \mid \tilde{T})) = v_\theta(a \mid T^*) + h(a_{-n})$. Therefore if a_n^* is the unique maximizer of $v_\theta(a^*(T^*) \mid T^*))$, it is also the unique maximizer of $v_\theta(a \mid \tilde{T}))$, showing outcome equivalence of $\tilde{T}$ and T^*. Since the transfer is equal to the sum collected, it is budget-balanced (and in particular, deficit equivalent) by construction. To show political feasibility, notice that

$$\delta_\theta(\tilde{T}, T^0; 0) = p^* \cdot \underbrace{\gamma(a^*(T^0))}_{=1} - \min\{\theta(b-c), p^*\} \geq 0$$

As shown in the proof of Proposition 2, we have $\frac{\partial \delta_\theta(T^*, T^0; \eta)}{\partial \eta} \geq 0$, which ensures $\delta_\theta(\tilde{T}, T^0; \eta) \geq 0$ for all $\eta > 0$. □

A.4 Theorems 1 and 2

Proof. Let $v_n(a \mid T) = w_n(a) + T_n(a)$. Denote $a^*(T^*) = a^*(T^*)$ and $a(T^0) = a^0$. Since T^* is politically feasible under $\eta = 1$, we have $w_n(a^*) + \hat{T}(a_n^*) - w_n(a^0) - T^0(a^0) \geq 0$ for at least β share of the population.

Let $\overline{a}_n = \arg\max_{a_n} w(a_n, a_{-n}^0) + T^*(a_n)$. We have

$$\delta_n(T^*, T^0; \eta) = w_n(\overline{a}_n, a_{-n}^0) + T^*(\overline{a}_n) - w_n(a^0) - T^0(a^0) + \eta[w_n(a^*) + \hat{T}(a_n^*) - w_n(\overline{a}_n, a_{-n}^0) - \hat{T}(\overline{a}_n)]$$

Let $\tilde{T}(a) = T^*(a_n) + \tilde{f}(a_{-n})$ where $\tilde{f}(a_{-n}) = k_n \cdot \mathbf{1}\{a_{-n} = a_{-n}^0\}$. We characterize k below in the case of Theorem 1 and 2. For now we emphasize that k_n does not depend on agent n's actions. Notice that by quasi-linearity, the best response is the same under T^* and $\tilde{T}$.

Take some agent for which $\delta_n(T^*, T^0; 1) \geq 0$. Notice that for any $k \geq 0$, we have $\delta_n(\tilde{T}, T^0; \eta) \geq \delta_n(T^*, T^0; \eta)$. There are two possible cases: the equilibrium effect can be either positive or negative. If the equilibrium effect is (weakly) negative, then $\delta_n(T^*, T^0; \eta)$ is (weakly) decreasing in η. Therefore we have

$$\delta_n(\tilde{T}, T^0; \eta) \geq \delta_n(T^*, T^0; \eta) \geq \delta_n(\hat{T}, T^0; 1) \geq 0,$$

and if T^* is politically feasible under T^0, so is $\tilde{T}$.

If the equilibrium effect is positive, then $\delta_n(\hat{T}, T^0; \eta)$ is increasing in η, and so it suffices to look at the case where $\eta = 0$. We have

$$\delta_n(\tilde{T}, T^0; 0) = w_n(\overline{a}_n, a_{-n}^0) + \hat{T}(\overline{a}_n) + \tilde{f}(a_{-n}^0) - w_n(a^0) - T^0(a^0)$$

Therefore, for any

$$k \geq w_n(a^0) + T^0(a^0) - w_n(\overline{a}_n, a_{-n}^0) - \hat{T}(\overline{a}_n), \tag{1}$$

the perceived benefit is positive, i.e., k_n must be equal to or larger than the direct effect.

We can therefore choose for all n

$$k_n = k = \max_m w_m(a^0) + T^0(a^0) - w_m(\overline{a}_m, a^0_{-m}) - \hat{T}(\overline{a}_m)$$

which only requires knowledge of the support.

To prove theorem 2, notice that since $\overline{a}_n$ is a best response, we must have

$$w_n(\overline{a}_n, a^0_{-n}) + T^*(\overline{a}_n) \geq w_n(a^0) + T^*(a^0_n).$$

Let $\tilde{T}(a^0, a) = T^*(a_n) + k_n(a^0_n) \cdot \mathbf{1}\{a_{-n} = a^0_{-n}\}$ where $k_n(a^0_n) = T^0_n(a^0) - T^*(a^0_n)$. Notice that whether k_n is paid or not only depends on other players' current actions, and its size only depends on past actions, which ensures outcome equivalence.

To show political feasibility, we use the same argument as before and notice that

$$T^0_n(a^0) - T^*(a^0_n) = w_n(a^0) + T^0(a^0) - w_n(a^0) - T^*(a^0_n) \geq w_n(a^0) + T^0(a^0) - w_n(\overline{a}_n, a^0_{-n}) - T^*(\overline{a}_n)$$

Last, to characterize 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

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

as needed. □